Qualitative and Quantitative Analysis of Scientific and Scholarly Communication

Nikolay K. Vitanov

Science Dynamics and Research Production

Indicators, Indexes, Statistical Laws and Mathematical Models

Springer

Qualitative and Quantitative Analysis of Scientific and Scholarly Communication

Series editors

Wolfgang Glänzel, Katholieke Univeristeit Leuven, Leuven, Belgium
Andras Schubert, Hungarian Academy of Sciences, Budapest, Hungary

More information about this series at http://www.springer.com/series/13902

Nikolay K. Vitanov

Science Dynamics and Research Production

Indicators, Indexes, Statistical Laws and Mathematical Models

Springer

Nikolay K. Vitanov
Institute of Mechanics
Sofia
Bulgaria

and

Max-Planck Institute for the Physics of
Complex Systems
Dresden
Germany

ISSN 2365-8371 ISSN 2365-838X (electronic)
Qualitative and Quantitative Analysis of Scientific and Scholarly Communication
ISBN 978-3-319-41629-8 ISBN 978-3-319-41631-1 (eBook)
DOI 10.1007/978-3-319-41631-1

Library of Congress Control Number: 2016944335

© Springer International Publishing Switzerland 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To my parents and teachers, who helped me to find my way through the mountains and valleys of life.

Preface

He who sees things grow from the beginning will have the best view of them

Aristotle

There is a variety of books on the topic of the "science of science," books, that are devoted to the social and economic aspects of science [1–8]; books devoted to innovation and technological change [9–11]; books devoted to the study of models of science dynamics [12–14]; books devoted to studies in the area of scientometrics, bibliometrics, informetrics, webometrics, scientometric indicators and their applications [15–36]; and especially books devoted to citations and citation analysis [37, 38]. The goal of this book is different from those of most of the books mentioned above, because this book is designed as an introductory textbook with elements of a handbook. Its goal is to introduce the reader to two selected areas of the science of science: (i) indicators and indexes for assessment of research production and (ii) statistical laws and mathematical models connected to science dynamics and research production. The introduction is from the point of view of applied mathematics (i.e., no proofs of theorems are presented).

In the course of time, science becomes more and more costly to produce, and because of this, the dynamics of research organizations and assessment of research production are receiving increasing attention. As a consequence of the increasing costs, many national funding authorities are pressed by the governments for better assessment of the results of their investment in scientific research. And this pressure tends to increase. Because of this, interest in objectively addressing the quality of scientific research has increased greatly in recent years. One observes an increase in the frequency of the formation and action of various groups for quality assessment of scientific research of individuals, departments, universities, systems of institutes, and even nations.

Mathematics may provide considerable help in the assessment of complex research organizations. Numerous indicators and indexes for the measurement of performance of researchers, research groups, research institutes, etc. have been

developed. Numerous models and statistical laws inform us about specific modalities of the evolution of scientific fields and research organizations. We shall discuss below some of these indicators, indexes, statistical laws, and mathematical models.

Let us consider the potential readers of this book from the point of view of their knowledge about science dynamics and the tools for evaluation of research production. We shall see in Chap. 4 that rankings often lead to a power-law distribution and to an effect called the concentration–dispersion effect: If we have components of some organization, and these components own units, then often large numbers of units are concentrated in a small percentage of the components (concentration), and the remaining units are dispersed among the remaining larger number of components (dispersion). Let us assume that this effect is valid for the readers of this book (the components) with respect to their knowledge about science dynamics (measured in units of research articles read on this subject). Then there may be a concentration of much knowledge about dynamics of science and features of research production in a small group of highly competent readers. The concentration–dispersion effect helps us to identify target groups of readers as follows.

- **Target group 1**: *Readers who want to understand the dynamics of research organizations and assessment of research production but don't have knowledge about the dynamics of such organizations and/or about the tools for assessment of research production*.
  This group is very important, since every researcher and every manager of a research organization was a member of this group at least at the beginning of his/her career. In order to make this book more valuable for this group of readers, we discuss a large number of topics on a small number of pages, and the level of mathematical difficulty is kept low. The presence of numerous references allows us to achieve this degree of compactness.
- **Target group 2**: *Readers who (i) have some knowledge in the area of theory of science dynamics, (ii) have some practice in the assessment of research, and (iii) want to increase their knowledge about science dynamics and assessment of research*.
  This group of intermediate size is quite important, since large number of researchers and managers belong to it. I hope that the part of the book devoted to models will be of interest to the practitioners, and that the discussions of concepts and results from their practical implementation will be of interest to theoreticians.
- **Group 3**: *Very experienced researchers and practitioners in the areas of science dynamics and assessment of research production*.
  This relatively small group of researchers is very competent and has much knowledge. I hope, however, that this book will also be of interest to such readers as a collection of tools and concepts about the evaluation of research production and the dynamics of research organizations, and as an applied mathematics point of view on the features of such organizations.

The positioning of this book as an introduction to the large field of the mathematical description of science dynamics and to quantitative assessment of research production determined the choice of the concepts and models discussed and led to the following features:

- A relatively large number of mathematical models, concepts, and tools are discussed. The goal of this is to provide the reader with an impression and basic knowledge about the huge field of models of science dynamics and about the even larger field of research on indicators and indexes for assessment of research production. Nevertheless, the number of discussed models is small in comparison to the number of existing models. Thus many classes of models, e.g., network models of research structures, are not discussed in detail. This is compensated by numerous references.
- The focus of the book is on the quantitative description of science dynamics and on the quantitative tools for assessment of research production. Because of this, a significant mathematical arsenal, especially from the area of probability theory and the theory of stochastic systems, was used. Nevertheless, many complicated mathematical models were omitted, but after studying the material of the book, the interested reader should have no difficulty in understanding even the most complicated models.
- About 1,200 references are included in the book. This allowed me to keep the size of the book compact, using the feature of references as a compressed form of research information. By means of the numerous references, the reader may quickly obtain a large quantity of additional information about the corresponding topic of interest directly from sources that represent the original points of view of experienced researchers.

The book consists of three parts. The first part of the book is devoted to a brief introduction to the complexity of science and to some of its features. The triple helix model of a knowledge-based economy is described, and scientific competition among nations is discussed from the point of view of the academic diamond. The importance of scientometrics and bibliometrics is emphasized, and different features of research production and its evaluation are discussed. A mathematical model for quantification of research performance is described.

The second part of the book contains a discussion of the indicators and indexes of research production of individual researchers and groups of researchers. It is hard to find an alternative to peer review if one wants to evaluate the quality of a paper or the quality of scientific work of a single researcher. But if one has to evaluate the research work of collectives of researchers from some department or institute, then one may need additional methodology, such as a methodology for analysis of citations and publications. The building blocks of such methodology as well as selected indicators and indexes are described in this book, and many examples for the calculation of corresponding indexes are presented. In such a way, the reader may observe the indexes "in action," and he/she can get a good impression of their strengths and weaknesses. An important goal of this part is to serve as a handbook of useful indicators and indexes. Nevertheless, some discussion about features

of indexes is presented. Special attention is devoted to the Lorenz curve and to the definition of sizes of different scientific elites on the basis of this curve.

The third part of the book is devoted to the statistical laws and mathematical models connected to research organizations, and the focus is on the models of research production connected to the units of information (such as research publication) and to units of importance of this information (such as citations of research publications). Numerous non-Gaussian statistical power laws of research production and other features of science are discussed. Special attention is devoted to the application of statistical distributions (such as the Yule distribution, Waring distribution, Poisson distribution, negative binomial distribution) to modeling features connected to the dynamics of research publications and their citations. In addition, deterministic models of science dynamics (such as models based on concepts of epidemics and other Lotka–Volterra models) and models based on the reproduction–transport equation and on a master equation, etc., are discussed.

Several concluding remarks are summarized in the last chapter of the book.

In the process of writing of a book, every author uses some resources and discusses different aspects of the text with colleagues. I would like to thank the Max-Planck Institute for the Physics of Complex Systems in Dresden, Germany, where I was able to use the scientific resources of the Max-Planck Society. In fact, two-thirds of the book was written in Dresden. I would like to thank personally Prof. Holger Kantz, of MPIPKS, for his extensive support during the writing of the book, as well as Prof. Peter Fulde for extensive advice about practical aspects of science dynamics and research management. I would like to thank also two COST Actions: TD1210 "Analyzing the dynamics of information and knowledge landscapes—KNOWeSCAPE" and TD1306 "PEERE" for the possibility of numerous discussions with leading scientists in the area of scientometrics and evaluation of scientific performance. I would like thank Dr. Zlatinka Dimitrova and Kaloyan Vitanov for countless discussions on different questions connected to the book and for their help in the preparation of the manuscript. Many thanks to the Springer team and especially to Dr. Claus Ascheron for their excellent work in the process of preparation of the book. Finally, I would like to thank the (wise) anonymous reviewer, who advised me on how to arrange the text. That was useful indeed.

Sofia and Dresden Nikolay K. Vitanov

References

1. J.D. Bernal, *The Social Function of Science* (The MIT Press, Cambridge, MA, 1939)
2. V.V. Nalimov, *Faces of Science* (ISI Press, Philadelphia, 1981)
3. G. Böhme, N. Stehr (eds.), *The Knowledge Society* (Springer, Netherlands, 1986)
4. M. Gibbons, C. Limoges, H. Nowotny, S. Schwartzman, P. Scott, M. Throw, *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies* (Sage Publications, London, 1994)

5. E. Mansfield, *Industrial Research and Technological Innovation: An Econometric Analysis* (Norton, New York, 1968)
6. W. Krohn, E.T. Layton, Jr., P. Weingart, *The Dynamics of Science and Technology* (Reidel, Dordrecht, 1978)
7. M. Hirooka. *Innovation Dynamism and Economic Growth. A Nonlinear Perspective* (Edward Elgar Publishing, Cheltenham, UK, 2006)
8. P.A.A. van den Besselaar, L.A. Leydesdorff, *Evolutionary Economics and Chaos Theory: New Directions in Technology Studies* (Frances Pinter Publishers, 1994)
9. H. Grupp (ed.), *Dynamics of Science-Based Innovation* (Springer, Berlin, 1992)
10. L. Girifalco, *Dynamics of Technological Change* (Van Nostrand Reinhold, New York, 1991)
11. H. Etzkowitz, *The Triple Helix: University-Industry-Government Innovation in Action* (Routledge, New York, 2008)
12. A.I. Yablonskii, *Mathematical Methods in the Study of Science* (Nauka, Moscow, 1986) (in Russian)
13. H. Small, *Bibliometrics of Basic Research* (National Technical Information Service, 1990)
14. A. Scharnhorst, K. Börner, P. van den Besselaar (eds.), *Models for Science Dynamics* (Springer, Berlin, 2012)
15. L. Leydesdorff, *The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications* (DSWO Press, Leiden, 1995)
16. E. Garfield, *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities* (Willey, New York, 1979)
17. D. de Solla Price, *Little Science, Big Science* (Columbia University Press, New York, 1963)
18. A. Andres, *Measuring Academic Research. How to Undertake a Bibliometric Study* (Chandos, Oxford, 2009)
19. S.D. Haitun, *Scientometrics: State and Perspectives* (Nauka, Moscow, 1983) (in Russian)
20. S.D. Haitun, *Quantitative Analysis of Social Phenomena* (URSS, Moscow, 2005) (in Russian)
21. I.K. Ravichandra Rao, *Quantitative Methods for Library and Information Science* (Wiley-Eastern. New Delhi, 1983)
22. A.F.J. van Raan (ed.), *Handbook of Quantitative Studies of Science and Technology* (North-Holland, Amsterdam, 1988)
23. Y. Ding, R. Rousseau, D. Wolfram (eds.), *Measuring Scholarly Impact* (Springer, Cham, 2014)
24. L. Egghe, R. Rousseau, *Introduction to Informetrics: Quantitative Methods in Library, Documentation, and Information Science* (Elsevier, Amsterdam, 1980)
25. M. Callon, J. Law, A. Rip, *Mapping of the Dynamics of Science and Technology* (McMillan, London, 1986)
26. L. Egghe, *Power Laws in the Information Production Process: Lotkaian Informetrics* (Elsevier, Amsterdam, 2005)
27. D. Wolfram, *Applied Informatics for Information Retrieval Research* (Libraries Unlimited, Westport, CT, 2003)
28. M. Thelwall, *Introduction to Webometrics: Quantitative Web Research for the Social Sciences* (Morgan & Claypool, San Rafael, CA, 2009)
29. K. Fisher, *Changing Landscapes of Nuclear Physics: A Scientometric Study* (Springer, Berlin, 1993)
30. T. Braun, E. Bujdoso, A. Schubert, *Literature of Analytical Chemistry: A Scientometric Evaluation* (CRC Press, Boca Raton, FL, 1987)
31. P. Ingwersen, *Scientometric Indicators and Webometrics and the Polyrepresentation Principle in Information Retrieval* (ESS Publications, New Delhi Bangalore, India, 2012)
32. B. Cronin, C.R. Sugimoto, *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (MIT Press, Cambridge, MA, 2014)
33. T. Braun, W. Glänzel, A. Schubert, *Scientometrics Indicators. A 32 Country Comparison of Publication Productivity and Citation Impact* (World Scientific, London, 1985)

34. H.F. Moed, W. Glänzel, U. Schmoch (eds.), *Handbook of Quantitative Science and Technology Research* (Springer Netherlands, 2005)
35. P. Vinkler, *The Evaluation of Research by Scientometric Indicators* (Chandos, Oxford, 2010)
36. M.A. Akoev, V.A. Markusova, O.V. Moskaleva, V.V. Pislyakov, *Handbook of Scientometrics: Indicators for Development of Science and Technology* (University of Ural Publishing, Ekaterinburg, 2014) (in Russian)
37. B. Cronin, *The Citation Process. The Role and Significance of Citations in Scientific Communication* (Taylor Graham, London, 1984)
38. H. Moed, *Citation Analysis in Research Evaluation*. (Springer, Netherlands, 2005)

Contents

Part I
Science and Society. Research Organizations and Assessment of Research

In this part, we present a minimum amount of basic knowledge needed for understanding indexes and mathematical models from the following two parts of the book. This part contains one chapter, which begins with a discussion of the complexity of science: science is considered an open system that needs numerous inflows in order to remain in an organized state. In addition, two important concepts connected to science are described. The triple helix concept shows the place of science and academic research in the modern knowledge-based economy. The second concept (academic diamond) is closely connected to the important question of competition and especially to scientific competition among nations.

The text continues by presenting basic information about assessment of research production. The discussion begins on a technical level from process indicators and continues to latent variables and scales of measurements. The non-Gaussian nature of many processes in science and research is emphasized, since this has implications for the methodology of modeling research dynamics and for the methodology for assessment of research production. Further, a minimum basic knowledge about scientometrics, bibliometrics, informetrics, and webometrics is presented, and an impression about the quantities that may be used in the process of research evaluation is given. The role of knowledge landscapes for the study of research systems is briefly discussed. The importance of the study of research publications and their citations for the assessment of research is emphasized. A method for quantification of research performance (based on qualitative and quantitative input information) is presented.

Chapter 1
Science and Society. Assessment of Research

Dedicated to Derek John de Solla Price and to all Price award winners whose contributions established scientometrics, bibliometrics and informertics as important and fast developing branches of the modern science.

Abstract Science is a driving force of positive social evolution. And in the course of this evolution, research systems change as a consequence of their complex dynamics. Research systems must be managed very carefully, for they are dissipative, and their evolution takes place on the basis of a series of instabilities that may be constructive (i.e., can lead to states with an increasing level of organization) but may be also destructive (i.e., can lead to states with a decreasing level of organization and even to the destruction of corresponding systems). For a better understanding of relations between science and society, two selected topics are briefly discussed: the Triple Helix model of a knowledge-based economy and scientific competition among nations from the point of view of the academic diamond. The chapter continues with a part presenting the minimum of knowledge necessary for understanding the assessment of research activity and research organizations. This part begins with several remarks on the assessment of research and the role of research publications for that assessment. Next, quality and performance as well as measurement of quality and latent variables by sets of indicators are discussed. Research activity is a kind of social process, and because of this, some differences between statistical characteristics of processes in nature and in society are mentioned further in the text. The importance of the non-Gaussianity of many statistical characteristics of social processes is stressed, because non-Gaussianity is connected to important requirements for study of these processes such as the need for multifactor analysis or probabilistic modeling. There exist entire branches of science, *scientometrics, bibliometrics, informetrics, and webometrics*, which are devoted to the quantitative perspective of studies on science. The sets of quantities that are used in scientometrics are mentioned, and in addition, we stress the importance of understanding the inequality of scientific achievements and the usefulness of knowledge landscapes for understanding and evaluating research performance. Next, research production

© Springer International Publishing Switzerland 2016

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1_1

and its assessment are discussed in greater detail. Several examples for methods and systems for such assessment are presented. The chapter ends with a description of an example for a combination of qualitative and quantitative tools in the assessment of research: the English–Czerwon method for quantification of scientific performance.

1.1 Introductory Remarks

The word *science* originates from the Latin word *scientia*, which means knowledge. Science is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the Universe. Modern science is a discovery as well as an invention. It is a discovery that Nature generally acts regularly enough to be described by laws and even by mathematics; and it required invention to devise the techniques, abstractions, apparatus, and organization for exhibiting the regularities and securing their law-like descriptions [1, 2]. The institutional goal of science is to expand certified knowledge [3]. This happens by the important ability of science to produce and communicate scientific knowledge. We stress especially the communication of new knowledge, since communication is an essential social feature of scientific systems [4]. This social function of science has long been recognized [5–9].

Research is creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of humans, culture, and society, and the use of this stock of knowledge to devise new applications [10]. Scientific research is one of the forms of research. Usually, modern science is connected to research organizations. In most cases, the dynamics of these organizations is nonlinear. This means that small influences may lead to large changes. Because of this, the evolution of such organizations must be managed very carefully and on the basis of sufficient knowledge on the laws that govern corresponding structures and processes. This sufficient knowledge may be obtained by study of research structures and processes. Two important goals of such studies are (i) adequate modeling of dynamics of corresponding structures and (ii) design of appropriate tools for evaluation of production of researchers.

This chapter contains the minimum amount of knowledge needed for a better understanding of indicators, indexes, and mathematical models discussed in the following chapters. We consider science as an open system and stress the dissipative nature of research systems. Dissipativity of research systems means that they need continuous support in the form of inflows of money, equipment, personnel, etc. The evolution of research systems is similar to that of other open and dissipative systems: it happens through a sequence of instabilities that lead to transitions to more (or less) organized states of corresponding systems.

Science may play an important role in a national economic system. This is shown on the basis of the Triple Helix model of a knowledge-based economy. Competition is an important feature of modern economics and society. Competition has many faces,

and one of them is scientific competition among nations. This kind of competition is connected to the academic diamond: in order to be successful in globalization, a nation has to possess an academic diamond and use it effectively.

In order to proceed to the methods for quantitative assessment of research and research organizations and to mathematical models of science dynamics, one needs some basic information about assessment of research. A minimum of such basic information is presented in the second part of the chapter. The discussion begins with remarks about quality and measurement of processes by process indicators. Measurement can be qualitative and quantitative, and four kinds of measurement scales are described. The discussion continues with remarks on the non-Gaussianity that occurs frequently as a feature of social processes. Research also has characteristics of a social process, and many components and processes connected to research possess non-Gaussian statistical characteristics.

If one wants to measure research, one needs quantitative tools for measurement. Scientometrics, bibliometrics, and informetrics provide such tools, and a brief discussion of quantities that may be measured and analyzed is presented further in the text. In addition, another useful tool for analysis of research and research structures, the knowledge landscape, is briefly discussed. Next, research production is discussed in more detail. Special attention is devoted to publications and citations, since they contain important information that is useful for assessment of research production. The discussion continues with remarks on methods and systems for assessment of research and research organizations. Tools for assessment of basic research as well as the method of expert evaluation and several systems for assessment of research organizations applied in countries from continental Europe are briefly mentioned. The discussion ends with a description of the English–Czerwon method for quantification of performance of research units, which makes it possible to combine qualitative and quantitative information in order to compare results of research of research groups or research organizations.

1.2 Science, Technology, and Society

Knowledge is our most powerful engine of production
Alfred Marshall

Science, innovation, and technology have led some countries to a state of developed societies and economies [11–16]. Thus science is a driving force of positive social evolution, and the neglect of this driving force may turn a state into a laggard [17]. Basic research is an important part of the driving force of science. This kind of research may have large economic consequences, since it produces scientific information that has certain characteristic features of goods [18] such as use value and value. The *use value* of scientific information is large if the obtained scientific information can be applied immediately in practice or for generation of new information. One indicator for the measure of this value is the number of references of the corre-

sponding scientific publication. The *value* of scientific information is large when it is original, general, coherent, valid, etc. The value of scientific information is evaluated usually in the "marketplace" such as scientific journals or scientific conferences.

The lag between basic research and its economic consequences may be long, but the economic impact of science is indisputable [19, 20]. This is an important reason to investigate the structures, laws, processes, and systems connected to research [21–26]. The goals of such studies are [27]: better management of the scientific substructure of society [28–30], increase of effectiveness of scientific research [31–34], efficient use of science for rapid and positive social evolution. The last goal is connected to the fact that science is the main factor in the increase of productivity. In addition, science is a sociocultural factor, for it directly influences the social structures and systems connected to education, culture, professional structure of society, social structure of society, distribution of free time, etc. The societal impact of science as well as many aspects of scientific research may be measured [35–43].

Science is an information-producing system [44, 45]. That information is contained in scientific products. The most important of these products are scientific publications, and the evaluation of results of scientific research is usually based on scientific publications and on their citations. Scientific information is very important for technology [46–48] and leads to the acceleration of technological progress [49–59]. Science produces knowledge about how the world works. Technology contains knowledge of some production techniques. There are knowledge flows directed from the area of science to the area of technology [60, 61]. In addition, technological advance leads to new scientific knowledge [62], and in the process of technological development, many new scientific problems may arise. New technologies lead also to better scientific equipment. This allows research in new scientific fields, e.g., the world of biological microstructures. Advances in science may reduce the cost of technology [63–66]. In addition, advances in science lead to new cutting-edge technologies, e.g., laser technologies, nanoelectronics, gene therapy, quantum computing, some energy technologies [67–74]. But the cutting-edge technologies do not remain cutting-edge for long. Usually, there are several countries that are the most advanced technologically (technology leaders), and the cutting-edge technologies are concentrated in those countries. And those countries generally possess the most advanced research systems.

In summary, what we observe today is a scientifically driven technological advance [75–81]. *And in the long run, technological progress is the major source of economic growth.*

The ability of science to speed up achievement of national economic and social objectives makes the understanding of the dynamics of science and the dynamics of research organizations an absolute necessity for decision-makers. Such an understanding can be based on appropriate systems of science and technology indicators and on tools for measurement of research performance [82–87]. Because of this, science and technology indicators are increasingly used (and misused) in public debates on science policy at all levels of government [88–96].

1.3 Remarks on Dissipativity and the Structure of Science Systems

The following point of view exists about the evolution of open systems in thermodynamics [97, 98]:

> *The evolution of an open thermodynamic system is a sequence of transitions between states with decreasing entropy (increasing level of organization) with an initial state sufficiently far from equilibrium. If the parameters of such systems change and the changes are large enough, the system becomes unstable, and there exists the possibility that some fluctuation of the parameters may push the system to a new state with smaller entropy. Thus the transition takes place through an instability.*

This type of development may be observed in scientific systems too. This is not a surprise, since scientific systems are open (they interact with a complex natural and social environment), and they are able to self-organize [99]. In addition, crises exist in these systems, and often these crises are solved by the growth of an appropriate fluctuation that pushes the scientific system to a new state (which can be more or less organized than the state before the crisis). Hence instabilities are important for the evolution of science, and it is extremely important to study the instabilities of scientific (and social) systems [100–102]. The time of instability (crisis) is a critical time, and the regime of instability is a critical regime. The exit from this time and this regime may lead to a new, more organized, and more efficient state of the system or may lead to degradation and even to destruction of the system.

1.3.1 Financial, Material, and Human Resource Flows Keep Science in an Organized State

Dissipative structures: *In order to keep a system far from equilibrium, flows of energy, matter, and information have to be directed toward the system. These flows ensure the possibility for self-organization, i.e., the sequence of transitions toward states of smaller entropy (and larger organization). The corresponding structures are called dissipative structures, and they can exist only if they interact intensively with the environment. If this interaction stops and the above-mentioned flows cease to exist, then the dissipative structures cannot exist, and the system will end at a state of thermodynamic equilibrium where the entropy is at a maximum and organization is at a minimum.*

Science structures are dissipative. In order to exist, they need inflows of information (since scientific information becomes outdated relatively fast), people (since the

scientists retire or leave and have to be replaced), money (needed for paying scientists, for building and supporting the scientific infrastructure), materials (for running experiments, machines, etc.), etc. The weak point of the dissipative structures is that they can be degraded or even destroyed by decreasing their supporting flows [103]. In science, this type of development to retrograde states may be observed when the flows of financial and material support decrease and flows of information decrease or cease.

1.3.2 Levels, Characteristic Features, and Evolution of Scientific Structures

Researchers act in two directions: (i) they produce new knowledge and information [104, 105] and decrease the disorder as current knowledge become better organized; (ii) the work of researchers leads to new problems and the possibility for new research directions and thus opens the way to new states with an even higher level of organization. By means of these actions, researchers influence the structure of science. There exist three levels and four characteristic features of the scientific structure [106]. The three levels are:

1. *Level of material structure*: Here are the scientific institutes, material conditions for scientific work, etc.
2. *Level of social structure*: This includes the scientists and other personnel as well as the different kinds of social networks connected to scientific organizations.
3. *Level of intellectual structure*: This includes the structures connected to scientific knowledge and the field of scientific research. There are differences in the intellectual structures connected to the social sciences in comparison to the intellectual structures connected to the natural sciences.

The four characteristic features of the scientific structure are:

1. *Dependence on material, financial, and information flows*. These flows are directed mainly to the material levels of the scientific structure. They include the flows of money and materials that are needed for the scientific work. But there are also flows to other levels of the scientific structure. An important type of such flows is motivation flows. For example, there exist (i) *psychological motivation flow*: connected to the social level of the scientific structure. This motivation flow is needed to support each scientist to be an active member of scientific networks and to be an expert in the area of his or her scientific work; (ii) *intellectual motivation flow*: connected to the intellectual level of the scientific structure. This flow supports scientists to learn constantly and to absorb the newest scientific information from their research area.
2. *Cyclical behavior of scientific productivity*. At the beginning of research in a new scientific area, there are many problems to be solved, and scientists deal with them (highly motivated, for example, by the intellectual motivation flow

and possibly by material flows that the corresponding wise national government assigns to support the research in this area). In the course of time, the simple scientific problems are solved, and what remains are more complex unsolved problems. The corresponding scientific production (the number of publications, for example) usually decreases. Some scientists change their field of research, and then a new scientific area or subarea may arise in this new field of research.

3. *Homeostatic feature*.

> *Homeostasis is the property of a system to regulate its variables in such a way that internal conditions remain stable and relatively constant.*

This feature of science is supported by the system of education, the set of traditions and institutional norms, the books and other material and intellectual tools that ensure the translation of knowledge from one generation of scientists to the next, etc. All this contributes to the stable functioning of scientific systems and helps them to overcome unfavorable environmental conditions.

4. *Limiting factors*. Limiting factors can be (i) material factors that decrease the intensity of work of the scientific organizations (such as decreased funding, for example); (ii) factors connected to decreasing the speed of the process of exchange of scientific information (closing access to an important electronic scientific journal, for example); (iii) factors that decrease the speed of obtaining new scientific results (for example, the constant pressure to increase the paperwork of scientists).

Scientific structures evolve. This evolution is connected to the evolution of scientific research [107–109]. Usually, the evolution of scientific structures has four stages: normal stage, network stage, cluster stage, specialty stage. Institutional forms of research evolve, for example, as follows. At the normal stage, these forms are informal; then small symposiums arise at the network stage. At the cluster stage, the symposiums evolve to formal meetings and congresses, and at the specialty stage, one observes institutionalization (research groups and departments at research institutes and universities). Cognitive content evolves too. At the normal stage, a paradigm is formulated. At the network stage, this paradigm is applied, and in the cluster stage, deviations from the paradigm (anomalies) are discovered. Then at the specialty stage, one observes exhaustion of the paradigm, and the cycle begins again by formulation of a new paradigm.

Now let us consider a more global point of view on research systems and structures and let us discuss briefly two additional aspects connected to these systems:

- The place of research in the economic subsystem of society from the point of view of the Triple Helix model of the knowledge-based economy;
- Relations among different national research systems: we discuss the competition among these systems from the point of view of the concept of the academic diamond.

1.4 Triple Helix Model of the Knowledge-Based Economy

Research priorities should be selected by taking into account primarily the requirements of the national economics and society, traditions and results previously attained, possible present and future human and financial potential, international relationships, trends in the world's economic and social growth, and trends of science.
Peter Vinkler

The Triple Helix model of the knowledge-based economy defines the main institutions in this economy as university (academia), industry, and government [110–119]. The Triple Helix has the following basic features:

1. A more prominent role for the university (and research institutes) in innovation, where the other main actors are industry and government.
2. Movement toward collaborative relationships among the three major institutional spheres, in which innovation policy should be increasingly an outcome of interaction rather than a prescription from government.
3. Any of the three spheres may take the role of the other, thus performing new roles in addition to their traditional function. This taking of nontraditional roles is viewed as a major source of innovation.

Organized knowledge production adds a new coordination mechanism in social systems (knowledge production and control) in addition to the two classical coordination mechanisms (economic exchanges and political control). In the Triple Helix model, the economic system, the political system, and the academic system are considered relatively autonomous subsystems of society that operate with different mechanisms. In addition to their autonomy, however these subsystems are interconnected and interdependent. There are amendments in the model of the Triple Helix, and even models of the helix exist with more than three branches [120].

The Triple Helix model allows for the evolution of the branches of the helix. At the beginning of operation of the Triple Helix:

1. Industry operates as a concentration point of production.
2. Government operates as the source of contractual relations and has to be a guarantor for stable interactions and exchange.
3. The academy operates as a source of new knowledge and technology, thus generating the base for establishing a knowledge-based economy.

With increasing time, the place of academia (universities and research institutes) in the helix changes. Initially, the academy is a source of human resources and knowledge, and the connection between academia and industry is relatively weak. Then academia develops organizational capabilities to transfer technologies, and instead of serving only as a source of new ideas for existing firms, academia becomes a source of new firm formation in the area of cutting-edge technologies and in advanced areas of science. Academia becomes a source of regional economic development, and this leads to the establishment of new mechanisms of economic activity and

community formation (such as business incubators, science parks, and different kinds of networks between academia and industry). Government supports all this by its traditional regulatory role in setting the rules of the game and also by actions as a public entrepreneur.

The Triple Helix model is a useful model that helps researchers, managers, et al. to imagine the place of research structures in the complex structure of modern economics and society. Let us mention that the Triple Helix can be modeled on the basis of the evolutionary "lock-in" model of innovations [121] connected to the efforts of adoption of competing technologies [122, 123]. And various concepts from time series analysis such as the concept of mutual information [119] can be used to study the Triple Helix dynamics.

1.5 Scientific Competition Among Nations: The Academic Diamond

It is not enough to do your best. You must know what to do and then do your best
W. Edwards Deming

Globalization creates markets of huge size, and every nation wants to be well represented at these markets with respect to exports of goods, etc. This can happen if a nation has competitive advantages. One important such advantage is the existence of effective national research and development (R & D) systems. Let us note that the scientific production by researchers, research groups, and countries is an object of absolute competition regardless of possible poor equipment, low salaries, or lack of grants for some of the participants in this competition. From this point of view, the evaluation of scientific results may be regarded as unfair if one compares scientists from different nations [4]. Poor working conditions for scientists is clearly a competitive disadvantage to the corresponding nation. In order to export high-tech production, the scientific and technological system of a nation has to work smoothly and be effective enough. A nation that has such a system and uses it effectively for cooperation [124, 125] and competition has a competitive advantage in the global markets. And in order to have such a system, a country should invest wisely in the development of its scientific system and in the processes of strengthening the connection between the national scientific, technological, and business systems and structures [126–130]. In particular, the four parts of the so-called academic diamond [131] should be cultivated.

Each of the four parts of the academic diamond is connected to the other three parts. The parts are:

1. Factor conditions: *human resources* (quantity of researchers, skills levels [132], etc.), *knowledge resources* (government research institutes, universities, private research facilities, scientific literature, etc.), *physical and basic resources* (land, water and mineral resources, climatic conditions, location of the country, proxim-

ity to other countries with similar research profiles, size of country, etc.), *capital resources* (government funding of scientific structures and systems, cost of capital available to finance academia, private funding for research projects, etc.), *infrastructure* (quality of life, attractiveness of country for skilled scientists, telecommunication systems, etc.).
2. Strategy, structure, and rivalry: *goals and strategies of the research organizations* (research profile, positioning and key faculties or research areas, internationalization path in terms of staff, campuses, and student body, etc.), *local rules and incentives* (salaries, promotion system, incentives for publication, etc.), *local competition* (number of research universities, research institutes, research centers, existing research clusters, territorial dynamics of scientific organizations, etc.).
3. Demand conditions: *public and private sectors* (demand for training and job positions for researchers, etc.), *student population* (trained students), *other academics in country and abroad* (active research scientists outside the government research institutes and universities).
4. Related and supporting industries: *publication industry*, *information technology industry*, *other research institutions*.

In addition, the academic diamond has two more components: *chance* and *government*. There are different aspects of chance connected to the research organizations. If we consider chance as the *possibility for something to happen*, then some countries have elites that ensure a good chance with respect to the positive development of science and technology. Government may contribute to the development of scientific and technological systems of a country. This contribution can be made through appropriate politics with respect to (higher) education; government research institutes; basic research [133, 134]; funding of research and development; economic development; etc.

1.6 Assessment of Research: The Role of Research Publications

Research is an important process in complex scientific systems. Research production is a result of this process that can be assessed. Quantitative assessment of research (at least of publicly funded basic research) has increased greatly in the last decade [135–138]. Some important reasons for this are economic and societal [134]: constraints on public expenditures, including the field of research and development; growing costs of instrumentation and infrastructure; requirements for greater public accountability; etc. Another reason is connected to the development of information technologies, bibliometrics, and scientometrics in the last fifty years. Several goals of quantitative assessment of research are [4] to obtain information for granting research projects; to determine the quantity and impact of information production for monitoring research activities; to analyze national or international standing of research organizations and countries' organizations for scientific policy; to obtain information for personnel decisions; etc.

In addition to the rise of quantitative assessment of research, one observes a process of the increasing use of mathematics in different areas of knowledge [139]. This process also concerns the field of knowledge about science. In the process of human evolution, more and more scientific facts have been accumulated, and these facts have been ordered by means of different methods that include also methods of mathematics. In addition, the use of mathematics (which means also the use of mathematical methods beyond the simplest statistical methods) is important and much needed for supporting decisions in the area of research politics.

Many mathematical methods in the area of assessment of research focus on the study of research publications and their citations. This is because publications are an important form of the final results of research work [140–142]. There is a positive correlation between the number of research publications and the meaning that society attaches to the scientific achievements of the corresponding researcher. There exists also a positive correlation between the number of a researcher's publications and the expert evaluation of his/her scientific work [143]. Senter [144] mentions five factors that may positively influence the research productivity of a researcher:

1. Education level: has important positive impact on productivity;
2. Rank of the scientist: has immediate positive impact on scientific productivity;
3. Years in service: positive impact on productivity but more modest in comparison to the impact of education and rank;
4. Influence of scientist on its research endeavor: positive impact but modest in comparison with the above three factors;
5. Psychological factors: usually they have small effect on productivity (if the problems that influence the psychological condition of the research are not too big).

In recent years, the requirements on the quality of research have increased. Because of this, we shall discuss briefly below several characteristics of quality, performance, quality management systems, and performance management systems, since they are important for the assessment of the quality of the results of basic and applied research [145–148].

1.7 Quality and Performance: Processes and Process Indicators

Scientific research and its product, scientific information, is multidimensional, and because of this, the evaluation of scientific research must also be multidimensional and based on quantitative indexes and indicators accompanied by qualitative tools of analysis. One important characteristic of research activity is its quality, because the performance of any organization is connected to the quality of its products [149–153]. A simple definition of quality is this: *Quality is the ability to fulfill a set of requirements with concrete and measurable actions.* The set of requirements can include social requirements, economic requirements, productive requirements, and specific scientific requirements. The set of requirements depends on the stakeholders' needs

and on the needs of producers. These needs should be fulfilled effectively, and an important tool for achieving this is a quality management system. In order to manage quality, one introduces different quality management systems (QMS), which are sets of tools for guiding and controlling an organization with respect to quality aspects of human resources; working procedures, methodologies and practices; technology and know-how.

Research production is organized as a set of processes. A simple definition of a process is as follows: *A process is an integrated system of activities that uses resources to transform inputs into outputs* [149]. We can observe and assess processes by means of appropriate indicators. *An indicator is the quantitative and/or qualitative information on an examined phenomenon (or process or result), which makes it possible to analyze its evolution and to check whether (quality) targets are met, driving actions and decisions* [154]. Let us note that we do not need simply to use some indicators. We have to identify the indicators that properly reflect the observed process. These indicators are called *key performance indicators*.

The main functions of indicators are as follows.

1. **Communication**. Indicators communicate performance to the internal leadership of the organization and to external stakeholders.
2. **Control**. Indicators help the leadership of an organization to evaluate and control performance of the corresponding resources.
3. **Improvement**. Indicators show ways for improvement by identifying gaps between performance and expectations.

Indicators supply us with information about the state, development, and performance of research organizations. Performance measurements are important for taking decisions about development of research organizations [155]. In general, performance measurements supply information about meeting the goals of an organization and about the state of the processes in the organization (for example, whether the processes are in control or there are some problems in their functioning). In more detail, the performance measurement supplies information about the *effectiveness of the processes*: the degree to which the process output conforms to the requirements, and about *efficiency of the processes*: the degree to which the process produces the required output at minimal resource cost. Finally, the performance measurements supply information about the need for process improvement.

1.8 Latent Variables, Measurement Scales, and Kinds of Measurements

Latent features of the studied objects and subjects often are the features we want to measure. One such feature is the scientific productivity of a researcher [156, 157]. Latent features are characterized by latent variables. Latent variables may reflect real characteristics of the studied objects or subjects, but a latent variable is not directly measurable. The indicators are what we measure in practice, e.g., the number of

publications or the number of citations. *Many latent variables can be operationally defined by sets of indicators. In the simplest case, a latent variable is represented by a single indicator. For example, the production of a researcher may be represented by the number of his/her publications. If we want a more complete characterization of the latent variables, we may have to use more than one indicator for their representation, e.g., one has to avoid (if possible) the reduction of representation of a latent variable to a single indicator. Instead of this, a set of at least two indicators should be used.*

A measurement means that certain items are compared with respect to some of their features. There are four scales of measurement:

1. **Nominal scale**: Differentiates between items or subjects based only on their names or other qualitative classifications they belong to. Examples are language, gender, nationality, ethnicity, form. A quantity connected to the nominal scale is *mode*: this is the most common item, and it is considered a measure of central tendency.
2. **Ordinal scale**: Here not only are the items and subject distinguished, but also they are ordered (ranked) with respect to the measured feature. Two notions connected to this scale are *mode* and *median*: this is the middle-ranked item or subject. The median is an additional measure of central tendency.
3. **Interval scale**: For this scale, distinguishing and ranking are available too. In addition, a degree of difference between items is introduced by assigning a number to the measured feature. This number has a precision within some interval. An example for such a scale is the Celsius temperature scale. The quantities connected with the interval scale are *mode*, *median*, *arithmetic mean*, *range*: the difference between the largest and smallest values in the set of measured data. Range is a measure of dispersion. An additional quantity connected to this kind of scale is *standard deviation*: a measure of the dispersion from the (arithmetic) mean.
4. **Ratio scale**: Here in addition to distinguishing, ordering, and assigning a number (with some precision) to the measured feature, there is also estimation of the ratio between the magnitude of a continuous quantity and a unit magnitude of the same kind. An Example of ratio scale measurement is the measurement of mass. If a body's mass is 10 kg and the mass of another body is 20 kg, one can say that the second body is twice as heavy. If the temperature of a body is 20 °C and the temperature of another body is 40 °C, one cannot say that the second body is twice as warm (because the measure of the temperature in degrees Celsius is a measurement by interval scale and not by ratio scale. The measure of temperature by a ratio scale is the measure in kelvins.

   In addition to all quantities connected to the interval scale of measurement, for the ratio scale of measurement one has the following quantities: *geometric mean*, *harmonic mean*, *coefficient of variation*, etc.

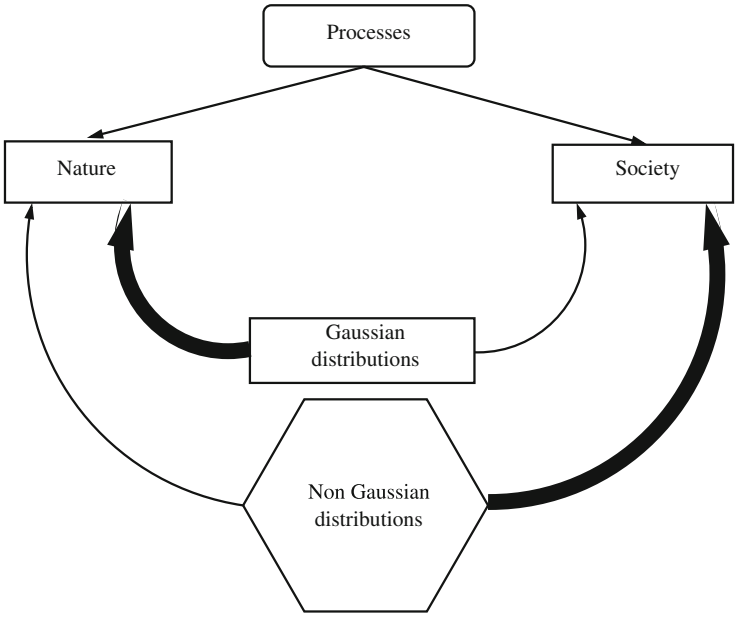

Fig. 1.1 Gaussian distributions are much used for description of natural systems and structures. Many distributions used for describing social systems and structures are non-Gaussian

With respect to the four scales, there are the following two kinds of measurements:

1. **Qualitative measurements**: measurements on the basis of nominal or ordinal scales.
2. **Quantitative measurements**: measurements on the basis of interval or ratio scales.

Before the start of a measurement, a researcher has to perform:

1. qualitative analysis of the measured class of items or subjects in order to select features that are appropriate for measurement from the point of view of the solved problems;
2. choice of the methodology of measurement.

After the measurements are made, it is again time for qualitative analysis of the adequacy of the results to the goals of the study: some measurement can be adequate for one problem, and other measurements can be adequate for another problem. The adequacy depends on the choice of the features that will be measured.

1.9 Notes on Differences in Statistical Characteristics of Processes in Nature and Society

Let us assume that measurements have led us to some data about a research organization of interest. Research systems are also social systems, and because of this, we have to know some specific features of these systems and especially the characteristics connected to the possible non-Gaussianity of the system.

A large number of processes in nature and society are random. These processes have to be described by random variables. If x is a random variable, it is characterized by a probability distribution that gives the probability of each value associated with the random variable x arising. Probability distributions are characterized by a probability distribution function $P(x \leq X)$ or probability density function $p(x) = dP/dx$.

If we want to study the statistical characteristics of some population of items, we study statistical characteristics of samples of the population. *We have to be sure that if the sample size is large enough, then the results will be close to the results that would be obtained by studying the entire population.*

For the case of a normal (Gaussian) distribution, the central limit theorem guarantees this convergence. For the case of non-Gaussian distributions, however, there is no such guarantee.

Let us discuss this in detail. We begin with the central limit theorem. The central limit theorem of mathematical statistics is the cornerstone of the part of the world described by Gaussian distributions. It is connected to the moments of a probability distribution $p(x)$ with respect to some value X:

$$M^{(n)} = \int dx\ (x - X)^n p(x). \tag{1.1}$$

The following two moments are of interest for us here:

1. The first moment ($n = 1$) with respect to $X = 0$: this is the mean value $\overline{x}$ of the random variable;
2. The second moment ($n = 2$) with respect to the mean ($X = \overline{x}$): dispersion of the random variable (denoted also by σ^2).

The central limit theorem answers the following question. We have a population of items or subjects characterized by the random variable x. We construct samples from this population and calculate the mean $\overline{x}$. If we take a large enough number of samples, then what will be the distribution of the mean values of those samples?

The central limit theorem states that if for the probability density function $p(x)$, the **finite mean and dispersion** exist, then the distribution of the mean values converges to the Gaussian distribution as the number of samples increases. The distributions that have this property are called **Gaussian**.

But what will be the situation if a distribution does not have the Gaussian property (for example, the second moment of this distribution is infinite)? Such distributions exist [158–160]. *They are called non-Gaussian distributions, and some of them play an important role in mathematical models of social systems, and in particular in the models connected to science dynamics. There exists a theorem (called the Gnedenko–Doeblin theorem) that states the central role of one distribution in the world of non-Gaussian distributions.* **This distribution is called the Zipf distribution**. Non-Gaussian distributions (and the Zipf distribution) will be discussed in Part III of this book.

Most distributions that arise in the natural sciences are Gaussian. Many distributions that arise in the social sciences are non-Gaussian (Fig. 1.1). Such distributions arise very often in the models of science dynamics [161, 162]. *We do not claim that only Gaussian distributions are observed in the natural sciences and that the distributions that are observed in the social sciences are all non-Gaussian. Non-Gaussian distributions arise frequently in the natural sciences, and Gaussian distributions exist also in the social sciences. The point is that the dominant number of continuous distributions observed in the natural sciences are Gaussian, and many distributions observed in the social sciences are non-Gaussian* [163].

Many distributions in the social sciences are non-Gaussian. Several important consequences of this are as follows.

1. **Heavy tails**. The tails of non-Gaussian distributions are larger than the tails of Gaussian distributions. Thus the probability of extreme events becomes larger, and the moments of the distribution may depend considerably on the size of the sample. Then the conventional statistics based on the Gaussian distributions may be not applicable.
2. The limit distribution of the sample means for large values of the mean is proportional (up to a slowly varying term) to the Zipf distribution (and not to the Gaussian distribution). **This is the statement of the Gnedenko–Doeblin theorem**.
3. In many natural systems, the distribution of the values of some quantity is sharply concentrated around its mean value. Thus one can perform the transition from a probabilistic description to a deterministic description. This is not the case for non-Gaussian distributions. There is no such concentration around the mean, and because of this, *a probabilistic description is appropriate for all problems of the social sciences in which non-Gaussian distributions appear.*

There exist differences between the objects and processes studied in the natural and social sciences. Several of these differences are as follows.

1. **The number of factors**. The objects and processes studied in the social sciences usually depend on many more factors than the objects and processes studied in the natural sciences. Let us connect this to the non-Gaussian distributions in the social sciences [164]. Let y be a variable that characterizes the influences on the studied object. Let $n(y)dy$ be the number of influences in the interval $(y, y + dy)$. Then $n(y)$ is the distribution of the influences. In order to define (a discrete) factor, we separate the area of values of y into subareas each of width Δy. Then if the area of values of y has length L, the number of factors will be $L/\Delta y$. Thus $n(y)$ has now the meaning of a distribution of factors. This distribution is Gaussian in most cases in the natural sciences and non-Gaussian in many cases of the social sciences. As we have mentioned above, the non-Gaussian distributions are not very concentrated around the mean value as compared to the Gaussian distributions. In other words, many more factors have to be taken into account when one analyzes items or subjects that are described by non-Gaussian distributions. *Thus the analysis of many kinds of social objects or processes must be a multifactor analysis.*
2. **Dominance of parameters**. In the case of systems from the natural sciences, usually there are several dominant latent parameters. *In the case of social systems, usually there is no dominant latent parameter. The links among parameters are weak, and in addition, many latent parameters can be important.*
3. **Subjectivity of the results of measurements**. The measurements in the study of social problems must be made very carefully. The main reasons for this are as follows: the measured system often cannot be reproduced; the researchers can easily influence the measurement process; the measurement can be very complicated.
4. **Mathematics should be applied with care**. The quantities that obey the laws of arithmetic are additive. There are two kinds of measurement scales that are used in the social sciences, and only one of them leads to additive quantities in most cases (i.e., to quantities that can be successfully studied by mathematical methods): closed measurement scales and open measurement scales. The *closed measurement scales* have a maximum upper value. Such a scale is, for example, the scale of school-children's grades. **Closed scales may lead to nonadditive quantities.** The *open measurement scales* do not have a maximum upper value. **Open scales lead in most cases to additive quantities**. The measurement scales in the natural sciences are mostly open scales. Thus mathematical methods are generally applicable there. Open scales must be used also in the social sciences if one wants to apply mathematical methods of analysis successfully. The application of mathematical methods (developed for analysis of additive quantities) to nonadditive quantities may be useless. One can also use closed measurement scales, of course. The results of these measurements, however, have to be analyzed mostly qualitatively.

1.10 Several Notes on Scientometrics, Bibliometrics, Webometrics, and Informetrics

The term scientometrics was introduced in [44]. Scientometrics was defined in [44] as *the application of those quantitative methods which are dealing with the analysis of science viewed as an information process*. Thus fifty years ago, scientometrics was restricted to the measurement of science communication. Today, the area of research of scientometrics has increased. This can be seen from a more recent definition of scientometrics:

> *Scientometrics is the study of science, technology, and innovation from a quantitative perspective* [165–170].

In several more words, by means of scientometrics one analyzes the quantitative aspects of the generation, propagation, and utilization of scientific information in order to contribute to a better understanding of the mechanism of scientific research activities [171]. The research fields of scientometrics include, for example, production of indicators for support of policy and management of research structures and systems [172–177]; measurement of impact of sets of articles, journals, and institutes as well as understanding scientific citations [178–189]; mapping scientific fields [190–192]. Scientometrics is closely connected to bibliometrics [193–201] and webometrics [202–210]. The term bibliometrics was introduced in 1969 (in the same year as the definition of scientometrics in [44]) as *application of mathematical and statistical methods to books and other media of communication* [211]. Thus fifty years ago, bibliometrics was used to study general information processes, whereas (as noted above) scientometrics was restricted to the measurement of scientific communication. Bibliometrics has received much attention [212–215], e.g., in the area of evaluation of research programs [216] and in the area of analysis of industrial research performance [217]. Today, the border between scientometrics and bibliometrics has almost vanished, and the the terms scientometrics and bibliometrics are used almost synonymously [218]. The rapid development of information technologies and global computer networks has led to the birth of webometrics. Webometrics is defined as *the study of the quantitative aspects of the construction and use of information resources, structures, and technologies on the Web, drawing on bibliometric and informetric approaches* [209, 210]. Informetrics is a term for *a more general subfield of information science dealing with mathematical and statistical analysis of communication processes in science* [219, 220]. Informetrics may be considered an extension of bibliometrics, since informetrics deals also with electronic media and because of this, includes, e.g., the statistical analysis of text and hypertext systems, models for production of information, information measures in electronic libraries, and processes and quantitative aspects of information retrieval [221, 222].

Many researchers have made significant contributions to scientometrics, bibliometrics, and informetrics. We shall mention several names in the following chapters.

Let us mention here the name of Eugene Garfield, who started the *Science Citation Index* (SCI) in 1964 at the Institute for Scientific Information in the USA. SCI was important for the development of bibliometrics and scientometrics and was a response to the information crisis in the sciences after World War II (when the quantity of research results increased rapidly, and problems occurred for scientists to play their main social role, i.e., to produce new knowledge). SCI used experience from earlier databases (such as Shepard's citations [223, 224]). In 1956, Garfield founded the company Eugene Garfield Associates and began publication of *Current Contents*, a weekly containing bibliographic information from the area of pharmaceutics and biomedicine (the number of covered areas increased very rapidly). In 1960, Garfield changed the name of the company to *Institute of Scientific Information*. Let us note that the success of the Current Contents was connected to the use of Bradford's law for "scattering" of research publications around research journals (Bradford's law will be discussed in Chap. 4 of the book) [225]. According to the Bradford's law, the set of publications from some research area can be roughly separated into three subsets: a small subset of core journals, a larger subset of journals connected to the research area, and a large set of journals in which papers from the research area could occur. Bradford's law was used in the selection of journals contributing to the multidisciplinary index SCI. In the following years, the SCI and ISI became the world leaders in the area of scientific information. This position remained unchallenged for almost fifty years, even after the rise of the Internet.

Below we consider three topics from the area of scientometrics that are of interest for our discussion. These topics are:

1. Quantities that may be analyzed in the process of study of research dynamics;
2. Inequality of scientific achievements;
3. Knowledge landscapes.

1.10.1 Examples of Quantities that May Be Analyzed in the Process of the Study of Research Dynamics

Below we present a short list of some quantities, kinds of time series, and other units of data that may be used in the process of assessment of research and research organizations. The list is as follows.

1. Time series for the number of published papers in groups of journals (for example in national journals).
2. Time series for the total number and for the percentage of coauthored papers [226]. Coauthorship is an important phenomenon, since the development of modern science is connected to a steady increase in the number of coauthors, especially in the experimental branches of science. Coauthorship contributes to the increase of the length of an author's publication list, and this length is important for the quality of research [227], for a scientific career, and for the process of approval of research projects.

The percentage of coauthored publications varies in the different sciences. In the social sciences, it is very low, and in the natural sciences it can reach 90 % and even more. There are interesting notes of Price and Rousseau with respect to coauthorship [228, 229]. Price notes that important factors for the growth of coauthorship of publications are (i) the expansion of the material base of scientific research, e.g., new equipment stimulates coauthorship; (ii) in times of expansion, the number of very good scientists increases at a slower rate than the number of scientists. In such conditions, the most productive authors increase their productivity further by becoming leaders of scientific collectives. In these collectives, scientists can be found who want to have publications but are unable to publish alone (because they are inexperienced PhD students, for example). Let us note here that in recent years, one observes frequently the phenomenon of hyperauthorship (a very large number of coauthors of a publication) [230].

3. Network analysis of coauthorship groups [231–240] and especially detection of dense and very productive coauthorship networks: "invisible colleges" [241–246]. An invisible college has a core and periphery. The core usually consists of researchers from the same research structure or from a few research structures, e.g., from the same research institute or from several universities where productive groups exist.
4. Cluster analysis of research publications [247, 248].
5. Time series for the number of patents and discoveries. What can be expected in times of fast growth of the number of scientific discoveries is that their period of doubling is about ten years [249].
6. Distribution of publications among research organizations [250].
7. Distribution of patents and discoveries among the countries from a group of countries (for example, EU countries or the entire world).
8. Statics and dynamics of landscapes of scientific discoveries and engineering patents for different scientific or engineering fields.
9. Time series for the number of scientists (in a country). When a country's research structures grow, one may expect doubling of the number of researchers every fifteen years. When the scientific structure becomes mature, the growth slows and may come to a halt.
10. Territorial distribution of scientists—national and international [251, 252]. Distribution of scientists with respect to their qualifications.
11. Dynamics of the age structure of scientists at the national level and comparison of the dynamics among countries from a group of countries.

Other kinds of quantities are connected to another important characteristic of research work: the citations of research publications [253–260]. One may analyze:

1. Time series for citations of individual scientists, scientific groups, or scientific organizations [261–268]. We note that the number of citations depends on the number of researchers who work in the corresponding scientific area [267], and there can be also negative citations of the publications of a researcher. Citation analysis allows us to identify different categories of researchers such as identity-creators and image-makers [269]. The number of citations depends on the rate

of aging of research information [270]. This rate of aging may be different for different scientific disciplines.
2. Distribution of journals with respect to the citations of the papers in these journals (the impact factor is one possible indicator that can be constructed on the basis of such studies [271]).
3. Distribution of scientific organizations with respect to the citations of the publications of the organization. One has to be very careful here, since in some areas of science there are many more citations than in other areas of science.
4. Citation networks [272–276]. Usually there are subnetworks of leading scientists in some scientific areas, and every leading scientist cites predominantly the other leading scientists. The nonleading scientists cite the leading scientists much more than other nonleading scientists.
5. Distribution of scientists with respect to the number of citations of their publications. Here different possibilities exist, e.g., the study of the distribution of citations of the most cited papers of scientists from a scientific group or scientific organization or the study of the distribution of the number of citations of the papers that contribute to the h-factors or g-indexes of the researchers from the assessed research groups or research organizations.
6. Distribution of publications of a scientific group or scientific organization with respect to the number of citations they have.
7. Distribution of citations among scientific fields [277, 278].
8. Distribution of the time interval between appearance of a publication and its first citation.
9. Landscapes of citations [279, 280] with respect to scientific discipline; countries; kind of publications; research organizations in a country, etc.
10. Distributions of self-citations for scientific disciplines and in research groups and research organizations.

In addition, one may analyze other characteristics of science dynamics such as interdisciplinarity of scientific journals on the basis of the betweenness centrality measure used in social networks analysis [281]; aging of scientific literature [282, 283]; dynamics of scientific communication [284], etc.

1.10.2 Inequality of Scientific Achievements

Different researchers have different scientific achievements. Many factors influence the achievement of individual researchers or group of researchers. If we consider individual researchers, four main factors may be considered [218]: the subject matter; the author's age; the author's social status; the observation period. Experienced researchers usually have larger scientific production and larger scientific achievements in comparison to the newcomers without research experience. Chemists usually have larger research production than mathematicians. An established professor

with high social status usually has a much larger network of collaborations and contributes to more scientific achievements in comparison to a young researcher without such social status.

One of the tasks of the assessment of research organizations is the measurement of the inequality of scientific achievements of researchers and research collectives. One may use the notion of Coulter [285] that the distribution of some characteristics of research productivity is the division of the units of these characteristics among the components of the corresponding structure of the research organization. Inequality then may be defined as variation of the above division. Observation of large research groups shows that researchers are distributed usually within three classes with respect to a part of their production (the part that can be measured by the number of authored and coauthored papers in the international databases such as ISI Web of Science or SCOPUS): a small class of very productive scientists; a large class of very unproductive scientists; a large class of scientists who fall between the above two extreme classes.

In the study of inequality in research organizations, one may use not only quantitative methods but also qualitative methods such as expert evaluations [286–288], surveys, and content analysis. We shall discuss numerous indexes of inequality of scientific achievements in the following chapters and especially in Chap. 4. Now we shall focus our attention on the concept of knowledge landscape.

1.10.3 Knowledge Landscapes

An understanding of the evolution of research organizations requires research complementary to mathematical investigations [289–302]. Very useful tool for such research are knowledge maps [303–325] and knowledge landscapes [326–333]. They may be used for identification of potential collaborators [334]; for creation of document-level maps of research fields [335]; for international comparisons of research systems [336]; for modeling of science [337]; etc.

The concept of knowledge landscape is as follows: *Describe the corresponding field of science or technology through a function of parameters such as height, weight, size, technical data, etc. Then a virtual knowledge landscape can be constructed from empirical data in order to visualize and understand innovation and other processes in science and technology.*

One example of a technological knowledge landscape can be given by the function $E = E(S, v)$, where E are the expenses for developing a new car, S and v being the size and velocity of the car. The landscape is constructed as the values of S and v are plotted on the horizontal axes, and the values of E are plotted on the vertical axis.

Knowledge landscapes may be used for evaluation and tracking evolution of research structures and systems [338]. Two selected examples are:

1. **Application of knowledge landscapes for evaluating national research strategies**. The national research structures can be considered to be groups of researchers who compete for scientific results following optimal research strategies. The efforts of each research structure or the efforts of each country become visible, comparable, and measurable by means of appropriate landscapes connected, for example, to the number of publications. The aggregate research strategies of a country can thereby be represented by the distribution of publications in the various scientific disciplines. In so doing, within a two-dimensional space, i.e., axes being the scientific disciplines and number of publications, different countries occupy different locations. Various political discussions can follow, and evolution strategies invented thereafter. In addition, one can track scientific areas of strategic importance on the basis of journal mappings [339] or can construct global maps of science [340] that can be very useful as elements of national research decision support systems.
2. **Landscapes for research evaluation based on scientific citations**. Citations are important in the retrieval and evaluation of information in scientific communication systems [341–344]. This is based on the objective nature of the citations as components of a global expert evaluation system, as represented by the Science Citation Index. Thus the importance of the citation landscapes increases in the process of formation of a research policy. One example of this is personnel management decisions, which influence individual research careers or investment strategies.

1.11 Notes on Research Production and Research Productivity

Researchers produce numerous kinds of items as a result of their work: research publications, technical reports, patents, etc. This research production may be counted, and the corresponding numbers can be divided by units of time (month, year, etc.). The obtained numbers are characteristics of the researcher's research productivity. The values of the characteristics of the research productivity depend on the considered time interval. If the time interval for calculation of the characteristics of the research productivity is the same as the career length of the researcher, then the characteristics of the research productivity coincide with the characteristics of the research production of the researcher.

Research structures and their system of functioning are extremely important for research productivity [345–349]. Research organizations provide the material conditions for research work [350], but they also support research environment that may stimulate or may influence negatively the research work [351]. A part of this environment are the interactions among researchers. These interactions influence the external motivation for research work (external means that the source of motivation is rooted in the actions of other persons). In addition, the above interactions

influence the internal motivation for research work (internal means that the source of motivations is rooted in the ideas of the researcher). It is observed that the more productive researchers search for and establish more connections with colleagues. Thus the presence of at least one productive researcher in any of the small units of a research organization (laboratories, sections, or chairs) creates more stimuli for research work.

Research productivity is age-dependent [352–355]. In most cases, the research productivity of an individual researcher decreases as he/she ages and comes close to retirement age. In most cases, research productivity of a research organization decreases with increasing average age of the researchers in the organization (if the average age is large enough). There exist other factors that may affect research productivity [356]. Eleven of these factors [357] are persistence, resource adequacy (adequate funds for research, adequate equipment, etc.), access to literature, professional commitment, intelligence, initiative, creativity, learning capability, concern for advancement, stimulative leadership, and external orientation (adequate contacts with superior researchers and participation in seminars and conferences).

Research production is an important quantity [358] with a complex structure. There exist research collectives, and these research collectives produce publications (but not only publications). Resources are spent for organizing researchers into structures and for supporting a system of functioning for those structures. One result of the actions of researchers is the set of their research publications, which is an important quantitative measure of the scientific productivity. If additional analysis of the content of the publications is made, the set of publications can be used also as a qualitative measure of research productivity. The connection between the research structures and the set of publications is relatively complicated for two reasons: there are scientists who don't have publications, and some publications are produced by more than one author. It is very interesting [143] that s large number of researchers (about 50 %) do not have publications. This will be further discussed in Chap. 4. But the absence of publications does not mean that the corresponding researchers are lazy or incompetent. It may mean that their contribution to the publications is not large enough for their names to be included in the list of coauthors of the publication. *In order to evaluate the scientific production of such researchers, one may have to use units that are smaller than a research publication. Such a smaller unit may be, for example, the time spent for support of research work, the number of collaborations with researchers who are authors of publications, etc.* The current units used widely for measurement of scientific performance are (i) *unit for information*: research publication (scientific paper); (ii) *unit for impact of the unit for information*: a citation of the research publication. It seems that these units are too crude or at least they are too large to measure the research performance of large classes of researchers. Then for the low-productive (in terms of research publications) researchers, one may think about other units for measurement of their performance, and research papers and their citations may be used to measure performance of the top class of researchers. The research community is greatly interconnected, and it needs all kind of "workers": full "workers" who produce scientific publications and partial "workers" who support the full "workers" in the process of producing the scientific publications.

Statistical methods are much used for evaluation of research work [359]. One has to be careful in applying such methods, because the number of components of a scientific organization may be large, but it is not very large in comparison with the number of molecules in a glass of water, for example. In addition, in contrast to molecules (which are weakly connected), the elements of a research organization may be strongly connected, and these connections are an important part of their structure and system of working. Thus one has to be careful about the direct application of methods from thermodynamics (such as temperature or entropy) to scientific systems and organizations. Such methods may be useful, but they reflect only some of the properties of the research organizations, since a research organization is not only a statistical collection of elements each of which is in random motion. For example, in addition to the random processes in research organizations, there exist also deterministic processes that may lead to the rise of complex research structures.

This book is focused on research publications and their citations. Because of this, we shall discuss below in more detail the importance of research publications for assessment of research production. The importance of citations of research publications for assessment of research production will be demonstrated in Chap. 2, where we shall discuss numerous indexes for evaluation of research production constructed on the basis of citations of research publications.

Research production has the form of books, monographs, reports, theses, articles in journals, etc. For the purposes of scientometric studies, papers published in refereed scientific journals seems to be the most suitable unit (at least today and for some time in the future). Usually, scientific papers are not subdivided into smaller units (this may cause some problems, since as the scientific contribution of large groups of research personnel is not enough to put the names of the corresponding researchers as coauthors of research articles). Other elements used in the evaluation of research are, e.g., the number of citations of a paper, the number of references, the number of coauthors of the paper, etc. The number of publications of a researcher may be used as an indicator of latent characteristics of researchers and scientific organizations such as prestige of the researcher; prestige of the research organization of the researcher; contribution to science of researchers or research organization; productivity of researcher and research organization; eliteness of the researcher or research organization. The speed of increase of the number of research publications can be used as an indicator for the currency of a scientific research area; perspectives of a scientific research area; phase of development of a research group or organization (at the initial phase of development, there is a rapid increase in the number of publications. Then the speed of increase of publications falls, and finally one observes maturity of the research group or organization with an almost constant number of research publications per year).

A significant number of publications are published jointly, and the tendency for joint publications increases with the increasing complexity of scientific research. But how to count joint publications? There are many ways to do this. The three most popular of them are [360]:

1. **Normal count**: Every author of a publication has received full credit. For an example, if a publication has four authors, it counts as one publication for every author [266].
2. **Straight count**: Only the first author receives credit for the publication. This count discriminates against the second and subsequent authors.
3. **Adjusted count**: Every author of a publication receives an equal fraction of the total credit of one unit [361]. For example, if a publication has four authors, every author is given one-fourth credit. By this measure, the relative contribution of every author to the article is ignored.

Several additional formulas for counting joint publications are as follows (N is the number of coauthors in all relationships below and S_n is the score assigned to the nth author)

- HCM-count (Howard–Cole–Maxwell) [362]

$$S_n = \frac{(3/2)^{N-n}}{\sum_{n=1}^{N}(3/2)^{n-1}}. \tag{1.2}$$

- EKW-count (Ellwein–Khahab–Waldman) [363]

$$S_n = \frac{b^{n-1}}{\sum_{n=1}^{N} b^{n-1}}. \tag{1.3}$$

  In [363], $b = 0.8$.
- LV-count (Lukovits–Vinkler) [364]

$$S_1 = \frac{N+1}{2NF}; \quad S_n = \frac{n+T}{2nFT}, \tag{1.4}$$

  where

$$T = \frac{100}{A}; \quad F = \frac{1}{2}\left(\frac{1}{N} + \frac{N-1}{T} + \sum_{n=1}^{N}\frac{1}{n}\right)$$

  and A is the authorship threshold (percentage as the lowest share of contribution to a paper: five or ten percent of total credit).
- TG-count (Trueba–Guerrero) [365]

$$S_n = \frac{2(2N-n+2)}{3N(N+1)}(1-f) + C_n f, \tag{1.5}$$

where f is the share for crediting favored authors (usually favored authors are the first, the second, and the last author) $0 < f < 1$; C_n is the rate of favoring the nth author $\sum_{i=1}^{N} C_n = 1$.

There are additional systems for distribution of scores among coauthors [366, 367]. The problem of coauthorship will be discussed again in Chap. 2 in connection with the h-index.

The following additional indicators based on publications may be used for assessment of publication activity of single researchers, research groups, or research organizations:

1. Distribution of number of publications in research organizations or in research groups of a research organization.
2. Distribution of publications in research journals.
3. Distribution of researchers with respect of number of publications for the three kinds of counts mentioned above.
4. Distribution of publications with respect to gender of the authors.
5. Distribution of publication with respect to language.
6. Distribution of publications with respect to their kind (articles, papers in conference proceedings, book chapters, books, etc.).
7. Distribution of publications with respect of the kind of research work (theoretical publications, experimental publications, reviews, etc.).

Finally, let us note here an interesting effect occurring when funding is linked to publication counts [368]. In this case, the publication numbers may jump dramatically, but with the highest percentage increase in the lower-impact journals. And the jump is larger as a percentage in universities than in government research institutes, which could mean that there is an unused research potential in universities, whereas research institutes already produce many publications and can't increase their percentage by as much as is the case for universities.

1.12 Notes on the Methods of Research Assessment

1.12.1 Method of Expert Evaluation

Assessment of the production and productivity of researchers can be made by expert evaluation of the work of any researcher, e.g., for the last three or five years. The method of expert evaluation judges mainly the quality of the work of a researcher. One variant of realization of this method is to use a commission of five to seven experts. These experts evaluate a researcher with respect to two criteria: contribution of researchers to the corresponding area of science (external criterion) and importance and usefulness of researchers for his/her scientific organization (internal criterion). The commission of experts ranks researchers with respect to the above two criteria.

After the ranking, the ranks of the researchers can be converted to points according to some appropriate conversion schema [143]. Evaluated researchers are divided into groups, since the productivity of a researcher depends on the scientific environment (conditions of work and relationships among the scientists from the scientific group to which the scientists belong) and on its status in the scientific organization. There are five groups of researchers:

1. Researchers without doctoral degree who perform assistant work in the research or applied units of the scientific organization.
2. Researchers without doctoral degree who perform research work in the applied units of the research organization.
3. Researchers without doctoral degree who perform research work in the research units of the research organization.
4. Researchers with doctoral degree who perform research work in the applied units of the research organization.
5. Researchers with doctoral degree who perform research work in the research units of the research organization.

There can be subgroups of these groups. For example, the researchers with doctoral degrees who perform research work in the research units of the research organization can be assistant professors, associate professors, and full professors.

Expert evaluation may be a part of the complex evaluation of a researcher. Such a complex evaluation may contain, for example [369]:

1. Evaluation of production by number of publications, number of internal and external scientific reports, number of developed methods, discovered effects or other significant achievements of the researcher.
2. Further evaluation of production by total number of pages of publications, reports, manuscripts, etc.
3. Evaluation of importance of produced knowledge by number of citations. An important problem here is the scaling of citations from different years or different subject categories [370, 371].
4. Evaluation by prestige of the journals and publishing houses that published the articles, book chapters, or books of the scientist.
5. Evaluation on the basis of obtained national and international awards.
6. Expert evaluation of the significance of the achievements for solving real problems.
7. Expert evaluation on the influence of the scientist on the basis of the following categories: influence in the area of research, influence in the corresponding scientific discipline, influence in other scientific disciplines.
8. Evaluation of influence by the number of colleagues from the scientific organization who think that this researcher is a high-quality scientist.
9. Further evaluation of influence on the basis of number of publications that are judged as important by the colleagues, and number of times each of the important publications is pointed to as important by the colleagues of the researcher.

In addition to the above evaluation, there may also be an evaluation of the distribution of the time spent by researcher as follows [143]: time for the main scientific work (time for the research work of the scientist, time for scientific research of general interest that leads to the solution of large classes of problems, time for scientific research of special interest that leads to the solution of specific problems, time for transfer of scientific research for improving products or processes or for obtaining new kinds of products or processes); time for directing research work of other researchers; time for collaboration with other researchers; time for consultations and expert evaluations; time for pedagogical work; time for administrative work (time for internal administrative work within the scientific unit (laboratory, section or chair), time for communication with the higher levels of the administrative hierarchy, time for relations with other scientific groups and clients). The list of evaluations may be enlarged, e.g., by numerous indexes presented in Chap. 2.

1.12.2 Assessment of Basic Research

Assessment of basic research is a problem for research administrators, since there are no simple measures of the contribution to scientific knowledge made by researchers. Many partial indicators exist, and each of them accounts for one (or several) factors that influence the basic research and is influenced by other factors that are not connected to the basic research. Thus in order to obtain reliable results, one has to minimize the influence of the factors that are not connected to research [372]. This can be done on the basis of the concept according to which basic research is considered as a process with inputs, process body (scientific production), and outputs [373]. The elements of the process of basic research are

1. **Inputs**: stock of scientific knowledge and existing techniques; financial resources; institutional scientific resources (scientific instruments, skilled personal, etc.); recruited personnel (untrained students, etc.); environmental conditions (such as diverse natural influences).
2. **Scientific production**: conceptual, experimental and technical work of scientists; support work by engineers, technicians, etc.; dissemination of research results; education work for development and reproduction of scientific skills (training young scientists, etc.); administrative support work (including organizing adequate inputs).
3. **Outputs**: scientific contributions to the discipline, new techniques and new scientific knowledge; scientific contributions to other areas of science; educational contribution: trained students, PhD students, and scientists; economic contribution (such as engineers and workers with increased skills for industry, technological spin-offs, commercial benefits for equipment suppliers, etc.; cultural contributions.

In order to minimize the influence of factors not connected to basic research, Martin and Irvin [372] have proposed *the method of converging partial indicators* for assessment of basic research. This method is based on the following five principles:

1. Indicators are applied to research groups rather than to the individual scientists;
2. Citation-based indicators are seen as reflecting the impact (and not the quality or importance) of the research work [374, 375];
3. A range of indicators is used, each of which focuses on different aspects of the group performance;
4. As far as possible, indicators are applied to a matched group (comparing like with like principle)
5. As the indicators used have an imperfect or partial nature, only in those cases where they yield convergent results can it be assumed that the influence of peripheral factors has been kept small. In these cases, it can be assumed that the indicators provide a reliable estimate of the contribution of the different groups to scientific progress.

An algorithm for scientometric assessment that may be used also for assessment of basic research has been proposed by Moravcsic [376]. This algorithm includes

1. specifying the purpose of the assessment;
2. specifying the system to be assessed;
3. deciding on the level of assessment;
4. setting criteria;
5. selecting methods and specifying indicators for each criterion;
6. determining the links among the components within and outside the system (scientific political issues, type of the subject field and activity, etc.);
7. carrying out measurements;
8. interpreting the results obtained;
9. drawing conclusions of the assessment.

The evaluators should combine the scientometric assessment and peer reviews [377, 378] in order to obtain a complete picture of the evaluated research group or research organization [379–384].

Applications of indicators for assessment of basic research is related to certain problems. Several of these problems are as follows [372]: (i) the contribution of each publication to the scientific knowledge is different; (ii) publication rates are different for different research fields. Four important problems connected to indicators based on citation analysis are [385, 386]: (i) technical problems with databases such as authors with identical names, variation of names, incomplete coverage of journals; (ii) variation of citation rate of a paper during its lifetime; (iii) Presence of critical citations or halo-effect citations (Halo effect means that a researcher's overall impression about other researchers influences the observer's feelings and thoughts about the properties of their publications); (iv) self-citations and in-house citations. Finally, there are problems connected to peer evaluation, e.g.: (i) individuals evaluate scientific contributions on the basis of their cognitive and social status (which can be

quite different); (ii) perceived implication of results of one's own center and competitors may affect evaluation; (iii) conformist assessments (for example, triggered by halo effect or by lack of knowledge about the contributions of different centers).

1.12.3 Evaluation of Research Organizations and Groups of Research Organizations

Evaluation of research and technology programs [387], research organizations, and groups of research organizations becomes increasingly important especially when limited resources for research should be distributed. Below we mention three examples of such systems used in continental Europe.

1. *The SEP system of The Netherlands*
   SEP (Standard Evaluation Protocol) has two key objectives: (i) to improve research quality based on an external peer review, including scientific and societal relevance of research, research policy and management; (ii) to be accountable to the board of the research organization, and toward funding agencies, government, and society. SEP has two levels: evaluation of the research institute as a whole, and evaluation of specific research groups or programs. The criteria for evaluation are quality, productivity, (social) relevance, and vitality and feasibility. Each of these criteria contains several subcriteria.
2. *AERES system (France)*
   The Evaluation Agency for Research and Higher Education (AERES) (www.aeres-evaluation.com) is an independent administrative authority whose task is to evaluate French research organizations and institutions, research and higher education institutions, scientific cooperation foundations and institutions as well as the French National Research Agency. Usually each year, 25 % of all institutions are evaluated, and thus the national evaluation cycle is four years. The evaluation criteria of AERES are: scientific quality and output; academic reputation and drawing power; interactions with the social, economic, and cultural environment; organization and life of the institution; involvement in training by research; strategy and scientific prospects for the next period.
3. *Evaluation of national research policy: OECD indicators*
   Various systems of indicators for evaluation of national research policy can be constructed (for example, see [388, 389]). One example is the system of eight categories of indicators applied by OECD for benchmarking of national research policies [390]. These categories of indicators are: human resources in research and technology development; public and private investment in research and technology development; scientific and technological production; impact of research and technology development on economic competitiveness and employment; human resources, knowledge creation; transmission and application of knowledge; innovation finance, output, and market.

1.13 Mathematics and Quantification of Research Performance. English–Czerwon Method

Egghe [391] discusses performance in twelve diverse information production systems. Below, we shall focus our attention on a mathematical method for quantification of research performance: the English-Czerwon method that may be used for assessment of research performance within the scope of the following kinds of information production systems discussed in [391]: papers—citations system; authors—publications system; authors—citations system. The English–Czerwon method [392] is an interesting example of a combination of evaluation on the basis of quantitative indicators with evaluation on the basis of peer review. Its description is as follows.

Suppose there are k research units, and let the performance of each unit for the nth year of evaluation be denoted by $p_i(n) \geq 0$. The performance has two components,

$$p_i(n) = o_i(n) + s_i(n), \tag{1.6}$$

where:

- $o_i(n)$ is the "objective" part of the performance, calculated on the basis of quantitative indicators;
- $s_i(n)$ is the "subjective" part of performance, calculated on the basis of the judgment of all evaluated units about the research of the ith unit for the nth year of evaluation.

There are two variants of the method: (i) weighting without accounting for the current performance; (ii) weighting with accounting for the current performance.

1.13.1 Weighting Without Accounting for the Current Performance

In this variant of the method, the "subjective" part of the performance is calculated as follows:

$$s_i(n) = \sum_{j=1}^{k} q_{ij} w_j(n), \tag{1.7}$$

where the weight $w_j(n)$ is defined as $w_j(1) = o_j(1)$ for the first year of evaluation, and

$$w_j(n) = o_j(n) + \frac{\sum_{l=1}^{n-1} p_j(l)}{n-1} \tag{1.8}$$

for the years $n > 1$. We note that in this variant of the methodology, the performance for the current year $p_i(n)$ is not taken into account. The methods for calculation of the parameters o_j, q_{ij}, and $w_j(n)$ will be discussed in a separate paragraph below.

1.13.2 *Weighting with Accounting for the Current Performance*

The performance in the current year $p_i(n)$ is taken into account by defining the weight as follows:

$$\overline{w}_j(n) = \frac{1}{n}\sum_{l=1}^{n} p_j(l). \tag{1.9}$$

Let us substitute (1.9) in (1.6). Taking into account (1.7), we obtain an equation for $p_i(n)$ as follows:

$$p_i(n) = o_i(n) + \sum_{j=1}^{k} q_{ij}\overline{w}_j(n) =$$

$$o_i(n) + \sum_{j=1}^{k} q_{ij}\frac{p_j(1) + \cdots + p_j(n-1)}{n} + \sum_{j=1}^{k} q_{ij}\frac{p_j(n)}{n}. \tag{1.10}$$

We have the following relationship from (1.9):

$$p_j(1) + \cdots + p_j(n-1) = (n-1)\overline{w}_j(n-1). \tag{1.11}$$

The substitution of (1.11) in (1.10) leads to the following equation for $p_i(n)$:

$$p_i(n) = o_i(n) + \sum_{j=1}^{k} q_{ij}\frac{(n-1)\overline{w}_j(n-1) + p_j(n)}{n}. \tag{1.12}$$

We note that (1.12) defines a system of equations for $p_i(n)$, and this system still has to be solved.

The solution of (1.12) can be presented in matrix form. Let I be the identity matrix (which contains 1 in all diagonal positions and 0 elsewhere) and Q the matrix whose elements are the evaluations q_{ij}. Let the vectors $\mathbf{p}(n)$, $\mathbf{o}(n)$, and $\overline{\mathbf{w}}(n-1)$ have components $p_i(n)$, $o_i(n)$, and $\overline{w}_i(n-1)$ respectively. Then the solution of (1.12) is

$$\mathbf{p}(n) = (nI - Q)^{-1}[n\mathbf{o}(n) + (n-1)Q\overline{\mathbf{w}}(n-1)]. \tag{1.13}$$

In order to use this solution, we have to impose an additional requirement on the matrix Q. This requirement has to ensure positive values of $p_i(n)$. The requirement

is that the eigenvector of Q with positive elements have an eigenvalue less than n. Then each eigenvalue is less than n (Frobenius–Perron theorem), and the time series expansion of $(nI - Q)^{-1}$ (where the exponent -1 means inversion of the matrix $(nI - Q)$) converges to a matrix with nonnegative matrix elements. This will ensure positive values of $p_i(n)$. For small values of the elements of Q, we have the approximation

$$(nI - Q)^{-1} \approx \frac{I}{n} + \frac{Q}{n^2}, \tag{1.14}$$

and the solution (1.13) for $\mathbf{p}(n)$ becomes

$$\mathbf{p}(n) \approx \mathbf{o}(n) + \frac{Q\mathbf{o}(n) + (n-1)Q\overline{\mathbf{w}}(n-1)}{n}. \tag{1.15}$$

1.13.3 How to Determine the Values of Parameters

In order to use the two variants of the English–Czerwon method, we have to determine the values $o_i(n)$ of the "objective" part of the performance as well as the values of the evaluation coefficients q_{ij} from the "subjective" part of the performance. The simplest way to set the values of o_i is just to consider the value of one indicator: the number of citations of the papers of the research organization for the current year, for example. Of course, more that one indicator can be incorporated in o_i by means of appropriate weighting.

The values of q_{ij} can be set as follows. We note that these values have to be small (in order to satisfy the assumption for small Q on which basis we obtained the relationship (1.15) above). The determination of the values can be made as follows:

1. **Evaluation of performance by ranking research organizations.**
   One takes several (five to ten) leading scientists from each research organization and asks each of them to rank the research organizations with respect to their performance. The ranking is between rank 1 and rank k. Then the average rank $\overline{r}_{ij}$ is calculated by averaging the assigned ranks to the ith organization from the scientists of the jth organization:

$$\overline{r}_{ij} = \frac{1}{L}\sum_{j=1}^{L} r_{ij}, \tag{1.16}$$

   where L is the number of evaluating scientists.
2. **Calculation of q_{ij}.**
   The q_{ij} are calculated as follows:

$$q_{ij} = \frac{k - \overline{r}_{ij}}{M}, \tag{1.17}$$

where M is a large number (of order of $5k$ or larger), which ensures that the values of q_{ij} are small.

3. **Obtaining** $p_i(n)$.
   The $p_i(n)$ are then obtained on the basis of (1.15). The final step is to perform a normalization

$$\overline{p}_i(n) = \frac{p_i(n)}{\sum_{j=1}^{k} p_i(n)}, \tag{1.18}$$

   and one can rank the institutions with respect to the values of $\overline{p}_i$ for the corresponding year.

1.14 Concluding Remarks

At the end of Part I of this book, the reader already may have a impression about the complexity of science and research organizations; about the importance of science for society; about the features of research production and non-Gaussianity of some statistical characteristics of quantities used for assessment of research; about quantities, methods, and systems used for assessment of research and research organizations. Thus the reader is prepared to move to the world of quantities and models used for the study of science dynamics and for assessment of research, researchers, and research organizations.

References

1. J.L. Heilbron (ed.), *The Oxford Companion to the History of Modern Science* (Oxford University Press, New York, 2003)
2. Science. Wikipedia, the free encyclopedia
3. R.K. Merton, *The Sociology of Science: Theoretical and Empirical Investigations* (The University of Chicago Press, Chicago, 1973)
4. P. Vinkler, *The Evaluation of Research by Scientometric Indicators* (Chandos, Oxford, 2010)
5. J.D. Bernal, *The Social Function of Science* (The MIT Press, Cambridge, MA, 1939)
6. J.D. Bernal, *Science and Industry in the Nineteenth Century* (Routhledge, New York, 1953)
7. G. Böhme, N. Stehr, The growing impact of scientific knowledge on social relations, in *The Knowledge Society*, ed. by G. Böhme, N. Stehr (eds.) (Springer, Netherlands, 1986), pp. 7–29
8. R. Whitley, *The Intellectual and Social Organization of the Sciences* (Oxford University Press, Oxford, 2000)
9. R. Whitley, J. Gläser (eds.), *The Changing Governance of the Sciences* (Springer, Dordrecht, 2007)
10. *Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development*, 6th. edn. (OECD, 2002)
11. A. Kaufmann, F. Tödtling, Science-industry interaction: the importance of boundary-crossing between systems. Res. Policy **30**, 791–804 (2001)
12. S. Weinberger, The evolving science of war. Nature **505**, 156–157 (2014)

13. R. Lidskog, G. Sundqvist, When does science matter? International relations meets science and technology studies. Glob. Environ. Polit. **15** (in press, 2015). doi:10.1162/GLEP_a_00269
14. F. Narin, J. Davidson Frame. The growth of Japanese science and technology. Science **245**(4918), 600 (1989)
15. C.S. Wagner, I. Branmakulam, B. Jackson, A. Wong, T. Yoda, *Science And Technology Collaboration: Building Capacity in Developing Countries* (RAND, 2001). MR-1357.0-WB
16. B. Parker, *The Physics of War: From Arrows to Atoms* (Prometheus Books, New York, 2014)
17. E.R. Gantman, Economic, linguistic, and political factors in the scientific productivity of countries. Scientometrics **93**, 967–985 (2012)
18. M.E.D. Koenig, Information policy—the mounting tension (value additive versus uniquely distributable "public good"). J. Inf. Sci. **21**, 229–231 (1995)
19. P.E. Stephan, The economics of science. J. Econ. Lit. **34**, 1199–1235 (1996)
20. P.E. Stephan, *How Economics Shapes Science* (Harvard University Press, Cambridge, MA, 2012)
21. J.N. Cummings, S. Kiesler, Organization theory and the changing nature of science. J. Organ. Des. **3**, 1–16 (2014)
22. M. Hirooka, Nonlinear dynamism of innovation and business cycles. J. Evolut. Econ. **13**, 549–576 (2003)
23. B.R. Martin, The evolution of science policy and innovation studies. Res. Policy **41**, 1219–1239 (2012)
24. F. Narin, K.S. Hamilton, D. Olivastro, Linkage between agency-supported research and patented industrial technology. Res. Eval. **5**, 183–187 (1995)
25. B.R. Martin, P. Nightingale, A. Yegros-Yegros, Science and technology studies: exploring the knowledge base. Res. Policy **41**, 1182–1204 (2012)
26. M. Hirooka, *Innovation Dynamism and Economic Growth: A Nonlinear Perspective* (Edward Elgar Publishing, Cheltenham, UK, 2006)
27. P. van den Besselaar, K. Börner, A. Scharnhorst. Science policy and the challenges for modeling science. p.p. 261 – 266 in A. Scharnhorst, K. Börner, P. van den Besselaar (eds.) Models for science dynamics (Springer, Berlin, 2012)
28. A. Smith, A. Stirling, F. Berkhout, The governance of sustainable socio-technical transitions. Res. Policy **34**, 1491–1510 (2005)
29. A. Stirling, A general framework for analyzing diversity in science, technology and society. J. R. Soc. Interface **4**, 707–719 (2007)
30. F. Berkhout, A. Smith, A. Stirling, Socio-tecgnical regimes and transition contexts, pp. 48–75 in *System Innovation and the Transition to Sustainability* ed. by B. Elzen, F.W. Geels, K. Green (Edward Elgar, Chentelham, 2004)
31. D.P. Gaver, V. Srinivasan, Allocating resources between research and development: a macro analysis. Manage. Sci. **18**, 492–501 (1972)
32. A.D. Bender, E.B. Pyle III, W.J. Westlake, B. Douglas, Simulation of R&D investment strategies. Omega **4**, 67–77 (1976)
33. S. Bretschneider, Operations research contributions to evaluation of R&D projects, pp. 122–153 in *Evaluating R&D Impacts: Methods and Practice* ed. by B. Bozeman, J. Melkers (Springer, US, 1993)
34. R.L. Schmidt, Recent progress in modelling R&D project-selection process. IEEE Trans. Eng. Manage. **39**, 189–201 (1992)
35. L. Bornmann, Measuring the societal impact of research. EMBO Rep. **13**, 673–676 (2012)
36. C.S. Wagner, The elusive partnership: science and foreign policy. Sci. Public Policy **29**, 409–417 (2002)
37. M.A. Rappa, K. Debackere, Technological communities and the diffusion of knowledge. R&D Manage. **22**, 209–220 (1992)
38. M.A. Rappa, K. Debackere, Technological communities and the diffusion of knowledge: a replication and validation. R&D Manage. **24**, 355–371 (1994)
39. L. Bornmann, What is societal impact of research and how can it be assessed? A literature survey. J. Am. Soc. Inf. Sci. Technol. **64**, 217–233 (2013)

40. A. Verbeek, K. Debackere, M. Luwej, E. Zimmermann, Measuring progress and evolution in science and technology I: the multiple uses of bibliometric indicators. Int. J. Manag. Rev. **4**, 179–211 (2002)
41. A. Verbeek, K. Debackere, M. Luwej, E. Zimmermann, Measuring progress and evolution in science and technology II: the multiple uses of technometric indicators. Int. J. Manag. Rev. **4**, 213–231 (2002)
42. L. Bornmann, W. Marx, How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparison. Scientometrics **98**, 211–219 (2014)
43. A.F.J. van Raan, Measurement of central aspects of scientific research: performance, interdisciplinarity, structure. Meas.: Interdiscip. Res. Perspect. **3**, 1–19 (2005)
44. V.V. Nalimov, G.M. Mulchenko, *Naukometriya* (Nauka, Moscow, 1969). (in Russian)
45. C. Michels, U. Schmoch, The growth of science and database coverage. Scientometrics **93**, 831–846 (2012)
46. F. Narin, K.S. Hamilton, D. Olivastro, The increasing linkage between US technology and public science. Res. Policy **26**, 317–330 (1997)
47. J. Anderson, K. Williams, D. Seemungal, F. Narin, D. Olivastro, Human genetic technology: exploring the links between science and innovation. Technol. Anal. Strat. Manag. **8**, 135–156 (1996)
48. F. Narin, D. Olivastro, Status report: linkage between technology and science. Res. Policy **21**, 237–249 (1992)
49. M. Mayer, Tracing knowledge flows in innovation systems. Scientometrics **54**, 193–212 (2002)
50. F. Narin, E. Noma, R. Perry, Patents as indicators of corporate technological strength. Res. Policy **16**, 143–155 (1987)
51. M. Mayer, Patent citation analysis in a novel field of technology: an exploration of nano-science and nano-technology. Scientometrics **51**, 163–183 (2001)
52. A. Verbeek, K. Debackere, M. Luwel, P. Andries, E. Zimmermann, F. Deleus, Linking science to technology: using bibliographic references in patents to build linkage schemes. Scientometrics **54**, 399–420 (2002)
53. S. Bhattacharya, H. Kretdchmer, M. Mayer, Characterizing intellectual spaces between science and technology. Scientometrics **58**, 369–390 (2003)
54. A.L. Porter, A. Thomas Roper, T.W. Mason, F.A, Rossini, J. Banks, *Forecasting and Management of Technology* (Wiley, New York, 1991)
55. R.J. Watts, A. Porter, Innovation forecasting. Technol. Forecast. Soc. Chang. **56**, 25–47 (1997)
56. M. Meyer, Measuring science-technology interaction in the knowledge-driven economy: the case of small economy. Scientometrics **66**, 425–429 (2006)
57. J. Dryden, Quantifying technological advance: S&T indicators at the OECD—challenges for the 1990s. Sci. Public Policy **19**, 281–290 (1992)
58. E.C.M. Noyons, A.F.J. van Raan, H. Grupp, U. Schmoch, Exploring the science and technology interface: Inventor-author relations in laser medicine research. Res. Policy **23**, 443–457 (1994)
59. A. Mathieu, M. Mayer, B. van Pottelberghe, de la Potterie, Turning science into business: a case study of a major European Research University. Sci. Public Policy **35**, 669–679 (2008)
60. M. Mayer, Tracing knowledge flows in innovation systems—an infometric perspective of future research on science-based innovation. Econ. Syst. Res. **14**, 323–344 (2002)
61. A. Klitkou, S. Nyagaard, M. Mayer, Tracking techno-science networks: a case study of fuel cells and related hydrogen technology. Scientometrics **70**, 491–518 (2007)
62. A.L. Porter, M.J. Detampel, Technology opportunities analysis. Technol. Forecast. Soc. Chang. **49**, 237–255 (1995)
63. J. McNerney, J. Doyne Farmer, S. Redner, J.E. Trancik, Role of design complexity in technology improvement. Proc. Natl. Acad. Sci. USA **108**, 9008–9013 (2011)
64. G.S. McMillan, F. Narin, D.L. Deeds, An analysis of the critical role of public science in innovation: the case of biotechnology. Res. Policy **29**, 1–9 (2000)

65. D. Hicks, A. Breitzman, K. Hilton, F. Narin, Research excellence and patented innovation. Sci. Public Policy **27**, 310–320 (2000)
66. F. Narin, A. Breitzman, Inventive productivity. Res. Policy **24**, 507–519 (1995)
67. M. Mayer, Does science push technology? Patents citing scientific literature. Res Policy **29**, 409–434 (2000)
68. A. Hullman, M. Meyer, Publications and patents in nanotechnology. Scientometrics **58**, 507–527 (2003)
69. B.G. van Vianen, H.F. Moed, A.F.J. van Raan, An exploraton of the science base of recent technology. Res. Policy **19**, 61–81 (1990)
70. M. Ioannidis, A. Vatikalos, Cutting-edge information and telecommunication technologies meet energy: Energy management systems and smart web platforms, pp. 153 – 162 in *Energy-Efficient Computing and Networking*, ed. by N. Hatziargyriou, A. Dimeas, T. Tomtsi, A. Weidlich (Springer, Berlin, 2011)
71. Z.S. Tao, L. Rui, Z. Xia, H.C. Hua, W.Y. Quan, The emerging cutting-edge of virus research. Viral Proteomics. Science China: Life Sci. **65**, 502–512 (2011)
72. C.S. Wagner, S.W. Popper, Identifying critical technologies in the United States: a review of the federal effort. J. Forecast. **22**, 113–128 (2003)
73. M.J. Jackson, *Micro and Nanomanufacturing* (Springer, New York, 2007)
74. D.N. Weil, *Economic Growth*, 3rd edn. (Pearson, Boston, 2013)
75. L. Leydesdorff, M. Mayer, The decline of university patenting and the end of the Bayh-Dole effect. Scientometrics **83**, 355–362 (2010)
76. M. Sanders, Scientific paradigms, entrepreneurial opportunities and cycles in economic growth. Small Bus. Econ. **28**, 339–354 (2007)
77. H. Grupp (ed.), *Dynamics of Science-Based Innovation* (Springer, Berlin, 1992)
78. G.M. Grossman, E. Helpman, *Innovation and Growth in the Global Economy* (The MIT Press, Cambridge, MA, 1993)
79. R.J. Barro, X. Sala-i-Martin, Technological diffusion, convergence, and growth. J. Econ. Growth **2**, 1–26 (1992)
80. R.E. Lucas Jr., *Lectures on Economic Growth* (Harward University Press, Cambridge, MA, 2002)
81. L. Girifalco, *Dynamics of Technological Change* (Van Nostrand Reinhold, New York, 1991)
82. H. Moed, R. de Druin, T.H. van Leeuwn, New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. Scientometrics **33**, 381–422 (1995)
83. H.F. Moed, W.J.M. Burger, J.G. Frankfort, A.F.J. van Raan, The use of bibliometric data for the measurement of university research performance. Res. Policy **14**, 131–149 (2002)
84. A.F.J. van Raan, Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics **62**, 133–143 (2005)
85. H. Moed, Measuring China's research performance using the science citation index. Scientometrics **53**, 281–296 (2002)
86. A.F.J. van Raan, Measuring science pp. 19–50 in Handbook of Quantitative Science and Technology Research, ed. by H.F. Moed, W. Glänzel, U. Schmoch (Springer, Netherlands, 2005)
87. T.N. van Leeuwen, M.S. Visser, H.F. Moed, T.J. Nederhof, A.F.J. van Raan, The Holly Grail of science policy: exploring and combining bibliometric tools in search for scientific excellence. Scientometrics **57**, 281–296 (2002)
88. J.A.D. Holbrook, Why measure science. Sci. Public Policy **19**, 262–266 (1992)
89. H. Legler, G. Licht, A. Spielkamp, *Germany's Technological Performance. A Study on Behalf of the German Federal Ministry of Education and Research* (Physica-Verlag, Berlin, 2000)
90. E.J. Rinia, Scientometric studies and their role in research policy of two research councils in the Netherlands. Scientometrics **47**, 363–378 (2000)
91. K. Fealing, J. Lane, J. Marburger III, S. Shipp (eds.), *The Science of science policy* (Stanford University Press, Stanford, CA, A handbook, 2011), p. 2011

92. L. Bornmann, H.-D. Daniel, Does the *h*-index for ranking of scientists really work? Scientometrics **65**, 391–392 (2005)
93. I. Feller, Performance measurement and the governance of American academic science. Minerva **47**, 323–344 (2009)
94. L. Leydesdorrf, L. Bornmann, Integrated impact indicators compared with impact factors: an alternative research design with policy implications. J. Am. Soc. Inf. Sci. Technol. **62**, 2133–2146 (2011)
95. L. Bornmann, L. Leydesdorrf, R. Mutz, The use of percentiles and percentile rank classes in the analysis of bibliometric data: opportunities and limits. J. Infometr. **7**, 158–165 (2013)
96. J.A.D. Holbrook, Basic indicators of scientific and technological performance. Sci. Public Policy **19**, 267–273 (1992)
97. D. Kondepudi, I. Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures* (Wiley, New York, 1998)
98. G. Nicolis, I. Prigogine, *Self-Organization in Non-equilibrium Systems* (Wiley, New York, 1977)
99. E.C.M. Noyons, A.F.J. van Raan, Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research. J. Am. Soc. Inf. Sci. **49**, 68–81 (1998)
100. P.A.A. van den Besselaar, L.A. Leydesdorff, Evolutionary Economics and Chaos Theory: New Directions in Technology Studies (Frances Pinter Publishers, 1994)
101. A. Stirling, Science, precaution, and the politics of technological risk. Ann. N. Y. Acad. Sci. **1128**, 95–110 (2008)
102. A. Smith, A. Stirling, The politics of social-ecological resilience and sustainable sociotechnical transitions. Ecol. Soc. **15**, Art. No. 11 (2010)
103. J. Pfeffer, G.R. Salancik, *The External Control of Organizations: A Resource Dependence Perspective* (Stanford University Press, Stanford, CA, 2003)
104. M. Gibbons, C. Limoges, H. Nowotny, S. Schwartzman, P. Scott, M. Throw, *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies* (Sage Publications, London, 1994)
105. L. Hessels, H. van Lente, Re-thinking new knowledge production: a literature review and a research agenda. Res. Policy **37**, 740–760 (2008)
106. T.R. Blackburn, Information and ecology of scholars. Science **181**, 1141–1146 (1971)
107. N.C. Mullins, *Theories and Theory Groups in Contemporary Sociology* (Harper & Row, New York, 1973)
108. E. Jimenez-Contreras, F. de Moya-Anegon, E.D. Lopez-Cozar, The evolution of research activity in Spain: the impact of the national commission for the evaluation of research activity (CNEAI). Res. Policy **32**, 123–142 (2003)
109. B.M. Gupta, C.R. Karisiddappa, Modelling the growth of literature in the area of theoretical population genetics. Scientometrics **49**, 321–355 (2000)
110. H. Etzkowitz, L. Leydesdorff, The Triple Helix: University-industry-government relations: a laboratory for knowledge based economic development. EASST Rev. **14**, 14–19 (1995)
111. L. Leydesdorff, H. Etzkowitz, Emergence of aTriple Helix of university-industry-government relations. Sci. Public Policy **23**, 279–286 (1996)
112. L. Leydesdorff, H. Etzkowitz, The Triple Helix as a model for innovation studies. Sci. Public Policy **25**, 195–203 (1998)
113. H. Etzkowitz, L. Leydesdorff, The endless transition: a 'Triple Helix' of university industry government relations. Minerva **36**, 203–208 (1998)
114. H. Etzkowitz, L. Leydesdorff, The dynamics of innovation: from National Systems and 'Mode 2' to a Triple Helix of university-industry-government relations. Res. Policy **29**, 109–123 (2000)
115. L. Leydesdorff, G. Zawdie, The Triple Helix perspective of innovation systems. Technol. Anal. Strat. Manag. **22**, 789–804 (2010)
116. L. Leydesdorff, The knowledge-based economy and the Triple Helix model. Annu. Rev. Inf. Sci. Technol. **1**, 365–417 (2010)

117. H. Etzkowitz, *The Triple Helix: University-Industry-Government Innovation in Action* (Routledge, New York, 2008)
118. L. Leydesdorff, E. Perevodchikov, A. Uvarov, Measuring Triple-Helix synergy in the Russian innovation system at regional, provincial, and national levels. J. Assoc. Inf. Sci. Technol. **66**, 1229–1238 (2015)
119. L. Leydesdorff, The mutual information of university-industry-government relations: an indicator of the Triple Helix dynamics. Scientometrics **58**, 445–467 (2003)
120. L. Leydesdorff, The Triple Helix, Quadruple Helix, ..., and N-tuple of helixes: explanatory models for analyzing the knowledge-based economy? J. Knowl. Econ. **3**, 25–35 (2012)
121. L. Leydesdorff, The Triple Helix: an evolutionary model of innovations. Res. Policy **29**, 243–255 (2000)
122. W.B. Arthur. Competing technologies. pp. 590–607, in *Technical Change and Economic Theory*, ed. by G. Dosi, C. Frenman, R. Nelson, G. Silverberg, L. Soete (Pinter, London, 1988)
123. W.B. Arthur, Competing technologies, increasing returns, and lock-in by historical events. Econ. J. **99**, 116–131 (1989)
124. M. Zitt, E. Bassecoulard, Y. Okubo, Shadows of the past in international cooperation: collaboration profiles of the top five producers of science. Scientometrics **47**, 627–657 (2000)
125. D.A. King, The scientific impact of nations. What different countries get for their research spending. Nature **430**, 311–316 (2004)
126. M.E. Porter, *The Competitive Advantage of Nations* (Basingstoke, New York, Palgrave MacMillan, 1990)
127. R.M. May, The scientific wealth of nations. Science **275**, 793–795 (1977)
128. A.L. Porter, J. David Roessner, X.-Y. Jin, N.C. Newman, Measuring national 'emerging technology' capabilities. Sci. Public Policy **29**, 189–200 (2002)
129. J.-Y. Choung, H.-R. Hwang, National systems of innovation: Institutional linkages and performances in the case of Korea and Taiwan. Scientometrics **48**, 413–426 (2000)
130. P. Zhou, L. Leydesdorff, The emergence of China as leading nation in science. Res. Policy **35**, 83–104 (2006)
131. A.-W. Harzing, A. Giroud, The competitive advantage of nations: An application to academia. J. Infometr. **8**, 29–42 (2014)
132. T.-E. Sandberg Hannsen, F. Jørgensen, The value of experience in research. J. Infometr. **9**, 16–24 (2015)
133. B.R. Martin, J. Irvine, Assessing basic research: some partial indicators of scientific progress in radio astronomy. Res. Policy **12**, 61–90 (1983)
134. B.R. Martin, The use of multiple indicators in the assessment of basic research. Scientometrics **36**, 343–362 (1996)
135. D. Hicks, Performance-based university research funding systems. Res. Policy **41**, 251–261 (2012)
136. S. Hornbostel, S. Böhmer, B. Klingsporn, J. Neufeld, M. von Ins, Funding of young scientist and scientific excellence. Scientometrics **79**, 171–190 (2009)
137. M. Lamont, *How Professors Think: Inside the Curious World of Academic Judgment* (Harvard University Press, Cambridge, MA, 2009)
138. D. Hicks, J.S. Katz, Equity and excellence in research funding. Minerva **49**, 137–151 (2011)
139. V.V. Nalimov, *Faces of Science* (ISI Press, Philadelphia, 1981)
140. L. Esterle, M. Zitt. Observation of scientific publications in astronomy/astrophysics. pp. 91–109 in *Organizations and Strategies in Astronomy*, ed. by A. Heck (Kluwer, Dordrecht, 2000)
141. M. Crosland, Scientific credentials: record of publications in the assessment of qualifications for election to the French Académie des Sciences. Minerva **19**, 605–631 (1981)
142. G.J. Feist, Quantity, quality, and depth of research as influences on scientific eminence: is quantity most important? Creat. Res. J. **10**, 325–335 (1997)
143. D.C. Pelz, F.M. Andrews, *Scientists in Organizations. Productive Climates for Research and Development* (Wiley, New York, 1966)

144. R. Senter Jr., A causal model of productivity in a research facility. Scientometrics **10**, 307–328 (1986)
145. T. Luukkonen, B. Stahle, Quality evaluations in the management of basic and applied research. Res. Policy **19**, 357–368 (1990)
146. B. Kim, H. Oh, An effective R&D performance measurement system: survey of Korean R&D researchers. Omega **30**, 19–31 (2002)
147. P. Dahler-Larsen, Constitutive effects of performance indicators. Public Manag. Rev. **16**, 969–986 (2014)
148. P. Dahler-Larsen, *The Evaluation Society* (Stanford Business Books, Stanford, CA, 2012)
149. ISO 9000:2000, *Quality Management Systems—Fundamentals and Vocabulary* (ISO, Geneva)
150. S. Helmin, Scientific quality in the eyes of the scientists. A questionnaire study. Scientometrics **27**, 3–18 (1993)
151. J.M. Pastor, L. Serrano, I. Zaera, The research output of European higher education institutions. Scientometrics **102**, 1867–1893 (2015)
152. U. Schmoch, T. Schubert, Are international co-publications an indicator for quality of scientific research? Scientometrics **74**, 361–377 (2008)
153. D. Cutlača, D. Babić, I. Živković, D. Štrbac, Analysis of qualitative and quantitative indicators of SEE countries scientific output. Scientometrics **102**, 247–265 (2015)
154. F. Franceschini, M. Galetto, D. Maisano, *Management by Measurement* (Springer, Berlin, 2007)
155. F. Welter, S. Schröder, I. Leisten, A. Richert, S. Jeschke, Scientific performance indicators - empirical results from collaborative research centers and clusters of excellence in Germany, pp. 203–220 in *Automation, Communication and Cybernetics in Science and Engineering 2013/2014*, ed. by S. Jeschke, I. Insenhardt, F. Hees, K. Henning (Springer International Publishing, Switzerland, 2014)
156. S.D. Haitun, The problem of indicator-latent relationship in metric models I: statement and general solution. Scientometrics **23**, 335–351 (1992)
157. S.D. Haitun, The problem of indicator-latent relationship in metric models II: metric models with a priori latent assignment. Scientometrics **24**, 221–235 (1992)
158. G. Samorodnitsky, M.S. Taqqu, Non-Gaussian Random Processes. Stochastic Models with Infinite Variance. (Chapmann & Hall, Boca Raton, 1994)
159. F.E. Beth, J. Kallsen, T. Meyer-Brandis, A non-Gaussian Ornstein—Uhlenbeck process for electricity spot price modeling and derivatives pricing. Appl. Math. Finan. **14**, 153–169 (2007)
160. O.E. Bandorff-Nielsen, N. Sheppard, Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. J. R. Stat. Soc. B **63**, 167–241 (2001)
161. S.D. Haitun, The "rank-distortion" effect and non-Gaussian nature of scientific activities. Scientometrics **5**, 375–395 (1983)
162. S.D. Haitun, Stationary scientometric distributions. II. Non-Gaussian nature of scientific activities. Scientometrics **4**, 89–104 (1982)
163. M.G. Kendall, Natural law in the social sciences. J. R. Stat. Soc. A **124**, 1–16 (1961)
164. S.D. Haitun, *Scientometrics: State and Perspectives* (Nauka, Moscow, 1983). (in Russian)
165. L. Leydesdorff, *The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications* (DSWO Press, Leiden, 1995)
166. W. Glänzel, U. Schoepflin, Little scientometrics, big scientometrics.. and beyond? Scientometrics **30**, 375–384 (1994)
167. L. Leydesdorff, S. Milojevich, Scientometrics. **1208**, 4566 (2012)
168. L. Bornmann, L. Leydesdorff, Scientometrics in a changing research landscape. EMBO Rep. **15**, 1228–1232 (2014)
169. A. Schubert, Scientometrics: the research field and its journal, pp. 179–195 in *Organizations and Strategies in Astronomy II*, ed. by A. Heck (Kluwer, Dordrecht, 2001)
170. L. Leydesdorff, P. van den Besselaar, Scientometrics and communication theory: towards theoretically informed indicators. Scientometrics **38**, 155–174 (1997)
171. T. Braun, E. Bujdodo, A. Schubert, *Literature of Analytical Chemistry: A scientometric evaluation* (CRC Press, Boca Raton, FL, 1987)

172. W. Glänzel, A. Schubert, A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics **56**, 357–367 (2003)
173. M. Zitt, E. Bassecoulard, Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. Ethics Sci. Environ. Polit. **8**, 49–60 (2008)
174. T. Braun, A. Schubert, Scientometric versus socio-economic indicators. Scatter plots for 51 countries. 1978–1980. Scientometrics **13**, 3–9 (1988)
175. F. Narin, M.B. Albert, V.M. Smith, Technology indicators and strategic planning. Sci. Public Policy **19**, 369–381 (1992)
176. E. Bassecoulard, M. Zitt, Indicators in a research institute: a multi-level classification of scientific journals. Scientometrics **44**, 325–345 (1999)
177. A. Schubert, S. Zsindely, T. Braun, Scientometric analysis of attendance at international scientific meetings. Scientometrics **5**, 177–187 (1983)
178. A. Schubert, W. Glänzel, T. Braun, Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985. Scientometrics **16**, 3–478 (1989)
179. T. Braun, W. Glänzel, A. Schubert, Publication productivity: from frequency distributions to scientometric indicators. J. Inf. Sci. **16**, 37–44 (1990)
180. T.A. Brooks, Private acts and public objects: an investigation of citer motivations. J. Am. Soc. Inf. Sci. **36**, 223–229 (1985)
181. M.J. Moravcsik, P. Murugesan, Some results on the function and quality of citations. Soc. Stud. Sci. **5**, 86–92 (1975)
182. T.A. Brooks, Evidence of complex citer motivations. J. Am. Soc. Inf. Sci. **37**, 34–36 (1986)
183. D.W. Aksnes, Characteristics of highly cited papers. Res. Eval. **12**, 159–170 (2003)
184. D.W. Aksnes, G. Sivertsen, The effect of highly cited papers on national citation indicators. Scientometrics **59**, 213–224 (2004)
185. B. Cronin, The need of a theory of citing. J. Doc. **37**, 16–24 (1981)
186. B. Cronin, The Citation Process. *The Role and Significance of Citations in Scientific Communication* (Taylor Graham, London, 1984)
187. B. Cronin, Norms and functions in citation: the view of journal editors and referees in psychology. Soc. Sci. Inf. Stud. **2**, 65–78 (1982)
188. V. Cano, Citation behavior: classification, utility, and location. J. Am. Soc. Inf. Sci. **40**, 284–290 (1989)
189. A. Schubert, The web of scientometrics. Scientometrics **63**, 3–20 (2002)
190. F. Janssens, J. Leta, W. Glänzel, B. de Moor, Towards mapping library and information science. Inf. Process. Manag. **42**, 1614–1642 (2006)
191. H. Small, K.W. Boyack, R. Klavans, Identifying emerging topics in science and technology. Res. Policy **43**, 1450–1467 (2014)
192. H. Small, Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics **87**, 373–388 (2011)
193. F. Narin, Bibliometrics. Annu. Rev. Inf. Sci. Technol. 35–58 (1977)
194. M.J. Kurtz, J. Bollen, Usage bibliometrics. Annu. Rev. Inf. Sci. Technol. **44**, 1–64 (2010)
195. H.D. White, K.W. McCain, Bibliometrics. Annu. Rev. Inf. Sci. Technol. **24**, 119–186 (1989)
196. F. Narin, Evaluative bibliometrics. Computer Horizons, Inc. Project No. 704R (1996)
197. B. Cronin, Bibliometrics and beyond: some thoughts on the web-based citation analysis. J. Inf. Sci. **27**, 1–7 (2001)
198. J. Nicolaisen, The scholarliness of published peer reviews: a bibliometric study of book reviews in selected social science fields. Res. Eval. **11**, 129–140 (2002)
199. A.J. Nederhof, A.F.J. van Raan, A bibliometric analysis of six economics research groups: A comparison with peer review. Res. Policy **22**, 353–368 (1993)
200. A.J. Nederhof, Bibliometric monitoring of research performance in the social sciences and the humanities: A review. Scientometrics **66**, 81–100 (2006)
201. F. Narin, Patent bibliometrics. Scientometrics **30**, 147–155 (1994)
202. L. Björneborn, P. Ingwersen, Towards a basic framework for webometrics. J. Am. Soc. Inf. Sci. Technol. **55**, 1216–1227 (2004)

203. M. Thelwall, L. Vaughan, L. Björnedborn, Webometrics. Annu. Rev. Inf. Sci. Technol. **39**, 81–135 (2005)
204. L. Björneborn, P. Ingwersen, Perspectives of webometrics. Scientometrics **50**, 65–82 (2001)
205. C. Borgman, J. Furner, Scholarly communication and bibliometrics. Annu. Rev. Inf. Sci. Technol. **36**, 3–72 (2002)
206. M. Thelwall, Introduction to webometrics: quantitative web research for the social sciences. Synth. Lect. Inf. Concepts, Retr., Serv. **1**, 1–116 (2009)
207. M. Thelwall, Bibliometrics to webometrics. J. Inf. Sci. **34**, 605–621 (2008)
208. T.C. Almind, P. Ingwersen, Informetric analyses of the World Wide Web: methodological approaches to 'webometrics'. J. Doc. **53**, 404–426 (1997)
209. L. Björneborn, Small-world link structures across an academic Web space: a library and information science approach. Doctoral dissertation (Royal School of Library and Information Science, Copenhagen, Denmark, 2004)
210. P. Ingwersen, L. Björneborn, Methodological issues of webometric studies, pp. 339–369 in *Handbook of Quantitative Science and Technology Research*, ed. by H.F. Moed, W. Glänzel, U. Schmoch (Kluwer, New York, 2004)
211. A. Pritchard, Statistical bibliography or bibliometrics? J. Doc. **24**, 348–349 (1969)
212. K. Debackere, W. Glänzel, Using a bibliometric approach to support research policy making: the case of the Flemish BOF-key. Scientometrics **59**, 253–276 (2004)
213. T.N. van Leeuwen, M.S. Visser, H.F. Moed, T.J. Nederhof, A.F. van Raan, The Holy Grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. Scientometrics **57**, 257–280 (2003)
214. D.W. Aksnes, R.E. Taxt, Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. Res. Eval. **13**, 33–41 (2004)
215. A. Andres, Measuring Academic Research. How to Undertake a Bibliometric Study (Chandos Publishing, Oxford, 2009)
216. F. Narin, Bibliometric techniques in the evaluation of research programs. Sci. Public Policy **14**, 99–106 (1987)
217. F. Narin, R.P. Rozek, Bibliometric analysis of the US pharmaceutical industry research performance. Res. Policy **17**, 139–154 (1988)
218. W. Glaänzel, Bibliometrics as a research field. *A course on theory and application of bibliometric indicators* (Ungarische Akademie der Wissenschaften, Budapest, 2003)
219. V.I. Gorkova, Informetrics (quantitative methods in scientific and technical information). Itogi Nauki i Tekhniki. Ser. Informatika **10**, 328 (1988). (in Russian)
220. M.S. Galyavieva, On the formation of the concept of informtrics (Review). Sci. Tech. Inf. Process. **40**, 89–96 (2013)
221. C.S. Wilson, Informetrics. Annu. Rev. Inf. Sci. Technol. **34**, 107–247 (1999)
222. J. Bar-Ilan, Informetrics at the beginning of the 21st century: a review. J. Informetr. **2**, 1–52 (2008)
223. W.C. Adair, Citation indexes for scientific literature? Am. Doc. **6**, 31–32 (1955)
224. E. Garfield, Citation indexes for science. Science **122**(3159), 108–111 (1955)
225. E. Garfield, The Mystery of the transposed journal lists—wherein Bradford's law of scattering is generalized according to Garfield's law of concentration. Curr. Contents **17**, 222–223 (1971)
226. O. Persson, Studying research collaboration using co-authorships. Scientometrics **36**, 363–377 (1996)
227. M.A. Abbas, Weighted indexes for evaluating the quality of research with multiple authorship. Scientometrics, **88**, 107–131 (2011)
228. D. De Solla Price, *Little Science, Big Science* (Columbia University Press, New York, 1963)
229. R. Rousseau, Why am I not cited or, why are multi-authored papers more cited than others? J. Doc. **48**, 79–80 (1992)
230. B. Cronin, Hyperauthorship: a postmodern perversion or evidence of a structural shift in scholarly communication practices? J. Am. Soc. Inf. Sci. Technol. **52**, 558–569 (2001)
231. C.S. Wagner, L. Leydesdorff, Network structure, self-organization, and the growth of international collaboration in science. Res. Policy **34**, 1608–1618 (2005)

232. A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations. Physica A **311**, 590–614 (2002)
233. W. Glänzel, A. Schubert, *Analysing scientific networks through co-authorship, pp. 257–276 in Handbook of Quantitative Science and Technology Research*, ed. by H.F. Moed, W. Glänzel, U. Schmoch (Springer, Netherlands, 2005)
234. A. Schubert, T. Braun, International collaborations in the sciences 1981–1985. Scientometrics **19**, 3–10 (1990)
235. A. Prarianes-Rodriguez, C. Olmeda-Gomez, F. Moya-Anegon, Detecting, identifying and visualizing research groups in co-authorship networks. Scientometrics **82**, 307–319 (2010)
236. A. Bookstein, H. Moed, M. Yitzahki, Measures of international collaboration in scientific literature: Part I. Inf. Process. Manag. **42**, 1408–1421 (2006)
237. S. Lehmann, B. Lautrup, A.D. Jackson, Citation networks in high energy physics. Phys. Rev. E **68**, Art. No. 026113 (2003)
238. A. Bookstein, H. Moed, M. Yitzahki, Measures of international collaboration in scientific literature: Part II. Inf. Process. Manag. **42**, 1422–1427 (2006)
239. M.J. Mulkay, G.N. Gilbert, S. Woolgar, Problem areas and research networks in science. Sociology: J. Brit. Soc. Assoc. **9**, 187–203 (1975)
240. C.S. Wagner, Measuring the network of global science: comparing international co-authorships from 1990 to 2000. Int. J. Technol. Glob. **1**, 185–208 (2005)
241. C.S. Wagner, The New Invisible College: Science for Development (The Brookings Institution, 2008)
242. D. Crane, *Invisible Colleges: Diffusion of Knowledge in Scientific Communities* (The University of Chicago Press, Chicago, 1972)
243. D.J. de Solla Price, D.B. Beaver, Collaboration in an invisible college. Am. Psychol. **21**, 1011–1018 (1966)
244. A. Zuccala, Modeling the invisible college. J. Am. Soc. Inf. Sci. Technol. **57**, 152–168 (2006)
245. A.A. Zuccala, Revisiting the invisible college: a case study of the intellectual structure and social process of singularity theory research in mathematics. Ph.D. thesis, University of Toronto, 2004
246. B. Cronin, Invisible colleges and information transfer. A review and commentary with particular reference to the social sciences. J. Doc. **38**, 212–236 (1982)
247. H. Small, B.G. Griffith, The structure of scientific literatures I: identifying and graphic specialities. Sci. Stud. **4**, 17–40 (1974)
248. B.G. Griffith, H.G. Small, J.A. Stonehill, S. Dey, The structure of scientific literatures II: toward a macro- and microstructure for science. Soc. Stud. Sci. **4**, 339–365 (1974)
249. P. Auger, *Tendances Actuelles de la Recherche Scientifique* (UNESCO, 1961)
250. H.F. Moed, F. de Moya-Anegon, C. Lopez-Illescas, M. Visser, Is concentration of university research associated with better research performance? J. Infometr. **5**, 649–658 (2011)
251. H.F. Moed, G. Halevi, A bibliometric approach to tracking international scientific migration. Scientometrics **101**, 1987–2001 (2014)
252. H.F. Moed, M. Aisati, A. Plume, Studying scientific migration in Scopus. Scientometrics **94**, 929–942 (2013)
253. E. Garfield, Citation analysis as a tool in journal evaluation. Science **178**, 471–479 (1972)
254. E. Garfield, Citation indexing for studying science. Nature **227**, 669–671 (1970)
255. L. Bornmann, H.-D. Daniel, What do citation counts measure? A review of studies on citing behavior. J. Doc. **64**, 45–80 (2008)
256. R. Plomp, The highly cited papers of professors as an indicator of a research group's scientific performance. Scientometrics **29**, 377–393 (1994)
257. H. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents. J. Am. Soc. Inf. Sci. **24**, 265–269 (1973)
258. B. Cronin, H. Snyder, H. Atkins, Comparative citation rankings of authors in monographic and journal literature: a study of sociology. J. Doc. **53**, 263–273 (1997)
259. L. Bornmann, R. Mutz, C. Neuhaus, H.-D. Daniel, Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics Sci. Environ. Polit. **8**, 93–102 (2008)

260. L. Bornmann, H.-D. Daniel, The citation speed index: a useful bibliometric indicator to add to the *h*-index. J. Infometr. **4**, 444–446 (2010)
261. L. Leydesdorff, O. Amsterdamska, Dimensions of citation analysis. Sci. Technol. Hum. Values **15**, 305–335 (1990)
262. H. Moed, *Citation Analysis in Research Evaluation* (Springer, Netherlands, 2005)
263. W. Glänzel, B. Dchlemmer, B. Thijs, Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. Scientometrics **58**, 571–586 (2003)
264. H.F. Moed, M. Vriens, Possible inaccuracies occurring in citation analysis. J. Inf. Sci. **15**, 95–107 (1989)
265. H.F. Moed, Citation analysis of scientific journals and journal impact measures. Curr. Sci. **89**, 1990–1996 (2005)
266. B. Cronin, K. Overfelt, Citation—based auditing of academic performance. J. Am. Soc. Inf. Sci. **45**, 61–72 (1994)
267. G. Lewison, The frequencies of occurrence of scientific papers with authors of each initial letter and their variation with nationality. Scientometrics **37**, 401–416 (1996)
268. H.F. Moed, New developments in the use of citation analysis in research evaluation. Arch. Immunol. Ther. Exp. **57**, 13–18 (2009)
269. B. Cronin, D. Shaw, Identity-creators and image-makers: using citation analysis and thick description to put authors in their place. Scientometrics **54**, 31–49 (2002)
270. L. Egghe, R. Rousseau, Aging, obsolescence, impact, growth, and utilization: definitions and relations. J. Am. Soc. Inf. Sci. **51**, 1004–1017 (2000)
271. E. Archambault, V. Lariviere, History of the journal impact factor: contingencies and consequences. Scientometrics **7**, 635–649 (2009)
272. N. Scibata, Y. Kajikawa, K. Matsushima, Topological analysis of citation networks to discover the future core articles. J. Assoc. Inf. Sci. Technol. **68**, 872–882 (2007)
273. E. Otte, R. Rousseau, Social network analysis: a powerful strategy, also in the information sciences. J. Inf. Sci. **28**, 441–453 (2002)
274. M. Girvan, M.E.J. Newman, Community structure in social and biological networks. PNAS **99**, 7821–7826 (2002)
275. N. Shibata, Y. Kajikawa, Y. Takeda, K. Matsushima, Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation **28**, 758–775 (2008)
276. N.P. Hummon, P. Dereian, Connectivity in a citation networks: the development of DNA theory. Soc. Netw. **11**, 39–63 (1989)
277. M. Zitt, E. Bassecoulard, Delineating complex scientific fields by an hybrid lexical-citation method: an application to nanosciences. Inf. Process. Manag. **42**, 1513–1551 (2006)
278. M. Zitt, S. Ramanana-Rahary, E. Bassecoulard, Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalization. Scientometrics **63**, 373–401 (2005)
279. H. Small, Tracking and predicting growth areas in science. Scientometrics **68**, 595–610 (2006)
280. H. Small, Paradigms, citations, and maps of science: a personal history. J. Am. Soc. Inf. Sci. Technol. **54**, 394–399 (2003)
281. L. Leydesdorff, Betweeness centrality as an indicator of the interdisciplinarity of scientific journals. J. Am. Soc. Inf. Sci. Technol. **58**, 1303–1319 (2007)
282. W. Glänzel, U. Schoepflin, A bibliometric study on ageing and reception processes of scientific literature. J. Inf. Sci. **21**, 37–53 (1995)
283. T. Pollmann, Forgetting and the ageing of scientific publications. Scientometrics **47**, 43–54 (2000)
284. I. Rafols, L. Leydesdorff, Contents-based and algorithmic classification of journals: perspectives on the dynamics of scientific communication and indexer effects. J. Am. Soc. Inf. Sci. Technol. **60**, 1823–1835 (2009)
285. P.B. Coulter, *Measuring Inequality* (Westview Press, Boulder, CO, 1989)
286. M.E.D. Koenig, Determinants of expert judgment of research performance. Scientometrics **4**, 361–378 (1982)

287. M.E.D. Koenig, Bibliometric indicators versus expert opinion in assessing research performance. J. Am. Soc. Inf. Sci. **34**, 136–145 (1983)
288. L. Langfeldt, Decision-making and sources of bias. Expert panels evaluating research. Res. Eval. **13**, 51–62 (2004)
289. A. Bryman, *Quantity and Quality in Social Research* (Unwin Hyman, London, 1988)
290. L. Leydesdorff, Various methods for the maping of science. Scientometrics **11**, 295–324 (1987)
291. L. Leydesdorrf, T. Schank, Dynamic animations of journal maps: indicators of structural changes and interdisciplinary developments. J. Am. Soc. Inf. Sci. Technol. **59**, 1810–1818 (2008)
292. L. Leydesdorff, I. Rafols, Local emergence and global diffusion of research technologies: an exploration of patterns of network formation. J. Am. Soc. Inf. Sci. Technol. **62**, 846–860 (2011)
293. E. Bassencoulard, A. Lelu, M. Zitt, Mapping nanosciences by citation flows: a preliminary analysis. Scientometrics **70**, 859–880 (2007)
294. H.P.F. Peters, A.F.J. van Raan, Representations by direct multidimensional scaling. Co-word-based science maps of chemical engineering. Part I. Res. Policy **22**, 23–45 (1993)
295. H.P.F. Peters, A.F.J. van Raan, Representations by combined clustering and multidimensional scaling, Co-word-based science maps of chemical engineering. Part II. Res. Policy **22**, 47–71 (1993)
296. M. Zitt, R. Barre, A. Sigogneau, F. Laville, Territorial concentration and evolution of science and technology activities in the European Union: A descriptive analysis. Res. Policy **28**, 545–562 (1999)
297. C. Chen, Visualising semantic spaces and author co-citation networks in digital libraries. Inf. Process. Manag. **35**, 401–420 (1999)
298. L. Kay, N. Newman, J. Youtie, A.L. Porter, I. Rafols, Patent overlay mapping: visualizing technological distance. J. Assoc. Inf. Sci. Technol. **65**, 2432–2443 (2014)
299. C. Chen, CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. **57**, 359–377 (2006)
300. M. Zitt, S. Ramanana-Rahary, E. Bassencoulard, F. Laville, Potential science - technology spillovers in regions: an insight on geographic co-location of knowledge activities in the EU. Scientometrics **57**, 295–320 (2003)
301. C. Chen, *Information Visualization: Beyond the Horizon* (Springer, London, 2006)
302. L. Leydesdorff, Clusters and maps of science journals based on bi-connected graphs in Journal Citations Reports. J. Doc. **60**, 371–427 (2004)
303. P. van den Besselaar, G. Heimeriks, Mapping research topics using word-reference co-occurrences: a method and an exploratory case study. Scientometrics **68**, 377–399 (2006)
304. K. Börner, D.E. Poley, *Visual Insights: A Practical Guide to Making Sense of Data* (MIT Press, Cambridge, MA, 2014)
305. K. Börner, *Atlas of Knowledge: Anyone can Map* (MIT Press, Cambridge, MA, 2014)
306. K. Börner, T.N. Theriault, K.W. Boyack, Mapping science introduction: past, present and future. Bull. Am. Soc. Inf. Sci. Technol. **41**, 12–16 (2015)
307. F. de Moya-Anegon, B. Vargas-Quesada, V. Herrero-Solana, Z. Chinchilla-Rodriguez, E. Corera-lvarez, F.J. Munoz-Fernande, A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics **61**, 129–145 (2004)
308. R.R. Braam, H.F. Moed, A.F.J. van Raan, Mapping of science by combined co-citation and word analysis I. Structural aspects. JASIS **42**, 233–251 (1991)
309. R.R. Braam, H.F. Moed, A.F.J. van Raan, Mapping of science by combined co-citation and word analysis II. Dynamical aspects. JASIS **42**, 252–266 (1991)
310. A.M. Zoss, K. Börner, Mapping interactions within the evolving science of science and innovation policy community. Scientometrics **91**, 631–644 (2011)
311. B. Vargas-Quesada, F. de Maoya-Anegon, *Visualizing the Structure of Science* (Springer, Berlin, 2007)

312. K.W. Boyack, R. Klavans, K. Börner, Mapping the backbone of science. Scientometrics **64**, 351–374 (2005)
313. K.W. Boyack, D. Newman, R.J. Duhon, R. Klavans, M. Patek, J.R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner, Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. PloS One **6**, e18029 (2011)
314. G. Heimeriks, M. Hoerlesberger, P. van den Besselaar, Mapping communication and collaboration in heterogeneous research networks. Scientometrics **58**, 391–413 (2003)
315. K. Börner, Plug-and-play macroscopes. Commun. ACM **54**, 60–69 (2011)
316. Y.W. Chen, S. Fang, K. Börner, Mapping the development of scientometrics: 2002–2008. J. Libr. Sci. China **3**, 131–146 (2011)
317. K. Börner, C. Chen, K.W. Boyack, Vizualizing knowledge domains. Annu. Rev. Inf. Sci. Technol. **37**, 179–255 (2003)
318. F. de Moya-Anegon, B. Vargas-Quesada, Z. Chinchilla-Rodriguez, E. Corera-lvarez, F.J Munoz-Fernandez, V. Herrero-Solana, Visualizing the marrow of science. J. Am. Soc. Inf. Sci. Technol. **58**, 2167–2179 (2007)
319. R.J.W. Tijsen, A.F.J. van Raan, Mapping changes in science and technology bibliometric co-occurrence analysis of the R&D literature. Eval. Rev. **18**, 98–115 (1994)
320. R. Klavans, K. Boyack, Toward a consensus map of science. J. Am. Soc. Inf. Sci. Technol. **60**, 455–476 (2009)
321. A. Quirin, O. Cordon, J. Santamaria, B. Vargas-Quesada, F. de Moya-Anegon, A new variant of the pathfinder algorithm to generate large visual science maps in cubic time. Inf. Process. Manag. **4**, 1611–1623 (2008)
322. K.W. Boyack, Mapping knowledge domains: characterizing PNAS. PNAS **101**(Supplement 1), 5192–5199 (2004)
323. H. Small, A SCI-MAP case study: building a map of AIDS research. Scientometrics **30**, 229–241 (1994)
324. P. van den Besselaar, L. Leydesdorff, Mapping change in scientific specialities: a scientometric reconstruction of the developing of artificial intelligence. J. Am. Soc. Inf. Sci. **47**, 415–436 (1996)
325. A. Perianes-Rodriguez, C. Olmeda-Gomez, F. Moya-Anegon, Detecting, identifying and visualizing research groups in co-authorship networks. Scientometrics **82**, 307–319 (2010)
326. S. Wright, The roles of mutation, inbreeding, crossbreeding and selection in evolution, in *Proceedings of the Sixth International Congress on Genetics* vol. 1, pp. 356–366 (1932)
327. H. Small, Update on science mapping: creating large document spaces. Scientometrics **38**, 275–293 (1997)
328. H. Small, A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system. Scientometrics **41**, 125–133 (1998)
329. L. Bornmann, L. Waltman, The detection of "hot regions" in the geography of science—a visualization approach by using density maps. J. Informetr. **5**, 547–553 (2011)
330. L. Leydesdorff, I. Rafols, A global map of science based on the ISI subject categories. J. Am. Soc. Inf. Sci. Technol. **60**, 348–362 (2009)
331. I. Rafols, A.L. Porter, L. Leydesdorff, Science overlay maps: a new tool for research policy and library management. J. Am. Soc. Inf. Sci. Technol. **61**, 1871–1887 (2010)
332. C. Wagner, L. Leydesdorff, Mapping the network of global science: comparing international co-authorships from 1990 to 2000. Int. J. Technol. Glob. **1**, 185–208 (2005)
333. E.C.M. Noyons, A.F.J. van Raan, Advanced mapping of science and technology. Scientometrics **41**, 61–67 (1998)
334. K. Boyack, Using detailed maps of science to identify potential collaborators. Scientometrics **57**, 27–44 (2008)
335. R. Klavans, K.W. Boyack, Using global mapping to create more accurate document-level maps of research fields. J. Am. Soc. Inf. Sci. Technol. **62**, 1–18 (2011)
336. M. Zitt, S. Ramanana-Rahary, E. Bassecoulard, Correcting glasses help fair comparisons in international science landscape: country indicators as a function of ISI database delineation. Scientometrics **56**, 259–282 (2003)

337. K.W. Boyack, R. Klavans, Creation of a highly detailed, dynamic, global model and map of science. J. Am. Soc. Inf. Sci. Technol. **65**, 670–685 (2014)
338. A. Poter, I. Rafols, Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics **81**, 719–745 (2009)
339. L. Leydesdorff, S. Cozzen, P. van den Besselaar, Tracking areas of strategic importance using scientometric journal mapings. Res. Policy **23**, 217–229 (1994)
340. L. leydesdorff, S. Carley, I. Rafols, Global maps of science based on the new Web-of-Science categories. Scientometrics **94**, 589–593 (2013)
341. A. Scharnhorst, Citation networks, science landscapes and evolutionary strategies. Scientometrics **43**, 95–106 (1998)
342. G. Krampen, R. Becker, U. Wahner, L. Montada, On the validity of citation counting in science evaluation: content analyses of references and citations in psychological publications. Scientometrics **71**, 191–202 (2007)
343. S.M. Lawani, A.E. Bayer, Validity of citation criteria for assessing the influence of scientific publications: new evidence with peer assessment. J. Am. Soc. Inf. Sci. **34**, 59–66 (1983)
344. L. Egghe, Mathematical theory of citation. Scientometrics **43**, 57–62 (1990)
345. Y.A. Shenhav, Y. Haberfield, The various faces of scientific productivity: a contingency analysis. Qual. Quant. **22**, 365–380 (1988)
346. B.F. Reskin, Scientific productivity and the reward structure of science. Am. Sociol. Rev. **42**, 419–504 (1977)
347. P.D. Allison, Inequality and scientific productivity. Soc. Stud. Sci. **10**, 163–179 (1980)
348. S. Cole, T.J. Phelan, The scientific productivity of nations. Minerva **37**, 1–23 (1999)
349. S. Lee, B. Bozeman, The impact of research collaboration on scientific productivity. Soc. Stud. Sci. **35**, 673–702 (2005)
350. L. Meitzer, Scientific productivity in organizational settings. J. Soc. Issues **12**, 32–40 (1956)
351. B.A. Jacob, L. Lefgren, The impact of research grant funding on scientific productivity. J. Public Econ. **95**, 1168–1177 (2011)
352. A. Bonaccorsi, C. Daraio, Age effects in scientific productivity. Case of Italian National Research Council (CNR). Scientometrics **58**, 49–90 (2003)
353. D.K. Simonton, Creative productivity and age: a mathematical model based on a two-step cognitive process. Dev. Rev. **4**, 77–111 (1984)
354. A. Diamond, An economic model of the life-cycle research productivity of scientists. Scientometrics **6**, 189–196 (1984)
355. S. Kyvik, Age and scientific productivity. Differences between fields of learning. High. Educ. **19**, 37–55 (1990)
356. P. Seglen, D. Aksnes, Scientific productivity and group size: a bibliometric analysis of Norwegian microbiological research. Scientometrics **49**, 125–143 (2000)
357. A. Ramesh Babu, Y.P. Singh, Determinants of research productivity. Scientometrics **43**, 309–329 (1998)
358. K. Jaffe, M. Caicedo, M. Manzaranes, M. Gil, A. Rios, A. Florez, C. Montoreano, V. Davilla, Productivity in physical and chemical science predicts the future economic growth of developing countries better than other popular indices. PLOS ONE, e66239 (2013)
359. A. Bonaccorsi, C. Daraio, A robust nonparametric approach to the analysis of scientific productivity. Res. Eval. **12**, 47–69 (2003)
360. D. Lindsey, Production and citation measures in the sociology of science: the problem of multiple authorship. Soc. Stud. Sci. **10**, 145–162 (1980)
361. H.F. Moed, Bibliometric indicators reflect publication and management strategies. Scientometrics **47**, 323–346 (2000)
362. G.S. Howard, D.A. Cole, S.E. Maxwell, Research productivity in psychology based on publication in the journals of the American Psychological Association. Am. Psychol. **42**, 975–986 (1987)
363. L.B. Ellwein, M. Khachab, R.H. Waldman, Assessing research productivity: evaluating journal publication across academic departments. Acad. Med. **64**, 319–325 (1989)

364. I. Lukovits, P. Vinkler, Correct credit distribution: a model for sharing credit among coauthors. Soc. Indic. Res. **36**, 91–98 (1995)
365. F.J. Trueba, H. Guerrero, A robust formula to credit authors for their publications. Scientometrics **60**, 181–204 (2004)
366. L.B. Dizon, M.S.M. Sadorra, Patterns of publication by the staff of an international fisheries research center. Scientometrics **32**, 67–75 (1995)
367. P. Vinkler, Evaluation of the publication activity of research teams by means of scientometric indicators. Curr. Sci. **79**, 602–612 (2000)
368. L. Butler, What happens when funding is linked to publication counts? pp. 389–405 in. H.F. Moed, W. Glänzel, U. Schmoch, *Handbook of Quantitative Science and Technology Research* (Springer, Netherlands, 2005)
369. G.G. Dyumenton, *Networks of Scientific Communications and Organization of Fundamental Research* (Nauka, Moscow, 1987) (in Russian)
370. G. Abramo, T. Cicero, C.A. D'Angelo, How important is choice of the scaling factor in standardizing citations? J. Informetr. **6**, 645–654 (2012)
371. G. Abramo, T. Cicero, C.A. D'Angelo, Revisiting the scaling of citations for research assessment. J. Informetr. **6**, 470–4479 (2012)
372. B.R. Martin, J. Irvine, Assessing basic research. Some partial indicators of scientific progress in radioastronomy. Res. Policy **12**, 61–90 (1983)
373. M.J. Moravcsik, Progress report on quantification of science. J. Sci. Ind. Res. **36**, 195–203 (1977)
374. D. Lindsey, Using citation counts as a measure of quality in science: measuring what's measurable rather than what's valid. Scientometrics **15**, 189–203 (1989)
375. D. Lindsey, The corrected quality ratio: a composite index of scientific contribution to knowledge. Soc. Stud. Sci. **8**, 349–354 (1978)
376. M.J. Moravcsik, Some contextual problems of science indicators. pp. 11 – 30 in *Handbook of Quantitative Studies of Science and Technology*, ed. by A.F.J. Van Raan (Elsevier, Amsterdam, 1988)
377. P. Abelson, Mechanisms for evaluating scientific information and the role of peer review. J. Am. Soc. Inf. Sci. **41**(3), 216–222 (1990)
378. G. Abramo, C.A. D'Angelo, Evaluating research: from informed peer review to bibliometrics. Scientometrics **87**, 499–514 (2011)
379. A. Pouris, Evaluating academic science institutions in South Africa. J. Am. Soc. Inf. Sci. **40**, 269–372 (1989)
380. R. Miller, The influence of primary task on R&D laboratory evaluation: a comparative bibliometric analysis. R&D Manag. **22**, 3–20 (1992)
381. P. Vinkler, General performance indexes calculated for research institutes of the Hungarian Academy of Sciences based on scientometric indicators. Scientometrics **41**, 185–200 (1998)
382. T.N. van Leeuwen, L.J. van der Wurff, A.F.J. van Raan, The use of combined bibliometric methods in research funding policy. Res. Eval. **10**, 195–201 (2001)
383. A. Watson, UK research funding—Universities raise their game, but the money doesn't flow. Science **294**, 2448–2449 (2001)
384. H.F. Moed, UK research assessment exercises: informed judgments on research quality or quantity? Scientometrics **74**, 153–161 (2008)
385. M.H. MacRoberts, B.R. MacRoberts, Problems of citation analysis: a critical review. J. Am. Soc. Inf. Sci. **40**, 342–349 (1989)
386. J. Nicolaisen, Citation analysis. Ann. Rev. Inf. Sci. Technol. **41**, 609–642 (2007)
387. L. Georghiou, D. Roessner, Evaluating technology programs: tools and methods. Res. Policy **29**, 657–678 (2000)
388. M. Marzolla, Quantitative analysis of the Italian national scientific qualification. J. Infometr. **9**, 285–316 (2015)
389. K. Rørstad, D.W. Aksness, Publication rate expressed by age, gender and academic position—A large scale analysis of Norwegian academic staff. J. Infometr. **9**, 317–333 (2015)

390. R. Barre, S&T indicators for policy making in a changing science-society relationship, pp. 115–131 in *Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems*, ed. by H.F. Moed, W. Glänzel, U. Schmoch (Springer, Netherlands, 2005)
391. L. Egghe, Performance and its relation with productivity in Lotkaian systems. Scientometrics **81**, 567–585 (2009)
392. H. English, H.-J. Czerwon, Quantification of the performance of research units: a simple mathematical model. Res. Polit. **19**, 477–480 (1990)

Part II
Indicators and Indexes for Assessment of Research Production

Publications are an important outcome of scientific research, since they contain the knowledge produced by a researcher or group of researchers. Citations of research publications are an important measure of the impact of these publications. But how are publications and their citations connected to the quality of research? High-quality research may remain unrecognized for years or even decades, but in most cases, high-quality research publications obtain many citations in a short time after publication. At the same time, review papers and methodological papers from the same scientific area may be highly cited, too. All the above shows that the quality of a research product is a complex quantity that may be assessed by multidimensional analysis based on qualitative tools combined with quantitative indicators and indexes.

Quantitative indicators and indexes of research production are discussed in this part of the book. The part consists of two chapters. Chapter 2 contains description of commonly used indexes for assessment of research production. These indexes are based mainly on the citations of research publications of the evaluated researcher(s). The chapter begins with several general remarks about indicators and indexes and about assessment of research production on the basis of a researcher's publications and citations of those publications. Then the most popular index of recent years, the h-index of Hirsch, is discussed in detail together with a description of variants of the h-index, many- h-like indexes, as well as indexes complementary to the h-index. After that, another popular index, the g-index of Egghe, is discussed. A description of numerous other indexes follows, e.g., p-index, IQ_p-index, A-index, R-index, PI-indexes, indexes of personal success of a researcher, etc. Finally, several indexes are mentioned in connection with the rapidly growing research area devoted to research networks.

Chapter 3 is devoted mainly to additional indexes for assessment of research production of groups of researchers. This chapter contain indexes that usually require the assessed groups to be separated into components that contain some units of interest, e.g., the components may be the researchers from a research group and the units may be the research publications or citations of research publications. The following groups of indexes are described: simple indexes; indexes for deviation from a simple tendency; indexes for difference; indexes for concentration; indexes for imbalance and fragmentation; indexes for dissimilarity, coherence, and diversity; indexes of advantage; indexes based on the concept of entropy; the Lorenz curve and

associated indexes; the RELEV method and associated indexes for assessment of scientific research; indicators and indexes for comparison of research communities in different countries; indicators for leadership; indicators and indexes for assessment of national scientific production: impact factor, intermediacy index, SJR, etc. Finally, an example of the application of the Lorenz curve in a geometric approach for detecting research elites is discussed.

Chapter 2
Commonly Used Indexes for Assessment of Research Production

Dedicated to Eugene Garfield: a pioneer in the world of scientometrics and bibliometrics
16 September 2015

Abstract In this chapter, selected indicators and indexes (constructed on the basis of research publications and/or on the basis of a set of citations of these publications) are discussed. These indexes are frequently used for assessment of production of individual researchers. The chapter begins with several general remarks about indicators and indexes used in scientometrics. Then the famous h-index of Hirsch, its variants, and indexes complementary to the h-index are discussed. Next the g-index of Egghe as well as the i_n-indexes are described. The h-index, g-index, and i_n-indexes may provide a minimum of information for the quantitative part of assessment of the production of a researcher. Numerous indexes are described further in the text such as the m-index, p-index, IQ_p-index, A-index, R-index. The discussion of indexes continues with a discussion of indexes for the success of a researcher. In addition, a short list of indexes for quantitative characterization of research networks and their dynamics is presented.

2.1 Introductory Remarks

The research area connected to (i) construction of indexes for assessment of research production and (ii) the study of the properties of these indexes is very large and continues to grow. One could write entire books devoted to indicators and indexes. Below, we shall devote about 100 pages to indexes and indicators for assessment of research production of individual researchers and groups of researchers (a group may contain researchers from a department, research institute, university, systems of research institutes, or even a national research community). In order to discuss indexes as much as possible in this small number of pages, the following strategy

© Springer International Publishing Switzerland 2016

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1_2

will be adopted: The corresponding indexes and indicators will be described briefly. Their characteristics (positive or negative) will not be discussed in much detail. Instead, examples for calculation of indexes for two (actually existing) researchers from the same research field are presented. The reader may observe how each new index enlarges the knowledge of the evaluator about the characteristics of research production and about differences between the two researchers. In addition, numerous references are presented where the strengths and weaknesses of the indexes are discussed by competent researchers. We stress again the introductory character of this book. Researchers who want to study the characteristics of scientometric indexes in more detail may need another more extended approach, including, for example, the calculation of the indexes for various available databases; the study of relationships among indexes; methodologies for rescaling indexes calculated for different time intervals, etc. One such possible approach is presented and followed by Vinkler [1].

In Chaps. 2 and 3, the following practically oriented classification is adopted with respect to the indicators and indexes:

1. Commonly used indicators and indexes for evaluation of research mainly of individual researchers, Chap. 2. The indexes discussed are based mainly on citations obtained by the publications written by the evaluated researchers [2–4];
2. Additional indicators and indexes for evaluation of research of groups of researchers, Chap. 3. Indicators and indexes considered in this chapter will be connected both to the research publications of the evaluated group of researchers and to the citations of these publications.

Any of the above two classes of indexes and indicators may contain as subclasses the classes of indexes and indicators according to Vinkler [1, 5, 6], who proposed the following classification of indexes with respect of the number of sets they represent:

1. *Gross indexes (indicators)*: these refer to the measure of a single scientometric aspect of evaluated systems represented by a single scientometric set with a single hierarchical level. The gross indexes (indicators) may be represented by the following relationship:

$$G = \sum_{k=1}^{N} w_k i_k, \tag{2.1}$$

where i_k is the kth item in the corresponding set, and w_k is the respective weight. An example of a gross indicator is the number of publications of a research group published for the period of evaluation (bibliometric size of the research group).

Another example of a complex index connected to publications is the RPR-index (research potential realized index) [7]. Let N be the number of papers published in a journal (or the number of papers authored by a researcher, research group, research institute, etc.). Let N_c be the number of cited papers published in the journal (or the number of cited papers of the researcher, research group, etc.).

Then

$$\mathrm{RPR} = \frac{N_c}{N} = \frac{1}{N}\sum_{i=1}^{N} w_i, \tag{2.2}$$

where w_i is equal to 1 if the ith paper is cited, and equal to 0 if the ith paper is not cited.

2. *Complex indexes (indicators)*: these refer to two or more sets or to a single set with more than a single hierarchical level. For the case of two sets $\{A\}$ and $\{B\}$, these indexes may be represented by the relationship

$$C = f(A, B), \tag{2.3}$$

where f is an appropriate function acting on the two sets. An example of a relationship for a complex index is

$$C^* = \frac{\sum_{i=1}^{N_A} w_{ai} a_i}{\sum_{i=1}^{N_B} w_{bi} b_i}, \tag{2.4}$$

where w_{ai} and w_{bi} are respective weighting factors. An example of a complex index is the impact factor (number of citations obtained by a journal for some time period divided by the number of published papers in the journal for this time period) [8–29]:

$$G_i = \frac{C_i}{P_{i-1} + P_{i-2}}, \tag{2.5}$$

where C_i is the number of citations obtained in the year i by the papers published in a journal in the years $i-1$ and $i-2$ (the number of these papers is P_{i-1} and P_{i-2}). The impact factor introduced by Garfield stimulated many researchers to construct such kinds of indexes [30–40].

3. *Composite indexes (indicators)*: these consist of several gross or complex indexes (indicators), usually with some weighting factors, and each representing a special aspect of the evaluated system. The relationship for this kind of index (indicators) is

$$D = \sum_{i=1}^{N} w_i \left(\frac{a_i}{\sum_{i=1}^{M} a_i} \right), \tag{2.6}$$

where M is the number of evaluated research groups, $N \leq M$, and w_i is the respective weighting factor. An example of such an indicator is given in the RELEV method, which will be discussed in the next chapter of the book.

2.2 Peer Review and Assessment by Indicators and Indexes

Non est ponenda pluralitas sine necessitate.
(Do not introduce more arguments than are necessary.)
William of Ockham

In order to increase the effectiveness of scientific research, various officials often implement concepts such as the "value for money" concept in science [41]. Such concepts must be used carefully, for they can lead to unexpected side effects (e.g., soon after monkeys learned the concept of money, the first prostitute monkey appeared [42]). The above remarks lead to a practical question: How to measure "value" in science? In order to perform such measurements, policymakers increasingly use quantitative data [43]. On the basis of statistical analysis of these data, one may construct indexes to measure research activity. Such an approach has been applied for more than a century. In recent decades, activity around the construction of indexes and the study of their properties have been concentrated in several branches of science, e.g., in scientometrics and several related branches for the case of indexes for evaluation of research production. Assessment of research organizations is an important element of the process of research management and implementation of research policy [44, 45]. Administrators of science have two main instruments for evaluation of research organizations:

1. **Peer review**: evaluation of work by one or more people of similar competence as the producers of the work (peers) [46–66]. The main problem with this instrument is to find competent evaluators.
2. **Sets of indicators and indexes**: this instrument may lead to quick, easy, and inexpensive evaluation of research performance [67–69]. The main problem here is that if the indicators and indexes are inappropriate, then the result of evaluation will not be adequate.

Competition at different levels (from individuals to countries) has led to demand for comparative indicators for scientific and other achievements [70]. In addition, indicators and indexes may also be used for other purposes, e.g., for measuring growth of science [71]. Such types of indicators and indexes will be discussed in Chap. 3.

2.3 Several General Remarks About Indicators and Indexes

The number of indicators applied in evaluations should be reduced to the possible lowest but still sufficient number of indicators.
Peter Vinkler

There are different points of view on indicators and indexes [5, 72, 73]. Below, the indicators and indexes will be understood from the point of view of statistics [74], i.e.,

> **Indicator**: *an observed value of a variable, or in other words, a sign of the presence or absence of the concept being studied.*

Several indicators can be aggregated into a single index. Thus from the point of view of statistics, an index is

> **Index**: *a composite statistic—a measure of changes in a representative group of individual data points, or in other words, a compound measure that aggregates multiple indicators. Indexes summarize and rank specific observations.*

Below we present four classifications of indicators. The first classification of indicators of scientific research is

1. **Input indicators**: they are characteristics of the inputs of scientific organizations such as equipment; spent money; employed personnel.
2. **Output indicators**: they are characteristics of the results and outcomes of the research process. This class of indicators will be of interest for us below.

The second classification is:

1. **Absolute indicators**: they refer to one particular characteristic of research activity (number of articles published, money spent, number of citations, etc.).
2. **Relative indicators**: they refer to the relationship between two or more aspects such as number of articles per research group or the number of citations per paper.

Relative indicators often are more useful for research evaluation.

The third classification of indicators is from the point of view of the type of research. From this point of view, there are three classes of indicators:

1. **Basic research indicators**: These indicators are connected mainly to basic-research scientific papers and their citations.
2. **Experimental development indicators**: These indicators are connected mainly to patents and their citations.
3. **Applied research indicators**: These indicators are intermediate between the above two classes of indicators. They can be connected with applied research papers and their citations as well as with patents and their citations.

The fourth classification of indicators is from the point of view of the size of social systems and structures they measure. From this point of view, there exist the following classes of indicators [75–77]:

1. **Microindicators**: indicators connected with individuals; indicators connected with research groups; indicators connected with status/target groups.

2. **Mesoindicators**: indicators connected with university departments and university institutes; indicators connected with universities, research institutes, and funding agencies; indicators connected with academic fields; indicators connected with research and grant programs; indicators connected with cross-sectional fields.
3. **Macroindicators**: indicators connected with scientific policies; indicators connected with national research and development systems; indicators connected with global developments.

Below, we shall focus on indexes and indicators connected to research publications and their citations. Usually these indexes are statistical functions defined on sets of bibliometric elements and units, and because of this relative complexity, there are requirements on the indexes, e.g., the indexes must be *valid*, i.e., we have to be sure that we really measure what we are intending to measure. Any publication assessment method has to cover the amount of scientific information (e.g., number of scientific papers) produced by the evaluated researcher or group of researchers [1]; the acknowledgement of the published results (e.g., the number of citations) [78–84]; eminence of the publication channels. When used carefully, publication and citation data [85–87] are meaningful for measuring scientific output and its impact on the course of scientific research. The number of publications that a research group produces may represent its scientific production and its contribution to the generation of new knowledge (but be careful about duplication and the number of coauthors [88]. A scientist with famous collaborators may be highly cited. But this is not a sufficient condition for assessing a large contribution to the advancement of science).

Publications usually contain new facts, new hypotheses, new theories or theorems, new explanations, or new syntheses of existing facts. This is a contribution to science, and the number of citations of the above information is a measure of the contribution to the advancement of research in the corresponding scientific field. But this indicator also must be used carefully. The number of citations depends on research area (chemists are usually much more cited than the mathematicians); number of collaborators and their position in the various scientific networks and systems, etc. [89]. Publications and citations are connected to the visibility of individual researchers and research collectives. But not all publications are equally visible. Visibility depends on the place of publication; on the language of publication; on the scientific field; on the current "fashion" in scientific research; on the presence of publications in international scientific databases, etc. Thus visibility as a characteristic for evaluation of researchers and research groups and organizations must be used with care.

Publications are an important channel for communication of scientific results. And the number of publications may be a quantitative measure of scientific production. In general, one can consider two criteria for evaluation of research production on the basis of publications:

1. *External criteria:* number of articles, books, patents, etc. published by the scientist.
2. *Internal criteria:* number of preprints, number of given seminars, number of written internal reports, etc.

To some extent, citations are a measure of the value of the scientific production of the corresponding scientist [30, 90]. In addition, citations are an important measure of the influence of the scientific production of the researcher. The citation is regarded as the scientometric unit of impact of scientific information. Higher scientific impact is revealed by a larger number of citations.

The indexes of research performance usually depend on the size of the analyzed data set. This is especially interesting for indexes connected to citation data [91], since it is often assumed that the level of excellence of a scientist is a function of his/her full citation record. An interesting fact in [91] is that at least fifty papers are needed in order to obtain a conclusion about the long-term scientific performance of two scientific authors and to discriminate between them on the basis of an appropriate one-dimensional (single) index of scientific performance. This means that citation-based one-dimensional indicators and indexes of research performance have to be used for discrimination between mature scientists (such as candidates for a professorship who have produced fifty or more papers) and not between young researchers. And if one wants to discriminate between scientists who have produced fewer than fifty papers each, one should use a multidimensional indicator (a set of indicators).

It is more difficult to evaluate individual researchers and to compare their achievements in comparison to evaluation and comparison of achievements of research groups. The reasons for such difficulty is the smaller sets of publications and citations and the increasing importance of nonscientific factors such as age, position, education, personal connections, etc. Thus in addition to the numerous indexes used in the evaluation, one should also use qualitative evaluation methods. Below we shall discuss many indexes for characterization of the results of the work of individual researchers. And almost all of them will be connected to the citations of publications of a researcher, since a citation may be considered a unit of impact of the information produced by the researcher.

2.4 Additional Discussion on Citations as a Measure of Reception, Impact, and Quality of Research

Citations are usually used to measure the reception of research results obtained by the corresponding research community. Discussion about the use of citation-based indicators intensified when bibliometric indicators were not only begun to be used for monitoring national or institutional research performance, but when they also became components of formulas for the funding of scientific research [92]. In the area of research management and science policy, citations are often used to measure the impact of research publications, or they even become a measure of the quality of the corresponding publication. The validity of such an approach depends on the number of citations. If a publication is very highly cited, then its impact is high, and it may be that its quality is also good. But if an article is not so highly cited, then is it of low quality, or is its impact low? Such a determination cannot

be made immediately and without further investigation. Thus the use of the number of citations as a measure of impact or quality is not unproblematic, since there are many limitations, biases, or shortcomings connected to citation analysis [30, 93, 94]. Nevertheless, citations remain an important form of scientific information within the framework of documented science communication [95]. Not all citations are given, however, because of the quality of the cited paper [96]. Weinstock (in *Current Contents* # 12, 23 June 1971, reprint from [96]) gives some (fifteen, in fact) reasons for using citations:

1. Paying homage to pioneers.
2. Giving credit for related work.
3. Identifying methodology, equipment, etc.
4. Providing background reading.
5. Correcting one's own work.
6. Correcting the work of others.
7. Criticizing previous work.
8. Substantiating claims.
9. Alerting to forthcoming work.
10. Providing leads to poorly disseminated, poorly indexed, or uncited work.
11. Authenticating data and classes of facts: physical constants, etc.
12. Identifying original publications in which an idea or concept was discussed.
13. Identifying original publications or other work describing an eponymic concept or term (as, e.g., Hodgkin's disease, Pareto's Law, Friedels–Crafts reaction).
14. Disclaiming work or ideas of others.
15. Disputing priority claims of others.

As we can see, for example, item 5 from the above list is certainly not connected to the quality of the cited work.

In addition to individual citations, there are many cases in which larger sets of citations have to be assessed. This will be one of the subjects of the next chapter of this book. Here we shall mention just that such sets of citations may be influenced by citation cliques (which are able to filter information sources), and numerous self-citations may be presented in the set of citations of an individual researcher or in the set of citations of a group of researchers. Thus the individual citations as well as set of citations and especially the frequency of such citations hardly may be considered a measure of the quality of the cited work. Citations may give us information about the impact of the work of the researchers, and the self citations may give us some information too: a lack of self-citations over a longer period may indicate lack of originality in research. The presence of many self-citations may indicate a significant record of publication activity of the corresponding researcher or group of researchers.

There exist quantitative evaluations on the amount of self-citations. The study [97] led to the result that in the area of basic research, the average number of self-citations is about 20 % of the number of citations. Another estimate [98] obtained percentages between 10 and 30 %. These estimates are for synchronous self-citations (The rate of synchronous self-citations is calculated as the citations to oneself relative to the total number of references). Another possible rate for self-citations is the diachronous

rate (number of self-citations divided by the total number of citations received) [99]. Synchronous and diachronous self-citation rates can be calculated for individual scientists, groups of scientists, journals, etc. Glänzel and coauthors [100] even obtained a square root law $f(k) \approx (k + 1/4)^{1/2}$ between the number of self-citations f and the number of foreign citations k (foreign citations are the non-self-citations). This law shows that the self-citations and foreign citations are not independent, i.e., the self-citations may be an essential part of scientific communication.

Before beginning our discussion on the indexes used for assessment of research, let us note again that citation patterns are much influenced by subject characteristics. And the subject characteristics are different in different research fields, e.g., in chemistry and mathematics. Because of this, one should not use citations for cross-field comparison without appropriate normalization.

2.5 The *h*-Index of Hirsch

The *h*-index of Hirsch has become very popular in recent years [101–118]. Because of this, it is much discussed and modeled [119–133]. The *h*-index is defined as follows. Let us suppose that a certain scientist has N research publications. Let us rank these publications by decreasing number of the number of citations (The most cited paper is on the top of the list; second in the list is the second most cited paper, etc. The least cited paper is at the bottom of the list).

A scientist has h-index equal to H if the top H of his/her N publications from the ranked list have at least H citations each.

The *h*-index is the solution of the equation

$$r = C(r), \tag{2.7}$$

where $C(r)$ is the number of citations of the rth publication from the ranked list or articles of the researcher. We note that the other publications of the researcher will have no more than h citations each.

The *h*-index [134] was introduced on the basis of the intention to measure simultaneously the quality and quantity of scientific output. The *h*-index was introduced also because of the disadvantages of other bibliometric indicators, such as total number of papers (it does not account for the quality of scientific publications); total number of citations (this number may be disproportionately affected by participation in a single publication; large influence of a certain class of papers (the methodological papers that propose new techniques, methods, or approximations typically generate many citations); many publications with few citations each).

The main reason for the popularity of the *h*-index is its simplicity [135]. The *h*-index has been calculated also for journals, topics, etc. [136–141]. Let us note

the interesting research on correlations between the h-index and thirty-seven other similar indexes [142]. Several of these indexes will be described below.

Assuming that a researcher publishes a constant number of papers each year and that each published paper receives a constant number of citations per year (and this for each subsequent year), Hirsch [134] obtained two relationships when the publication time (which is approximately equal to the length of the scientific career of the scientist) is not too small. The relationships are

- *Relationship between total number of citations N and the Hirsch index h*,

$$N(t) \approx Ah^2(t). \tag{2.8}$$

- *Relationship between Hirsch index and the time t* (in years of research career),

$$h(t) \approx bt, \tag{2.9}$$

where A and b are some appropriate constants that can be different for different scientists; A has values between 3 and 5, and by b Hirsch classifies the scientists as

  - *successful*: $b = 1$;
  - *outstanding*: $b = 2$;
  - *unique*: $b = 3$.

Soon after its definition, the h-index was generalized to the h_α-index [143].

A scientist has h_α-index equal to H_α if the top H_α of his/her N publications from the ranked list have at least αH_α citations each.

If $\alpha = 1$, then $h_\alpha = h$. The h_α index has the following properties:

$$\lim_{\alpha \to 0} h_\alpha \sim p; \quad \lim_{\alpha \to \infty} h_\alpha \sim c, \tag{2.10}$$

where p is the number of papers published by the scientist that have been cited at least once and c is the number of citations of the most cited paper published by the scientist (these numbers can be called p-indicator and c-indicator).

2.5.1 Advantages and Disadvantages of the h-Index

The h-index is simple to calculate, and it encourages the performance of research work that is highly visible (and may be of high quality). In addition, the h-index is a measure of a combination of two important characteristics of research production: the number of publications and the citation impact of those publications. The h-index compares established scientists from the same scientific field. It does not discriminate

much among the average scientists in the field. If a researcher has published many highly visible papers, then his/her h-index may increase with the accumulation of citations even if he or she no longer publishes.

When using the h-index for evaluation of research production, one should keep in the mind that the h-index doesn't account for the typical number of citations in different scientific fields or for the typical number of citations in different journals. In addition, the h-index doesn't account for the number of authors of a paper. The index favors scientific fields with large numbers of researchers working in the field. Moreover, the index favors scientific fields with larger sizes of research groups working in the field. The h-index is bounded by the total number of publications: it favors scientists with a longer career. Scientists who have written a small number of papers but have important discoveries are at a disadvantage.

The h-index doesn't account for the place of the scientist in the author list of the paper. In addition, the h-index does not account for authorship without authorization (the name of a researcher is put in the list of the authors without his/her knowledge or permission). The h-index can be manipulated through self-citations [144–148]. h-index doesn't account for the context of citations. A citation can be made in a positive context, but a citation can also be made in a negative context. And some citations can be more significant for the citing paper. Finally, the h-index doesn't account for the citation bias connected to the review papers.

The h-index is an attempt to achieve a balance between scientific productivity and quality of scientific production [91]. This index, however, assumes an equality between incommensurable quantities: number of papers and number of citations of a paper. A more general relationship between these two quantities could be

$$r^{\alpha} = \beta C(r). \tag{2.11}$$

For the case of the h-index, $\alpha = \beta = 1$, and perhaps this is one of the simplest possible choices of the parameters α and β [149].

Let us note an interesting effect related to the h-index: the h-bubble [150]. This effect is connected to the rapidly increasing number of citations gained by the authors who first began to study the characteristics of the h-index [151]. It is assumed that this fast growth forms a bubble like a stock market bubble. The question is whether after the bubble there will be a crash. The future will answer this question.

In order to give some simple examples for calculation of the h-index and of some of the indexes described in the chapter below, we shall consider data about citations of the fifty most-cited publications for two actually existing researchers from the research area of applied mathematics. The ranked numbers of citations (the number of citations of the most-cited publication is listed first) data are as follows

1. **Researcher A (49 years old, 117 publications, 1375 citations)**:
   93, 73, 67, 65, 59, 44, 43, 42, 38, 36, 36, 35, 34, 33, 33, 32, 29, 29, 29, 28, 27, 27, 26, 23, 23, 21, 21, 21, 20, 20, 20, 19, 18, 17, 16, 15, 15, 13, 11, 10, 10, 8, 8, 7, 7, 6, 6, 6, 6, 5.
2. **Researcher B (63 years old, 260 publications, 1562 citations)**:
   113, 65, 58, 51, 49, 42, 41, 37, 36, 34, 31, 27, 27, 25, 24, 24, 23, 23, 22, 20, 18,

17, 17, 16, 16, 16, 16, 16, 14, 14, 14, 14, 13, 12, 11, 11, 11, 11, 11, 10, 10, 10, 10, 10, 9, 9, 9, 9, 9, 9.

The h-index of researcher A is $h_A = 23$. The h-index of researcher B is $h_B = 20$. Thus the younger researcher has a larger h-index. The value of a single index, however, is not enough for comparison of the characteristics of the research production of the two researchers. Below, the values of additional indexes will be calculated. In such a way, an evaluator may obtain a table of values of appropriate indexes, and this table may be used for the quantitative part of the assessment of the research production. Such a table for our two researchers will be presented below.

2.5.2 Normalized h-Index

One can consider a normalized Hirsch index

$$h^* = \frac{h}{N}, \tag{2.12}$$

where h is the Hirsch index of the researcher and N is the number of the researcher's publications. The value h^* (for large enough N) is closer to an intensive quantity in comparison to the extensive quantity h that in most cases increases in the course of a scientific career.

For our two researchers, the normalized h-index has the following values: $h_A^* = 0.1965$; $h_B^* = 0.0769$. Note that h_A^* is more than twice h_B^*. This is a serious difference that can give us a hint about the effectiveness of the two researchers with respect to the impact of the research information they produce.

Another normalization of the h-index was proposed in [152]. This normalized index is equal to the square of the h-index divided by the total number of the authorships of the papers (sum of the number of authors for all papers from the set of papers) that determine the h-index of the researcher. The idea of normalization of the h-index was developed further in [153] by construction of the MII-index. This index was constructed for institutions but can also be used for evaluation of a group of researchers who have written a sufficiently large number of papers. The definition of the MII-index is

$$\mathrm{MII} = \frac{h}{10^{\alpha} N^{\beta}}, \tag{2.13}$$

where

- h: the h-index of the scientist;
- N: number of papers published by the scientist;
- α: intercept of the line describing the dependence of the h-index on the number of publications in the $\log_{10}$-scale;
- β: slope of the line describing the dependence of the h-index on the number of publications in the $\log_{10}$-scale.

Construction of the $h(N)$ line is as follows. For the ith member of the group of scientists (or for each institution from the group of evaluated institutions), one plots on a log-log plot the point with coordinates (N_i, h_i). The resulting points are fitted by a regression line

$$\log_{10} h_i = \alpha + \beta \log_{10} N_i + \varepsilon_i, \tag{2.14}$$

and in such a way, one determines α and β.

The MII-index is constructed for comparison of the quality of research of institutions of different sizes. It can be applied also to a group of researchers with different productivities. A value of MII that is larger than 1 means that the corresponding researcher from the research group of interest performs better than the average in terms of its h-index. The MII-index can also be used for evaluation of performance of a research institute in a large enough group of institutes from the same research area.

2.5.3 Tapered h-Index

The tapered h-index [154] is an extension of the h-index introduced in order to account for the citations of all papers of a researcher (and not only for the h papers that are cited at least h times). The definition of the index is as follows:

$$h_T = \sum_{j=1}^{N} h_T(j), \tag{2.15}$$

where $h_T(j)$ is the score for the jth paper in the ranked list (with respect to citations) of the publications of the researcher. In other words, we assume that the researcher has N publications ranked by the number of citations $n_1 \geq n_2 \geq \cdots \geq n_N$. The number $h_T(j)$ is determined as follows:

$$\begin{aligned} h_T(j) &= \frac{n_j}{2j-1}, \ n_j \leq j, \\ h_T(j) &= \frac{j}{2j-1} + \sum_{i=j+1}^{n_j} \frac{1}{2i-1}, \ n_j > j. \end{aligned} \tag{2.16}$$

The tapered h-index is larger than the h-index and is an additional characteristic that can be used to evaluate production (and impact of this production) of researchers.

We leave the calculation of the tapered h-indexes for the top cited fifty publications of our two researchers to the interested reader. The contributions of the first five publications that are not included in the h-index to the tapered h-index of the two researchers are:

- **Researcher A**: 23/47, 23/49, 21/51, 21/53, 21/55;
- **Researcher B**: 18/41, 17/43, 17/45, 16/49, 16/51.

2.5.4 Temporally Bounded h-Index. Age-Dependent h-Index

In the temporally bounded version of the h-index, one counts the citations of the articles for some time interval (for the last five years, for example), and then one makes a list in which we rank the papers with respect to the number of these citations.

A scientist has a temporally bounded h-index H if the top H of his/her N papers from the list have at least H citations each for some time interval (for example, for the last five years).

The temporally bounded h-index allows a comparison between the impacts of the papers of scientists working in the same scientific area. For our two researchers, the temporally bounded h-index for their citations for the last five years is:

- **Researcher A**: $h_A^{temp} = 19$ (1041 citations for the last five years);
- **Researcher B**: $h_B^{temp} = 12$ (724 citations for the last five years).

The h-index can be made age-dependent. The classic h-index is the solution of the equation (2.7). We can think about an appropriate inclusion of the time in the h-index in order to compensate for the length of the scientific career of younger scientists. One possibility is as follows. Let

$$C_r^* = C(r)/a_r, \tag{2.17}$$

where a_r are the ages of the rth paper from the ranked list. Let us perform a ranking $C^*(r)$ of the papers with respect to the values of C_r^* Then we can define the age-dependent h-index as the point of intersection of the straight line $y = r$ and the curve $y = C^*(r)$, i.e., as the unique solution of

$$r = C^*(r) \tag{2.18}$$

2.5.5 The Problem of Multiple Authorship. $\overline{h}$-Index of Hirsch and gh-Index of Galam

Frequently, a publication has several coauthors [155–158]. Coauthorship can be used as a measure of scientific collaboration. On the basis of the observation of the coauthorship pattern, one can conclude that scientific collaboration has increased greatly during recent decades at different levels of aggregation, e.g., at the level of individual authors; at the level of collaboration between sectors such as universities, research institutes, and industry; and at the level of international collaboration.

There are many reasons why researchers collaborate. One list of such reasons is as follows [159]:

1. Access to expertise.
2. Access to equipment, resources, or "stuff" one doesn't have.
3. Improved access to funds.
4. To obtain prestige or visibility; for professional advancement.
5. Efficiency: multiplies hands and minds; easier to learn the tacit knowledge that goes with a technique.
6. To make progress more rapidly.
7. To tackle "bigger" problems (more important, more comprehensive, more difficult, global).
8. To enhance productivity.
9. To get to know people, to create a network, like an "invisible college".
10. To retool, learn new skills or techniques, usually to break into a new field, subfield, or problem.
11. To satisfy curiosity, intellectual interest.
12. To share the excitement of an area with other people.
13. To find flaws more efficiently, reduce errors and mistakes.
14. To keep one more focused on research, because others are counting on one to do so.
15. To reduce isolation, and to recharge one's energy and excitement.
16. To educate (a student, graduate student, or oneself).
17. To advance knowledge and learning.
18. For fun, amusement, and pleasure.

The classic version of the h-index does not account for multiple authorship [160]. Because of this, Hirsch [161] defined another index, called the $\overline{h}$ index, as follows: *A scientist has index $\overline{h}$ if $\overline{h}$ of his/her papers belong to his/her $\overline{h}$ core. A paper belongs to the $\overline{h}$ core of a scientist if it has $\geq \overline{h}$ citations and in addition belongs to the $\overline{h}$-core of each of the coauthors of the paper*. The $\overline{h}$-index shows one way to deal with multiple authorship in the process of evaluation of a researcher's scientific production. Another way has been proposed by Galam [162], who introduced the gh-index as follows. Let us consider the function $g(r, k)$ that describes the fraction of the publication assigned to the rth author in the list of authors for a publication that has k coauthors. Then $\sum_{r=1}^{k} g(r, k) = 1$. If an author has authored and coauthored T publications, then the fraction of publications that is assigned to this author will be

$$T_g = \sum_{i=1}^{T} g_i(r, k). \tag{2.19}$$

If the ith paper of the above set of T papers has n_i citations, then the fraction of citations that will be assigned to the rth author will be $n_i g_i(r, k)$. Then the fraction of citations that will be assigned to the investigated author will be

$$N_g = \sum_{i=1}^{T} n_i g_i(r,k). \tag{2.20}$$

There are different proposals for the form of the function $g(r,k)$. Several of them are:

- *Egalitarian allocation*: $g(r,k) = \frac{1}{k}$ [163];
- *Arithmetic allocation*: $g(r,k) = \frac{2(k+1-r)}{k(k+1)}$ [164];
- *Geometric allocation*: $g(r,k) = \frac{2^{1-r}}{2(1-2^{-k})}$ [165], etc.

Galam proposed an allocation with bonuses for the first and for the last author of a publication as follows. Let a publication have k coauthors. We consider the decreasing arithmetic series $k, k-1, \ldots, 2, 1$ and two bonuses: δ for the first author and μ for the last author. Let us call them the bonus of the hard worker (the first author) and the bonus of the boss (usually the last author). The sum of the above arithmetic series and of the two bonuses is $S_k = \frac{k(k+1)}{2} + \delta + \mu$. Then the function $g(r,k)$ becomes

$$\begin{aligned} g(1,k) &= \frac{k+\delta}{S_k}; \\ g(k,k) &= \frac{k-1+\mu}{S_k}; \\ g(r,k) &= \frac{k-r}{S_k}; \end{aligned} \tag{2.21}$$

and $g(1,k)$ and $g(k,k)$ are defined when the publication has more than two coauthors, and $g(r,k)$ is defined only when the publication has at least three coauthors.

The final step is to set the values of the bonuses δ and μ. These values have to be set by consensus. Possible relationships are $\delta = 2\mu$; $\delta = 3\mu + 1$; etc. [162]. After setting the values of the bonuses, one can calculate the effective number of citations $n_{\text{eff}}(i)$ of the ith paper by

$$n_{\text{eff}}(i) = n_i g_i(r,k), \tag{2.22}$$

and then one can calculate the gh-index simply by calculating the h-index for the set of $n_{\text{eff}}(i)$, $i = 1, \ldots, N$. The gh-index obtained in such a way has smaller value compared to h (the two indexes are equal only if the scientist has no coauthors for any publication).

Finally, let us note one more index that has to deal with the problem of coauthorship [166]. This index is called the *P-index* of a researcher. Its definition is

$$P = \sum_{k=1}^{K} A_k^* J_k, \tag{2.23}$$

where

- J_k: journal impact factor of the journal where the kth paper of the researcher was published;
- A_k^*: A^*-index of the kth paper of the researcher.

The A^*-index is defined as follows. Let an article of the researcher have n coauthors that can be separated into $m \leq n$ groups and in each of these groups, the coauthors have the same credit (say c_i for the ith group of coauthors). The value of A^* for a coauthor from the group i is then

$$A^*(i) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{\sum_{k=1}^{j} c_k}. \tag{2.24}$$

If no coauthors claim an equal contribution, then $m = n$, $c_i = 1$, and

$$A^*(i) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{j}. \tag{2.25}$$

2.5.6 The m-Index

The m-index has been proposed in [167]. In order to define it, one needs to know about the *Hirsch core*. In the process of calculation of the Hirsch index, the papers of the scientists of interest are ranked with respect to the number of citations each of them has obtained. The papers from the ranked list whose rank is less than or equal to h build the Hirsch core of the ranked list of the paper. Then the m-index is *the median number of citations received by the papers in the Hirsch core.*

The m-index focuses on the impact of publications with the highest citation counts and is a characteristic of the quality of the production of the evaluated scientist taken from the core of his/her most cited scientific production. The m-index for our two researchers is approximately:

- **Researcher A**: $m_A \approx 38.8$;
- **Researcher B**: $m_B \approx 34.6$.

The word approximately above was used because the multiplication of the number of citations for both researchers leads to very large number that is represented only approximately by the simplest calculators. So the m-index usually can be calculated only approximately for researchers whose h-factor is relatively large (e.g., greater than 15).

2.5.7 *h*-Like Indexes and Indexes Complementary to the Hirsch Index

One can define central area indexes and central interval indexes [168]. These indexes are connected to the Hirsch index and supply information complementary to the information obtained on the basis of the h-index. For example, the central area index of radius j is defined as follows:

$$A_j = (h-j)c_{h-j} + \sum_{i=h-j+1}^{h+j} c_i; \quad j = 1, \ldots, h-1, \tag{2.26}$$

where h is the h-index and c_i are the citations received by the ith-ranked publication of the scientist ($c_1 \geq c_2 \geq \cdots \geq c_n$ for a scientist who has n publications). The idea of this index is to reduce one of the negative effects of the Hirsch index, which penalizes authors with heavy tails in their citation distribution. The central area index for such authors increases faster in comparison to the central area index for authors whose least-cited papers have a small number of citations.

Generalizations of the h- and g-indexes are presented in [169], and the robustness of the corresponding set of indexes is investigated. The result is that the most robust of them is the h-index, which is most insensitive to the extreme values of the corresponding citation distribution. At the expense of this, the h-index has quite low discriminating power (many scientists with different citation distributions can have the same h-index of their citations).

In [170], Egghe develops further an idea of Glänzel and Schubert [171] about characteristic scores and scales (CSS). The original idea is to determine, on the rank-order citation distribution, a sequence of points ε_k, $k = 1, 2, \ldots$, and σ_k, $k = 1, 2, \ldots$, where the k are some ranks of papers and $\sigma_k = \gamma(\varepsilon_k)$ are the corresponding characteristic scores, i.e., the number of citations to the paper of rank $r = k$. The function $\gamma(r)$ gives the number of citations of the paper of rank k, and it is called the rank-order frequency function. In other words, let $\sigma_1 = \mu$ be the average number of citations of the paper authored by a scientist. Let us discard all papers with fewer citations than σ_1. The average number of citations of the remaining papers is $\sigma_2 > \sigma_1$. Let us remove the papers with fewer citations than σ_2. The average number of citations of the remaining papers is $\sigma_3 > \sigma_2$. This process can be continued (as long as set of remaining papers is not empty).

CSS is a set of indexes that characterize the distribution of citations of a scientist. As a multicomponent characteristic, it has the advantages of supplying evaluators with more information in comparison to the use of a single indicator (which is the average number of citations of the papers of the scientist. It can be shown that $\sigma_k = \mu^k$ for the case in which $\gamma(r)$ satisfies Lotka's law $\gamma(t) \propto r^{\alpha}$ [172] (we shall discuss the Lotka's law in greater detail in Part III of this book). The idea of Egghe in [170] is to base the characteristic scores and scales on the h-index instead of on the average number of citations. For the case of validity of the Lotka's law, this leads

to a sequence of values

$$h_k = \frac{-\sum_{j=0}^{k-1} h_j + \left[\left(\sum_{j=1}^{k-1} h_j\right)^2\right]^{1/2}}{2}, \tag{2.27}$$

where $h_0 = h$ and $h_1 = h\frac{\sqrt{5}-1}{2}$ (h is the value of the h-index). Of course, the CSS can also be based on other indexes (on the g-index, for example).

Finally, let us discuss three recently introduced indexes that are complementary to the h-index [173, 174]. These indexes are called the perfectionism index (PIX), the extreme perfectionism index (EPIX), and the academic trace. Our notation is different from that in [173] in order not to confuse these indexes with the productivity indexes (PI) that will be discussed below.

Let us assume a researcher who has published p papers, and these publications have been cited C times. We recall that the h-index of the researcher separates his or her publications into two groups: the core (the h publications that are cited at least h times) and the tail (the other $p - h$ publications). Let the number of citations of the publications from the core be C_H and the number of citations of publications from the tail be C_T ($C = C_h + C_T$). We define the following two quantities:

- $C_E = C_H - h^2$: this quantity accounts for the eventual large number of citations in the core area;
- $C_{TC} = h(p-h) - C_T$: this quantity penalizes researchers who wrote many papers that are not much cited (the mass producers).

Then the perfectionism index is

$$\mathrm{PIX} = \kappa h^2 + \lambda C_E - \nu C_{TC}, \tag{2.28}$$

where κ, λ, and ν are real numbers.

In order to define the extreme perfectionism index, we need also

- $C_{IC} = \overset{*}{\sum} p - C_i$,

where $\overset{*}{\sum}$ means summation over all publications whose number of citations C_i is less than (the number of publications) p. Then the extreme perfectionism index is

$$\mathrm{EPIX} = \kappa h^2 + \lambda C_E + \mu C_T - \nu C_{IC}, \tag{2.29}$$

where κ, λ, μ, and ν are real numbers. The value of these numbers must be fixed, and the proposal to do this from [173] is just to set all of them to 1 (or to set some of them to 1 and the others to 0). Let us set the values of the parameters to 1. Then from (2.28), we obtain

$$\mathrm{PIX} = C + h(h - p). \tag{2.30}$$

If a researcher has 65 publications with 900 citations and h-index equal to 20, then PIX $= 0$. If the researcher has 1000 citations, then PIX $= 100$. If the researcher has 500 citations, then PIX $= -400$. The classification of the influential scientists and mass producers is:

1. If a researcher has PIX > 0, then he/she is an influential scientist;
2. If a researcher has PIX < 0, then he/she is a mass producer.

The academic trace index is defined as follows [174]:

$$T = \frac{h^2}{p} + \frac{C_T^2}{C} + \frac{C_E^2}{C} - \frac{p_0^2}{p}, \tag{2.31}$$

where

- p_0: number of publications that are not cited.

Another interesting index complementary to the h-index is defined in [152]. This index is

$$h_I = \frac{h^2}{N_a^{(T)}}, \tag{2.32}$$

where h is the h-index and $N_a^{(T)}$ is the total number of authors of the h-core of the corresponding author (multiple author occurrences in different papers is counted, e.g., if an author is coauthor in k papers, then he/she is counted k times).

It is claimed in [152] *that the h_I index rank plots collapse into a single curve. This is an important property, since in such a case, on the basis of the h_I-index, one can compare scientists from different scientific fields.*

Dorogovtsev and Mendes [175] note that the use of only the h-index for assessment of research may lead to a reshaping of research behavior: misleading citation-based targets may substitute for the real aims of scientific research: strong results. If h is the value of the h-index and C is the number of citations of the articles of a researcher, then the region of small values of the relationship $h/\sqrt{C}$ (i.e., the region where the researcher has a small number of very good articles that are highly cited) is occupied by outstanding researchers [105]. An interesting conclusion in [175] is that for given C, the h-index usually decreases with increasing $\langle c \rangle = C/N$ (i.e., with increasing mean number of citations per paper, the h index decreases). Thus *it seems that the h-index favors modestly performing scientists and punishes stronger researchers with a large mean number of citations per paper*. In order to make a better ranking of evaluated scientists on the basis of a single metric, the o-index was proposed in [175]:

$$o = \sqrt{\tilde{m}h}, \tag{2.33}$$

where h is the value of the h-index of the researcher and $\tilde{m}$ is the number of citations of the most cited paper of the same researcher. The motivation for such an index

is that $\tilde{m}$ accounts for the best result of the researcher, and h accounts for his/her persistence and diligence. In order to relate the o-index to the number of citations C of the researcher and to the mean number of citations per paper $\langle c \rangle$, one may use the following estimates: $h \sim \sqrt{C}$, and the mean number of citations per paper is between $n_1 = C$ and $n_2 = C/N$. Thus one may assume that $\tilde{m} \sim C/\sqrt{N}$ ($\tilde{m}^2 \sim n_1 n_2$). Then

$$o \sim C^{3/4} N^{-1/4} = C^{1/2} \langle c \rangle^{1/4}. \tag{2.34}$$

Thus the o-index should grow with the average number of citations per paper.

The o-index considered above grows much faster with the number of citations than with the average number of citations per paper. If we want to put more weight on the average number of citations per paper, we can generalize the o-index as follows:

$$o_{\alpha,\beta} = h^{\alpha} \tilde{m}^{\beta}. \tag{2.35}$$

Then

$$o_{\alpha,\beta} \sim C^{\alpha/2+\beta} N^{-\beta/2} = C^{(\alpha+\beta)/2} \langle c \rangle^{\beta/2}. \tag{2.36}$$

The o-index from [175] is

$$o = o_{1/2,1/2}. \tag{2.37}$$

Let $\beta > 0$. Then if $-\beta < \alpha < 0$, we have $\alpha + \beta < \beta$, which ensures a large weight of $\langle c \rangle$. For example, let $\alpha = -\beta + \delta$, where $\delta > 0$. Then

$$o_{\alpha,\beta} \sim C^{\delta}/2 \langle c \rangle^{\beta/2}. \tag{2.38}$$

If δ is 0 or very close to 0, then the contribution of C to the index could be very small.

Many other variants of h-indexes and h-like indexes exist. Let us note several of them:

1. The two-sided h-index [176], which accounts for the papers and citations out of the Hirsch core and allows comparison of researchers with the same values of the h-index.
2. The self-citations correction to the h-index [177] and to the g-index [178] (for discussion of the g-index, see Sect. 3.5).
3. Multidimensional extension of the h-index [179].
4. Successive h-indexes [180].
5. h-type index of coauthor partnership ability [181].
6. q^2-index uses the number and impact of papers in the Hirsch core [182],

$$q^2 = \sqrt{hm}, \tag{2.39}$$

where h is the Hirsch index and m denotes the median number of citations received by papers in the h-core of the corresponding set of articles (this is the m-index discussed above). The q^2-index is designed to supply a more global view of the

scientific production of researchers, since it is based on two indices that describe different dimensions of the research output: the h-index describes the number of papers (quantitative dimension) in a researcher's productive core, while the m-index is connected to the impact of research output.

7. The hg-index, which is the geometric mean of the product of the h-index and g-index [183, 184]:

$$hg = \sqrt{h \times g}. \tag{2.40}$$

The value of the hg-index is between the value of the h-index of Hirsch and g-index of Egghe: $h \leq hg \leq g$.

2.6 The g-Index of Egghe

Another very popular index based on the number of citations of the publications of a researcher is the g-index [185–188]. Let us make an ordered list of the papers of a researcher, and the order criterion is the number of citations: the most-cited paper is at the top of the list, the second-most-cited paper is at place 2 of the list, and the least-cited paper is at the bottom of the list. Then:

> *The g-index is the largest natural number g such that the top g articles received (together) at least g^2 citations.*

The g-index accounts for the number of citations of the highly cited papers of a scientist. The citations from higher-cited papers are used to bolster lower-cited papers. Because of this, the value of the g-index is at least equal to the value of the h-index, and in most cases, the g-index has a larger value than the h-index of the corresponding scientist.

The g-index can be generalized as follows [143]. The g-index above is restricted to integer values. One can define a g^*-index that is not restricted to integer values. Let x_i, $i = 1, \ldots, N$, be the number of citations of the ith article of a researcher ordered in such a way that $x_1 \geq x_2 \geq \cdots \geq x_N$. Let $x(u)$ be a function that approximates the values of the sequence x_i. Then one can define a continuous version of the g-index:

$$g^* = \max\{\mathrm{u} \mid \int_0^{\mathrm{u}} \mathrm{dv}\ \mathrm{x(v)} \geq \mathrm{u}^2\}. \tag{2.41}$$

The g^*-index is connected to the g-index as follows: $g \leq g^* < g + 1$. g^*-index can be generalized further. One can define the g^*_α-index as follows:

$$g^*_\alpha = \max\{\mathrm{u} \mid \int_0^{\mathrm{u}} \mathrm{dv}\ \mathrm{x(v)} \geq \alpha \mathrm{u}^2.\} \tag{2.42}$$

It is clear that when $\alpha = 1$, g^*_α reduces to g^*. In addition,

$$\lim_{\alpha \to 0} g^*_\alpha \sim s; \quad \lim_{\alpha \to \infty} g^*_\alpha \sim c, \tag{2.43}$$

where $s = \sum_{i=1}^{N} x_i$ is the total number of citations of all papers published by the scientist and c is the c-indicator defined above i devoted to the h_α-index.

The g-index for our two researchers is as follows:

- **Researcher A**: $g_A = 33$;
- **Researcher B**: $g_B = 30$.

Let us note that the larger values of the g-index are more difficult to reach. Researcher B has 946 citations of his 31 most-cited publications; 960 citations of his 32 most-cited publications and 973 citations of his 33 most-cited publications. In order to reach a g-index of 31, he will need an additional 15 $(29 - 14)$ citations of his top-cited 31 publications. In order to reach a g-index of 32 after reaching $g = 31$, he will need an additional 49 $(63 - 14)$ citations of his top-cited 32 publications. Finally, in order to reach $g = 33$ from $g = 32$, he will need an additional 52 $(65 - 13)$ citations of his top-cited 33 publications.

The g-index can be temporally bounded. *The temporally bounded g-index is the largest natural number g such that the top g articles received (together) at least g^2 citations for some time interval (for example, for the last five years)*. The temporally bounded g-index allows for a comparison between the impacts of the papers of scientists working in the same scientific area. The g-index can be modified in order to account for multiauthorship of publications [189, 190].

Similar to the gh-index discussed above, one can obtain also a gg-index on the basis of the effective citations of the papers of the scientists as calculated by (2.22). There is a discussion as to whether the h-index or g-index is better [191]. Our experience shows that each of the two indexes gives a piece of information about the performance of researchers, and these pieces of information are not the same. Thus we recommend the use of both indexes together. For example, if one has to evaluate established researchers from the same research area of the natural sciences (on the occasion of competition for some award or some high academic position), then the set of the h-index and g-index is a good choice for a minimum set of indexes that can give an initial impression about the quantitative aspects of the results of the scientific work of the candidates.

2.7 The i_n-Index

This index simply counts the number of papers of the scientist that are cited more than n times. For example, the i_{10} index (used in Goggle Scholar) counts the number of papers that are cited more than ten times. There are two versions of this index:

Nonbounded i_n-index: *This index counts the number of papers of the scientist that are cited n times for the time of the scientist's entire scientific career.*

and

Temporally bounded i_n-index: *This index counts the number of papers of the scientist that are cited n times for some time interval (for example, for the last five years).*

The temporally bounded i_n-index allows a comparison between the impacts of the papers of scientists working in the same scientific area. *The combination of the h-index, g-index, and several i_n indexes is another candidate for a set of indexes that may give a good initial impression about the quantitative aspects of the production of the evaluated researchers.*

The i_n indexes for our two researchers are as follows:

- **Researcher A**: $i_{100} = 0$; $i_{50} = 5$; $i_{30} = 16$; $i_{10} = 41$;
- **Researcher B**: $i_{100} = 1$; $i_{50} = 4$; $i_{30} = 11$; $i_{10} = 44$.

Interesting is the temporally bounded i_{10} index for the two researchers for the last five years. It is:

- **Researcher A**: $i_{10}^{temp} = 36$;
- **Researcher B**: $i_{10}^{temp} = 16$,

which shows that many more units of scientific information of researcher A (36 publications) are recognized as relatively important in comparison with the units of research information (16 publications) of researcher B. But the longer research career of researcher B has led to a larger value of his non-temporally bounded i_{10}-index. With respect to the i_{30} and i_{50} indexes, researcher A has already an advantage (despite the shorter research career). Researcher B still has a lead with respect to i_{100}.

2.8 *p*-Index. IQ_p-Index

The *p*-index was introduced by Prathap [192, 193] on the basis of the exergy indicator

$$X = k^2 P, \tag{2.44}$$

where P is the number of papers published by a scientist and $k = C/P$ is the ratio of the number of citations C of the P papers published by the scientist. The *p*-index is defined on the basis of the indicator X as follows:

$$p = X^{1/3} = \left(k^2 P\right)^{1/3}. \tag{2.45}$$

The p-index is designed as a joint measure of publication–citation activity of a researcher. The values of this index for our two researchers are

- **Researcher A**: $p_A = 25.28$;
- **Researcher B**: $p_B = 21.09$.

The larger value of the p-index for researcher A is due to his better ratio between obtained citations and research publications. This ratio participates at power 2 in the index and compensates for the twice larger number of publications of researcher B.

The IQ_p-index was introduced in [194] to measure the impact of a researcher along two dimensions: production (output, which is measured by the number of publications) and quality (measured by the number of citations). In order to define this index, one has to introduce a quantity called estimated citations E. It is defined as

$$E = \frac{ca(p+1)}{2}, \tag{2.46}$$

where

- a: age of the researcher;
- p: number of papers written by the researcher;
- c: correction factor reflecting the citations an average article receives in a particular research area. The value of c is based on the weighted aggregate journal impact factor of the top three subject categories in which the person has been cited.

Then $IQ_p = QP$, where Q and P are the quality and production components of the index, defined as follows:

$$Q = \frac{C}{E}; \quad P = p\frac{E/p}{p + E/p}, \tag{2.47}$$

where C is the number of citations of the papers written by the scientist and the production P is measured by the number of adjusted papers [194]. The result is

$$IQ_p = \frac{C}{p + \frac{ac(p+1)}{2p}} \tag{2.48}$$

Note that the value of this index depends on the manner of counting citations and publications.

Let us calculate the IQ_p index for our two researchers. We shall avoid the unknown quantity c in the following manner. For researcher B, we shall assume $c = 1$. This will correspond to 1562 citations/260 publications. Then the value of c for researcher A will be (1375 citations/117 publications)/(1562 citations/260 publications) = 1.956. Then for the two researchers, the values of the index are as follows:

- **Researcher A**: $IQ_p^A = 8.316$;
- **Researcher B**: $IQ_p^B = 5.356$.

The IQ_p index assigns about a 60 % greater impact of researcher A in comparison to researcher B.

2.9 *A*-Index and *R*-Index

The equation for the *A*-index is [195]

$$A = \frac{1}{h}\sum_{i=1}^{h} C_i, \tag{2.49}$$

where

- h: the value of the h-index for the evaluated scientist.
- C_i: number of citations for the ith paper from the list of ranked papers connected with the h-index.

The *A*-index may be sensitive to the number of citations of highly cited papers. It can happen as follows. Let us suppose two scientists: Alain and Paul. The h-index of Paul is larger than the h-index of Alain. But the most-cited papers of Alain are much more frquently cited than the papers of Paul. Then it can happen that the *A*-index of Alain has a larger value than the *A*-index of Paul.

Because of the above, one often uses an additional index called the *R*-index (*R* is used because the index contains a square root). Its equation is

$$R = \sqrt{\sum_{i=1}^{h} C_i} = \sqrt{A \cdot h}, \tag{2.50}$$

where

- h: the value of the h-index for the evaluated scientist.
- C_i: number of citations for the ith paper from the list of ranked papers connected with the h-index.

The square root of the sum used in *R* leads to the consequence that the values of the index are not very large. In addition, there is no division by h, as in the case of *A*, and nevertheless, the values of the two indexes do not differ much.

The *R*-index never decreases. This happens even if the corresponding scientist has ended his or her publication activity. One way to deal with this is to define an age-dependent *R*-index. The equation for this index is [195]

$$R^* = \sqrt{\sum_{i=1}^{h} \frac{C_i}{a_i}}, \tag{2.51}$$

where

- h: the value of the h-index for the evaluated scientist.
- C_i: number of citations for the ith paper from the list of ranked papers connected with the h-index.

- a_i: age of the ith article.

On the basis of the R-index, a dynamic h-type index can be defined [196]. This index is

$$d_h(T) = R(T)\nu_h(T), \tag{2.52}$$

where $R(T)$ is the R-index, equal to the square root of the sum of all citations received by articles belonging to the h-core at time T, and $\nu_h(T)$ is the h-velocity at time T,

$$\nu_h(T) = \frac{dh}{dt} \mid_{t=T} = \lim_{t\to 0} \frac{h(T+t) - h(T)}{t}. \tag{2.53}$$

The definition of d_h contains three time-dependent elements: the size and contents of the h-core; the number of citations received; and the h-velocity. According to [196], the time $T = 0$ should be chosen not at the beginning of the researcher's career but five to ten years from the current moment of time (if the corresponding career is long enough). Then the function $h(T)$ should be fitted for determination of $\nu_h(T)$. There are several estimates of $h(T)$ [123, 197]. The estimate of Egghe [123] is

$$h(t) = [P_\infty C(t)^{\alpha-1}]^{1/\alpha}, \tag{2.54}$$

where $C(t)$ is the continuous citation distribution function; P_∞ is the number of publications at $t = \infty$; $\alpha > 1$ is the (Lotka) exponent for the citation function. Then

$$d_h(T) = R(T)\left[P_\infty(\alpha-1)C(t)^{\alpha-2}\frac{dC}{dt}\right]\frac{[P_\infty C(t)^{\alpha-1}]^{(1-\alpha)/\alpha}}{\alpha}. \tag{2.55}$$

The values of the A-index and of the R-index for our two researchers are

- **Researcher A**: $A_A = 40.6$; $R_A = 30.561$;
- **Researcher B**: $A_B = 38.6$; $R_B = 27.784$.

The values of the A-index reflect the fact that the number of citations per publication from the h-core of researcher A is larger than the corresponding number of citations per publication of researcher B. The values of the R-index reflect the fact that the number of citations for the publications from the h-core of researcher A is larger than the number of citations for the publications from the h-core of researcher B.

Let us end here the calculation of various indexes connected to the research production of the researchers A and B. We can summarize the obtained results as follows. We have calculated values only for a small number of the indexes discussed in this chapter. As an exercise, the interested reader may enlarge the table with the values of additional indexes. As one may see from Table 2.1, the values of the indexes give us compact quantitative information about the research production of researchers, and on the basis of the values of the indexes, we can compare the researchers. Such an evaluation should be made on the basis of a sufficiently large number of values

Table 2.1 Values of various indexes calculated for researchers A and B

Index	Researcher A	Researcher B	Advantage researcher A	Advantage researcher B
Research publications	117	260		+
citations	1375	1562		+
h-index	23	20	+	
Temporary bounded h-index (last 5 years)	19	12	+	
Normalized h-index	0.1965	0.0769	+	
g-index	33	30	+	
i_{100}-index	0	1		+
i_{50}-index	5	4	+	
i_{30}-index	16	11	+	
i_{10}-index	41	44		+
Temporally bounded i_{10}-index (last 5 years)	36	16	+	
m-index	38.8	34.6	+	
p-index	25.28	21.09	+	
IQ_p-index	8.316	5.356	+	
A-index	40.6	38.6	+	
R-index	30.561	27.784	+	

of indexes. And in addition to quantitative evaluation, qualitative evaluation (peer review, etc.) of research production of researchers should be made.

Now let us continue the discussion of the indexes.

2.10 More Indexes for Quantification of Research Production

2.10.1 Indexes Based on Normalization Mechanisms

1. *Index* B_1

   For a set of n papers, this index is defined as [198–201]

$$B_1 = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} e_i}, \tag{2.56}$$

where

- c_i: number of citations of the ith publication ($i = 1, 2, \ldots$);
- e_i: expected number of citations of the ith publication.

The expected number of citations e_i given the field and the year of publication is the average number of citations of all papers published in the same field and in the same year.

2. *Index* B_2
This index is defined as [198]

$$B_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i}{e_i}. \tag{2.57}$$

We note that the above two indexes should be used carefully for evaluation of sets of papers that are published too soon, since then, the expected number of citations e_i can have a relatively large difference in the values for different years.

2.10.2 PI-Indexes

The popularity of the Hirsch index is due in great part to the fact that it is a composite index, because its value depends not only on the number and distribution of citations over journal papers but also on the number of papers. One of the problems of the h-index is that it is not appropriate for analysis of publication performance of scientists with a relatively small number of publications. Such a situation can arise in mathematics, for example. There are highly cited scientists with a relatively small number of publications. As additional indexes for quantification of results of scientific production in such cases, one can use the *PI* indexes [202]

$$PI(\log) = \ln(pC^3), \tag{2.58}$$

where

- P: number of journal papers of the scientists;
- C: total number of citations obtained by the journal papers;

$$\begin{aligned} PI(C) &= 0.01(P + 2C); \\ PI(2C) &= 0.01(P + 1.5C + 2C_{3P}); \\ PI(3C) &= 0.01(P + 1.3C + 3C_{3P}); \end{aligned} \tag{2.59}$$

where P and C are as above and C_{3P} are the citations of the three most cited papers of the scientists.

One can imagine other kinds of PI indexes. For example,

$$PI_k = \ln(C_{kP})/(k), \tag{2.60}$$

where C_{kP} are the citations of the most-cited k papers, etc.

Another interesting kind of productivity index was introduced by Phelan [203, 204]. It is well suited for research fields in which the most important contributor is generally listed as the first author. In such fields, production might be better measured by an index that weights both first-author publications and citations. Such an index is

$$PI_i = \left(\frac{p_i c_i}{\sum\limits_k p_k c_k} \right)^{1/2}, \tag{2.61}$$

where p_i equals the total number of first-authored publications and c_i equals the total number of citations from first-authored publications. The sum is over all k first authors of papers in the research field or subfield of interest. The value of PI_i can be multiplied by 100 for ease of reference.

Vinkler [205] proposed also the index

$$\pi = 0.01 C_S \tag{2.62}$$

where C_S is the number of citations obtained from S of the most-cited papers of the researcher. The number S is obtained as follows. One takes all publications (whose number is P, for example) and ranks them with respect to the number of citations they have obtained. Then $S = \sqrt{P}$.

2.10.3 Indexes for Personal Success of a Researcher

The h-index is not a proper quantity by which to compare scientists from different scientific fields, because of different citing behavior, different numbers of scientists working in different scientific fields, etc. Wu [206] proposed a field-independent index of the personal success of a researcher as follows:

$$F = \frac{1}{K} \sum_{k=1}^{K} \sum_{i \in k; i \in N} \frac{C_i(t)}{D_k(t)} \tag{2.63}$$

where

- $k = \{1, \ldots, K\}$: index for numbering of subject categories in which the author has published;

- $i = \{1, \ldots, N\}$: index for numbering of published papers;
- $C_i(t)$: number of citations received up to some year of interest by the ith paper, published in the year t.
- $D_k(t)$: the average number of citations received up to the year of interest by all papers in the same publication year t as paper i and belonging to the same category k as the paper i.

Another kind of success index (s-index) was proposed in [207–210]. It is connected to the indicator called the NSP (number of successful papers) [211]. *From the point of view of NSP, a paper is successful if it has received more citations than the number of references in the list of references of the the paper*. The concept of a successful paper is refined further in the case of the success index. The paper i of a researcher is successful if its citations c_i are more numerous than the corresponding comparison term CT_i specific for the ith paper. In this case, the ith paper receives the score $\mathrm{sc}_i = 1$. If the paper is not successful with respect to CT_i, the ith paper receives the score $\mathrm{sc}_i = 0$. The s-index is the sum of the scores sc_i,

$$s = \sum_{i=1}^{p} \mathrm{sc}_i. \tag{2.64}$$

The question is how to constrict CT_i. Two possible constructions are [209] these:

- The average (or median) number of references made by the articles published in the same journal and year of the publication concerned.
- The average (or median) number of references made/received by a sample of publications representing the "neighborhood" of the publication concerned.

The success index s can be connected, for example, to the h-index and to the g-index. Let all CT_i equal χ. Then the success index can be written as

$$s(\chi) = \int_{\chi}^{\infty} dj\, f(j), \tag{2.65}$$

where $f(j)$ can be connected to an information-production process as follows. An information-production process has sources (for example, publications) that produce items (which are citations when publications are the sources). Then $f(j)$ is the density of the sources in item-density j. Let the size frequency function of the sources be a decreasing power law

$$f(j) = \frac{C}{j^{\alpha}};\ C > 0;\ \alpha \geq 1;\ j \geq 1 \tag{2.66}$$

(this power law is called Lotka's law and will be discussed in detail in Chap. 4). Then the success index is

$$s(\chi) = \frac{C^*}{\chi^{\alpha-1}};\quad C^* = \frac{C}{\alpha - 1}. \tag{2.67}$$

From the definition of success index $s(\chi)$, one can easily see that:

- If $\chi = h$, then the success index $s(\chi)$ is equal to the h-index of Hirsch;
- If

$$\chi = h\left(\frac{\alpha - 2}{\alpha - 1}\right)^{1/\alpha}, \tag{2.68}$$

then the success index $s(\chi)$ is equal to the g-index of Egghe.

Often, the personal success of a researcher is connected to his/her publication strategy. The publication strategy of a researcher can be characterized by two indexes: the PS-index (publication strategy index) [67, 201, 212] and the RPS-index (relative publication strategy index). These indexes use the impact factor of Garfield (see Sect. 3.16 from the next chapter). The indexes are defined as follows.

Publication strategy index

$$\mathrm{PS} = \left(\sum_{i=1}^{N} n_i G_i\right) / \left(\sum_{i=1}^{N} n_i\right), \tag{2.69}$$

where

- N: number of journals where the papers of the evaluated researcher (or evaluated research group) are published;
- n_i: number of papers published in the ith journal;
- G_i: impact factor of the ith journal.

The PS-index gives interesting additional information about the publication practices of the evaluated researchers. The index can be applied for monitoring the publication channels used by the evaluated researcher or group of researchers. Since researchers from different research fields use different channels, the value of the PS-index may depend greatly on the bibliometric characteristics of the research field. Because of this, the PS-index should be applied for comparison of sets of papers of authors working in similar research fields.

Relative publication strategy index

The RPS-index is calculated on the basis of the PS-index as follows:

$$\mathrm{RPS} = \frac{\mathrm{PS}}{G_m}, \tag{2.70}$$

where PS is the value of the PS-index and $G_m = \frac{1}{K}\sum_{i=1}^{K} G_i$ is the mean of the impact factors of some reference set of K journals.

2.10.4 Indexes for Characterization of Research Networks

The theory of networks [213, 214] (and especially its branch devoted to social networks) has already many applications in different areas of research. Here are several examples:

- biology [215, 216];
- epidemic spreading [217–219];
- crowd analysis [220];
- human dynamics and community detection [221–224];
- collaboration networks [225–232];
- consensus formation and agreement dynamics [233–235];
- study of spatial structures [236–238];
- structure and evolution of the Internet [239–241];
- rumor spreading [242, 243].

Network theory has also been applied to the area of study of dynamics of research structures and evaluation of research production [244–246]. We expect that in the course of the time, the number of these applications will grow steadily. Below, we give several examples of the use of concepts of network theory in the area of science dynamics and evaluation of research production.

- Schubert, Korn, and Telcs [247] have constructed two indexes of Hirsch type to characterize properties of networks of scientists. The basic concept of these indexes is the *degree h-index* of a network. This index is defined as follows:

> A network has a degree h-index of h if not more than h of its nodes have degree not less than h.

On the basis of the degree h-index of a network, two indexes have been constructed in [247]:

- Degree h-index of paper h_p: here the nodes of the network are the papers published in a journal, and the links are between papers that share at least one common author. Such a network of papers has degree h-index h_p if h_p is the largest number of papers in the network that have degree at least h_p.
- Degree h-index of authors h_A: in this case, the nodes of the network are the authors who publish in a journal. Links of the network are between authors that coauthored at least one paper in the studied journal. In such a case, the network of authors has degree h-index h_A, which is the largest number of authors in the network who have a degree at least h_A.

Networks are important for dynamics of science and scientific production [248, 249] (for example, an important element of scientific structure and processes is the collaboration networks or the networks connected to the citation of the results of scientific research). This importance is a factor for the increase in research on scientific networks and for the introduction of new indexes and indicators connected to these networks [250–254]. Let us mention several indexes and indicators for the reader's information.

- Network centrality in social networks has been much discussed since the famous paper of Freeman [255]. Network centrality refers to indicators and indexes that identify the most important vertices within a graph, connected to a certain (in our case scientific) network. An example of such an index is the C-index and its derivates [256]. This index presents a network centrality measure for collaborative competence. Another network centrality measure is given by the l-index (lobby index) [257, 258].
- Ausloos [259] measures the impact of the research of a scientist by means of his/her scientific network performance and defines the coauthor core in this network analogously to the core of papers defined by the h-index.

2.11 Concluding Remarks

A significant part of the discussion is this chapter was devoted to the h-index of Hirsch: to its strengths and weaknesses and to numerous h-like indexes and indexes complementary to the h-index. The reason for this is the popularity and widespread use of this index. Numerous other indexes are also discussed, and they may help evaluators to perform the quantitative part of the assessment of research production of individual researchers. It was demonstrated on the basis of data about citations of two researchers from the area of applied mathematics that these indexes may also provide useful information for the comparison of research production of researchers.

The indexes discussed above, e.g., the h-index, may also be calculated for research groups and departments as well as for research institutes, universities, and even for research communities of countries. Thus the indexes discussed in Chap. 2 may also be used for assessment of research production of groups containing many researchers.

We have applied numerous indexes above in the text in order to assess the research production of two researchers. The bibliometric analyses might go far beyond such computation and direct comparison of values of indexes. Analysis of links and relations in research networks and especially in copublication networks, analysis of citation impact, etc., may require a multidimensional approach and advanced data-analytical techniques such as cluster analysis or other data-analysis approaches that allow a simultaneous analysis of quantitative relationships among several variables. One example of an index that characterizes relations between two sets is the Jaccard

index [260, 261]. Let us suppose we have two sample sets A and B. If A and B are both empty, one sets the Jaccard index $J(A, B) = 0$. Otherwise,

$$J(A, B) = \frac{|\, A \cap B \,|}{|\, A \cup B \,|} = \frac{|\, A \cap B \,|}{|\, A \,| + |\, B \,| - |\, A \cap B \,|}. \tag{2.71}$$

The values of the Jaccard index are between 0 (inclusive) and 1 (inclusive). One can define the Jaccard distance

$$d_J(A, B) = 1 - J(A, B). \tag{2.72}$$

An example of a bibliometric application of the Jaccard index is as follows. Let us suppose we have a list of references, and let A and B be two sample sets of references from this list containing n_A and n_B references. Let n_{AB} be the number of references that are present in both lists A and B. Then the Jaccard index for the two sample sets is

$$J = \frac{n_{AB}}{n_A + n_B - n_{AB}}. \tag{2.73}$$

If the two lists of references are identical, then $J = 1$, and the corresponding Jaccard distance is $d_J = 0$. If the two lists of references are completely different (no references appear in both lists), then $J = 0$ and $d_J = 1$.

The results of multidimensional bibliometric analyses can be presented very effectively by various kinds of maps and landscapes, and because of this, the importance of these kinds of visualization techniques is increasing continuously. Additional indexes characterizing relations among sets of units will be described in the next chapter.

References

1. P. Vinkler, *The Evaluation of Research by Scientometric Indicators* (Chandos, Oxford, 2010)
2. J. King, A review of bibliometric and other science indicators and their role in research evaluation. J. Inf. Sci. **13**, 261–276 (1987)
3. R. Todorov, W. Glänzel, Journal citation measures: a concise review. J. Inf. Sci. **14**, 47–56 (1988)
4. H.F. Moed, Bibliometric indicators reflect publication and management strategies. Scientometrics **47**, 323–346 (2000)
5. P. Vinkler, An attempt for defining some basic categories of scientometrics and classifying the indicators of evaluative scientometrics. Scientometrics **50**, 539–544 (2001)
6. P. Vinkler, An attempt of surveying and classifying bibliometric indicators for scientometric purposes. Scientometrics **13**, 239–259 (1988)
7. L.M. Raisig, Mathematical evaluation of the scientific serial. Science **131**, 1417–1419 (1960)
8. E. Garfield, *Citation Indexing—its Theory and Application in Science, Technology, and Humanities* (Wiley, New York, 1979)
9. E. Garfield, Citation analysis as a tool in journal evaluation. Science **178**, 471–479 (1972)
10. E. Garfield, Journal impact factor: a brief review. Can. Med. Assoc. J. **161**, 979–980 (1999)
11. M. Zitt, H. Small, Modifying the journal impact factor by fractional citation weighting: the audience factor. J. Am. Soc. Inf. Sci. Technol. **59**, 1856–1860 (2008)

12. M. Zitt, Citing-side normalization of journal impact: a robust variant of the audience factor. J. informetr. **4**, 392–406 (2010)
13. M. Chew, E.V. Villanueva, M.B. van der Weyden, Life and times of the impact factor: retrospective analysis of trends for seven medical journals (1994–2005) and their Editors' views. J. R. Soc. Med. **100**, 142–150 (2007)
14. E. Garfield, The impact factor and using it correctly. Unfallchirurg **101**, 413–414 (1998)
15. E. Garfield, The history and meaning of the journal impact factor. J. Am. Med. Assoc. **295**, 90–93 (2006)
16. M. Zitt, The journal impact factor: Angel, devil, or scapegoat? A comment on J. K. Vanclay's article 2011. Scientometrics **92**, 485–503 (2012)
17. S.J. Bensman, Garfield and the impact factor. Annu. Rev. Inf. Sci. Technol. **41**, 93–155 (2007)
18. I. Marshakova-Shaikevich, The standard impact factor as an evaluation tool of science fields and scientific journals. Scientometrics **35**, 283–290 (1996)
19. R. Rousseau, Median and percentile impact factors: a set of new indicators. Scientometrics **63**, 431–441 (2005)
20. N. De Bellis, *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics* (Scarecrow Press, Lanham, MD, 2009)
21. G. Abramo, C.A. D'Angelo, F. Di Costa, Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? Scientometrics **84**, 821–833 (2010)
22. J.M. Campanario, Empirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation window. Scientometrics **87**, 189–204 (2011)
23. G. Buela-Casal, I. Zych, What do the scientists think about the impact factor? Scientometrics **92**, 281–292 (2012)
24. P. Ingwersen, The calculation of Web impact factors. J. Doc. **54**, 236–243 (1998)
25. P. Ingwersen, The pragmatics of a diachronic journal impact factor. Scientometrics **92**, 319–324 (2012)
26. M.R. Elkins, C.G. Maher, R.D. Herbert, A.M. Moseley, C. Sherrington, Correlation between the journal impact factor and three other journal citation indices. Scientometrics **85**, 81–93 (2010)
27. M.C. Calver, J.S. Bradley, Should we use the mean citations per paper to summarise a journal's impact or to rank journals in the same field? Scientometrics **81**, 611–615 (2009)
28. L. Leydesdorff, T. Opthof, Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. J. Am. Soc. Inf. Sci. Technol. **61**, 2365–2369 (2010)
29. M. Amin, M. Mabe, Impact factors: use and abuse. Perspect. Publ. **1**, 1–6 (2000)
30. H.F. Moed, *Citation Analysis in Research Evaluation* (Springer, Berlin, 2005)
31. B.K. Sen, Normalized impact factor. J. Doc. **48**, 318–329 (1992)
32. H.F. Moed, T.N. van Leeuwen, Improving the accuracy of Institute for Scientific Information's journal impact factors. J. Am. Soc. for Inf. Sci. **46**, 461–467 (1995)
33. H.F. Moed, T.N. van Leeuwen, J. Reedijk, Towards appropriate indicators of journal impact. Scientometrics **46**, 575–589 (1999)
34. B.K. Sen, K. Shailendra, Evaluation of recent scientific research output by a bibliometric method. Scientometrics **23**, 31–46 (1992)
35. A.M. Ramirez, E.O. Garcia, J.A. Del Rio, Renormalized impact factor. Scientometrics **47**, 3–9 (2000)
36. P.O. Seglen, Why the impact factor of journals should not be used for evaluating research. Br. Med. J. **314**, 498–502 (1997)
37. N. Sombatsompop, T. Markpin, W. Yochai, M. Saechiew, An evaluation of research performance for different subject categories using Impact Factor Point Average (IFPA) index: Thailand case study. Scientometrics **65**, 293–305 (2005)
38. R. Mansilla, E. Köppen, G. Cocho, P. Miramontes, On the behavior of journal impact factor rank-order distribution. J. informetr. **1**, 155–160 (2007)
39. P. Vinkler, Ratio of short term and long term impact factors and similarities of chemistry journals represented by references. Scientometrics **46**, 621–633 (1999)

40. W. Glänzel, H.F. Moed, Journal impact measures in bibliometric research. Scientometrics **53**, 171–193 (2002)
41. Value for money. Editorial. Nat. Mater. **8**, 535 (2009)
42. S.J. Dubner, S.D. Levitt, *Monkey Business* (The New York Times Magazine, 6 May 2005)
43. E. Geisler, *The Metrics of Science and Technology* (Quorum Books, Westport, CT, 2000)
44. B.R. Martin, J. Irvine, Assessing basic research. Some partial indicators of scientific progress in radio astronomy. Res. Policy **12**, 61–90 (1983)
45. B.R. Martin, The use of multiple indicators in the assessment of basic research. Scientometrics **36**, 343–362 (1996)
46. L. Bornmann, Scientific peer review. Annu. Rev. Inf. Sci. Technol. **45**, 197–245 (2011)
47. H.-D. Daniel, *Guardians of Science: Fairness and Reliability of Peer Review*. (VCH, Weinheim, 1993)
48. E. Rinia, T. van Leeuwen, H. van Vuren, A.F. van Raan, Comparative analysis of a set of bibliometric indicators and central peer review criteria. Res. Policy **27**, 95–107 (1998)
49. J.M. Campanario, Peer review for journals as it stands today—Part 1. Sci. Commun. **19**, 181–211 (1998)
50. J.M. Campanario, Peer review for journals as it stands today—Part 2. Sci. Commun. **19**, 277–306 (1998)
51. M. Reinhart, Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. Scientometrics **81**, 789–809 (2009)
52. A. Ragone, K. Mirylenka, F. Casati, M. Marchese, On peer review in computer science. Scientometrics **97**, 317–356 (2013)
53. S. Cole, L. Rubin, J.R. Cole, *Peer Review in the National Science Foundation. Phase One of a Study* (National Academy Press, Washington D.C., 1978)
54. J.R. Cole, S. Cole, *Peer review in the National Science Foundation. Phase Two of a Study* (National Academy Press, Washington D.C., 1981)
55. F. Godlee, T. Jefferson, *Peer Review in Health Sciences* (BMJ Books, London, 1999)
56. L. Butler, I. McAllister, Metrics or peer review? Evaluating the 2001 UK research assessment exercise in political science. Polit. Stud. Rev. **7**, 3–17 (2009)
57. T. Luukkonen, Conservatism and risk-taking in peer review: emerging ERC practices. Res. Eval. **21**, 48–60 (2012)
58. W.G.G. Benda, T.C.E. Engels, The predictive validity of peer review: a selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. Int. J. Forecast. **27**, 166–182 (2010)
59. L. Bornmann, H.-D. Daniel, Selection of Research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions. Scientometrics **63**, 297–320 (2005)
60. S. McKay, Social policy excellence—peer review or metrics? Analyzing the 2008 research assessment exercise in social work and social policy and administration. Soc. Policy Adm. **46**, 526–543 (2012)
61. L. Allen, C. Jones, K. Dolby, D. Lynn, M. Walport, Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. PLoS One **4**(6), e5910 (2009)
62. L. Bornmann, H.-D. Daniel, Selecting scientific excellence through committee peer review—A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. Scientometrics **68**, 427–440 (2006)
63. L. Bornmann, H.-D. Daniel, Selecting manuscripts for a high-impact journal through peer review: a citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. J. Am. Soc. Inf. Sci. Technol. **59**, 1841–1852 (2008)
64. L. Bornmann, G. Wallon, A. Ledin, Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. PloS One **3**, e3480 (2008)

65. V. Bence, C. Oppenheim, The influence of peer review on the research assessment exercise. J. Inf. Sci. **30**, 36–347 (2004)
66. L. Bornmann, H.-D. Daniel, Reliability, fairness and predictive validity of committee peer review. B.I.F. Futura **19**, 7–19 (2004)
67. A. Schubert, T. Braun, Relative indicators and relational charts for comparative assessment of publication output and citation impact. Scientometrics **9**, 281–291 (1986)
68. R. Plomp, The highly cited papers of professors as an indicator of a research group's scientific performance. Scientometrics **29**, 377–393 (1994)
69. Y. Elkana, J. Lederberg, R.K. Merton, A. Thackray, H. Zuckerman. *Toward a Metric of Science: the Advent of Science Indicators* (Wiley, New York, 1978)
70. C. Freeman, L. Soete, Developing science, technology and innovation indicators: what can we learn from the past. Res. Policy **38**, 583–589 (2009)
71. G.N. Gilbert, Measuring the growth of science: a review of indicators of scientific growth. Scientometrics **1**, 9–34 (1978)
72. T. Braun, W. Glänzel, A. Schubert, *Scientometrics Indicators. A 32 Country Comparison of Publication Productivity and Citation Impact* (World Scientific, London, 1985)
73. A.F.J. van Raan, Measuring science, in *Handbook of Quantitative Science and Technology Research*, eds. by H.F. Moed, W. Glänzel, U. Schmoch (Kluwer, Dordrecht, 2004), pp. 19–50
74. E. Babie, *The Practice of Social Research*, 13th edn. (Wadsworth, Australia, 2012)
75. W. Glänzel, B. Thijs, The influence of author self-citations on bibliometric macroindicators. Scientometrics **59**, 281–310 (2004)
76. S. Hinze, W. Glänzel. Scientometric indicators in use: An overview. Presentation at European summer school of scientometrics, Berlin (2013). http://www.scientometrics-school.eu/images/2_1_13Hinze.pdf
77. B. Thijs, W. Glänzel, The influence of author self-citations on bibliometric meso-indicators. The case of European universities. Scientometrics **66**, 71–80 (2006)
78. M.H. MacRoberts, B.R. MacRoberts, Problems of citation analysis: a critical review. J. Am. Soc. Inf. Sci. **40**, 342–349 (1989)
79. M.H. MacRoberts, B.R. MacRoberts, Problems of citation analysis. Scientometrics **36**, 435–444 (1996)
80. L.I. Meho, C.R. Sugimoto, Assessing the scholarly impact of information studies: a tale of two citation databases—Scopus and Web of Science. J. Am. Soc. Inf. Sci. Technol. **60**, 2499–2508 (2009)
81. R.N. Kostoff, Citation analysis of research performer quality. Scientometrics **53**, 49–71 (2002)
82. R.N. Kostoff, W.L. Martinez, Is citation normalization realistic? J. Inf. Sci. **31**, 57–61 (2005)
83. M.H. Medoff, The efficiency of self-citations in economics. Scientometrics **69**, 69–84 (2006)
84. K. Kousha, M. Thelwall, An automatic method for extracting citations from Google Books. J. Assoc. Inf. Sci. Technol. **66**, 309–320 (2015)
85. N.L. Geller, J.S. de Cani, R.E. Davies, Lifetime-citation rates to compare scientists' work. Soc. Sci. Res. **7**, 256–345 (1978)
86. N.L. Geller, J.S. de Cani, R.E. Davies, Lifetime-citation rates: a mathematical model to compare scientists' work. J. Am. Soc. Inf. Sci. **32**, 3–15 (1081)
87. C.S. Lin, M.H. Huang, D.Z. Chen, The influences of counting methods on university rankings based on paper count and citation count. J. Informetr. **7**, 611–621 (2013)
88. G. Abramo, C.A. D'Angelo, F. Rosati, The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in the life sciences. J. Informetr. **7**, 198–208 (2013)
89. L.M.A. Bettencourt, D.I. Kaiser, J. Kaur, Scientific discovery and topological transitions in collaboration networks. J. Informetr. **3**, 210–221 (2009)
90. D.W. Aksnes, A. Rip, Researchers' perceptions of citations. Res. Policy **38**, 895–905 (2009)
91. S. Lehman, A.D. Jackson, B.E. Lautrup, A quantitative analysis of indicators of scientific performance. Scientometrics **76**, 369–390 (2008)
92. W. Glänzel, K. Debackere, B. Thijs, A. Schubert, A concise review on the role of author self-citations in information science, bibliometrics and science policy. Scientometrics **67**, 263–277 (2006)

93. A.J. Chapman, Assessing research: citation-count shortcomings. Psychol.: Bull. Br. Psychol. Soc. **8**, 339–341 (1989)
94. P. Wouters, The citation culture. Academisch proefschrift ter verkrijging van de graad van doctor aan de Universiteit van Amsterdam (1999)
95. W. Glänzel, U. Schoepflin, A bibliometric study of reference literature in the sciences and social sciences. Inf. Process. Manag. **35**, 31–44 (1999)
96. N. Weinstock, Citation indexes, in *Encyclopedia of Library and Information Science* ed. by A. Kent, vol. 5 (Marcel Dekker, New York, 1971), pp. 16–41
97. E. Garfield, I.H. Sher, New factors in the evaluation of scientific literature through citation indexing. Am. Doc. **14**, 195–201 (1963)
98. M.H. MacRoberts, B.R. MacRoberts, Problems of citation analysis: a critical review. J. Am. Soc. Inf. Sci. **40**, 342–349 (1989)
99. S.M. Lawani, On the heterogeneity and classification of author self-citations. J. Am. Soc. Inf. Sci. **33**, 281–284 (1982)
100. W. Glänzel, B. Thijs, B. Schlemmer, A bibliometric approach to the role of author self-citations in scientific communication. Scientometrics **59**, 63–77 (2004)
101. S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera, *h*-index: a review focused in its variants, computation and standardization for different scientific fields. J. Informetr. **3**, 273–289 (2009)
102. P. Jasco, The pros and cons of computing *h*-index using Google Scholar. Online Inf. Rev. **32**, 437–452 (2008)
103. P. Jacso, The pros and cons of computing the h-index using Scopus. Online Inf. Rev. **32**, 524–535 (2008)
104. P. Jacso, Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. Libr. Trends **56**, 784–815 (2008)
105. S. Redner, On the meaning of the *h*-index. J. Stat. Mech.: Theory Exp. L03005 (2010)
106. E. Csajbok, A. Berhidi, L. Vasas, A. Schubert, Hirsch-index for countries based on Essential Science Indicators data. Scientometrics **73**, 91–117 (2007)
107. L. Bornmann, H.-D. Daniel, What do we know about *h*-index. J. Am. Soc. Inf. Sci. Technol. **58**, 1381–1385 (2007)
108. W. Glänzel, On the *h*-index—A mathematical approach to a new measure of publication activity and citation. Scientometrics **67**, 315–321 (2006)
109. L. Egghe, The influence of transformations on the *h*-index and the *g*-index. J. Am. Soc. Inf. Sci. Technol. **59**, 1304–1312 (2008)
110. W. Glänzel. On the opportunities and limitations of the *h*-index. Sci. Focus **1**, 10–11 (2006)
111. L. Bornmann, The state of *h*-index research. EMBO Rep. **10**, 2–6 (2009)
112. B. Cronin, L. Melo, Using the *h*—index to rank influential information scientists. J. Am. Soc. Inf. Sci. Technol. **57**, 1275–1278 (2006)
113. A. Schubert, Using the *h*-index for assessing single publications. Scientometrics **78**, 559–565 (2008)
114. W. Glänzel, *h*-index for price medalists. ISSI Newsl. **4**, 15–18 (2005)
115. R. Rousseau, Reflections on recent developments of the *h*-index and *h*-type of indices. COLLNET J. Sci. Inf. Manage. **2**, 1–8 (2008)
116. L. Egghe, I.K. Ravichandra Rao, Study of different *h*-indexes for groups of authors. J. Am. Soci. Inf. Sci. Technol. **59**, 1276–1281 (2008)
117. R. Rousseau, A case study: evolution of JASIS' Hirsch index. Sci. Focus **1**, 16–17 (2006)
118. L. Bornmann, R. Mutz, H.-D. Daniel, The *h*-index research output measurement: two approaches to enhance its accuracy. J. Informetr. **4**, 407–414 (2010)
119. L. Egghe, R. Rousseau, An infometric model for the Hirsch-index. Scientometrics **69**, 121–129 (2006)
120. R. Rousseau, The influence of missing publications on the Hirsch index. J. Informetr. **1**, 2–7 (2007)
121. M. Norris, C. Oppenheim, Peer review and the h-index: two studies. J. Informetr. **4**, 221–232 (2010)

122. M. Norris, C. Oppenheim, The h-index: a broad review of a new bibliometric indicator. J. Doc. **66**, 681–705 (2010)
123. L. Egghe, Dynamic *h*-index: the Hirsch index in function of time. J. Am. Soc. Inf. Sci. Technol. **58**, 452–454 (2007)
124. M. Henzinger, J. Sunol, I. Weber, The stability of the h-index. Scientometrics **84**, 465–479 (2009)
125. R. Burrows, Living with the h-index? Metric assemblages in the contemporary academy. Sociol. Rev. **60**, 355–372 (2012)
126. X. Hu, R. Rousseau, J. Chen, In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. J. Inf. Sci. **36**, 73–85 (2010)
127. L. Egghe, R. Rousseau, A *h*-index weighted by citation impact. Inf. Process. Manag. **44**, 770–780 (2008)
128. L. Egghe, Modelling successive *h*-indexes. Scientometrics **77**, 377–387 (2008)
129. J.E. Iglesias, C. Pecharroman, Scaling the h-index for different scientific ISI fields. Scientometrics **73**, 303–320 (2007)
130. R. Guns, R. Rousseau, Real and rational variants of the *h*-index and *g*-index. J. Informetr. **3**, 64–71 (2009)
131. L. Egghe, L. Liang, R. Rousseau, A relation between *h*-index and impact factor in the power-law model. J. Am. Soc. Inf. Sci. Technol. **60**, 2362–2365 (2009)
132. J. Bar-Ilan, Which *h*-index? A comparison of WoS Scopus and Google Scholar. Scientometrics **74**, 257–271 (2008)
133. L. Egghe, Mathematical theory of the *h*- and *g*-index in case of fractional counting of authorship. J. Am. Soc. Inf. Sci. Technol. **59**, 1608–1616 (2008)
134. J.E. Hirsch, An index to quantify an individual's scientific research output. PNAS **102**, 16569–16572 (2005)
135. B. Cronin, L.I. Meho, Using the *h*-index to rank influential information scientists. J. Am. Soc. Inf. Sci. Technol. **57**, 1275–1278 (2006)
136. T. Braun, W. Glänzel, A. Schubert, A Hirsch-type index for journals. Scientist **19**, 8 (2005)
137. A.W. Harzing, R. van der Wal, A Google Scholar h-index for journals: an alternative metric to measure journal impact in economics and business. J. Am. Soc. Inf. Sci. Technol. **60**, 41–46 (2009)
138. M.G. Banks, An extension of the Hirsch index: indexing scientific topics and compounds. Scientometrics **69**, 161–168 (2006)
139. A. Schubert, W. Glänzel, A systematic analysis of Hirsch-type indices for journals. J. Informetr. **1**, 179–184 (2007)
140. A.W.F. Edwards, System to rank scientists was pedaled by Jeffreys. Nature **437**(7061), 951 (2005)
141. F.Y. Ye, An investigation of mathematical models of the *h*-index. Scientometrics **81**, 493–498 (2009)
142. L. Bornmann, R. Mutz, S.E. Hug, H.-D. Daniel, A multilevel meta-analysis of studies reporting correlations between the *h* index and 37 different *h*-index variants. J. Informetr. **5**, 346–359 (2011)
143. N.J. van Eck, J. Waltman, Generalizing the h- and g-indices. J. Informetr. **2**(4), 263–271 (2008)
144. L. Engqvist, J.G. Frommen, New insights into the relationship between the h-index and self-citations? J. Am. Soc. Inf. Sci. Technol. **61**, 1514–1516 (2010)
145. J.H. Fowler, D.W. Aksnes, Does self-citation pay? Scientometrics **72**, 427–437 (2007)
146. M.H. Huang, W.Y.C. Lin, Probing the effect of author self-citations on h-index: a case study of environmental engineering. J. Inf. Sci. **37**, 453–461 (2011)
147. E. Gianoli, M.A. Molina-Montenegro, Insights into the relationship between the h-index and self-citations. J. Am. Soc. Inf. Sci. Technol. **60**, 1283–1285 (2009)
148. R. Costas, T.N. van Leeuwen, M. Bordons, Self-citations at the meso and individual levels: effects of different calculation methods. Scientometrics **82**, 517–537 (2010)
149. K.R. Dienes, Completing *h*. J. Informetr. **9**, 385–397 (2015)

150. R. Rousseau, C. Garcia-Zorita, E. Sanz-Casado, The *h*-bubble. J. Informetr. **7**, 294–300 (2013)
151. L. Zhang, B. Thijs, W. Glänzel, The diffusion of h-related literature. J. Informetr. **5**, 583–593 (2011)
152. P.D. Batista, M.G. Campiteli, O. Kinouchi, A.S. Martinez, Is it possible to compare researchers with different scientific interests? Scientometrics **68**, 179–189 (2006)
153. V. Sypsa, A. Hatzakis, Assessing the impact of biomedical research in academic institutions of disparate sizes. BMC Med. Res. Methodol. **9**, Article No. 33. doi:10.1186/1471-2288-9-33
154. T.R. Anderson, R.K.S. Hankin, P.D. Killworth, Beyond the Durfee square: enhancing the *h*-index to score total population output. Scientometrics **76**, 577–588 (2008)
155. N.T. Hagen, Harmonic allocation of authorship credit: source-level correction of bibliometric bias assures accurate publication and citation analysis. PLOS One **3**, e4021 (2008)
156. N.T. Hagen, Harmonic publication and citation counting: sharing authorship credit equitably—not equally, geometrically or arithmetically. Scientometrics **84**, 785–793 (2010)
157. X.Z. Liu, H. Fang, Fairly sharing the credit of multi-authored papers and its application in the modification of h-index and g-index. Scientometrics **91**, 37–49 (2012)
158. X.Z. Liu, H. Fang, Modifying h-index by allocating credit of multi-authored papers whose author names rank based on contribution. J. Informetr. **6**, 557–565 (2012)
159. D.B. de Beaver, Reflections on scientific collaboration (and its study): past, present, and future. Scientometrics **52**(2001), 365–377 (2001)
160. M. Schreiber, A modification of the *h*-index: the h_m-index accounts for multi-authored manuscripts. J. Informetr. **2**, 211–216 (2008)
161. J.E. Hirsch, An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics **85**, 741–754 (2010)
162. S. Galam, Tailor based allocations for multiple authorship: a fractional *gh*-index. Scientometrics **89**, 365–379 (2011)
163. D. de Solla Price, Multiple authorship. Science **212**(4498), 986–986 (1981)
164. G. van Hooydonk, Fractional counting of multiauthored publications: consequences for the impact of authors. J. Am. Soc. Inf. Sci. **48**, 944–945 (1997)
165. L. Egghe, R. Rousseau, G. van Hooydonk, Methods for acrediting publications to authors or countries: consequences for evaluation studies. J. Am. Soc. Inf. Sci. **51**, 145–157 (2000)
166. J. Stallings, E. Vance, J. Yang, M.W. Vanier, J. Liang, L. Pang, L. Dai, I. Ye, G. Wang, Determining scientific impact using a collaboration index. PNAS **110**, 9680–9685 (2013)
167. L. Bornmann, R. Mutz, H.-D. Daniel, Are there better indices for evaluation purposes than the *h*-index? A comparison of nine different variants of the *h*-index using data from biomedicine. J. Am. Soc. Inf. Sci. Technol. **59**, 830–837 (2008)
168. P. Dorta-Gonzales, M.-I. Dorta-Gonzales, Central indexes to the citation distribution: a complement to the *h*-index. Scientometrics **88**, 729–745 (2011)
169. W. Glänzel, A. Schubert, Hirsch-type characteristics of the tail of distributions. J. Informetr. **4**, 118–123 (2010)
170. L. Egghe, Characteristic scores and scales based on *h*-type indices. J. Informetr. **4**, 14–22 (2010)
171. W. Glänzel, A. Schubert, Characteristic scores and scales in assessing citation impact. J. Inf. Sci. **14**, 123–127 (1988)
172. L. Egghe, Characteristic scores and scales in a Lotkaian framework. Scientometrics **83**, 455–462 (2010)
173. A. Sidiropoulos, D. Katsaros, Y. Manolopoulos, Ranking and identifying influential scientists versus mass producers by the Perfectionism Index. Scientometrics **103**, 1–31 (2015)
174. F. Ye, L. Leydesdorff, The academic trace of the performance matrix: a mathematical synthesis of the *h*-index and the integrated impact indicator (I3). J. Assoc. Inf. Sci. Technol. **65**, 742–750 (2014)
175. S.V. Dorogovtsev, J.F.F. Mendes, Ranking scientists. Nat. Phys. **11**, 882–883 (2010)
176. M.A. Garzia-Perez, An extension of the *h*-index that covers the tail and the top of the citation curve and allows ranking researchers with similar *h*. J. Informetr. **6**, 689–699 (2012)

177. M. Schreiber, Self-citations corrections to the Hirsch index. Europhys. Lett. **78**, Art. No. 30002 (2007)
178. M. Schreiber, The influence of self-citation corrections on the Egghe's g-index. Scientometrics **76**, 187–200 (2008)
179. M.A. Garcia-Perez, A multidimensional extension to Hirsch's *h*-index. Scientometrics **81**, 779–785 (2009)
180. A. Schubert, Successive *h*-indices. Scientometrics **70**, 201–205 (2007)
181. A. Schubert, A Hirsch-type index of co-author partnership ability. Scientometrics **91**, 303–308 (2011)
182. F.J. Cabrerizo, S. Alonso, E. Herrera-Viedma, F. Herera, q^2-index: quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. J. Informetr. **4**, 23–28 (2010)
183. S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera, *hg*-index: a new index to characterize the scientific output of researchers based on the *h*-and *g*-indices. Scientometrics **82**, 391–400 (2009)
184. S. Moussa, M. Touzani, Ranking marketing journals using the Google-based *hg*-index. J. Informetr. **4**, 107–117 (2010)
185. L. Egghe, An improvement of the *h*-index: the *g*-index. ISSI Newsl. **2**, 8–9 (2006)
186. L. Egghe, Theory and practice of the g-index. Scientometrics **69**, 131–152 (2006)
187. L. Egghe, An econometric property of the *g*-index. Inf. Process. Manage. **45**, 484–489 (2009)
188. L. Egghe, The Hirsch-index and related impact measures. Annu. Rev. Inf. Sci. Technol. **44**, 65–114 (2010)
189. M. Schreiber, How to modify the g-index for multi-authored manuscripts. J. Informetr. **4**, 42–54 (2010)
190. M. Schreiber, A case study of the modified g-index: counting multi-author publications fractionally. J. Informetr. **4**, 636–643 (2010)
191. R. Costas, M. Bordons, Is *g*-index better than the *h*-index? An exploratory study at the individual level. Scientometrics **77**, 267–288 (2008)
192. G. Prathap, The 100 most prolific economists using the *p*-index. Scientometrics **84**, 167–172 (2010)
193. G. Prathap, The energy-exergy-entropy (or EEE) sequences in bibliometric assessment. Scientometrics **87**, 515–524 (2011)
194. J. Antonakis, R. Lalive, Quantifying scholarly impact: IQ_p versus the Hirsch index. J. Am. Soc. Inf. Sci. Technol. **59**, 956–969 (2008)
195. B.-H. Jin, L.-M. Liang, R. Rousseau, L. Egghe, The R- and AR-indices: complementing the h-index. Chin. Sci. Bull. **52**(6), 855–863 (2007)
196. R. Rousseau, F.Y. Ye, A proposal for a dynamic *h*-type index. J. Am. Soc. Inf. Sci. Technol. **59**, 1853–1855 (2008)
197. Q.L. Burrell, Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. Scientometrics **73**, 19–28 (2007)
198. L. Waltman, N.J. van Eck, T.N. van Leeuween, M.S. Visser, A.F.J. van Raan, Towards new crown indicator: an empirical analysis. Scientometrics **87**, 467–481 (2011)
199. R.E. de Bruin, A. Kint, M. Luwel, H.F. Moed, A study of research evaluation and planning. Res. Eval. **3**, 25–41 (1993)
200. T. Braun, W. Glänzel, United Germany: the new scientific superpower? Scientometrics **19**, 513–521 (1990)
201. H.F. Moed, R.E. de Bruin, T.N. van Leeuween, New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. Scientometrics **33**, 381–422 (1995)
202. P. Vinkler, Eminence of scientists in the light of the *h*-index and other scientometrics indicators. J. Inf. Sci. **33**, 481–491 (2007)
203. T.J. Phelan, A compendium of issues for citation analysis. Scientometrics **45**, 117–136 (1999)
204. T.J. Phelan, Is Australian educational research worthwhile? Aust. J. Educ. **44**, 175–194 (2000)

205. P. Vinkler, The π-index. A new indicator for assessing scientific impact. J. Inf. Sci. **35**, 602–612 (2009)
206. J. Wu, Investigating the universal distributions of normalized indicators and developing field-independing index. J. Informetr. **7**, 63–71 (2013)
207. F. Franceschini, M. Galetto, D. Maisano, L. Mastrogiacomo, The success-index: an alternative approach to the *h*-index for evaluating an individuals research output. Scientometrics **92**, 621–641 (2012)
208. F. Franceschini, M. Galetto, D. Maisano, L. Mastrogiacomo, Further clarifications about the success-index. J. Informetr. **6**, 669–673 (2012)
209. F. Francheschini, M. Galeto, M. Maisano, L. Mastrogiacomo, An infometric model for the success index. J. Informetr. **7**, 109–116 (2013)
210. L. Egghe, Impact coverage of the success index. J. Informetr. **8**, 384–389 (2014)
211. M. Kosmulski, Successful papers: a new idea in evaluation of scientific output. J. Informetr. **5**, 481–485 (2011)
212. P. Vinkler, Evaluation of some methods for the relative assessment of scientific publications. Scientometrics **10**, 157–177 (1986)
213. A.-L. Barabasi, *Linked. How Everything is Connected to Everything Else and Ehat it Means for Business, Science, and Everyday Life* (Basic Books, New York, 2014)
214. M. Newman, A.-L. Barabasi, D.J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, NJ, 2006)
215. A.-L. Barabasi, Z.N. Oltivai, Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**, 101–113 (2004)
216. H. Jeong, B. Tombor, R. Albert, Z.N. Oltivai, A.-L. Barabasi, The large-scale organization of metabolic networks. Nature **407**(6804), 651–654 (2000)
217. R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86**(14), 3200–3203 (2001)
218. D. Smilkov, C.A. Hidalgo, L. Kocarev, Beyond network structure: how heterogeneous susceptibility modulates the spread of epidemics
219. R. Pastor-Satorras, A. Vespignani, Epidemic dynamics and endemic states in complex networks. Phys. Rev. E **63**, Art. No. 0661117 (2001)
220. L. Isella, J. Stehle, A. Barrat, C. Cattuto, J.-F. Pinton, W. van den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks. J. Theor. Biol. **271**, 166–180 (2011)
221. A.-L. Barabasi, The origin of bursts and heavy tails in human dynamics. Nature **435**(7039), 207–211 (2005)
222. S. Fortunato, M. Brathelemy, Resolution limit in community detection. PNAS **104**, 36–41 (2007)
223. G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
224. A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis. Phys. Rev. E **80**, Art. No. 056117 (2009)
225. J.J. Ramasco, S.N. Dorogovtsev, R. Pastor-Satorras, Self-organization of collaboration networks. Phys. Rev. E **70**, Art. No. 036106 (2004)
226. A. Scharnhorst, M. Thelwall, Cytation and hyperlink networks. Curr. Sci. **89**, 1518–1523 (2005)
227. G. Ahuja, Collaboration networks, structural holes and innovation: a longitudinal study. Adm. Sci. Q. **45**, 425–455 (2000)
228. F. Havemann, A. Scharnhornst, bibliometrische netzwerke, in *Handbuch Netzwerkforschung*, eds. by C. Stegbauer, R. Häusling (Springer, Berlin, 2010), pp. 799–823
229. A. Pyka, A. Scharnhorst (eds.), *Innovation networks. New Approaches in Modeling and Analyzing* (Springer, Berlin, 2009)
230. A. Scharnhorst, Citation-networks, science landscapes and evolutionary strategies. Scientometrics **43**, 95–106 (1998)
231. J. Ortega, I. Aguillo, V. Cothey, A. Scharnhorst, Maps of the academic web in the European Higher Education Area—an exploration of visual web indicators. Scientometrics **74**, 295–308 (2007)

232. M.E.J. Newman, The structure of scientific collaboration networks. PNAS **98**, 404–409 (2001)
233. B. Kozma, A. Barrat, Consensus formation on adaptive networks. Phys. Rev. E **77**, Art. No. 016102 (2008)
234. S. Fortunato, V. Latora, A. Pluchino, A. Rapisarda, Vector opinion dynamics in a bounded confidence consensus model. Int. J. Mod. Phys. C **16**, 1535–1551 (2005)
235. L. Dall'Asta, A. Baroncheli, A. Barrat, V. Loreto, Agreement dynamics on small-world networks. Europhys. Lett. **73**, 969–975 (2006)
236. M. Barthelemy, Spatial networks. Phys. Rep. **499**, 1–101 (2011)
237. R. Albert, A.-L. Barabasi, Topology of evolving networks: local events and Universality. Phys. Rev. Lett. **84**, 5234–5237 (2000)
238. R. Albert, I. Albert, G.L. Nakarado, Structural vulnerability of the North American power grid. Phys. Rev. E **69**, Art. No. 025103 (2004)
239. R. Pastor-Satorras, A. Vespignani, *Evolution and structure of the internet: a statistical physics approach. Evolution and Structure of the Internet* (Cambridge University Press, Cambridge, 2004)
240. A. Reka, H. Jeong, A.-L. Barabasi, Diameter of the World Wide Web. Nature **401**, 130–131 (1999)
241. A. Vazquez, R. Pastor-Satorras, A. Vespignani, Large-scale topological and dynamical properties of the Internet. Phys. Rev. E **65**, Art. No. 066130 (2002)
242. Y. Moreno, M. Nekovee, A.F. Pacheco, Dynamics of rumor spreading in complex networks. Phys. Rev. E **69**, Art. No. 066130 (2004)
243. M. Nekovee, Y. Moreno, G. Bianconi, M. Marsili, Theory of rumor spreading in complex social networks. Phys. A **374**, 457–470 (2007)
244. E. Otte, R. Rousseau, Social network analysis: a powerful strategy, also for the information sciences. J. Inf. Sci. **28**, 441–453 (2002)
245. Y.-H. Eom, S. Fortunato, Characterizing and modeling citation dynamics. PLOS One **6**, e24926 (2011)
246. F. Raddichi, S. Fortunato, B. Markines, A. Vespignani, Diffusion of scientific credits and the ranking of scientists. Phys. Rev. E **80**, Art. No. 056103 (2009)
247. A. Schubert, A. Korn, A. Telcs, Hirsch-type indices for characterizing networks. Scientometrics **78**, 375–382 (2009)
248. F. Mali, L. Kronegger, P. Doreian, A. Ferligoj, Dynamics scientific co-authorship networks, in *Models of Science Dynamics*, eds. by A. Scharnhorst, K. Börner, P. van den Besselaar (Springer, Berlin, 2012), pp. 195–232
249. F. Raddichi, S. Fortunato, A. Vespigniani, Citation networks, in *Models of Science Dynamics*, eds. by A. Scharnhorst, K. Börner, P. van den Besselaar (Springer, Berlin, 2012), pp. 233–257
250. L. Egghe, R. Rousseau, Co-citation, bibliographic coupling and a characterization of lattice citation networks. Scientometrics **55**, 349–361 (2002)
251. Y. Ding, E.J. Yan, A. Frazho, J. Caverlee, PageRank for ranking authors in co-citation networks. J. Am. Soc. Inf. Sci. Technol. **60**, 2229–2243 (2009)
252. P. Chen, S. Redner, Community structure of the physical review citation network. J. Informetr. **4**, 278–290 (2010)
253. S.X. Zhao, R. Roussseau, F.Y. Ye, *h*-degree as a basic measure in weighted networks. J. Informetr. **5**, 668–677 (2011)
254. A. Abbasi, L. Hossain, L. Leydesdorf, Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. J. Informetr. **6**, 403–412 (2012)
255. L.C. Freeman, Centrality in social networks. Conceptual clarification. Soc. Netw. **1**, 215–239 (1978)
256. X. Yan, L. Zhai, W. Fan, C-index: a weighted network mode centrality measure for collaboration competence. J. Informetr. **7**, 223–239 (2013)
257. A. Korn, A. Schubert, A. Telcs, Lobby index in networks. Phys. A **388**, 2221–2226 (2009)
258. M.G. Campiteli, A.J. Holanda, L.D.H. Soares, P.R.C. Soles, O. Kinouchi, Lobby index as a network centrality measure. Phys. A **392**, 5511–5515 (2013)

259. M. Ausloos, A scientometric law about co-authors and their ranking: the co-author core. Scientometrics **95**, 895–909 (2013)
260. P. Jaccard, The distribution of the flora in the alpine zone. New Phytol. **11**, 37–50 (1912)
261. M. Levandowsky, D. Winter, Distance between sets. Nature **234**(5), 34–35 (1971)

Chapter 3
Additional Indexes and Indicators for Assessment of Research Production

Dedicated to the Max-Planck Society (a treasure for scientific information ensuring high-quality research) and to the MPIPKS (one of the places where the quality of research work of young researchers has increased enormously in a short time)

Abstract About forty-five indexes for assessment of research production of single researchers have been discussed in Chap. 2. These indexes are based mainly on citations of publications of the evaluated researcher. The indexes form Chap. 2 can be calculated also for groups of researchers. In addition to indexes from Chap. 2, other indexes useful for assessment of production of groups of researchers may be used. About ninety such indexes are discussed in this chapter. The indexes are grouped in the following classes: simple indexes; indexes for deviation from simple tendency; indexes for difference; indexes for concentration, dissimilarity, coherence, and diversity; indexes for advantage and inequality; indexes for stratified data; indexes for imbalance and fragmentation; indexes based on the concept of entropy; Lorenz curve and associated indexes. In addition, the set of indexes connected to the RELEV method for assessment of scientific research performance within public institutes as well as indicators and indexes for scientific research performance of nations and about comparing national scientific productions are discussed. Finally, we discuss briefly several journal citation measures as well as an example of an application of a geometric tool for detection of scientific elites in a group of institutes on the basis of Lorenz curves.

3.1 Introductory Remarks

Two modes of knowledge production may be considered [1, 2]: Mode 1 and Mode 2. Mode 1 of knowledge production is motivated by scientific knowledge alone, e.g., by fundamental research. In other words, Mode 1 of knowledge production is not connected to the search for applications of the obtained results. Mode 1 of knowledge

© Springer International Publishing Switzerland 2016

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1_3

production is founded on the separation of science into discrete disciplines (e.g., a researcher from one discipline may not bother about another discipline). In Mode 2 of knowledge production, multidisciplinary teams are brought together for short periods of time to work on specific problems in the real world for knowledge production. Mode 2 is closely connected to the project system of research, e.g., to how research funds are distributed among scientists and how scientists focus on obtaining these funds. In the case of Mode 1, the scientific knowledge production is carried out by actors who are distributed, yet proximate. In the case of Mode 2, knowledge production is distributed, and the actors are far apart. The notion of distribution may be considered in five proximity dimensions (cognitive, organizational, social, institutional, geographical) [3].

Mode 2 of knowledge production has been increasingly applied in the research systems of many countries. This shift of science toward Mode 2 of knowledge production has occurred because Mode 2 is considered to be more interdisciplinary, more heterogeneous, closer to social actors and contexts, and more susceptible to social critique [4]. Mode 2 is an important factor in the increasing importance of indicators and indexes for assessment of research production of groups of researchers, since Mode 2 is connected to actions of teams consisting of several research groups.

In this chapter, additional indicators and indexes for assessment of research production of groups of researchers are discussed. From the viewpoint of bibliometric methodology, one may make a distinction among three levels of aggregations [5]: *micro level*—publication output of individuals and research groups; *meso level*—publication output of institutions or studies of scientific journals; *macro level*—publication output of regions and countries and groups of countries. We discuss below indexes belonging mainly to the meso level and the macro level of aggregation. These indexes may be applied to any organization that has components, and these components possess some units. Components may be researchers from a research group; research groups from a research institute or faculty; research institutes belonging to groups of institutes, etc. Units may be publications, citations, patents, etc. The following groups of indexes will be discussed:

1. *Simple indexes*: index of quality of scientific output; annual impact index; MAPR-index; T-index; RPG-index; TPP-index; TIA-index.
2. *Indexes for deviation from simple tendency*: Schutz coefficient of inequality; Wilcox deviation from the mode; Nagel's index of equality; coefficient of variation.
3. *Indexes for differences between components*: Gini's mean relative difference; Gini's coefficient of inequality.
4. *Indexes for concentration, dissimilarity, coherence, and diversity*: Herfindahl–Hirschmann index of concentration; Horwat's index of concentration; RTS-index of concentration; diversity index of Lieberson; generalized Stirling diversity index; index of dissimilarity; generalized coherence index.
5. *Indexes of imbalance and fragmentation*: Index of imbalance of Taagepera; RT-index of fragmentation.

6. *Indexes based on the concept of entropy*: Theil's index of entropy; redundancy index of Theil; negative entropy index; expected information content of Theil.
7. *Lorenz curve and associated indexes*; Lorenz curve, index of Kuznets; Pareto diagram.
8. *Indexes for the case of stratified data*: Index of Gini for stratified data; index of Kuznets for stratified data; coefficient of variation for stratified data; index of Theil for stratified data.
9. *Indexes of advantage and inequality*: Index of net difference of Lieberson; index of average relative advantage; index of inequity of Coulter; proportionality index of Nagel.
10. *RELEV method for assessment of scientific research*: Indexes and indicators connected to the RELEV method.
11. *Indexes and indicators for comparison among scientific communities in different countries*.
12. *Indexes and indicators for efficiency of research production from the point of view of publications and patents*.
13. *Indexes for characteristics of scientific production of a nation*.
14. *Indicators for leadership*.
15. *Selected journal citation measures*: Impact factor, intermediacy index; SNIP indicator; SJR.

Many examples for calculation of these indexes are provided. Special attention is devoted to calculation of the values of indexes for the two extreme cases (when one component possesses all the units and when all components possess the same number of units). Finally, we shall discuss the important question for research elites on the basis of a geometric detection of kinds of scientific elites from the Lorenz curve of the publications written by groups of researchers.

3.2 Simple Indexes

We shall discuss two indexes connected to citations of production of a group of researchers: the index for the quality of scientific output and the annual impact index. The remaining indexes discussed are connected with characteristics of the research publications of the group. They include the mean annual percentage rate (MAPR) index, the doubling-time index, the relative publication growth (RPG), and indexes of total publication productivity and total institutional authorship.

3.2.1 A Simple Index of Quality of Scientific Output Based on the Publications in Major Journals

Let us consider a hypothetical group of researchers (research group, department, institute, etc.). The group of researchers produces some output that is cited. Let us

count the citations for some time period (say one year or several years). One may consider the index

$$Q_1 = \frac{N_m}{N}, \tag{3.1}$$

where

- N: total number of citations of the research output of the group of scientists;
- N_m: number of citations of the research group in major journals.

In order to use this index, we need a list of major journals. If we have such a list for the corresponding scientific area, then the evaluators of scientific performance can use Q_1 as an orientation for the quality of the research output of the scientific group. In addition, some further analysis of N_m can be made. It may happen that:

1. Almost all of the N_m citations are citations of the output of a single person or of a small number of persons from the group of scientists. In this case, we have a group with one or several scientific leaders.
2. The citations N_m are more or less spread evenly among the scientific productions of all members of the group. In this case, we have a scientific group with some (smaller or larger) degree of homogeneity with respect to the quality of scientific output.

Let us discuss two examples of calculation of index of quality. We consider two research groups. Each group consists of five researchers. The first group consists of only young researchers. The number of citations (N_m, N) for the members of this group are (10, 15); (20, 31); (14, 22); (35, 48); (55, 62). Thus for the entire group, $N_m = 134$ and $N = 178$. Then $Q_1^I \approx 0.75$. The second group contains two established researchers. The number of citations for the members of this group are (753, 1042); (554, 782); (80, 119); (41, 56); (12,16). Thus for the entire group, $N_m = 1440$ and $N = 2011$. Then $Q_1^{II} \approx 0.72$, i.e., the quality index of the scientific output of the two groups is almost the same. This example was especially designed in order to show again that evaluation and comparison of research groups based on a single index is insufficient: in the one-dimensional space of the values of the simple index of quality, the two groups of researchers are very close one to each other. In order to evaluate them properly, we need a higher-dimensional space, i.e., we need sets of values of various indexes. These sets may represent the coordinates of the research groups in the multidimensional space of values of the indexes (quantitative evaluation space). A larger dimension of this space means more indexes to be used, and an increase in the dimension of the quantitative evaluation space usually increases the corresponding distance between points corresponding to the research groups in the space. The larger distance between research groups in the quantitative evaluation space allows better comparison of their research results.

3.2.2 Actual Use of Information Published Earlier: Annual Impact Index

The annual impact index for the ith year of the papers published in the nth year ($n < i$) $\mathrm{AI}_{i,n}$ [6] is defined as follows:

$$\mathrm{AI}_{i,n} = \frac{C_{i,n}}{P_n}, \tag{3.2}$$

where

- $C_{i,n}$: number of citations received in year i by the papers published in year n;
- P_n: number of papers, published in year n.

Let us fix n. When i is close to n, the annual impact index may increase with increasing i. Usually, at some value of i, the index has its maximum value, and when i increases further, the value of the index begins to decrease (one factor for this decrease is the aging of the information contained in the papers published in the year n).

The index of the actual use of information helps us to see easily whether the research information produced by a group of researchers is useful for the research society. Let us demonstrate this. We consider two research groups. Research group A has six publications for 2010, and research group B has twelve publications for 2010. The quantity of information produced by research group B is larger than that produced by A. The sets of citations of the above publications for the period 2011–2015 of the two groups are as follows:

- **Research group A**: 3, 8, 17, 38, 60;
- **Research group B**: 2, 8, 21, 49, 94.

The corresponding values of the $AI_{i,n}$-index are (approximately)

- **Research group A**: 0.5, 1.33, 2.83, 6.33, 12;
- **Research group B**: 0.16, 0.66, 1.75, 4.08, 7.83.

Thus according to the $AI_{i,n}$-index, the impact of the information produced by research group A is larger (at least for the five-year period of evaluation 2011–2015).

3.2.3 MAPR-Index, T-Index, and RPG-Index

These indexes are characteristics of the set of publications produced by the evaluated group of researchers [7, 8]. The MAPR-index (mean annual percentage rate) is defined as

$$\mathrm{MAPR}_t = 100\left[\frac{1}{t}\sum_{i=1}^{t}\frac{P_i - P_{i-1}}{P_{i-1}}\right], \tag{3.3}$$

where

- t: length of the studied period (in years);
- P_i: number of papers written by the group of researchers in year i

For example, if the period of evaluation is five years, then

$$\mathrm{MAPR}_5 = 20\left[\frac{P_1 - P_0}{P_0} + \frac{P_2 - P_1}{P_1} + \frac{P_3 - P_2}{P_2} + \frac{P_4 - P_3}{P_3} + \frac{P_5 - P_4}{P_4}\right]. \quad (3.4)$$

Note that all the P_i should be different from 0. The MAPR-index can also be used for characterization of the evolution of the number of publications in a research field or in a journal or group of journals.

The MAPR-index easily detects the phases of increasing or decreasing research activity. Let us consider two research groups that are evaluated for a period of five years ($t = 5$). Group A is a newly established research group, and group B is a mature group in a research field that is slowly beginning to be exhausted. The number of publications of the two groups are

- **Research group A**: 3, 5, 5, 7, 8, 11;
- **Research group B**: 63, 64, 62, 60, 58, 60.

The values of the MAPR_5-index for the two research groups are

- **Research group A**: $\mathrm{MAPR}_5^A \approx 1.583$;
- **Research group B**: $\mathrm{MAPR}_5^B \approx -0.346$.

The values of the MAPR-index that are very close to 0 or negative are evidence of maturity or of problems in the corresponding research group. The nature of such problems may be further studied by other quantitative or qualitative tools.

The T-index (the doubling time) is defined as

$$T = \frac{1}{2}\frac{0.301(t-1)}{\ln(P_t) - \ln(P_1)}, \quad (3.5)$$

where

- P_1: number of papers in the starting year;
- P_t: number of papers in the tth year.

The T-index gives a good impression about the mean size of the expansion of the scientific information produced by the research group or the mean size of expansion of information in a given research field. Let us consider two research fields with T-indexes of seven years and fifteen years. The first field expands faster. Faster expansion (and small value of the T-index) is characteristic for new fields or for established fields after a large discovery is made. The T-index of a mature field has a large value.

For the two research groups discussed above, we obtain the following values of the doubling-time index:

- **Research group A**: $T_A \approx 1.84$;
- **Research group B**: $T_B \approx -12.28$.

The results show that the tempo of advancing of the research activity of the newly established group is good, whereas the negative value of the index shows a shrinking of research production in the mature group B. Let us note that it is good practice to include only publications in journals of some level (i.e., journals with an impact factor or journals with an SJR factor) in order to achieve greater objectivity regarding the information supplied by the MAPR-index and by the T-index.

The RPG-index (relative publication growth index) [9] is defined as

$$\mathrm{RPG}_j(T) = \frac{P_j}{Q_j}; \quad Q_j = \sum_{i=1}^{T=j-1} P_i, \tag{3.6}$$

where

- $T = j - 1$: period in which the published papers are counted ($T \geq 3$);
- j: year for which the index is calculated;
- P_i: number of published papers in the ith year of the period of interest.

The value of T can be five years, ten years, twenty years, etc. The value of the RPG-index for several databases of papers (Chemical Abstracts, Biological Abstracts, Science Citations Index, etc.) can be found in Table 4.2 of [8]. The RPG index calculated with appropriately selected time periods may give us information about the dynamic equilibrium between recent information and previously published information.

As defined above, we can calculate, for example, $\mathrm{RPG}_{11}(10)$ but not $\mathrm{RPG}_{11}(8)$. In order to be able to calculate the value of the last index, we have to redefine the index slightly as follows:

$$\mathrm{RPG}_j(T = j - k) = \frac{P_j}{Q_j}; \quad Q_j = \sum_{i=k}^{j-1} P_i, \tag{3.7}$$

where $1 \leq k \leq j - 1$.

The $\mathrm{RPG}_5(4)$-index for the two research groups discussed in the subsection for the MAPR-index has the values

- **Research group A**: $\mathrm{RPG}_5^A(4) \approx 0.44$;
- **Research group B**: $\mathrm{RPG}_5^B(4) \approx 0.246$.

The result shows that the rate of total publications growth for research group A is about twice that of the rate for research group B.

3.2.4 Total Publication Productivity, Total Institutional Authorship

The TPP-index (total publication productivity index) [8] compares the total information productivity of groups of researchers working in fields with similar bibliometric features. The definition of the index is

$$\mathrm{TPP}_T = \frac{p_T}{\kappa T}, \tag{3.8}$$

where

- T: period of evaluation;
- κ: mean number of researchers working in the research group in the period T;
- p_T: total number of scientific publications published by the members of the research group in the period T.

As publications, one may count journal papers (also in electronic form), or in principle one may count any kind of scientific publications except conference abstracts.

The value of the TPP-index can be greatly influenced by multiple authorship. Because of this, it is useful if the TPP-index is accompanied by the TIA-index (total institutional authorship index) [10]

$$\mathrm{TIA}_T = \frac{A_a(T)}{A_t(T)}, \tag{3.9}$$

where

- T; period of evaluation;
- $A_a(T)$: Number of authors attributed to the evaluated research group for the period T;
- $A_t(T)$: total number of authors of the publications published by the research group for the period T.

3.3 Indexes for Deviation from a Single Tendency

The concept of indexes for deviation from a single tendency is as follows. One has a numerical series. By a mathematical operation one defines a value that is called the standard value (standard) for the series (different definitions can lead to different standards). Each value of the series deviates from the standard value. The indexes are constructed on the basis of these deviations.

In general, the tendency of deviation from a standard value can change over time. Below, we shall discuss mostly deviations from time-independent quantities. We just note that one can also construct deviations from time-dependent quantities. One such index is the Przeworski index of instability [11].

An important type of deviation from the time-independent quantities is deviations (absolute or squared) from a central tendency. Among the absolute deviations from a central tendency we shall discuss the indicators called the Schutz coefficient of inequality and the Wilcox deviation from the mode.

3.3.1 Schutz Coefficient of Inequality

The equation for this index is [12]

$$I_1 = \frac{\sum_{i=1}^{K} \left(\frac{P_i}{\overline{P}} - 1 \right)}{K - 1}, \tag{3.10}$$

where the quantities are as follows:

- K: number of components.
- P_i: percentage of the total number of units possessed by the ith component.
- $\overline{P}$: average percentage ($\overline{P} = (1/K) \sum_{i=1}^{K} P_i$).

Let us illustrate the extreme values of I_1 in terms of scientists and papers. If all scientists possess the same number of papers (absolute equality), then $P_i = \overline{P}$ and $I_1 = 0$. The other extreme case is that one of the scientists possesses all the papers and the other $K - 1$ scientists have none. Then the denominator of I_1 has the value $K - 1$, and $I_1 = 1$.

Inequality is an important concept in the social sciences and economics [13, 14]. Many measures developed for measuring economic and social inequality [15] can be used for measurement of different aspects of inequality of research production of researchers. Some of these indexes will be discussed below.

3.3.2 Wilcox Deviation from the Mode (from the Maximum Percentage)

The equation for this index is [16]

$$I_2 = 1 - \frac{\sum_{i=1}^{K} (P_m - P_i)}{K - 1}, \tag{3.11}$$

where the quantities are as follows:

- K: number of the components.
- P_i: percentage of the total number of units possessed by the ith component.
- P_m: the maximum value among $P_1, \ldots, P_k$.

I_2 measures the extent to which the nonmodal components resemble the modal component. In our example about scientists and papers, P_m is the share of the most productive scientist (measured by the number of published papers). If all the scientists are as productive as the most productive one, then $I_2 = 1$. If the most productive scientist wrote all the papers and the other scientists wrote none, then $I_2 = 0$, which indicates a problem.

The index of Wilcox was developed for measurement in political science. There exist several more indexes proposed by Wilcox [17] for measurement of different aspect of public opinion. These indexes (as has been shown above) can be easily adapted for assessment of research production.

Let us now turn to the squared deviations from a central tendency. Here we shall discuss Nagel's index of equality.

3.3.3 Nagel's Index of Equality

The equation for this index is [18]

$$I_3 = 1 - \frac{\sum_{i=1}^{K} \left(N_i - \frac{N}{K}\right)^2}{\left(Z - \frac{N}{K}\right)^2}. \tag{3.12}$$

The quantities above are as follows:

- N_i: Number of units possessed by the ith component of the organization;
- N: Total number of units distributed among the components;
- K: Number of components of the organization;
- Z: The worst possible allocation of components in terms of equality. This worst possible allocation occurs when one of the components owns all of the units and the other components own nothing.

Let the worst possible allocation be realized (a single researcher wrote all the publications in the research group, and the other researchers have none). Then $I_3 = 0$. And if all researchers from the research group wrote the same (N/K) number of publications, then $I_3 = 1$. Thus very small values of Nagel's index of inequality are evidence for the presence of a small number of highly productive researchers in the research group.

We note that the value of Nagel's index is sensitive to the number of components of the system.

3.3.4 Coefficient of Variation

The equation for the coefficient of variation is [19]

$$I_4 = \frac{1}{\overline{U}} \sqrt{\frac{1}{K} \sum_{i=1}^{K} (U_i - \overline{U})^2}, \tag{3.13}$$

where

- K: number of components of the organization;
- U_i: number of units owned by the ith component;
- $\overline{U}$: average number of units owned by the system components.

The variation coefficient is obtained by division of the standard deviation of the data by the mean value of the units owned by a component of the organization. Let the organization be a research group. One extreme case occurs when one component of the organization (one member of the research group) has written all the publications. Then $U_1 = U_2 = \cdots = U_{K-1} = 0$ and $U_K = K\overline{U}$. Then $I_4 = \sqrt{K-1}$. The other extreme case occurs when all researchers have written the same number of publications (namely $\overline{U}$ publications). Then $I_4 = 0$. Thus the presence of large differences in the research production of the researchers from the group will lead to a significant deviation of I_4 from 0. Another index of this kind is the logarithmic variance

$$I_5 = \frac{1}{K} \sum_{i=1}^{K} (\ln U_i - \ln \overline{U})^2. \tag{3.14}$$

If all researchers from the research group wrote the same number of publications, then $I_5 = 0$. In the extreme case that one of the researchers from the group wrote all the publications, then $I_5 = (\ln(K))^2/K$. We note that the indexes I_4 and I_5 can be easily normalized in order to have values between 0 and 1. We now proceed to the group of indexes for differences between components. Such indexes include, for example, the two quantities used by Gini: Gini's mean relative difference and Gini's coefficient of inequality.

3.4 Indexes for Differences Between Components

3.4.1 Gini's Mean Relative Difference

Gini's mean relative difference [20, 21] is calculated as follows:

$$I_6 = 1 - \frac{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \mid P_i - P_j \mid}{K-1}, \tag{3.15}$$

where the quantities are

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component.

The values of I_6 are between 0 and 1. When one of the components possesses all units, the value of I_6 is 0, regardless of the number of components. When all components possess the same number of units, the value of I_6 is 1. Let the units be the publications written by the researchers from a research group. If one of the researchers wrote all the publications, then Gini's mean relative difference is 0. If all researchers wrote the same number of publications, then the value of the index is 1.

Gini's mean relative difference also has a continuous version [22], which was used for quantification of the speed of technological adoption in India. An extensive discussion on the measures of Gini and similar measures such as the Lorenz curve can be found in [23, 24].

We note that the values of I_6 do not correspond to expectations that might arise from the name of the index. One might expect that the value 0 will be assigned to the case in which no difference between researchers exists (all of them wrote the same number of publications). And for the extreme case (one researcher wrote all publications), the expectation for the value of the index is that it should be equal to 1. The real situation is exactly the opposite, and this is one factor that contributes to the popularity of the following index: I_7.

3.4.2 Gini's Coefficient of Inequality

Gini [20] preferred to use I_6, but in the course of time, Gini's coefficient of inequality

$$I_7 = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left| \frac{P_i - P_j}{K} \right| \tag{3.16}$$

become more popular. The quantities in I_7 are as follows:

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component.

Gini's coefficient is sensitive to the number of components K, and because of this, it is better when I_7 is used in organizations that have a large number of components K. If the number of components is small, then it is better to use I_6.

Let us calculate Gini's coefficient of inequality for several cases of groups of researchers and their research publications. If all researchers wrote the same number of publications, then $I_7 = 0$. If one of researchers wrote all publications, and other researchers wrote none, then $I_7 = K - 1$. (Thus the index can be normalized when one divides it by $K - 1$: $I_7^+ = I_7/(K - 1)$.) Let us now suppose we have a research group of five researchers and the percentage of publications they wrote is $P_1 = 0.15$;

$P_2 = 0.18$, $P_3 = 0.22$, $P_4 = 0.30$, $P_5 = 0.15$. Then the value of Gini's coefficient of inequality is $I_7 = 0.101$. Thus the value of I_7 is closer to 0 than to the maximum value of 4, which reflects the fact that the inequality with respect to the number of publications in the group of researchers is not very large.

Gini's coefficient is much used in economics, the social sciences, ecology, etc. [25–30]. An example of its use in the area of scientific research is for quantification of the concentration of scientific research and innovation [31].

3.5 Indexes of Concentration, Dissimilarity, Coherence, and Diversity

The next group of indexes are indexes for concentration and diversity. These indexes inform us how the quantities associated with research production (number of publications, number of citations, etc.) are concentrated among groups of researchers. An exploration of the concentration of research production reveals also fragmentation, diversity, coherence, and imbalance with respect to scientific production in research organizations. Diversity may be defined as the property of apportioning units into categories in any system [32]. Coherence may be defined as the property of relating categories via units. Coherence captures the extent to which the various parts in a system are directly connected via some relation. Diversity has the following three distinct attributes: (i) *variety*—number of categories into which the units are apportioned; (ii) *balance*—evenness of the distribution of units across categories; (iii) *disparity*—degree to which the categories of the units are different [4]. The diversity of a system increases not only with more categories (higher variety) and with a more balanced distribution (higher balance), but also if the units are allocated to more different categories (higher disparity). Coherence has the following attributes: (i) *density*—number of relations between categories; *intensity*—overall intensity of the relations in the system; (iii) *disparity*—degree to which the categories of the relations are different.

In the process of analysis of diversity, one may use units such as university, institute, faculty, department, article, researcher, and research topic such as an emergent technology. Some of these units may be connected to a small number of the corresponding items. Thus one may not have enough items for a robust statistical analysis, which may worsen the quality of the resulting measures.

3.5.1 Herfindahl–Hirschmann Index of Concentration

The equation for this index [33, 34] is

$$I_8 = \sum_{i=1}^{K} P_i^2, \tag{3.17}$$

where

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component.

This form of the index is insensitive to small values of P_i, since the square of a value that is close to 0 is quite a small number. The index I_8 has its maximum value of 1 when one of the components of the organization possesses all units (in the case of our example, when one of the scientists possesses all the papers). The minimum value of the index is $1/K$ when all the components possess an equal number of units (there is no concentration of papers). Thus the lower bound of the index depends on the number of components K. In order to avoid this and to bound I_8 between 0 and 1, one can use the following form of the index:

$$I_8^* = 1 - \frac{1 - I_8}{(1 - 1/K)}. \tag{3.18}$$

When the number of components (the number of researchers) is large, then $1/K$ is small, and one can use I_8. If, however, the number of components is small, then it is better to use I_8^*.

Let us calculate I_8 for the case of the group discussed above for the case of index I_7. The result is $I_8 = 0.2158$, which reflects the relatively small level of concentration of ownership of research publications in the evaluated research group.

The Herfindahl–Hirschmann index has been used for measurement of dominant power [35], for measuring concentration in portfolio management [36], etc. [37, 38].

3.5.2 *Horvath's Index of Concentration*

The equation for this index is [39]

$$I_9 = P_m + \sum_{i=2}^{K} P_i^2[1 + (1 - P_i)], \tag{3.19}$$

where

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component;
- P_m: percentage of the total number of units possessed by the modal component (the component that possesses the largest number of units).

Horvath's concentration index measures the influence of the largest component. In our example, the modal component consists of the researcher with the largest number of publications. The index is useful in cases in which one of the scientists dominates the group of scientists with respect to some quantity (for example the number of published papers). The index I_9 measures the change in the primacy of this researcher

within the group in the course of time. Let us illustrate this. We shall consider a research group of five researchers. At the beginning, one of the researchers possesses all the publications of the group, and the other (young) researchers have not written any publications. In this case, $I_9 = 1$. In two years, the situation changes. The experienced researcher still dominates with 90 % of the papers, but the other four researchers have also written some papers. Let the percentage distribution be 0.9, 0.04; 0.02; 0.02; 0.02. Then the value of the index is $I_9 = 0.95512$, which reflects the changes but still shows the dominance of the most experienced researcher from the evaluated research group.

3.5.3 RTS-Index of Concentration

This index was designed by Ray et al. [40, 41]. The equation for this index is

$$I_{10} = \left[\frac{\sum_{i=1}^{K} P_i^{\alpha} - K^{(1-\alpha)}}{1 - K^{(1-\alpha)}} \right]^{(1/\alpha)}, \tag{3.20}$$

where

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component;
- α: parameter.

A characteristic feature of this index is that it depends on the parameter α. For $\alpha = 0$, $I_{10} = 0$. For $\alpha = 1$, $I_{10} = 1$. As $\alpha \to \infty$, $I_{10} \to P_m$, where P_m is the modal share of units (the number of units of the largest possessor of units).

Indexes of concentration are quite useful in the evaluation of research groups. They can exhibit hidden problems, such as concentration of research publications in researchers who are at the end of their scientific career, which hints at a future decrease in research productivity of this research group.

3.5.4 Diversity Index of Lieberson

The equation for this index is [42]

$$I_{11} = \frac{1 - \sum_{i=1}^{K} P_i^2}{(1 - 1/K)}, \tag{3.21}$$

where

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component.

The index I_{11} is bounded between 0 and 1. Let us discuss a group of researchers and their research publications. If one of the researchers owns all publications, then $I_{11} = 0$, and if all researchers have written the same number of publications, then $I_{11} = 1$. As an example for application of the index of diversity, let us consider two research groups. Research group A consists of five researchers, and the percentages of research publications are as follows 0.3, 0.25, 0.2, 0.15, 0.1. Research group B consists of six researchers, and the percentages of research publications are 0.25, 0.2, 0.15, 0.15, 0.15, 0.1. The values of the index are as follows:

- **Research group A**: $I_{11}^{A} = 0.96875$;
- **Research group B**: $I_{11}^{B} = 0.984$.

Thus the diversity of the two research groups is almost the same, and the value of the index is close to 1, which hints at sufficient activity of all researchers from the evaluated research groups.

3.5.5 Second Index of Diversity of Lieberson

Let us consider two populations Q and R. Now we want to study the diversity between the populations with respect to some category. The equation for the index is [42]

$$I_{12} = 1 - \sum_{i=1}^{C} Q_i R_i, \tag{3.22}$$

where

- Q_i: proportion of the category in population Q;
- R_i: proportion of the category in population R;
- C: the number of categories.

The populations Q and R can be of any type. For example, they may be the populations of researchers in two research institutes. The category can be any nominal category of some attribute. For example, the attribute can be the age of researchers and the categories can be young researchers (up to age 40); intermediate-age researchers (40–60 years old), and mature researchers (over 60 years old).

The index I_{12} reaches its maximum value of 1 when the diversity between the two populations is maximal. This happens when, for example, all Q_i equal 0 and all R_0 are positive.

Let us consider one example. We have two research institutes from the same area (say physics). For institute A, the percentage of young researchers is 0.05, the

percentage of intermediate age researchers is 0.15, and the percentage of mature researchers is 0.8. In = institute B, the percentage of young researchers is 0.08, the percentage of intermediate-age researchers is 0.25, and the percentage of mature researchers is 0.67. The index of diversity of Lieberson for these two institutes is $I_{12} = 0.4325$.

The diversity index of Lieberson can be used for analysis of different kinds of networks [43], electoral competition [44], etc.

3.5.6 Generalized Stirling Diversity Index

Let us consider units of something (e.g., publications) distributed among N categories (e.g., categories connected to the ISI Web of Science). Let p_i be the proportion of the units in category i, and d_{ij} the distance between categories i and j. Then the generalized Stirling diversity index is [32]

$$S = \sum_{i,j(i\neq j)} (p_i p_j)^{\alpha} d_{ij}^{\beta}, \tag{3.23}$$

where α and β are parameters. In order to use this index, one has to choose appropriate categories and to assign units to each category. Then one has to construct adequate metrics for the distance d_{ij} and to set appropriate values of the parameters α and β. Often one chooses the density in the interval $0 < d_{ij} < 1$, and the choice of small values of β emphasizes the importance of distance for the studied problem.

Particular cases of the generalized Stirling diversity index are the Rao–Stirling diversity index ($\alpha = \beta = 1$) [45, 46]

$$S_{RS} = \sum_{i,j(i\neq j)} (p_i p_j) d_{ij}; \tag{3.24}$$

and the Simpson diversity index ($\alpha = 1; \beta = 0$)

$$S_S = \sum_{i,j(i\neq j)} (p_i p_j) = 1 - \sum_i p_i^2. \tag{3.25}$$

The Rao–Stirling index may be interpreted as the average cognitive distance between elements, as seen from the categorization, since it weights the cognitive distance d_{ij} over the distribution of elements across categories [4]. The Rao–Stirling diversity index can be added over scales (under some plausible assumptions) [47]. Then, for example, the diversity of a research institute is the sum of the diversities within each article it has published, plus the diversity between the articles. This interesting property leads to the possibility of measuring the diversity of large organizations in a modular manner.

3.5.7 Index of Dissimilarity

Let us have two groups of researchers that are classified with respect to some characteristic that has two possible values (for example, one group consists of researchers who have published papers, and the second group consists of researchers who have not published even a single paper). The equation for the index is

$$I_{13} = \frac{1}{2} \sum_{i=1}^{K} \mid G_{1i} - G_{2i} \mid, \tag{3.26}$$

where

- K: number of investigated research organizations;
- G_{1i}: proportion of components of the ith organization that can be characterized by the first value of the characteristics;
- G_{2i}: proportion of components of the ith organization that can be characterized by the second value of the characteristics.

Let us now consider two research groups. Research group A has ten members, and eight of them have publications. Research group B has fourteen members, and eleven of them have publications. In this case, $I_{13} = 0.015$. Let now two new PhD students join research group B. Thus it has sixteen members, and eleven of them have publications. The value of the index changes to $I_{13} = 0.1175$, which reflects the fact of increasing dissimilarity and diversity between the two groups of researchers.

In its original definition [48], I_{13} was defined as an index of segregation (for example, segregation of citizens of different skin color in some urban area).

3.5.8 Generalized Coherence Index

Let us consider units of something (e.g., publications) distributed among N categories (e.g., categories connected to the ISI Web of Science). Let p_i be the proportion of units in category i; I_{ij} the intensity of relations between categories i and j; and d_{ij} the distance between categories i and j. Let us suppose that we have constructed adequate metrics for distance and intensity. The generalized coherence index [4] is given by the equation

$$G = \sum_{ij(i \neq j)} I_{ij}^{\gamma} d_{ij}^{\beta}. \tag{3.27}$$

When $\gamma = \delta = 0$, the value of G is equal to M. For $\gamma = 1$ and $\delta = 0$, we obtain a measure of intensity

$$G_I = \sum_{ij(i \neq j)} I_{ij} = 1 - \sum_{i} I_{ii}, \tag{3.28}$$

and for $\gamma = \delta = 1$, we obtain a measure of coherence

$$G = \sum_{ij(i \neq j)} I_{ij} d_{ij}. \tag{3.29}$$

If the intensity of relations is defined as the distribution of relations (i.e., when I_{ik} is equal to p_{ik}), then the coherence from (3.29) may be interpreted as the average distance over the distribution of relations p_{ik}.

3.6 Indexes of Imbalance and Fragmentation

The next group of indexes consists of indexes of imbalance and fragmentation. From among these indexes, we shall discuss the index of imbalance of Taagepera and the RT-index of fragmentation.

3.6.1 Index of Imbalance of Taagepera

This index treats imbalance as a comparison of the size of the largest component with respect to the size of the next-largest one. The equation for the index is [49]

$$I_{14} = \frac{\sum_{i=1}^{K-1} \frac{(P_i - P_{i+1})}{i} - (\sum_{i=1}^{K} P_i^2)^2}{\sqrt{\sum_{i=1}^{K} P_i^2 - (\sum_{i=1}^{K} P_i^2)^2}}, \tag{3.30}$$

where the components of the organization are ranked in decreasing order with respect to the possessed units and

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component.

The index I_{14} is most sensitive to the size difference (called imbalance) between the two largest components of the organization. A larger difference leads to a larger value of I_{14}.

3.6.2 RT-Index of Fragmentation

The relationship for this index is [50]

$$I_{15} = 1 - \frac{\sum_{i=1}^{K} N_i(N_i - 1)}{N(N-1)}, \tag{3.31}$$

where

- K: number of components of the organization;
- N_i: total number of units possessed by the ith component;
- N: total number of units possessed by all components of the organization.

The index is designed as 1 minus a measure of concentration of units among the components of the organization. In our example, the concentration of all papers to the account of one scientist leads to $I_{15} = 0$. When the papers are uniformly distributed among the scientists, then I_{15} is roughly equal to $1 - 1/K^2$, and for a large number of components of the organization, this value is almost equal to 1. From the last sentences, it follows that one has to use I_{15} for evaluation of fragmentation in organizations that have a large enough number of components.

We stress the following characteristic of I_{15}. If two groups of researchers (each with some fragmentation with respect to the possession of their published papers) are combined into a single group, then I_{15} for the new group will have a larger value than the values for the two groups considered separately. In other words, when groups are combined, then I_{15} shows a greater fragmentation in the new group in comparison to the two groups that are combined.

3.7 Indexes Based on the Concept of Entropy

Most of the indexes discussed below have the useful properties of **aggregation** and **decomposition**. The decomposition property means that the corresponding measure (of inequality in research productivity, for example) for the entire population of researchers (of a research group, research institute, etc.) can be decomposed as a sum of measures within the subpopulations (within the sections of the institute). Aggregation means the opposite: the sum of the corresponding measures for the subpopulation gives the value of the measure for the entire population.

The concept of entropy is used in analyses of science dynamics [51]. In order to understand the indexes based on the concept of entropy, we need the following concepts:

- **Bit**: Let us have m alternatives and we have to choose one of them. The number of bits of information h needed to select one of these alternatives is defined as $m = 2^h$. Then $h = \log_2 m$. In other words, one bit of information is gained when the value of a specific random variable (a variable that can take the value 0 or 1 with equal probability) becomes known.
- **Entropy of a set of random variables**: Let us have a set of L random variables each of which has its own probability of occurrence p_i and its own information

of h_i bits. The entropy of the set equals the sum of the information values of all the individual variables, each weighted by the corresponding probability of occurrence:

$$H = \sum_{i=1}^{L} p_i h_i = \sum_{i=1}^{L} p_i \log_2(1/p_i) = -\sum_{i=1}^{L} p_i \log_2(p_i).$$

The maximum value of the entropy is obtained when all probabilities of occurrence are the same. When one of the probabilities of occurrence is close to 1 (and the others are close to 0), then H is close to 0.

3.7.1 Theil's Index of Entropy

The probabilities p_i discussed above can be interpreted as percentages of the total number of units possessed by the ith component. In such a way, the entropy can be used directly as a measurement of (scientific) inequality. The result is Theil's index of entropy. The equation for the index is [52–54]

$$I_{16} = -\sum_{i=1}^{K} P_i \log_2 P_i, \tag{3.32}$$

where

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component.

A larger value of I_{16} corresponds to greater equality in the group of components (which means that the differences among the numbers of published papers among the scientists from the studied group is not very large).

Let us calculate I_{16} for several cases for a group of researchers and their research publications. Let one of researchers own all of publications, and the other members of groups have written no publications. There will be a difficulty in calculating I_{16} if some of the researchers have no publications, but we can assume that the contribution of the corresponding term to the index is 0. Then $I_{16} = 0$. For the case that all researchers have written the same number of publications, the value of the index is $I_{16} = \log_2 K$. The last result shows that I_{16} can be rescaled as follows:

$$I_{16}^{*} = -\frac{\sum_{i=1}^{K} P_i \log_2 P_i}{\log_2 K}. \tag{3.33}$$

Let us suppose a group of four researchers and that the percentages of publications that they have written are 0.5, 0.3, 0.1, 0.1. Let us have another group of eight

researchers with percentages of publications 0.3, 0.15.0.15.0.15, 0.1, 0.1, 0.03, 0.02. The values of Theil's index of entropy are

- **Research group A**: $I_{16}^{*\,A} \approx 0.84$;
- **Research group B**: $I_{16}^{*\,B} \approx 0.89$,

which means that the level of equality in group B with respect to research publications is slightly greater than the equality in research group A.

Theil's index is much used in sociology [55] and in economics [56].

3.7.2 Redundancy Index of Theil

The equation for this index is [57, 58]

$$I_{17} = \log_2 K + \sum_{i=1}^{K} P_i \log_2 P_i, \tag{3.34}$$

where

- K: number of components of the organization
- P_i: percentage of the total number of units possessed by the ith component.

The index I_{17} is an index of concentration, since we subtract the absolute entropy from a certain constant value. This index can be normalized as follows:

$$I_{17}^{*} = \frac{\log_2 K + \sum_{i=1}^{K} P_i \log_2 P_i}{\log_2 K}. \tag{3.35}$$

For the two research groups studied by means of I_{17}^{*}, one obtains the following values of the normalized redundancy index of Theil:

- **Research group A**: $I_{17}^{*\,A} \approx 0.16$;
- **Research group B**: $I_{17}^{*\,B} \approx 0.11$,

which shows that the concentration of publications in research group A is greater than that of research group B.

3.7.3 Negative Entropy Index

The equation for this index is

$$I_{18} = \text{antilog}_2\left(-\sum_{i=1}^{K} P_i \log_2 P_i\right), \tag{3.36}$$

where

- K: number of components of the organization;
- P_i: percentage of the total number of units possessed by the ith component.

The antilog function is the inverse of the log function. In (3.36), we use 2 as the base of the log and antilog functions. In the original definition of the index [59], the base was 10.

In our examples about researchers and their publications, I_{18} measures the closeness in the values of the numbers of publications written by every researcher. The index can be normalized as follows:

$$I_{18}^{*} = \frac{\text{antilog}_2\left(-\sum_{i=1}^{K} P_i \log_2 P_i\right)}{K}. \tag{3.37}$$

3.7.4 Expected Information Content of Theil

Let us suppose that we have a message that *an a priori distribution* $\sum p_i$ *has turned into an a posteriori distribution* $\sum q_i$. The expected information content of this message is [60]

$$I = \sum_i q_i^2 \log \frac{q_i}{p_i}. \tag{3.38}$$

If the logarithm has base of 2, then I is expressed as bits of information. Leydesdorff [51] has used this index in order to study statistics of journals from the SCI Journal Citation Reports.

3.8 The Lorenz Curve and Associated Indexes

3.8.1 Lorenz Curve

In general, the Lorenz curve can be defined as follows [61, 62]. Let us assume a probability distribution $P = F(x)$ of some quantity (number of papers, number of citations, amount of money, etc.) owned by members of some class of people (such as scientists) and let x be normalized in such a way that its value is between 0 and 1. The inverse distribution of F is $x = F^{-1}(P)$, and the Lorenz curve is defined by

$$L(F) = \int_0^1 F^{-1}(P)dP. \tag{3.39}$$

Let us assume a group of K researchers, and suppose we are interested in constructing the Lorenz curve for the number of papers written by every scientist. Let us rank the scientists with respect to the number of papers written by them. Let n_i be the number of papers of the ith scientist from the ranked list (the ranking is made in such a way that $n_1 \leq n_2 \leq \ldots n_K$). Then the coordinates of the corresponding Lorenz curve are

$$F_i = \frac{i}{K}; \quad L_i = \frac{\sum_{j=1}^{i} n_i}{\sum_{i=1}^{K} n_i}. \tag{3.40}$$

The Lorenz curve is much used in research on income distributions [63, 64], land use [65], economic concentration [66], etc. [67]. The Lorenz curve is used in scientometrics for characterization of conjugate partitions [68], for measurement of relative concentration [69, 70], group preferences [71], distribution of publications [72], distribution of research grants [73], regional research evaluation [74], and university ranking [75].

3.8.2 *The Index of Gini from the Point of View of the Lorenz Curve*

The points (0, 0); (0, 1); (1, 0); (1, 1) determine a square in the (L, F)-plane. The diagonal of this square that connects (0, 0) and (1, 1) is called the line of absolute equality: all components of the organization possess the same number of units. In practice, there is no absolute equality, and in this case, the Lorenz curve is below the line of absolute equality. Then a region exists between the line of absolute equality and the Lorenz curve. The area of this region is connected to the index of Gini:

$$I_{19}^{\dagger} = 1 - 2\int_0^1 L(F)dF. \tag{3.41}$$

The discrete version of the index of Gini is closely connected to the Gini coefficient of inequality (I_7) discussed above. The difference is that the index of Gini is divided also by the mean number of units $\overline{U}$ owned by a system component:

$$I_{19} = \frac{1}{2K^2\overline{U}} \sum_{i=1}^{K} \sum_{j=1}^{k} \mid U_i - U_j \mid, \tag{3.42}$$

where

- K: number of components of the organization;
- U_i: number of units owned by the ith component;
- $\overline{U}$: average number of units owned by the system components.

If the components are ranked with respect to the units they own ($U_1 \geq U_2 \geq \cdots \geq U_K$), then the equation for the index of Gini is

$$I_{19} = 1 + \frac{1}{K} - \frac{2}{K^2 \overline{U}} \sum_{i=1}^{K} i U_i. \tag{3.43}$$

3.8.3 Index of Kuznets

The equation for this index is [19]

$$I_{20} = \frac{1}{2K\overline{U}} \sum_{i=1}^{K} \mid U_i - \overline{U} \mid . \tag{3.44}$$

where

- K: number of components of the organization;
- U_i: number of units owned by the ith component;
- $\overline{U}$: average number of units owned by the system components.

The index of Kuznets has a form that is similar to that of the index of Gini, discussed above. There is, however, a difference. In the case of the index of Gini, one compares each component to each other component with respect to the number of possessed units (papers, citations, or money). In the case of the index of Kuznets, the comparison is different: the number of units possessed by each component is compared to the mean number of possessed units.

3.8.4 Pareto Diagram (Pareto Chart)

The Pareto diagram, also called a Pareto chart, is famous in the area of econometrics [76, 77]. In general, it is constructed as follows. On the abscissa of the coordinate system one puts the logarithm of the number of units (number of citations, for example). On the ordinate, one puts the logarithm of the relative cumulative frequencies (of the number of scientists that have the corresponding number of citations).

It can happen (as happens often in econometrics) that some of the points are approximately on a straight line (Pareto line). Then the angle between the Pareto line and the abscissa (the coefficient α of Pareto) is a characteristic measure of the corresponding distribution.

3.9 Indexes for the Case of Stratified Data

In some cases, the empirical data are stratified into layers. For example, we know the number of researchers who have published between zero and five papers; then the number of researchers who have published between six and ten papers, etc. We do not know the distribution within the layers (e.g., we do not know how many scientists have written seven papers). In addition, it may happen that the sizes of the different layers are not the same.

There are equations for many of the indexes for the case of stratified data. For the indexes discussed above, some of the equations are as follows [19]:

- Index of Gini for stratified data. The equation is

$$I_{19}^{*} = \left[\sum_{i=1}^{M} \left(2 \sum_{j=1}^{i} P_i - P_j \right) P_i \frac{U_i}{\overline{U}}, \right] - 1 \tag{3.45}$$

  where

  – M: number of layers for the stratified data;
  – P_i: percentage of the total number of units possessed by the ith component;
  – U_i: number of units owned by the ith component;
  – $\overline{U}$: average number of units owned by the system components,

  where U_i are ordered as follows: $U_1 \leq U_2 \leq \cdots \leq U_M$.
- Index of Kuznets for stratified data. The equation is

$$I_{20}^{*} = \frac{1}{2} \sum_{i=1}^{M} P_i \mid \frac{U_i}{\overline{U}} - 1 \mid, \tag{3.46}$$

  where

  – M: number of layers for the stratified data;
  – P_i: percentage of the total number of units possessed by the ith component;
  – U_i: number of units owned by the ith component;
  – $\overline{U}$: average number of units owned by the system components.
- Coefficient of variation for stratified data. The equation is

$$I_{4}^{*} = \frac{1}{\overline{U}} \sqrt{\sum_{i=1}^{M} (U_i - \overline{U})^2 P_i}, \tag{3.47}$$

  where

  – M: number of layers for the stratified data;
  – P_i: percentage of the total number of units possessed by the ith component;

– U_i: number of units owned by the ith component;
– $\overline{U}$: average number of units owned by the system components.

The equation for the coefficient of logarithmic variance is

$$I_5^* = \sum_{i=1}^{M} \left(\ln \frac{U_i}{\overline{U}} \right)^2 P_i. \tag{3.48}$$

- Index of Theil. The equation is

$$I_{21} = \sum_{i=1}^{M} P_i \frac{U_i}{\overline{U}} \log_2 \left(\frac{U_i}{\overline{U}}, \right) \tag{3.49}$$

where

– M: number of layers for the stratified data;
– P_i: percentage of the total number of units possessed by the ith component;
– U_i: number of units owned by the ith component;
– $\overline{U}$: average number of units owned by the system components.

Up to now, we have discussed a group of researchers. When one has to compare several groups of researchers (for example, several institutes of physics belonging to a national research institution), one may use additional indexes. Some of them will be discussed below.

3.10 Indexes of Inequality and Advantage

3.10.1 Index of Net Difference of Lieberson

The equation for this index is [79]

$$I_{22} = \sum_{i=1}^{I} A_i \left(\sum_{j=1}^{i-1} B_j \right) - \sum_{i=1}^{I} B_i \left(\sum_{j=1}^{i-1} A_j \right), \tag{3.50}$$

where

- I: number of classes;
- i: a class of the ranked distribution of the classes;
- A_i: proportion of units of group A in the class i;
- B_i: proportion of units of group B in the class i;
- $\left(\sum_{j=1}^{i-1} A_j \right)$: cumulative percentage of units of group A ranked below class i;

- $\left(\sum_{j=1}^{i-1} B_j\right)$: cumulative percentage of units of group B ranked below class i.

Within the scope of our example about the researchers and their publications, the application of the index can be as follows, for example. Let us define $I = 6$ classes: between 0 and 10 papers; between 11 and 20 papers; between 21 and 30 papers; between 31 and 40 papers; between 41 and 50 papers; and over 50 papers. Let us define the two groups of researchers as follows:

- group A: young researchers up to 40 years old;
- group B: researchers over 40 years old.

Then I_{22} will measure the net difference between the young and mature researchers with respect to the six classes defined above (and connected to the number of papers written by a scientist).

The index of net difference of Lieberson can be used to investigate segregation [80]. In the area of scientific systems and structures, the index has been used, for example, for studying the distribution of scientific positions for women in Israel [81].

3.10.2 Index of Average Relative Advantage

The equation for this index is [82]

$$I_{23} = \sum_{i=1}^{I} \sum_{j=1}^{J} k_{ij} A_i B_j, \tag{3.51}$$

where

- A_i: proportion of units of group A in the class i;
- B_j: proportion of units of group B in the class j;

and k_{ij} is a coefficient that has values as follows:

- $k_{ij} = \frac{A_i - B_i}{A_i}$ if $A_i > B_i$;
- $k_{ij} = 0$ if $A_i = B_i$, $k_{ij} = \frac{A_i - B_i}{B_i}$ if $A_i < B_i$.

This index accounts for all possible pairwise combinations, and it weights them by a coefficient that is proportional to the relative magnitude of the advantage involved (where the advantage is understood as a larger share of the units of class A in comparison to the units of class B).

The index of average relative advantage has been used to study the advantages and disadvantages of social groups with respect to jobs, income, education, etc. But this index can also be used to study groups of researchers with respect to the characteristics of their scientific production (such as number of papers or number of citations).

Let us now consider two indexes of inequity. These indexes measure the deviation from uniformity in some distribution.

3.10.3 Index of Inequity of Coulter

The equation for this index is [83]

$$I_{24,\alpha} = \frac{[\sum_{i=1}^{K} \mid P_i - Q_i \mid^{\alpha}]^{(1/\alpha)}}{[(1 - \min(Q))^{\alpha} - (\min Q)^{\alpha} + \sum_{k=1}^{K} Q_i^{\alpha}]^{(1/\alpha)}}, \tag{3.52}$$

where

- P_i: the proportional share of a component;
- Q_i: the proportional share that should be received by the component with respect to the equity standard distribution;
- $\min(Q)$: the smallest value of Q;
- α: a value that is set by the investigator. The value of α determines the sensitivity of the index to concentration. Thus an appropriate choice of α makes the index sensitive not only to inequality but also to concentration to the degree that is desired by the investigator.

The inequality index of Coulter may be used in the analysis of possible locations of different facilities (including scientific facilities) [84].

3.10.4 Proportionality Index of Nagel

The equation for this index is [18]

$$I_{25} = 1 - \frac{\sum_{i=1}^{K}(P_i - A_i)^2}{\sum_{i=1}^{K}(Q_i - A_i)^2}, \tag{3.53}$$

where

- P_i: actual frequency distribution of the units to the components (proportion of units assigned to the ith component);
- A_i: distribution of units to the components in proportion to merit (standard distribution—shares that would occur if the units were distributed in proportion to an equity standard such as merit);
- Q_i: zero allocation (the most inequitable distribution of units to the components possible). Often the distribution treats the case in which one of the components owns all the units and the other components do not possess anything.

In our example, P_i is proportional to the number of publications of the ith researcher; A_i reflects the situation in which all researchers have the same number of publications. And the values of Q_i correspond to the situation that one of the researchers has written all the publications and the other researchers have written none.

The frequency of quantitative evaluations of national research systems has been increasing [85–93]. Because of this, we shall discuss below the following methods and sets of indicators and indexes for performing such evaluation: the RELEV method for assessment of scientific research performance within public institutes; indexes and indicators for comparison among scientific communities in different countries; efficiency of research production from the point of view of publications and patents, etc.

3.11 The RELEV Method for Assessment of Scientific Research Performance in Public Institutes

The RELEV method [94–97] assigns a single numerical value to the research performance of a research institute. With respect to this value, different institutes working on closely related research fields can be compared. The index provided by the RELEV method can be a useful addition to the basket of indexes that form the quantitative part of research evaluation in a system of research institutions. The definition of the index for the ith institution from the set of compared institutions is as follows:

$$\Omega_{\mathrm{RELEV}}(i) = 3 - X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i} + 2X_{6i} + X_{7i}, \tag{3.54}$$

where seven indexes connected to the evaluated n institutions are taken into account:

1. A: Index of public funds attributed to the research institutions, $(\alpha_1, \ldots, \alpha_n)$;
2. B: index of self-financing (funds attracted by the research institution in addition to the public funds), $(\beta_1, \ldots, \beta_n)$;
3. X: index of personnel in training (number of trained individuals), $\xi_1, \ldots, \xi_n$;
4. Δ: index of teaching activities of researchers (hours of teaching by the scientists), $\delta_1, \ldots, \delta_n$;
5. E: index of national publications (numbers of national publications), $\varepsilon_1, \ldots, \varepsilon_n$;
6. Φ: index of international publications (number of international publications), $\phi_1, \ldots, phi_n$;
7. Γ: patent index (number of patents), $\gamma_1, \ldots, \gamma_n$.

The indexes above can be calculated in two ways: per researcher; as the total number for the corresponding institution. Our experience shows that in most cases, it is more reasonable to calculate the above indexes per researcher.

Let $max_A, max_B, max_X, max_\Delta, max_E, max_\Phi, max_\Gamma$ be the maximum values of corresponding indexes in the set of evaluated institutions. Then

$$X_{1i} = \alpha_i / max_A;\ X_{2i} = \beta_i / max_B;\ X_{3i} = \xi_i / max_X;\ X_{4i} = \delta_i / max_\Delta;$$
$$X_{5i} = \varepsilon_i / max_E;\ X_{6i} = \phi_i / max_\Phi;\ X_{7i} = \gamma_i / max_\Gamma. \quad (3.55)$$

Some of the indexes can have larger weights, as, for example, the index X_{6i} connected to publications of international journals. Weight coefficients can be introduced for all indexes, and this is a main direction of work on adjustment of the RELEV method in evaluating institutions [95, 97]. In addition, the number of indexes can be increased or some of the indexes can be changed. This depends on the specifics of the evaluated institutions.

3.12 Comparison Among Scientific Communities in Different Countries

Countries can be compared with respect to different characteristics of their scientific communities. For this, one needs an appropriate system of indicators and indexes. Below, we shall present the methodology of an important comparison of the correlation between the structure of scientific research, scientometric indicators, and GDP of several countries from the EU and outside the EU [98].

The methodology is based on the following indicators and indexes of research production of the scientific community of a country:

1. **Journal paper citedness**

$$\text{JPC} = \frac{C}{P}, \quad (3.56)$$

where

- P: number of journal papers produced by the research community in a country for the time interval of interest;
- C: number of citations obtained by the researchers from the scientific community of a country for the time interval of interest.

2. **Relative subfield citedness**

$$\text{RW} = \frac{C}{P[C/P]_{st}}, \quad (3.57)$$

where

- P: number of journal papers produced by the research community in a country for the time interval of interest and for the research field of interest;
- C: number of citations obtained by the scientists from the scientific community of a country for the time interval of interest and for the scientific field of interest.

- $[C/P]_{st}$: Journal paper citedness for the corresponding field in the world (obtained by the data from a large database such as Web of Science or Scopus).

3. **Journal paper productivity**

$$\text{JPP} = \frac{P}{Pop}, \tag{3.58}$$

where

- P: number of journal papers of a country;
- Pop: population of the country in millions of people.

4. **Highly cited papers productivity**

$$\text{HCPP} = \frac{HCP}{Pop}, \tag{3.59}$$

where

- HCP: number of highly cited papers (ranking among the top 1 % most cited for their subject field and year of publication);
- Pop: population of the country in millions of people.

5. **Relative prominence index**

$$\text{RPI} = \left(\frac{P_c}{\sum P_c}\right) / \left(\frac{P}{\sum P}\right), \tag{3.60}$$

where

- $\frac{P_c}{\sum P_c}$: share of cited papers of a country within the total number of papers cited in the world;
- $\frac{P}{\sum P}$: share of journal papers of a country within the total number of papers in the world.

6. Specific impact contribution

$$\text{SIC} = \frac{C\%}{P\%}, \tag{3.61}$$

where

- $C\%$: percentage share of citations of a country within the total number of citations in the world;
- $P\%$: percentage share of a country in journal papers within the total number of papers in the world.

7. **Rate of highly cited researchers**

$$\text{RHCR} = \frac{HCR}{Pop}, \tag{3.62}$$

where

- HCR: number of researchers of a country in the top 1 % of the researchers most cited;
- Pop: population of the country in millions of people.

8. **Composite publication index**

$$\text{CPI} = w_1(JPP) + w_2(SIC) + w_3(HCPP), \tag{3.63}$$

where

- n: number of countries in the world;
- $w_1 = 1/\sum_{i=1}^{n} JPP_i$;
- $w_2 = 2/\sum_{i=1}^{n} SIC_i$;
- $w_3 = 3/\sum_{i=1}^{n} HCPP_i$

9. **Field structure difference index for country k in field i**

$$\text{FSD}_{k,i} = \frac{(P_{k,i} - P_{s,i})^2}{P_{s,i}}, \tag{3.64}$$

where

- $P_{k,i}$: is the percentage share of publications of country k in the ith scientific field.
- $P_{s,i}$: is the mean percentage share of the standard. As the standard one considers fourteen European Community member states (member states that are not from Eastern Europe) plus the USA and Japan.

10. **Mean structural difference index**

$$\text{MSD}_k = \frac{1}{F}\sum_{i=1}^{F} \frac{(P_{k,i} - P_{s,i})^2}{P_{s,i}}, \tag{3.65}$$

where i is the number of considered scientific subfields.

Vinkler [98] applied the above procedure to the EU countries and to several other countries. We note here that the differences among the countries are well exhibited by the values of the mean structural difference index. For example, the value of this index for Germany for 1995–2005 was 0.18; for the USA, 0.68; for the Czech Republic, 1.17; and for Bulgaria, 2.25. And while the Czech Republic has moved close to the fourteen West European countries, the structure of science in Bulgaria differs greatly from the standard (provided by the fourteen EU countries plus the USA and Japan).

3.13 Efficiency of Research Production from the Point of View of Publications and Patents

As countries become more developed, the ratio between paper production and patent production changes [99]. And the ratio between produced papers and produced patents normalized by population of the country can be considered an index of efficiency of the corresponding national research system. This methodology is developed further in [100]. An analysis of a country's efficiency (within some group of countries) can be made the basis of the following indexes:

1. **Patents–papers index**

$$E_1 = \frac{Pat}{Pap}, \tag{3.66}$$

where

- Pat: number of patents per one million inhabitants of the country;
- Pap: number of papers per one million inhabitants of the country.

2. **Expenditure efficiency index**

$$E_2 = \frac{GERD}{Pap}, \tag{3.67}$$

where

- Pap: number of papers written by the country's researchers;
- $GERD$: gross expenditure on research and development.

3. **Manpower efficiency index**

$$E_3 = \frac{Pap}{MP}, \tag{3.68}$$

where

- Pap: number of papers written by the country's researchers;
- MP: manpower (number of people participating in research activities).

4. **Patent expenditure efficiency index**

$$E_4 = \frac{GERD}{Pat}, \tag{3.69}$$

where

- Pat: number of patents obtained by the country's researchers;
- $GERD$: gross expenditure on research and development.

5. **Patent manpower efficiency index**

$$E_5 = \frac{Pat}{MP}, \tag{3.70}$$

where

- Pat: number of patents obtained by the country's researchers;
- MP: manpower (number of people participating in research activities).

An analysis of several countries performed in [100] shows low efficiency in publishing but high efficiency in patenting in the USA. This pattern is observed also for Germany, Japan, France, and Korea, and China is moving to join this club.

3.14 Indicators for Leadership

Indicators for leadership can be used to assess institutional and national publication activities. Klavans and Boyack [101] consider three kinds of indicators for leadership.

1 *Indicators for current leadership*: Current leadership indicators are connected to the count of the current research publications. These indicators refer to research groups, research institutions, or countries that lead in terms of numbers of papers published, particularly if attention is paid to the most current literature [102].
2. *Indicators for discovery leadership*: These indicators refer to research groups, research institutions, or countries that lead in terms of any of a number of impact measures, which are typically based on citation counts to older literature. For example, a nation with a larger fraction of highly cited papers in a particular field may be considered a discovery leader in the corresponding research field [103]. Other indicators for discovery leadership may be the total citations and fraction of the top one percent of highly cited papers for the earlier time period. One has to be careful, since citation levels can be artificially inflated due to self citations. Special attention should be given to negative citations, which may indicate problems in the corresponding research.
3. *Indicators for thought leadership*: These indicators are a bridge between current leadership and discovery leadership. Thought leadership is an activity measure that examines whether current papers are building on more recent discoveries or on older discoveries in a field. An indicator for thought leadership is the mean reference date in the list of references of the published articles. Thought leadership shows the research groups, institutions, or countries that are quick to follow recent discoveries, e.g., a research group with mean reference date 2012 is quicker to follow research discoveries in comparison to a research group whose mean reference date is 1999. A research group, research organization, or country is considered a thought leader if it is building on the more recent discoveries in its field. At the national policy level, the measure of thought leadership should be age of the scientific environments that the nation wants to pursue [101]. At

this level, the nations that are thought leaders fund mostly young research areas. But even in young research areas, there are discoveries that are of different ages. This is connected to thought leadership at the group (laboratory) level. At this level, where the choice of topic is given, the measure shifts to relative age. Thus when an area of science is targeted, the scientists from groups that are thought leaders focus on the most recent discoveries within this area. Then a country may be a thought leader in some research (i.e., the most recent research areas for this kind of research are funded), but the research groups in this country may not be thought leaders in the corresponding research (if they focus on discoveries that are not the latest in the corresponding research areas).

3.15 Additional Characteristics of Scientific Production of a Nation

Schubert and Braun [104] considered the following relative indexes of scientific production of researchers from different nations and scientific fields (the indexes can be applied also to scientific organizations within a country):

1. **Activity index**
   This index was proposed in [105] and further studied in [106]. It is defined as follows:
   $$AI = \frac{N_1}{N_2}, \tag{3.71}$$
   where
   - N_1: the given field's share in the country's publication output;
   - N_2: the given field's share in the world's publication output.

   $AI = 1$ means that the country's research effort in a given scientific field corresponds to the world average; $AI > 1$ means that the country's effort is greater than the world's average effort.
   Instead of the world average, one can use the average with respect to a set of countries of interest. In this case, the activity index becomes
   $$AI^* = \frac{N_1}{N_2^*}, \tag{3.72}$$
   where
   - N_1: the given field's share in the country's publication output;
   - N_2^*: the given field's share in the publication output of the selected set of countries.

   On the basis of the activity index, one can introduce the *relative specialization index*

$$RSI = \frac{AI - 1}{AI + 1}. \tag{3.73}$$

The relative specialization index has values from -1 to 1 inclusive. $RSI = -1$ means that there is no activity in the corresponding research field. $RSI = 1$ arises when no field other than the given one is active. Negative values of RSI indicate activity that is lower than the average activity. Positive values of RSI indicate activity that is higher than average activity. $RSI = 0$ means that the country's research effort in a given scientific field corresponds to the world average.
The relative specialization index gives evidence of the existence of four patterns in the national publication profiles of the countries of the world [5]:

- *The Western model*: the characteristic pattern of the developed Western countries with clinical medicine and biomedical research as dominating fields;
- *The Japanese model*: engineering and chemistry are dominant. This model is typical also for other developed Asian economies;
- *The former socialist countries model*: physics and chemistry are dominant. Such a model may be observed in the East-European countries, Russia, and China;
- *The bio-environmental model*: biology and earth and space sciences are dominant. Such a model is observed in Australia, South Africa, and some developing countries with relatively large territory and natural resources.

2. **Attractivity index**

$$AAI = \frac{N_3}{N_4}, \tag{3.74}$$

where

- N_3: the given field's share in the citations attracted by the country's publications;
- N_4: the given field's share in the citations attracted by all publications in the world.

This index can be reformulated to compare a country to a set of other countries:

$$AAI^* = \frac{N_3}{N_4^*}, \tag{3.75}$$

where

- N_3: the given field's share in the citations attracted by the country's publications;
- N_4^*: the given field's share in the citations attracted by all publications in the selected set of countries.

3. **Relative citation rate**
   This index is defined as

$$RCR = \frac{N_5}{N_6}, \tag{3.76}$$

   where

   - N_5: observed citation rate over all papers published by the given country in the given field;
   - N_6: observed citation rate over all papers published by the selected set of countries in the given field.

   Observed citation rate of a paper is the actual citation rate and *expected citation rate of a paper* is the average citation rate of the journal in which the paper has been published.

   $RCR > 1$ means that the papers produced by the scientists of a country in the scientific field of interest are more frequently cited than the standard citation rate, and $RCR < 1$ means that the papers are less frequently cited than expected (one reason for this (among many reasons) may be related to their quality).

On the basis of the activity and attractivity indexes, one can produce a *relational chart* of countries (or of scientific organizations in a country). The relational chart is produced as follows: The value of the activity index appears on the x-axis; and the value of the attractivity index appears on the y-axis. The diagonal is the line where the observed and expected citation rates match exactly. If a point corresponding to a country is below the diagonal (and far from the diagonal), this is a sign of problems. A significant distance of a point from the diagonal means that AI or AAI differ significantly for 0. There is a test to check whether the difference is significant [104]:

1. One calculates

$$t_{AI} = \frac{AI - 1}{\Delta_{AI}}; \quad t_{AAI} = \frac{AAI - 1}{\Delta_{AAI}}, \tag{3.77}$$

   where

$$\Delta_{AI} = AI\sqrt{1/N - 1/S}; \quad \Delta_{AAI} = AAI\sqrt{1/M - 1/T},$$

   and

   - N: number of country's publications in the given field;
   - M: number of country's citations in the given field;
   - S: number of country's publications in all scientific fields;
   - T: number of country's citations in all scientific fields;

2. if $t < 2$, the corresponding indicator does not differ significantly from 1 at a significance level of 0.95.

An analogous test can also be performed for the relative citation rate. First one calculates

$$t_{RCR} = \frac{RCR - 1}{\Delta_{RCR}}, \tag{3.78}$$

where

$$\Delta_{RCR} = \sqrt{RCR\frac{Q}{N}}$$

and

- N: country's publications in the given field;
- Q: solution of the equation $\frac{\ln Q}{Q-1} = -\frac{\ln f}{X}$, where X is the mean observed citation rate per publication and f is the fraction of uncited publications.

Then if $t_{RCR} < 2$, RCR does not differ significantly from 1 at a significance level of 0.95.

On the basis of the RCR index, one can introduce another index that rewards papers with RCR value larger than 1 and "punishes" papers with RCR smaller than 1 [107]. This index is just

$$RCR_2 = (RCR)^2. \tag{3.79}$$

We shall finish our discussion of production of researchers from a nation with a description of a set of indexes for measurement of scientific production [108] called FSS-indexes ("Fractional Scientific Strength" indexes). These indexes are based on a measurement of average yearly labor production of researchers at various levels of units (individual, field, discipline, entire organization, region, country). The FSS-indexes connect the salary of researchers with results of their research measured by publications and citations.

The FSS-indexes at different levels are

1. **Individual level**

$$FSS_R = \frac{1}{S_R}\frac{1}{t}\sum_{i=1}^{N} f_i \frac{c_i}{\overline{c}}, \tag{3.80}$$

where

- S_R: average yearly salary of researcher;
- t: number of years of work of researcher in the period of observation;
- N: number of publications of researcher in the period of observation;
- f_i: fractional contribution of researcher to publication i;
- c_i: citations received by the ith publication;
- $\overline{c}$: average number of publications received for all cited publications of the same year and subject category.

2. **Research field level**

$$FSS_F = \frac{1}{S_F} \sum_{i=1}^{N} f_i \frac{c_i}{\overline{c}}, \tag{3.81}$$

where

- S_F: total salary of the research staff (working in the corresponding research field) in the observed period;
- N: Number of publications of the above research staff in the period of observation;
- f_i: fractional contribution of researchers form evaluated group to publication i;
- c_i: citations received by the publication i;
- $\overline{c}$: average number of publications received for all cited publications of the same year and subject category.

3. **Department level**

$$FSS_D = \frac{1}{N_{RS}} \sum_{i=1}^{N_{RS}} \frac{FSS_{R_i}}{\overline{FSS_R}}, \tag{3.82}$$

where

- N_{RS}: number of researches in the department for the observed period;
- FSS_{R_i}: productivity of the ith researcher from the department for the observed period;
- $\overline{FSS_R}$: average national productivity of all productive researchers from the same scientific discipline.

4. Level of multifield units: Such units, for example, are universities or a system of research institutes or even the entire national research system. In this case,

$$FSS_U = \sum_{i=1}^{N_U} \frac{S_{SD_k}}{S_U} \frac{FSS_{SD_k}}{\overline{FSS_{SD_k}}}, \tag{3.83}$$

where

- S_U: total salary of the research staff of the multifield unit for the observed period;
- S_{SD_k}: total salary of the research staff from the observed unit that works in the scientific discipline k in the observed period of time.
- N_U: number of scientific disciplines in the observed unit;
- FSS_{SD_k}: labor productivity in the scientific discipline SD_k of the evaluated unit;
- $\overline{FSS_{SD_k}}$: weighted average of the research productivities in all other units of the kind of unit that is evaluated (of all other universities if the evaluated unit is a university)

The FSS-indexes could lead to quite interesting results for research units and countries where the salaries of researchers are low and their scientific production is not very low. Then it can happen that the effectiveness of the research units in such countries is very good.

3.16 Brief Remarks on Journal Citation Measures

Journal citation measures are much used in library science, research evaluation, etc. In research evaluation, the journal citation measures are applied at all levels: from evaluation of research of individual researchers to evaluation of national research performance. Because of this, we shall mention below several of these measures.

The first very successful journal citation measure was the *impact factor* [109]. The relationship for this index for a journal is

$$IF_n = \frac{c_n}{p_{n-2} + p_{n-1}}, \tag{3.84}$$

where

- c_n: number of citations obtained in the year n by the papers published in the journal in the years $n-1$ and $n-2$;
- p_{n-1}: number of papers published in the journal in the year $n-1$;
- p_{n-2}: number of papers published in the journal in the year $n-2$.

The impact factor is much used today, and it has various strengths such as stability, reproducibility, comprehensibility (the impact factor measures the frequency with which an average article published in a given journal has been cited in a particular year) and independence of the size of the journal (on the number of articles published in the journal per year). In order to be useful, the impact factor must be used carefully, e.g., the impact factors of journals must be used with great care for the purposes of comparison of production of researchers from different scientific areas. One should keep in mind, e.g., that a single measure might not be sufficient to describe citation patterns of scientific journals [5].

In analogy to the impact factor, one may also define the *intermediacy index*

$$II_n = \frac{c_n}{p_n}, \tag{3.85}$$

where

- c_n: number of citations obtained in year n by the papers published in the journal in year n;
- p_n: number of papers published in the journal in year n.

Another index is the *SNIP indicator* (source normalized impact per paper) [110]. The classic version of SNIP is defined as follows:

$$\text{SNIP} = \frac{\text{RIP}}{\text{RDCP}}, \tag{3.86}$$

where

- RIP (raw impact per paper): the RIP value of a journal is equal to the average number of times the journal's publications in the three preceding years were cited in the year of analysis. For example, if 200 publications appeared in a journal in the period 2012–2014 and if these publications were cited 600 times in 2015, then the RIP value of the journal for 2015 equals $600/200 = 3$. What is specific is that *in the calculation of RIP values, citing and cited publications are included only if they have the Scopus document type article, conference paper, or review*. The RIP indicator is similar to the journal impact factor, but the RIP indicator uses three instead of two years of cited publications and includes only citations to publications of selected document types.
- RDCP (relative database citation potential): RDCP is calculated as follows:

$$\text{RDCP} = \frac{\text{DCP}}{\text{m(DCP)}}, \tag{3.87}$$

  where

  – DCP (database citation potential): DCP is calculated as follows:

$$\text{DCP} = \frac{\sum_{i=1}^{n} r_i}{n}, \tag{3.88}$$

    where n is the number of publications in the subject field of the journal and r_i denotes the number of references in the ith publication to publications that appeared in the three preceding years in journals covered by the database.
  – m(DCP): the median DCP value of all journals in the database.

Finally, let us mention the *SJR*: Scimago journal rank, which is based on the transfer of prestige from a journal to another journal [111]. Prestige is transferred through the references that a journal makes to the rest of the journals and to itself. The SJR is calculated as follows:

$$\text{SJR}_j = \frac{1-d-e}{N} + \frac{e\,\text{Art}_i}{\sum_{j=1}^{N}\text{Art}_i} + d\sum_{j=1}^{N}\frac{C_{ji}\text{SJR}_j}{C_j}\frac{1-\left[\sum_{k\in\{\text{Dangling nodes}\}}\right]}{\sum_{h=1}^{N}\sum_{k=1}^{N}\frac{C_{kh}\text{SJR}_k}{C_k}} + d\left[\sum_{k\in\{\text{Dangling nodes}\}}\right]\frac{\text{Art}_i}{\sum_{j=1}^{N}\text{Art}_j}, \tag{3.89}$$

where

- C_{ij}: citations from journal j to journal i.
- C_j: number of references of journal j.
- N: number of journals.
- d: constant (usually equal to 0.85).
- e: constant (usually equal to 0.1).
- Art_i: number of articles in journal i.
- Dangling nodes: these are journals of the universe that do not have references to any other journal of the universe, although they can be cited or not. They constitute impasses in a graph, since from them it is impossible to jump to other nodes. In order to ensure that the iterative process is convergent, dangling nodes are virtually connected to all those of the universe, and its prestige is distributed between all the nodes proportionally to the number of articles of each.

On the basis of the SDJR, one can calculate another index specific to the ith journal:

$$\text{SJRQ}_i = \frac{\text{SJR}_i}{\text{Art}_i}. \tag{3.90}$$

The iterative procedure of calculation of the SJR involves the following three steps:

1. Initial assignment of the SJR: a default prestige is assigned to each journal. The calculation of the SJR is a converging process, so the initial values don't determine the final result (but the initial values influence the number of iterations needed).
2. Iteration process of calculation: departing from step 1, the computation is iterated to calculate the prestige of each journal based on the prestige transferred by the rest. The process ends when the variation of the SJR between two iterations is less than a limit fixed before the calculation process. The final result is the SJR of each journal.
3. Computation of SJRQ: After the computation of SJR of all journals, one divides the SJR by the number of articles published in the citation window. The result is the average prestige per article.

Another version of the SJR (the SJR2) is also available [112]. Let us note that a major drawback of the journal impact factor is its lack of field (subject) normalization, i.e., differences in citation volumes between different fields are not taken into account. SNIP belongs to indexes that are based on the idea that citations to publications should be normalized with respect to the length of the reference lists of the citing publications (sources). The source normalized indexes are based on the observation that the reference lists' lengths vary across fields. Source-normalized indexes do not require a field classification scheme. There are also indexes based on other ideas. An example is MNCS (mean normalized citation score) [113, 114], based on the approach to field normalization, in which a classification scheme is used (i.e., each publication is assigned to one or more of the fields of the scheme). In the case of

MNCS, citation scores of the target publications (e.g., the publications under evaluation) are compared to expected citation scores for publications in the fields to which the publications belong (these fields are the Thomson Reuters subject categories of journals).

3.17 Scientific Elites. Geometric Tool for Detection of Elites

Elites are very important parts of social structures [115–117]. There exist characteristic features of research organizations that lead to the formation of research elites. Usually a small number of researchers publish many papers and a small number of researchers are highly cited. These categories of researchers form some of the scientific elites. Elites are of great importance for the dynamics and evolution of scientific structures and systems. Because of this, scientific elites are the subject of intensive research [118–129].

There is a **square root law of Price** [130]: *half of the literature on a subject will be contributed by the square root of the total number of authors publishing in that area.*

Let $g(x)$ represent the probability of an author making x published contributions to a subject field. Then the mathematical formulation of the square root law of Price is [131]

$$\lim_{x_{\max}\to\infty}\left[\frac{\sum_{x=h}^{x_{\max}} x\ g(x)}{\sum_{x=1}^{x_{\max}} x\ g(x)}\right]=\frac{1}{2}, \tag{3.91}$$

where h is such that

$$\left[\sum_{x=1}^{x_{\max}} g(x)\right]^{1/2}=\sum_{x=h}^{x_{\max}} g(x). \tag{3.92}$$

Let the total number of authors in a scientific discipline be A. The law of Price can be generalized as follows [78]: A^{α} *authors will generate a fraction* α *of the total number of papers*. Then if $\alpha = 1/2$, one obtains the square root law of Price.

One can select groups of elite researchers on the basis of the law of Price. Another kind of possible rule for selecting an elite is the arithmetic $a\%/b\%$-rule: $a\%$ *of the papers are produced by* $b\%$ *of the scientists*. The most famous of these rules is the 80/20-rule: 80 % of the papers are produced by 20 % of the scientists. (Note that it is not necessary that $a + b = 100$.)

In the next chapter we shall discuss more of the theory of Price for scientific elites. This theory will lead us to the following conclusion: assuming the validity of the law of Lotka for scientific publications, one can obtain that the scientific elite consists of scientists whose number of publications is between $0.749\sqrt{i_{max}}$ and i_{max} publications (where i_{max} is the maximum number of publications written by a scientist from the corresponding group of scientists). And the size of this elite is about $\frac{0.812}{\sqrt{i_{max}}}$ of the size of the group of scientists. In this chapter we shall discuss another methodology for

determination of classes of scientific elites. This methodology is based on geometry and doesn't require validity of some law for scientific production. The corresponding measures will be obtained on the basis of the Lorenz curve for the ownership of scientific publications. As we have mentioned above, the Lorenz curve is an instrument for visualization of inequality in a population. It is very popular in the study of wealth distribution in a population [132–134]. Below, we shall be interested in the number of publications owned by researchers from some population (in our case, the population will consist of the members of a research institute). *We note that the measures of the sizes of the elites discussed below can be applied not only to populations of researchers but also to all populations that can be characterized by a Lorenz curve. Thus the methodology discussed below may be used to determine elites with respect to other characteristics of scientific production, such as the number of citations.*

3.17.1 Size of Elite, Superelite, Hyperelite, …

Let us consider the Lorenz curve shown in Fig. 3.1. Let us trace the diagonal from the point (0, 1) to the point (1, 0) in the (P, L)-plane. This diagonal crosses the Lorenz curve at a point with coordinates $(P_e, 1 - P_e)$. We shall consider the number $1 - P_e$

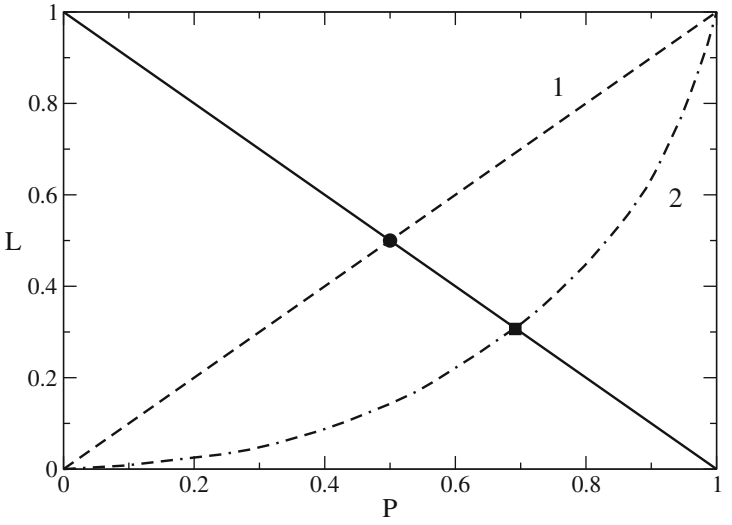

Fig. 3.1 Elite size measure by the Lorenz curve. The measure is the coordinate $1 - P_e$ of the cross point of the diagonal $(P, 1 - P)$ and the corresponding Lorenz curve. For the Lorenz curve marked by 1 (all scientists own the same number of papers), the cross point (*filled circle*) has coordinates (0.5, 0.5). In percentages, this is the 50/50-curve (nonelite distribution). For the Lorenz curve marked by 2 (corresponding to the situation at the Institute of Mechanics of the Bulgarian Academy of Sciences), the cross point (*filled square*) is (0.69, 0.31). In percentages, this is the 69/31 curve

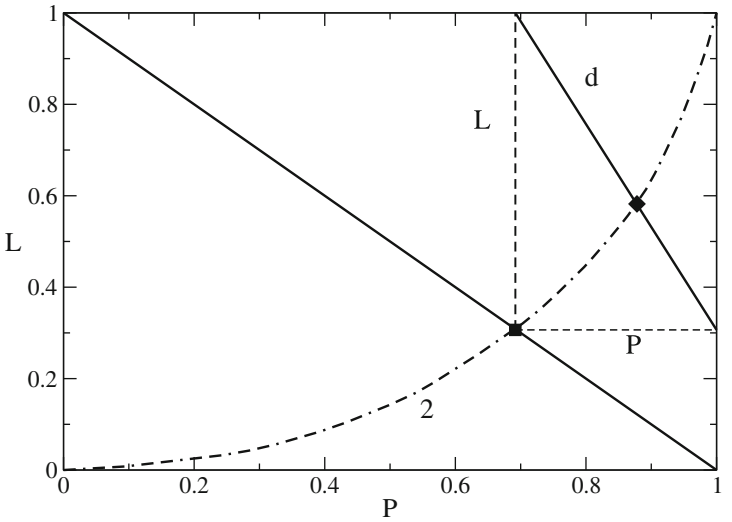

Fig. 3.2 The geometric measure for the scientific superelite by the Lorenz curve. The Lorenz curve marked by 2 is the same as in Fig. 3.1. One introduces a new Cartesian coordinate system with axes P^* and L^* and initial point that coincides with the point $(P_e, 1 - P_e)$ connected with the definition of the size of the scientific elite from Fig. 3.1. In this new coordinate system, the diagonal marked with d is plotted. The point (P_s, L_s) marked by a diamond gives the size and the production of the corresponding superelite. For the case of the Lorenz curve 2 (corresponding to the Institute of Mechanics of the Bulgarian Academy of Sciences), the coordinates of the point marked by a diamond are approximately $(P_s, L_s) = (0.88, 0.58)$, which means that the corresponding superelite consists of $1 - P_s = 0.12$, i.e., 12 % of the population of scientists owns $1 - L_s = 0.42$, i.e., 42 % of all papers. We recall here that the measure of the size of the elite from the previous figure tells us that the size of the elite of the institute was 31 % of the scientists, and this elite owns 69 % of the papers produced by the institute scientists

to be a measure of the size of the elite of the population corresponding to the Lorenz curve. Let us discuss this measure a bit further.

For the Lorenz curve corresponding to the case that all scientists own the same number of publications (in this case, the Lorenz curve is the diagonal that connects the points (0, 0) and (1, 1)), we have $P_e = 0.5$. We shall call such a curve a curve of class 50/50 (the elite has its maximum size). We can continue the construction of geometric measures one step further, and this will lead us to the concept of the scientific superelite. The procedure is illustrated in Fig. 3.2. The next step (definition of the superelite and its size) is geometrically analogous to the step that led us to the geometric measure of the size and production of the scientific elite. For this step, the initial point of the Cartesian coordinate system is not $(P, L) = (0, 0)$ but $(P, L) = (P_e, 1 - P_e)$, where P_e is the coordinate connected to the point corresponding to the geometric elite measure above (i.e., the point that is the intersection point of the Lorenz curve and the diagonal marked with a solid line in Fig. 3.2). Next we construct the axes P^* and L^* shown in Fig. 3.2. Finally, we plot the diagonal d shown

in Fig. 3.2, and the intersection point of this diagonal with the Lorenz curve gives us the geometric measure of the size and the production of the superelite. This point is marked with a diamond in Fig. 3.2, and its coordinates can be easily calculated. The coordinates of the point marked by a square (let us call it point E), which gives the size and production of the elite, are $E = (P_e, 1 - P_e)$. Then the coordinates of the point marked by a diamond (let us call it point S) are $S = (P_s, P_e \frac{P_s - P_e}{1 - P_e})$. For the case of the 61/39 curve marked by 2 in Fig. 3.2 and $P_s = 0.88$ measured by the intersection of the Lorenz curve and diagonal d, we obtain the coordinates of the point S to be approximately $S = (0.88, 0.58)$. In summary:

1. **Elite**: the coordinate P_e gives us information about the size and production of the scientific elite of the group of scientists described by the corresponding Lorenz curve.
2. **Superelite**: The coordinates P_e and P_s give us information about the size and production of the corresponding superelite.
3. **Hyperelite**: We can continue the process of construction of geometric measures starting now from the point S. What we shall obtain is the next point (let us call it H), which shall give us information about a smaller group of scientists called the hyperelite. The coordinates of this point will be $(P_h, \frac{P_h - P_s}{1 - P_s})$. Then the coordinates P_e, P_s, P_h will give us information about the size and production of the hyperelite.

The above geometric procedure may be continued further, and additional higher-order elites may be determined.

3.17.2 Strength of Elite

Next we can introduce a quantity that we shall call strength of the elite. Let us consider a geometric measure connected to the size and production of the elite. This measure is connected to the point E that has coordinates $(P_e, L_e = 1 - P_e)$. We define the strength of the elite as

$$s_e = \frac{1 - L_e}{1 - P_e} = \frac{P_e}{1 - P_e}. \tag{3.93}$$

We can define also the strength of the superelite. The coordinates of the point S connected to the size and production of the superelite are $S = (P_s, L_s)$. Then the strength of the superelite is defined as

$$s_s = \frac{1 - L_s}{1 - P_s} = \frac{1 - P_e \frac{P_s - P_e}{1 - P_e}}{1 - P_s} = \frac{1 - P_e(1 + P_s - P_e)}{(1 - P_e)(1 - P_s)}. \tag{3.94}$$

Finally, we can define the relative size of the superelite with respect to the size of the corresponding elite:

$$S_{se} = \frac{1 - P_s}{1 - P_e}. \tag{3.95}$$

Table 3.1 Parameters of the scientific elites and superelites of the studied institutes of the Bulgarian Academy of Sciences. $1 - P_e$: size of the scientific elite; $1 - L_e$: percentage of total number of papers owned by the members of the scientific elite; s_e: strength of the scientific elite; $1 - P_s$: size of the scientific superelite; $1 - L_s$: percentage of total number of papers owned by the members of the scientific superelite; s_e: strength of the scientific superelite. The studied institutes are from Bulgarian Academy of Sciences: Institute of Mathematics and Informatics (IMI); Institute of Mechanics (IMECH); Institute of Information and Communication Technologies (IIKT); Institute of Solid State Physics (ISSP); Institute of Electronics (IE); Institute of Optical Materials and Technologies (IOMT); Institute of Nuclear Research and Nuclear Energy (INRNE); Central Laboratory for Solar Energy and New Energy Sources (CLSENES)

Institute	$1 - P_e$ (%)	$1 - L_e$ (%)	s_e	$1 - P_s$ (%)	$1 - L_s$ (%)	s_s
IMI	34	64	1.88	14	38	2.71
IICT	30	70	2.33	12	40	3.33
IMECH	31	69	2.23	12	42	3.50
CLSENES	32	68	2.13	14	39	2.79
IOMT	35	65	1.86	16	35	2.19
IE	32	68	2.13	13	41	3.15
ISSP	34	66	1.88	14	39	2.79
INRNE	32	68	2.13	13	40	3.08

We note that the measures discussed above are different from the classic measures connected to the scientific elites.

Table 3.1 shows results about the size and production of the elites and superelites at the studied institutes of the Bulgarian Academy of Sciences. The sizes and productivities are very close, which means that in the size–production plane, the elites and the superelites form two clusters of researchers.

The elites at the mathematics and the physics institutes consist of about one-third *of the scientists, and these elites own about* two-thirds *of the scientific publications of the corresponding institute. The superelites consists of about* one-seventh *of the scientists, and they own about* two-fifths *of the scientific production. In addition, about two-thirds of the scientists do not belong to the scientific elites, and all these scientists own about one-third of the scientific production of the corresponding institute. Six-sevenths of the scientists do not belong to the superelite, and these scientists own about three-fifths of the scientific production of the corresponding institute.*

After selection of researchers that belong to elite, superelite, etc., one can study different characteristics of the selected groups of researchers. Here we shall mention only one of these characteristics: the age structure of the studied Bulgarian elites and superelites. Almost 80 % of the members of the superelites are of age 60 and older. In ten years, these scientists will no longer be staff scientists of the corresponding institute. Such people are also in the majority of the corresponding elites. The younger generation of scientists (ages between 40 and 60) is insufficiently represented in the scientific elites and scientific surepelites. Hence entire fields of national scientific research can be under the influence of aging researchers. This may have negative consequences, since many cases, the growth rate of research production is positive

and increases up to the ages about 30. After that age, the growth rate of research productivity usually begins to decrease [135]. This effect may not concern scientists belonging to superelites and hyperelites. And when such a researcher is no longer active, this is a great loss to the national research program in the corresponding research field.

References

1. M. Gibbons, C. Limoges, H. Nowotny, S. Schwartzman, P. Scott, M. Trow, *The new production of knowledge: the dynamics of science and research in contemporary societies* (Sage, London, 1994)
2. L.K. Hessels, H. van Lente, Re-thinking new knowledge production: a literature review and a research agenda. Res. Policy **37**, 740–760 (2008)
3. R.A. Boschma, Proximity and innovation: a critical assessment. Reg. Stud. **39**, 61–74 (2005)
4. I. Rafols, Knowledge integration and diffusion: measures and mapping of diversity and coherence, ed. by Y. Ding, R. Rousseau, D. Wolfram, *Measuring Scholarly Impact. Methods and Practice*. (Springer, Cham, 2014), pp. 169–192
5. W. Glänzel, *Bibliometrics as a research field: a course on theory and application of bibliometric indicators* (Ungarische Akademie der Wissenschaften, Budapest, 2003)
6. P. Brown, The half-life of the chemical literature. J. Am. Soc. Inform. Sci. **31**, 61–63 (1980)
7. R.E. Burton, R.W. Kebler, The "half-life" of some scientific and technical literatures. Am. Documentation **11**, 18–22 (1960)
8. P. Vinkler, *The Evaluation of Research by Scientometric Indicators* (Chandos, Oxford, 2010)
9. P. Vinkler, Publication velocity, publication growth and impact factor: an empirical model, ed by B. Cronin, H.B. Atkins. *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. ASIS Monograph Series (Information Today Inc, Medford, NJ, 2000), pp. 163–176
10. P. Vinkler, Research contribution, authorship and team cooperativeness. Scientometrics **26**, 213–230 (1993)
11. A. Przeworski, (Institutionalization of voting patterns or is mobilization the source of decay? Am. Polit. Sci. Rev. **69**, 49–67 (1975)
12. R.R. Shutz, On the measurement of income inequality. Am. Econ. Rev. **41**, 107–122 (1951)
13. A.B. Atkinson, On the measurement of inequality. J. Econ. Theory **2**, 244–263 (1970)
14. F.A. Cowell, *Measuring Inequality* (Oxford University Press, Oxford, UK, 2011)
15. P.D. Allison, Measures of inequality. Am. Sociol. Rev. **43**, 865–880 (1978)
16. A.R. Wilcox, Indices of qualitative variation and political measurement. Western Political Quart. **26**(2), 325–343 (1973)
17. A.L. Wilcox, *Indices of Qualitative Variation*. ORRN-TM-1919, (Oak Ridge National Laboratoty, Oak Ridge, Tennessee, 1967)
18. S.S. Nagel, *Public Policy: Goals, Means and Methods* (St. Martin Press, New York, 1984)
19. A.P. Lüthi, *Messung wirtschaflicher Ungleichheit.* Lecture Notes in Economic and Mathematical Systems No. 189 (Springer, Berlin, 1981)
20. C. Gini, *Variabilita e mutabilita* (Bologna, Italy, 1912)
21. L. Ceriani, P. Verme, The origins of Gini index: extracts from variabilita e Mutabilita (1912) by Corrado Gini. J. Econ. Inequality **10**, 421–443 (2012)
22. H.G.P. Jansen, Gini's coefficient of mean difference as a measure of adoption speed: theoretical issues and empirical evidence from India. Agric. Econ. **7**, 351–369 (1992)
23. I.I. Eliazar, I.M. Sokolov, Measuring statistical evenness: a panoramic overview. Phys. A **391**, 1323–1353 (2012)
24. S. Yitzhaki, E. Schechtman, *The Gini Methodology* (Springer, New York, 2013)

25. J.G. Rodriguez, R. Salas, The Gini coefficient: majority voting and social welfare. J. Econ. Theory **152**, 214–223 (2014)
26. B. Milanovic, A simple way to calculate the Gini coefficient, and some implications. Econ. Lett. **56**, 45–49 (1997)
27. C.J. Groves-Kirkby, A.R. Denman, P.S. Phillips, Lorenz curve and Gini coefficient: novel tools for analysing seasonal variation of environmental radon gas. J. Environ. Manage. **90**, 2480–2487 (2009)
28. J. Yang, X. Huang, X. Liu, An analysis of education inequality in China. Int. J. Educ. Dev. **37**, 2–10 (2014)
29. K. Kimura, A micro-macro linkage in the measurement of inequality: another look at the Gini coefficient. Qual. Quant. **28**, 83–97 (1994)
30. P.A. Rogerson, The Gini coefficient of inequality: a new interpretation. Lett. Spatial Resour. Sci. **6**, 109–120 (2013)
31. M.-H. Huang, H.-H. Chang, D.-Z. Chen, The trend in scientific research and technological innovation: a reduction of the predominant role of the U.S. in world research and technology. J. Infometrics **6**, 457–468 (2012)
32. A. Stirling, A general framework for analysing diversity in science, technology and society. J. Royal Soc. Interface **4**, 707–719 (2007)
33. A.O. Hirschman, *National Power and Structure of Foreign Trade* (University of California Press, Berkeley, CA, 1945)
34. O.C. Herfindahl, *Concentration in the steel industry*. Ph.D. Thesis, (Columbia University, 1950)
35. R. Linda, Competition policies and measures of dominant power, ed. by H.W. de Jorg. W.G. Shepherd, *Mainstreams in Industrial Organization* (Martinus Nijhoff Publishers, Dordrecht, 1986), pp. 287–307
36. G. Chammas, J. Spronk. Concentration measures in portfolio management, ed. by S. Greco, B. Bouchoin-Meunter, G. Colleti, M. Fedrizzi, B. Matarazzo, R.R. Yager, Advances in Computational intelligence, in *14th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, IPMU 2012, Catania, Italy, July 9–13, 2012, Proceedings, Part IV. (Springer, Berlin, 2012), pp. 94–103
37. A. Arlandis, E. Baranes, Interactions between network operators, content producers and internet intermediaries: empirical implications on network neutrality. Intereconomics **2**, 98–105 (2011)
38. W. Naude, R. Rossouw, Export diversification and economic performance: evidence from Brazil, China, India and South Africa. Econ. Change Restructuring **44**, 99–134 (2011)
39. J. Horvath, Suggestion for a comprehensive measure of concentration. South. Econ. J. **36**, 446–452 (1970)
40. J.L. Ray, D. Singer, Measuring the concentration of power in the international system. Sociol. Methods Res. **1**, 403–437 (1973)
41. R. Taagepera, J.L. Ray, A generalized index of concentration. Sociol. Methods Res. **5**, 367–383 (1977)
42. S. Lieberson, Measuring population diversity. Am. Sociol. Rev. **34**, 850–862 (1969)
43. L.A. Renzulli, H. Aldrich, Who can you turn to? The activation within core business discussion networks. Soc. Forces **84**, 323–341 (2005)
44. J.R. Bond, The influence of constituency diversity on electoral competition in voting for Congress 1974–1978. Legislative Stud. Quart. **8**, 201–217 (1983)
45. C.R. Rao, Diversity and dissimilarity coefficients: a unified approach. Theor. Popul. Biol. **21**, 24–43 (1982)
46. C. Ricotta, L. Szeidl, Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. Theor. Popul. Biol. **70**, 237–243 (2006)
47. L. Cassi, W. Mescheba, E. Turckheim, How to evaluate the degree of interdisciplinarity of an institution? Scientometrics **101**, 1871–1895 (2014)

48. O.D. Duncan, B. Duncan, A methodological analysis of segregation indexes. Am. Sociol. Rev. **20**, 210–217 (1955)
49. R. Taagepera, Inequality, concentration, imbalance. Polit. Methodol. **6**, 275–291 (1979)
50. D.W. Rae, M. Taylor, *The Analysis of Political Cleavages* (Yale University Press, New Haven, Conn, 1971)
51. L. Leydesdorff, Indicators of structural change in the dynamics of science: entropy statistics of the SCI Journal Citation Reports. Scientometrics **53**, 131–159 (2002)
52. H. Theil, The desired political entropy. Am. Polit. Sci. Rev. **63**, 521–525 (1969)
53. H. Theil, On the estimation of relationships involving qualitative variables. Am. J. Sociol. **76**, 103–154 (1970)
54. Y. Wang, Decomposing the entropy index of racial diversity: in search of two types of variance. Ann. Reg. Sci. **48**, 897–915 (2012)
55. K.D. Bailey, Sociological entropy theory: toward a statistical and verbal congruence. Qual. Quant. **18**, 113–133 (1983)
56. B. Raj, J. Koerts (eds.), *Henri Theil's Contributions to Economics and Econometrics*. Volume 2: Consumer demand analysis and information theory. (Kluwer, Dordrecht, 1992)
57. H. Theil, *Economics and Information Theory* (North Holland, Amsterdam, 1967)
58. D.F. Batten, *Spatial Analysis of Intercatting Economies* (Kluwer, Dordrecht, 1983)
59. K. Kesselman, French local politics: a statistical examination of grass roots consensus. Am. Polit. Sci. Rev. **60**, 963–974 (1966)
60. H. Theil, *Statistical Decomposition Analysis* (North Holland, Amsterdam, 1972)
61. J. Fellman, Lorenz curve. ed by M. Lovric. *International Encyclopedia of Statistical Science* (Springer, Berlin, 2011), pp. 760–761
62. D. Chotikapanich, *Modeling Income Distributions and Lorenz Curves* (Springer, New York, 2008)
63. E. Scalas, T. Radivoevic, U. Garibaldi, Wealth distribution and Lorenz curve: a finitary approach. J, Econ. Interact. Coordinartion (in press) (2015). doi:10.1007/s11403-014-0136-2
64. G. Warner, A Lorenz curve based index of income stratification. Rev. Black Polit. Econ. **28**, 41–57 (2001)
65. J. Tang, X. Wang, Analysis of land use structure based on Lorenz curves. Environ. Monit. Coord. **151**, 175–180 (2009)
66. O. Alonso-Villar, Measuring concentration: Lorenz curves and their decompositions. Ann. Reg. Sci. **47**, 451–475 (2011)
67. P. Suppes, Lorenz curves for various processes: a pluralistic approach to equity. Soc. Choice Welfare **5**, 89–101 (1988)
68. L. Egghe, Conjugate partitions in infometrics: Lorenz curves, h-type indices, Ferrer graphs and Durfee squares in a discrete and continuous setting. J. Infom. **4**, 320–330 (2010)
69. R. Rouseau, Measuring concentration: sampling design issues, as illustrated by the case of perfectly stratified samples. Scientometrics **28**, 3–14 (1993)
70. L. Egghe, R. Rousseau, Symmetric and asymmetric theory of relative concentration and applications. Scientometrics **52**, 261–290 (2001)
71. L. Egghe, R. Rousseau, How to measure own-group preference? A novel approach to a sociometric problem. Scientometrics **59**, 233–252 (2004)
72. R. Ketzer, K.F. Zimmermann, Publications: German scientific institutions on track. Scientometrics **80**, 231–252 (2009)
73. S. Shibayama, Distribution of academic research grants: a case of Japanese national research grant. Scientometrics **88**, 43–60 (2011)
74. B. Jarneving, Regional research and foreign collaboration. Scientometrics **83**, 295–320 (2010)
75. W. Halffman, L. Leydesdorf, Is inequality among universities increasing? Gini coefficients and the elusive rise of elite universities. Minerva **48**, 55–72 (2010)
76. T.J. Cleophas, A.H. Zwinderman, Pareto charts for identifying the main factors of multifactorial outcomes. ed. by T.J. Cleophas, A.H. Zwinderman. *Machine Learning in Medicine* (Springer, Berlin, 2014), pp. 101–106

77. S.H. Kan, *Metrics and Models in Software Quality Engineering* (Addison-Wesley, Boston, 2002)
78. L. Egghe, R.A. Rousseau, A characterization of distributions which satisfy Price's law and consequences for the laws of Zipf and Mandelbrot. J. Inform. Sci. **12**, 193–197 (1986)
79. S. Lieberson, Rank-sum comparisons between groups, ed. by D. Heise. *Sociological Methodology* (Jossey-Bass, San Francisco, 1976), pp. 276–291
80. S. Lieberson, An asymmetrical approach to segregation, ed. by C. Peach, V. Robinson, S. Smith. *Ethnic Segregation in Cities* (Croom Helm, London, 1981), pp. 61–82
81. N. Toren, V. Kraus, The effects of minority size on women's position in academia. Soc. Forces **65**, 1090–1100 (1987)
82. M. Fosset, J.S. Scott, The measurement of intergroup income inequality: a conceptual review. Social Forces **61**, 855–871 (1983)
83. P.B. Coulter, Measuring the inequity of urban public services. Policy Stud. J. **8**, 683–698 (1980)
84. M.T. Marsh, D.A. Schilling, Equity measurement in facility location analysis: a review and framework. Eur. J. Oper. Res. **74**, 1–17 (1994)
85. K. Barker, The UK research assessment exercise: the evolution of a national research evaluation system. Res. Eval. **16**, 3–12 (2007)
86. G. Falavigna, A. Manello, External funding, efficiency and productivity growth in public research: the case of the Italian National Research Council. Res. Eval. **23**, 33–47 (2014)
87. B.M. Coursey, A.N. Link, Evaluating technology-based public institutions: the case of radiopharmaceutical standards research at the National Institute of Standards and Technology. Res. Eval. **7**, 147–157 (1998)
88. G. Lewison, Evaluation of national biomedical research outputs through journal-based esteem measures. Res. Eval. **5**, 225–235 (1995)
89. C.M. Sa, A. Kretz, K. Sigurdson, Accountability, performance assessment, and evaluation: policy pressures and responses from research councils. Res. Eval. **22**, 105–117 (2013)
90. F. Xu, X.X. Li, W. Meng, W.B. Liu, J. Mingers, Ranking academic impact of world national research institutes014by the Chinese Academy of Sciences. Res. Eval. **22**, 337–350 (2013)
91. L. Georghiou, Research evaluation in European national science and technology systems. Res. Eval. **5**, 3–10 (1995)
92. N. Kastrinos, Y. Katsoulacos, Towards a national system of research evaluation in Greece. Res. Eval. **5**, 63–68 (1995)
93. C.-G. Yi, K.-B. Kang, Developments of the evaluation system of government-supported research institutes in Korean science and technology. Res. Eval. **9**, 158–170 (2000)
94. M. Coccia, A basic model for evaluation R&D performance: theory and application in Italy. R&D Manage. **31**, 453–464 (2001)
95. M. Coccia, Models for measuring the research performance and identifying the productivity of public research institutes. R&D Manage. **34**, 267–280 (2005)
96. M. Coccia, A scientometric model for the assessment of scientific research performance within public institutes. Scientometrics **65**, 307–321 (2005)
97. M. Coccia, Measuring performance of public research units for strategic change. J. Infometrics **2**, 184–194 (2008)
98. P. Vinkler, Correlation between the structure of scientific research, scientometric indicators and GDP in EU and non-EU countries. Scientometrics **74**, 237–254 (2008)
99. E. Albuquerque, Science and technology systems in less developed countries, ed. by H. Moed, W. Glaänzel, U. Schmoch. *Handbook of Quantitative Science and Technology Research* (Kluwer , Dordrecht, 2005), pp. 759–778
100. A. Basu, The Albuquerque model and efficiency indicators in national scientific productivity with respect to manpower and funding of science. Scientometrics **100**, 531–539 (2014)
101. R. Klavans, K. Boyack, Thought leadership: a new indicator for national and institutional comparison. Scientometrics **75**, 239–250 (2008)
102. A.F.J. van Raan, Statistical properties of bibliometric indicators: research group indicator distributions and correlations. J. Am. Soc. Inform. Sci. Technol. **57**, 408–430 (2006)

103. D.A. King, The scientific impact of nations. Nature **430**, 311–316 (2004)
104. A. Schubert, T. Braun, Relative indicators and relational charts for comparative assessment of publication output and citation impact. Scientometrics **9**, 281–291 (1986)
105. J.D. Frame, Mainstream Research in Latin America and the Caribbean. Interciencia **2**, 143–147 (1977)
106. R. Rousseau, L. Yang, Reflections on the activity index and related indicators. J. Infometrics **6**, 413–421 (2012)
107. P. Vinkler, Weighted impact on publications and relative contribution score. Two new indicators characterizing publication activity of countries. Scientometrics **14**, 161–163 (1988)
108. G. Abramo, C.A. D'Angelo, How do you define and measure research productivity? Scientometrics **101**, 1129–1144 (2014)
109. E. Garfield, *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities* (Wiley, New York, 1979)
110. H.F. Moed, Measuring contextual citation impact of scientific journals. J. Informetrics **4**, 265–277 (2010)
111. B. Gonzalez-Pereira, V.P. Guerrero-Bote, F. Moya-Anegon, A new approach to the metric of journals scientific prestige: the SJR indicator. J. Informetrics **4**, 379–391 (2010)
112. V.P. Guerrero-Bote, F. Moya-Anegon, A further step forward in measuring journals' scientific prestige: the SJR2 indicator. J. Informetrics **6**, 674–688 (2012)
113. L. Waltman, N.J. van Eck, T.N. van Leeuwen, M.S. Visser, A.J.F. van Raan, Towards a new crown indicator: an empirical analysis. Scientometrics **87**, 467–481 (2011)
114. L. Waltman, N.J. van Eck, T.N. van Leeuwen, M.S. Visser, A.J.F. van Raan, Towards a new crown indicator: some theoretical considerations. J. Informetrics **5**, 37–47 (2011)
115. R. Abzug, Community elites and power structure, ed. by R.A. Cnaan, C. Milofsky, *Handbook oF Community Movements and Local Organizations* (Springer, New York, 2008), pp. 89–101
116. J.S. Coleman, *Power and Structure of Society* (Norton, New York, 1974)
117. L. Trilling, Technological elites in France and the United States. Minerva **17**, 225–243 (1979)
118. N. Elias, H. Martins, R. Whitley, *Scientific Establishments and Hierarchies* (Reidel, Dordrecht, 1982)
119. M. Mulkay, The mediating role of the scientific elite. Soc. Stud. Sci. **6**, 445–470 (1975)
120. H. Best, U. Becker, *Elites in Transition*. Elite research in Central and Eastern Europe. (VS Verlag für Sozialwissenschaften, 1997)
121. H. Zuckerman, *Scientific Elites*. Nobel laureates in the United States. (Free Press, New York, 1977)
122. J.N. Parker, C. Lortie, S. Allesina, Characterizing a scientific elite: the social characteristics of the most highly cited scientists in environmental science and ecology. Scientometrics **85**, 129–143 (2010)
123. M. Davis, C. Wilson, Elite researchers in ophthalmology: aspects of publishing strategies, collaboration and multi-disciplinarity. Scientometrics **52**, 395–410 (2001)
124. E. Lazega, L. Mounier, M.-T. Jourda, R. Stofer, Organizational vs. personal social capital in scientists' performance: a multi-level network study of elite French cancer researchers (1996–1998). Scientometrics **67**, 27–44 (2006)
125. C. Cao, R.P. Suttmeier, China's new scientific elite: distinguished young scientists, the research environment and hopes for Chinese science. China Quart. **168**, 960–984 (2001)
126. R.S. Hunter, A.J. Oswald, B.G. Charlton, The elite brain drain. Econ. J. **119**, F231–F251 (2009)
127. G. Laudel, Migration currents among scientific elite. Minerva **43**, 377–395 (2005)
128. B. Golub, The Croatian scientific elite and its socio-professional roots. Scientometrics **43**, 207–229 (1998)
129. N.C. Mullins, Invisible colleges as scientific elites. Scientometrics **7**, 357–368 (1985)
130. D. De Solla Price, *Little Science, Big Science* (Columbia University Press, New York, 1963)
131. W. Glänzel, A. Schubert, Price distribution: an exact formulation of Price's 'square root law'. Scientometrics **7**, 211–219 (1985)

132. J.L. Gast, The estimation of the Lorenz curve and Gini index. Rev. Econ. Stat. **54**, 306–316 (1972)
133. N.C. Kakwani, Applications of Lorenz curves in economic analysis. Econometrica: J. Econometric Soc. **43**, 719–727 (1977)
134. A. Dragulescu, V.M. Yakovenko, Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. Phys. A **299**, 213–221 (2001)
135. A. van Heeringen, P.A. Dijkwel, The relationships between age, mobility and scientific productivity. Part II. Effect of age on productivity. Scientometrics **11**, 281–293 (1987)

Part III
Statistical Laws and Selected Models

This part of the book consists of three chapters. In Chap. 4, several famous statistical laws connected to research are discussed. The discussion begins with remarks about frequency and rank approaches to research production. Then the special status of the Zipf distribution in the world of non-Gaussian distributions (which are frequently observed in the process of statistical description of properties of research organizations, publications, and citations) is emphasized. The discussion continues with a description of power laws connected to research production. The following statistical laws are considered: the law of Lotka for scientific publications and the corresponding Pareto and Pareto II distributions; the law of Zipf and its extension (the law of Zipf–Mandelbrot); the law of Bradford for scientific journals. In addition, several important effects and statements from the area of research dynamics are described: the concentration-dispersion effect in science, the Matthew effect in science, the invitation paradox, and the Ortega hypothesis. Finally, several remarks about relationships between the discussed statistical laws are mentioned, and a more general point of view on power laws as informetric distributions is presented.

In Chap. 5, the discussion is focused on selected deterministic and probability models of dynamics of research organizations and dynamics of research publications and their citations. The models discussed are from the three main topics of interest: research publications, citations of research publications, and dynamics of research organizations, connected to the dynamics of publications and citations. In addition, the models are selected in such a way that the reader is supplied with information on important tools used in the area of modeling of research systems: epidemic models, birth and death stochastic processes, Yule distribution, Waring distribution, negative binomial distribution, Poisson distribution, mixed Poisson distribution, Gumbel distribution, Weibull distribution, GIGP distribution, generalized Zipf distribution, etc.

This part ends with a chapter containing several concluding remarks on research dynamics, research productivity, and the importance of mathematics in their understanding and description.

Chapter 4
Frequency and Rank Approaches to Research Production. Classical Statistical Laws

Dedicated to Lotka, Zipf, Pareto, Mandelbrot, Bradford, Price, and all others who contributed to the study of non-Gaussian effects in the research area of scientometrics

Abstract We discuss several classical statistical laws that are important for understanding characteristics of research production and for its assessment. The statistical laws are grouped in such a way that the two much-used statistical approaches for the study of research systems and especially for the study of research publications (frequency approach and rank approach) are appropriately addressed. We begin with some remarks on the frequency and rank approaches to distributions and discuss why the frequency approach is much used in the natural sciences and the rank approach is widely used in the social sciences. Then the stable non-Gaussian distributions are described, and their importance for statistical methodology of research dynamics is emphasized. The laws of Lotka, Pareto, Zipf, Zipf–Mandelbrot, and Bradford are discussed from the point of view of their application to describing different aspects of scientific production. In addition to the discussion of statistical laws, we discuss two important effects: the concentration–dispersion effect (which reflects the separation of the researchers into a small group of highly productive ones and a large group of researchers with limited productivity) and the Matthew effect in science (which reflects the larger attention to the research production of the highly ranked researchers). In addition, we mention the invitation paradox (many papers accepted in highly ranked journals are not cited as much as expected) and the Ortega hypothesis (the big discoveries in science are supported by the everyday hard work of ordinary researchers). At the end of the chapter we discuss more general questions; relationships between the statistical laws and power laws as informetric distributions.

© Springer International Publishing Switzerland 2016

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1_4

4.1 Introductory Remarks

The action of various "soft" laws may be observed in the area of research dynamics. An example of such a law is the principle of cumulative advantage formulated by Price [1]: *Success seems to breed success. A paper which has been cited many times is more likely to be cited again than one which has been little cited. An author of many papers is more likely to publish again than one who has been less prolific. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use.* Our attention is concentrated in this book on research publications as units of scientific information and on citations of research publications as units for impact of the corresponding scientific information. Below, we discuss several statistical power laws connected to research publications and their citations. We emphasize the fact that the discussed power laws should be considered statistical laws ("soft" laws), i.e., more as trends and not as laws that are similar to the "hard" laws of physics. Because of this, one could expect that deviations from the discussed power laws will occur in some real situations. There is a large amount of literature devoted to application of different power laws for modeling features of research dynamics [2–11], and this literature is a part of the literature devoted to the applications of power laws in different areas of science [12–16]. From the point of view of mathematics, the statistical laws connected to research publications and citations are very interesting, since these laws are described mathematically by the same kinds of relationships (hyperbolic relationships),[1] which is evidence of a general structural mechanism of research organizations and scientific systems [17].

The regularities discussed below describe a wide range of phenomena both within and outside of the information sciences. These regularities (called laws and named after the prominent researchers associated with them) were observed in many research fields in the last century. Below, we shall discuss mainly regularities connected to research publications. Let us note that the discussed statistical laws occur in many other areas, such as linguistics, business, etc (Figs. 4.1 and 4.2).

4.2 Publications and Assessment of Research

The pure and simple truth is rarely pure and never simple
Mark Twain

Research production is evaluated often by indicators and indexes connected to research publications [18–21]. There are interesting relationships connected to publications and their authors. These relationships are based on the existence of regularities in the publication activity of the authors of publications. The first relationship

[1] Hyperbolic relationships are relationships of type $m_i i^{\alpha} = \text{const}$. Such relationships are frequently observed in different areas of science such as biology and physics. They exist also in the area of mathematical modeling of structures, processes, and systems in the area of social sciences.

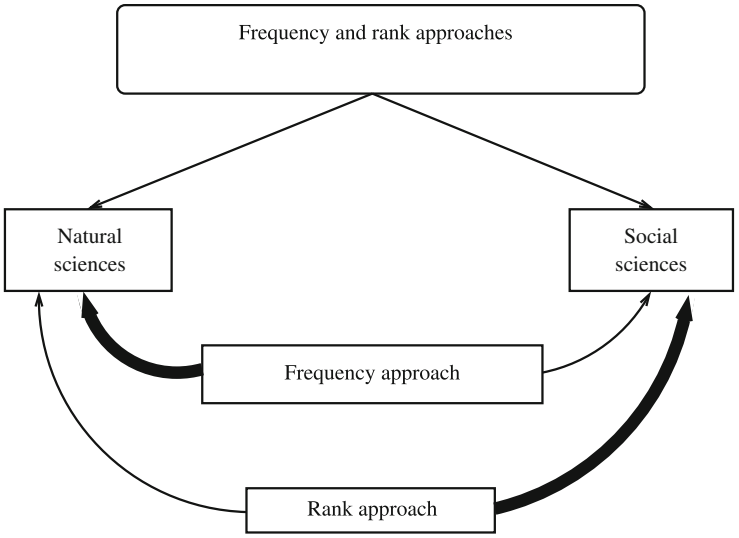

Fig. 4.1 The frequency approach is dominant in the natural sciences. The rank approach is much used in the social sciences

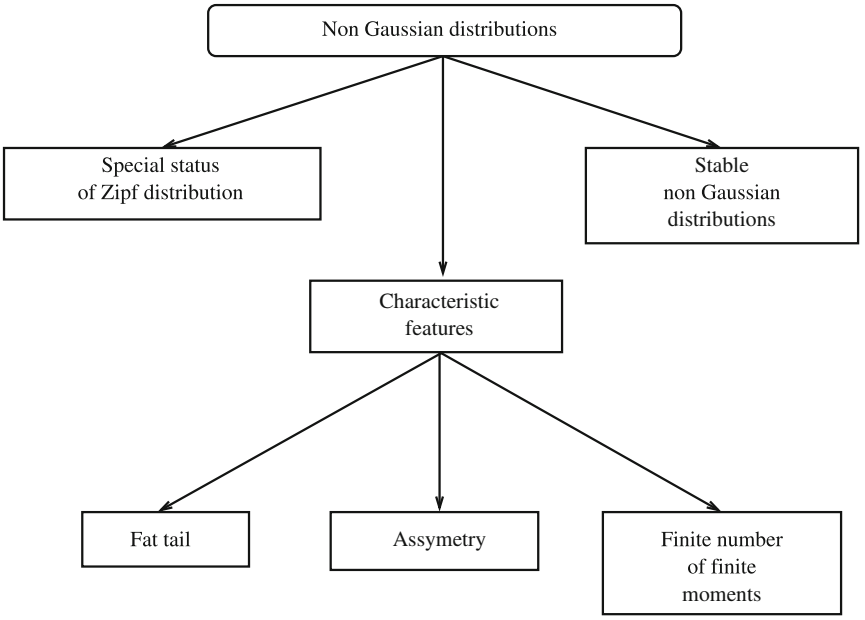

Fig. 4.2 The Zipf distribution has a special status in the world of non-Gaussian distributions (and this status is close to the status of the normal distribution in the world of Gaussian distributions). Non-Gaussian distributions have interesting features that have even more interesting consequences. Stable non-Gaussian distributions arise frequently in different areas of science

was discovered in 1926, when Alfred Lotka (the same Lotka known for the famous Lotka–Volterra equations in population dynamics) published an article [22] on the frequency distribution of scientific productivity determined from an index of *Chemical Abstracts*. The conclusion was that *the number of authors making n contributions is about* $1/n^2$ *of those making one contribution; and the proportion of all contributors who make a single contribution is about* 60 %.

Further discoveries of such kinds of relationships followed. In 1934, Bradford [23] published a study of the frequency distribution of papers over journals. Bradford's conclusion was that *if scientific journals are arranged in order of decreasing productivity on a given subject, they may be divided into a nucleus of journals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus when the numbers of periodicals in the nucleus and the succeeding zones will be as* $1 : b : b^2 : \ldots$. In 1949, Zipf [24] discovered a law in quantitative linguistics (with applications in bibliometrics). This law states that $rf = C$, where r is the rank of a word, f is the frequency of occurrence of the word, and C is a constant that depends on the analyzed text. As we shall see below, this relationship is connected to the relationships obtained by Lotka and Bradford. Zipf also formulated as interesting principle (of least effort) that serves to explain the above relationship: *a person …will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems…* [24]. In 1963, Price [25] formulated the famous square root law: *Half of the scientific papers are contributed by the top square root of the total number of scientific authors.*

Characteristics of research publications such as their number, type, and distribution are the most commonly applied indicators of scientific output, e.g., the production of a research group is measured often by its number of publications, and productivity is expressed often as the number of publications per person–year [26]. Researchers from different fields of science put different weights on different kinds of publications. Researchers from the natural sciences prefer to publish papers in refereed international journals with (possibly larger) impact factors. Researchers from the humanities prefer to publish results in book form rather than as articles. And researchers from the applied sciences publish their results very often as engineering research reports and patents.

Even within each of the above large fields of science, the weights of the different sorts of the dominant kinds of publications vary. Let us concentrate on the natural sciences and on publications in the form of articles. For a long time, articles have been classified as follows:

1. articles published in journals with impact factor (assigned by SCI (Science Citation Index)) [27–35]. The SCI journals are much cited, highly visible journals for which citation data are available;
2. articles published in journals without impact factor (non-SCI journals). Since the visibility of these journals is smaller compared to the visibility of the SCI journals, publication in non-SCI journals is unlikely to produce the same level of citation.

Because of the above facts, most researchers from the natural sciences have preferred to publish in SCI journals, since publication in such a journal is perceived as a mark of quality of the scientific research. Of interest is that this perception doesn't account for the citation levels, and an uncited article may also be considered a consequence of research of good quality.

Two statistical approaches are much used in the study of sets of research publications and citations: the frequency approach and the rank approach. Let us discuss some of their characteristic features.

4.3 Frequency Approach and Rank Approach: General Remarks

The frequency approach is based on analysis of the frequency of observation of values of a random variable. In the case of research publications, the frequency of observation of a value is the probability that a researcher has written x papers, and the random variable is the production of a researcher from the observed large group of researchers. Such an approach will lead us to the laws of Lotka and Pareto.

The rank approach is based on a preliminary ordering (ranking) of the subgroups (having the same value of the studied quantity) with respect to decreasing values of some quantity of interest. Then one can study the subgroups with respect to their rank. In our case, one can rank the researchers from a large group after building subgroups of researchers having the same number of publications. Such an approach will lead us to the laws of Zipf and Zipf–Mandelbrot. And when we rank the sources of information such as scientific journals, the rank approach will lead us to the law of Bradford. Let us stress here that *a general feature of the laws of Lotka, Pareto, Zipf, and Zipf–Mandelbrot is that these laws are expressed mathematically by hyperbolic relationships.*

The frequency approach and rank approach are appropriate for describing different regions of the distribution of research productivity. The rank approach (the law of Zipf, for example) is appropriate for describing the productivity of highly productive researchers, for which two researchers with the same number of papers rarely exist and the ranking can be constructed effectively. The frequency approach (the law of Lotka, for example) is appropriate for describing the productivity of not so highly productive researchers. This group may contain many researchers, and many of them may have the same number of publications. Because of this, they cannot be effectively ranked, but they can be investigated by statistical methods based on frequency of occurrence of different events (such as number of publications or number of citations).

If the maximum production (the number of publications, number of citations, etc.) of a member of a group of researchers is larger than the number of the members of the group, we may usually use the rank approach for characterization of the research production of these researchers. If the maximum production is much smaller than the number of the members of the group, we have to use the frequency approach.

The frequency and rank statistical distributions have differential and integral forms. Let us consider a large enough sample of items of interest for our study. Let the sample size be N. Let the values of the measured characteristics in the sample vary from x_{min} to x_{max}, and we separate this interval of M subintervals of size $\Delta = (x_{max} - x_{min})/(M)$. Then the differential form of the frequency distribution of x, denoted by $n(x)$ (where n is the frequency of values of x in the interval that contains x), satisfies the relationship

$$\sum_{x_{min}}^{x_{max}} n(x) = N. \tag{4.1}$$

The integral form of the frequency distribution is

$$f(x) = \frac{1}{N} \sum_{x_{min}}^{x} n(x^*). \tag{4.2}$$

The differential form of the rank distribution is

$$r = \sum_{x}^{x_{max}} n(x^*), \quad 1 \leq r \leq N, \tag{4.3}$$

and the integral form of the rank distribution is

$$R(r) = \sum_{1}^{r} x. \tag{4.4}$$

Above, the rank means the number of the position of the value x of the studied random variable when all values of the random variable are listed ordered by decreasing frequency $n(x)$.

Let us stress again that in the natural sciences, most of the probability distributions used are frequency distributions. In the social sciences, many of the probability distributions used are rank distributions. But why are frequency distributions dominant in the natural sciences and rank distributions frequently used in the social sciences?

The choice of type of distribution convenient for the statistical description of some sample depends on two factors [36]: the sample size and the value of x_{max}. The

frequency form of the probability distribution is convenient when the normalized frequency $n(x)/N$ is a good approximation of the probability density function. This happens when the frequencies $n(x)$ are large enough and

$$\frac{x_{max} - x_{min}}{\Delta} = M \ll N. \tag{4.5}$$

The corresponding condition for the application of the rank distribution is [36]

$$\frac{x_{min} + x_{max}}{\Delta} \gg 2, \tag{4.6}$$

which means that the rank distributions are more applicable when $\frac{x_{max}}{\Delta}$ is large.

For the case of data from the natural sciences, we usually have large values of N such that the condition (4.5) is satisfied much better that the condition (4.6). In addition, the value of x_{max} is usually not very large. Thus the frequency distributions are dominant. In the social sciences, N is usually not very small, and since the non-Gaussian distributions occur frequently the value of x_{max} is usually large. Thus the condition (4.6) is better satisfied than the condition (4.5), and the rank distributions are used much more than the frequency distributions.

4.4 The Status of the Zipf Distribution in the World of Non-Gaussian Distributions

There is a quotation that if a question is formulated appropriately, that is already half the answer. So let us formulate the question: Why is the status of the Zipf distribution in the world of non-Gaussian distributions almost the same as the status of The Normal distribution is just one distribution from the class of Gaussian distributions?

As we already know, because of the central limit theorem, the normal distribution plays a central role in the world of Gaussian distributions, which are the dominant distributions in the natural sciences. And we know that the non-Gaussian distributions occur frequently in the social sciences. Is there a non-Gaussian distribution that plays almost the same central role for non-Gaussian distributions? There is indeed such a distribution, and its name is the Zipf distribution.

The special status of the Zipf distribution is regulated by the Gnedenko–Doeblin theorem. This theorem [37–40] states that necessary and sufficient conditions (as $x \to \infty$) for convergence of normalized sums of identically distributed independent random variables to stable distributions different from the Gaussian distribution are

$$f(-x) \propto C_1 \frac{h_1(x)}{\mid x \mid^{\alpha^*}}; \quad 1 - f(x) \propto C_2 \frac{h_2(x)}{x^{\alpha^*}};$$
$$C_1 \geq 0; \quad C_2 \geq 0; \quad C_1 + C_2 > 0; \quad 0 < \alpha^* < 2, \tag{4.7}$$

where $f(x)$ is the integral frequency form of the corresponding distribution, C_1, C_2, and α^* are parameters, and h_1 and h_2 are slowly varying functions i.e., for all times $t > 0$,

$$\lim_{x \to \infty} \frac{h_k(tx)}{h_k(x)} = 1, \quad k = 1, 2. \tag{4.8}$$

In other words, the Gnedenko–Doeblin theorem states that the asymptotic forms of the non-Gaussian distributions converge to the Zipf distribution (up to a slowly varying function of x).

Let us stress the following.

1. Note the words "up to a slowly varying function." This means that some statistical distributions connected to research publications and citations may deviate from a power law relationship.
2. Note that in the Gnedenko–Doeblin theorem, $\alpha^* < 2$, and for $\alpha^* < 2$, the Zipf distribution is a non-Gaussian distribution. For $\alpha^* > 2$, the Zipf distribution is a Gaussian distribution.
3. When the sample sizes are infinite, the Gaussian distributions have finite moments, and many of the moments of the non-Gaussian distributions are infinite.
4. In practice, one works with finite samples. Then the moments of the Gaussian distributions and the moments of the non-Gaussian distributions may depend on the sample size.

In addition, we note that the statement of the Gnedenko–Doeblin theorem is about the asymptotic form of a non-Gaussian distribution. This has some consequences for the laws (of Lotka, Bradford, etc.) that we shall discuss below. These laws may be considered statistical relationships that are valid for larger sets. In other words, and in most cases (when the studied sets are not large enough), the laws discussed below should be considered trends and not strict rules. These laws are not like the exact 'hard' laws of the natural sciences. However, these laws are stricter than the 'soft' laws that can be found in many of the social sciences.

4.5 Stable Non-Gaussian Distributions and the Organization of Science

Let us recall some characteristic features of non-Gaussian distributions:

(1) **Their "heavy tail"** [41, 42]: This means, for example, that in a research organization there may exist a larger number of highly productive researchers than the normal distribution would lead one to expect.
(2) **Their asymmetry**: There exist many low-productive researchers and not so many high-productive researchers. We shall discuss below that another manifestation of this asymmetry is the concentration–dispersion effect: there is a concentration of productivity and publications at the right-hand side of the Zipf–Pareto distribution, and dispersion of scientific publications among many low-productive researchers at the left-hand side of the distribution.
(3) **They have only a finite number of finite moments**. For example, for the Zipf–Pareto law (with characteristic exponent α), there exist moments of order $n < \alpha$. And if $\alpha = 1$ (as in the case of many practical applications such as the law of Lotka), then there is no finite dispersion.

The nonexistence of the finite second moment violates an important requirement of the central limit theorem (namely the existence of a finite second moment), and thus some distributions do not converge to the normal distribution. Then there is a class of non-Gaussian distributions that describe another "nonnormal" world. And many social and economic systems belong to this world.

The infinite second (and often the infinite first) moment of non-Gaussian distributions means that the probability of large deviations increases, and if the first moment is infinite, then there is no concentration around some mean value.

An important class of non-Gaussian distributions is the class of stable non-Gaussian distributions. The definition of a stable distribution is [43, 44] this: *Suppose that $S_k = X_1 + \cdots X_k$ denotes the sum of k independent random variables, each with the same nondegenerative distribution P. The distribution P is said to be stable if the distribution of S_k is of the same type for every positive integer k. A random variable is called stable if its distribution has this property.*

The normal distribution is a stable distribution. Another class of stable distributions is the class of non-Gaussian distributions with infinite dispersion. And the asymptotic behavior (at $x \to \infty$) of all of these stable non-Gaussian distributions is $\propto \frac{1}{x^{1+\alpha}}$, i.e., convergence to the Zipf–Pareto law.

The origin of the Zipf–Pareto law as the limit distribution for the class of stable non-Gaussian distributions shows that the Zipf–Pareto law reflects fundamental aspects of the structure and operation of many complex organizations om biology, economics, society, etc.

Three stable distributions are known explicitly:

1. The distribution of Gauss (not of interest for us here).
2. The distribution
$$p(x) = \frac{1}{(2\pi)^{1/2}} x^{-3/2} \exp(-\frac{x}{2}), \tag{4.9}$$
which is connected to a large number of branching processes. At large x, the asymptotic behavior of this distribution is $p(x) \propto \frac{a}{x^{3/2}}$, where $a = (2\pi)^{-1/2}$.
3. The Cauchy distribution [45, 46] (known also as the Lorenz distribution or Breit–Wigner distribution):
$$p(x, x_0, \gamma) = \frac{1}{\gamma\pi \left[1 + \left(\frac{x-x_0}{\gamma},\right)^2\right]} \tag{4.10}$$
where
   - x_0: location parameter that specifies the position of the peak of the distribution;
   - γ: scale parameter that specifies the half-width at the half-maximum.

   Here we shall consider the standard Cauchy distribution $p(x, 0, 1)$, i.e.,
$$p(x) = \frac{1}{\pi} \frac{1}{1 + x^2}, \tag{4.11}$$
whose asymptotic form for large x is $p(x) \propto \frac{a}{x^2}$, where $a = 1/\pi$. We note that for this asymptotic form, we have $\alpha^* = \alpha + 1 = 2$, i.e., $\alpha = 1$. Thus the value of the exponent is the same as the value of the exponent for the law of Lotka for authors and their publications (see the next section). *In other words, the law of Lotka emerges as the asymptotic form of the standard Cauchy distribution.*

4.6 How to Recognize the Gaussian or Non-Gaussian Nature of Distributions and Populations

Usually for non-Gaussian distributions, the moments increase as the the sample size goes up [47]. According to the central limit theorem, the first two moments of Gaussian distributions are finite (which is not the case for the non-Gaussian distri-

butions). Thus the first criterion that a distribution may be Gaussian is *that we are able to express analytically the mean and the variance of the distribution in finite form via distribution parameters*. This is the case of the distributions of Gauss and Poisson, the lognormal distribution, logarithmic distribution, geometric distribution, negative binomial distribution, etc.

The second criterion is connected to the Gnedenko–Doeblin theorem discussed above. The criterion reads thus: *If we are able to determine the asymptotic type of a distribution* $f(x)$ *and these asymptotics* $(x \to \infty)$ *are*

$$f(x) \sim \frac{1}{x^{1+\alpha}}, \tag{4.12}$$

then for $\alpha < 2$, *the distribution is non-Gaussian, and for* $\alpha > 2$, *the distribution is Gaussian*.

The distributions that at large values of x have the form of a Zipf distribution may be called *Zipfian distributions*. If in (4.12) we have $\alpha = \infty$, then the corresponding distribution is non-Zipfian. The above-mentioned Gaussian distributions are all non-Zipfian distributions. From (4.12), one obtains

$$\lim_{x\to\infty} \frac{d}{d(\ln x)} f(x) = -(1+\alpha). \tag{4.13}$$

For the Gaussian non-Zipfian distributions, $\alpha = -\infty$.

Two distributions that will be much discussed in the next chapter are the (generalized) Waring distribution and the GIGP (generalized inverse Gauss–Poisson) distribution. It will be useful to know the values of the corresponding parameters for which these distributions are non-Gaussian and/or Zipfian. The GIGP distribution (called also Sichel distribution) is

$$f(x) = \frac{(1-\theta)^{\nu/2}}{K_\nu[\beta(1-\theta)^{1/2}]} \frac{(\beta\theta/2)x}{x!} K_{x+\nu}[\beta], \tag{4.14}$$

where $K_n[z]$ is the modified Bessel function of the second kind of order n and with argument z. The asymptotics of $f(x)$ when $x \to \infty$ are given by [47]

$$f(x) \sim \frac{\theta^x}{x^{1-\nu}}. \tag{4.15}$$

Then

$$\lim_{x\to\infty} \frac{d}{d(\ln x)} f(x) = -(1-\nu) + x\ln(\theta). \tag{4.16}$$

If $\theta = 1$, then as $x \to \infty$,

$$f(x) \sim \frac{1}{x^{1-\nu}}, \tag{4.17}$$

and ν has to be negative (since $f(x)$ has to yield a normalization). With negative ν and $\alpha = -\nu$, the GIGP distribution is from the class of Zipfian distributions. If $\theta = 1$ and $\nu < 2$, the GIGP distribution is Gaussian. If $\theta = 1$ and $\nu > 2$, the GIGP distribution is non-Gaussian. If $\theta < 1$, the GIGP distribution is a Gaussian non-Zipfian distribution. When $\beta = 0$ and $\nu = 0$, the GIGP distribution is reduced to the logarithmic distribution. Finally, when $\beta = 0$ and $\nu = 0$, the GIGP distribution is reduced to the negative binomial distribution.

The generalized Waring distribution and its particular cases will be much discussed in the next chapter. The generalized Waring distribution can be written in different mathematical forms. The form that expresses the distribution through the gamma and beta functions is

$$f(x) = \frac{\Gamma(a+c)}{B(a,b)\Gamma(c)} \frac{\Gamma(x+c)\Gamma(x+b)}{\Gamma(x+a+b+c)} \frac{1}{x!}. \tag{4.18}$$

The asymptotic behavior of this distribution as $x \to \infty$ is

$$f(x) \sim \frac{1}{x^{1+a}}. \tag{4.19}$$

Thus the generalized Waring distribution is a Zipfian distribution, and $\alpha = a$. if $a < 2$, the distribution is non-Gaussian. If $\alpha > 2$, the distribution is Gaussian.

In practice, one has to work with samples and calculate the moments of the corresponding distributions on the basis of the available samples. Thus the researcher has to observe the growth of the corresponding moments with increasing sample size N. In other words, one has to check the dependence of the mean and variance on N. If the dependence is negligible, then the corresponding population with large probability is a Gaussian one. If a dependence exists, then with large probability, the corresponding population is non-Gaussian.

4.7 Frequency Approach. Law of Lotka for Scientific Publications

The databases of scientific publications are an important final result of the activities of research organizations. And the production of research publications can be highly skewed. This means that in many research fields, a small number of highly productive researchers may be responsible for a significant percentage of all publications in the field.

Alfred Lotka (the same Lotka who is famous for the Lotka–Volterra model in populations dynamics [48–50]) investigated the database of the journal *Chemical*

Abstracts [22] and counted the number of scientists who wrote 1, 2, . . ., i_{max} papers. Lotka obtained the following relationship:

$$N_i = \frac{N_1}{i^2}, \tag{4.20}$$

where

- N_1: number of scientists who wrote one paper;
- N_i: number of scientists who wrote i papers.

Let us note that the law of Lotka doesn't consider the case $N_0 = 0$. This case may be considered on the basis of the Price distribution [51], which will be discussed in the next chapter within the scope of the discussion of the more general Waring distribution.

One may consider two variants of the law of Lotka [52–59] based on (4.20): a variant for the case of infinite productivity of the most productive scientist and a variant for the case of finite scientific productivity of the most productive scientist. Below we shall consider these two variants.

4.7.1 Presence of Extremely Productive Scientists: $i_{\max} \to \infty$

Let N^* be the number of all scientists. Then we can introduce the proportions of the scientists who wrote i papers as $p_i = \frac{N_i}{N^*}$. From (4.20), we have

$$N^* = \sum_{i=1}^{i_{max}} N_i \approx \sum_{i=1}^{\infty} N_i = N_1 \sum_{i=1}^{\infty} \frac{1}{i^2} = N_1 \frac{\pi^2}{6}. \tag{4.21}$$

Then

$$p_i = \frac{N_i}{N^*} = \frac{N_1/i^2}{N_1/\frac{\pi^2}{6}} = \frac{6}{\pi^2}\frac{1}{i^2} \approx \frac{0.608}{i^2} \approx \frac{0.6}{i^2}, \tag{4.22}$$

where $\sum_{i=1}^{\infty} p_i = 1$. Equation (4.22) presents the law of Lotka:

The proportion of scientists who wrote i publications is inversely proportional to i^2 (the square of the number of publications).

Let us stress that in order to investigate whether the law of Lotka is present in some database of scientific publications, we have to be sure that this database is large enough. Two additional remarks are in order here.

Remark 1 If we set $i = 1$ in the law of Lotka, we obtain that the minimally productive researchers (who wrote just one paper) constitute (at least) 60 % of the population of researchers. Then in a research organization, we can expect to find many researchers who have written a small number of papers (for a variety of reasons) and a small number of highly productive researchers.

Remark 2 In the general case, the exponent of the law of Lotka can be different from 2.

Another form of the law is

$$p_i = \frac{p_1}{i^{1+\alpha}}; \quad p_1 = \frac{1}{\zeta(1+\alpha)}, \tag{4.23}$$

where α is the characteristic exponent of the law, and $\zeta(1+\alpha)$ is the Riemann zeta function: $\left(\zeta(\mu) = \sum_{i=1}^{\infty} \frac{1}{i^{\mu}}\right)$. If $\alpha = 1$, then the exponent in the law of Lotka is 2. The form of the law of Lotka (4.23) is similar to the *differential frequency form* of the Zipf distribution:

$$p(x) = \frac{C}{x^{1+\alpha}}, \ 0 \le \alpha < \infty, \tag{4.24}$$

where C and α are parameters of the distribution. The Zipf distribution will be much discussed below. Let us note here that the *integral frequency form* of the Zipf distribution is

$$P(x) = \frac{C}{\alpha N}\left(\frac{1}{x_0^{\alpha}} - \frac{1}{x^{\alpha}}\right), \tag{4.25}$$

where N, x_0, α, and C are parameters of the distribution.

The law of Lotka has been much discussed in connection with data sets for the publication activities of different categories of researchers [60, 61].

4.7.2 i_{max} Finite: The Most Productive Scientist Has Finite Productivity. Scientific Elite According to Price

The productivity of scientists is finite: $i_{max} \neq \infty$. In order to account for this, we have to set corrections to the above formulas. As we shall see, these corrections are small, and because of this, one often uses the formulas derived on the basis of the assumption of infinite productivity of the most productive researcher.

The finite productivity corrections will be based on the relationship [62]

$$\sum_{k=1}^{i_{max}} \frac{1}{k^2} \approx \frac{\pi^2}{6} - \frac{1}{i_{max}}. \tag{4.26}$$

On the basis of this relation, the correction for the relationship (4.21) between the number of all researchers N^* and the number of researchers who have published one paper N_1 becomes

$$N^* = N_1 \left(\frac{\pi^2}{6} - \frac{1}{i_{max}} \right), \tag{4.27}$$

and the finite-size productivity correction of the proportion of researchers who have i publications becomes

$$p_i = \frac{N_i}{N^*} = \frac{6 i_{max}}{i^2 (\pi^2 i_{max} - 6)}. \tag{4.28}$$

Price defined the scientific elite as those researchers who have more than m publications, where m is such a number that the researchers who wrote more than m publications (the elite) possess the half the total number of publications of the group of researchers.

The result for the elite will be obtained on the basis of the following approximate relationship:

$$\sum_{i=1}^{i_{max}} \frac{1}{i} \approx \ln(n) + C_E, \tag{4.29}$$

where $C_E = 0.577\ldots$ is Euler's constant.

The number of publications of the subgroup of researchers in which every researcher has i publications is $P(i) = iN_i$. The entire group of researchers obeys Lotka's law for scientific production. Then

$$P(i) = iN_i = i\frac{N_i}{i^2} = \frac{N_1}{i}. \tag{4.30}$$

Then half the total number of publications of the group of researchers is

$$\frac{1}{2} \sum_{i=1}^{i_{max}} P(i) = \frac{1}{2} \sum_{i=1}^{i_{max}} \frac{N_1}{i} \approx \frac{1}{2} N_1 [\ln(i_{max}) + C_E]. \tag{4.31}$$

The number of researchers who have more than m publications is

$$\sum_{i=m}^{i_{max}} \frac{N_1}{i} = \sum_{i=1}^{i_{max}} \frac{N_1}{i} - \sum_{i=1}^{m} \frac{N_1}{i} \approx N_1 [\ln(i_{max}) - \ln(i)]. \tag{4.32}$$

From (4.31) and (4.32), one obtains

$$m = \exp\left(-\frac{C_E}{2} \right) \sqrt{i_{max}} \approx 0.749 \sqrt{i_{max}}. \tag{4.33}$$

Hence if the group of researchers have publications that obey the law of Lotka, then according to Price, the scientific elite consists of the researchers who have between $0.749\sqrt{i_{max}}$ and i_{max} publications.

What is the size of this elite?
The number of elite scientists is

$$N_e = \sum_{i=m}^{i_{max}} \frac{N_1}{i^2} \approx N_1\left(\frac{1}{m} - \frac{1}{i_{max}}\right). \tag{4.34}$$

The total number of scientists is given by (4.27). Thus the size of the elite of Price is

$$S_e = \frac{N_e}{N^*} = \frac{\pi(i_{max} - m)}{m(6i_{max} - \pi)}. \tag{4.35}$$

For the case of large maximum productivity i_{max},

$$S_e \approx \frac{\pi}{6m} = \frac{\pi}{6 \times 0.749\sqrt{i_{max}}} \approx \frac{0.812}{\sqrt{i_{max}}}. \tag{4.36}$$

Let $i_{max} = 250$. Then the size of the corresponding elite will be approximately 5 % of the size of the group of scientists. The research topic connected to scientific elites enjoys significant current interest, and that interest is very high especially for the study of highly cited researchers and publications [63–67].

4.7.3 The Exponent α as a Measure of Inequality. Concentration–Dispersion Effect. Ortega Hypothesis

According to the law of Lotka, the distribution of scientific production (the number of written publications) in a large enough group of researchers is determined by three parameters:

1. p_1: the percentage of minimally productive researchers;
2. i_{max}: the maximum productivity of a researcher from the group;
3. α: the exponent in the power law of Lotka.

If we fix one of the parameters, we can study the significance of one of the other parameters as a function of the third parameter. We are interested in the parameter α. Thus we fix i_{max} and discuss the relationship between α and p_1. From (4.23), one obtains

$$\frac{\partial p_1}{\partial \alpha} > 0, \tag{4.37}$$

which means that when α increases, the number of not very productive researchers increases too. At the same time, $i_{max} = \text{const}$, i.e., there is at least one highly productive researcher, but the number of highly productive researchers decreases with increasing α.

In other words, α is a measure of the stratification in a group of researchers with respect to the production characteristic called "number of published papers." And as α becomes larger, this stratification increases: there are more and more not very productive researchers and a smaller and smaller number of highly productive researchers.

The above stratification is one example of the concentration–dispersion effect.

Concentration–dispersion effect:
Two processes are simultaneously observed in organizations governed by hyperbolic laws: the concentration of units in a small number of components (formation of an elite) and dispersion of the rest of the units to many components of an organization.

The concentration–dispersion effect applied to our group of researchers means that there exists a small group of researchers that produce large number of publications, and there exists a large group of researchers who have only few publications each. In other words, we have to expect that most of the researchers will be not highly productive and that there will be small number of highly productive researchers. This doesn't mean that the research in the corresponding institution or country is not well organized. The periphery of low-productive researchers is a necessary part of the core–periphery structure, whereby the core contains a small number of highly productive researchers. One cannot try to eliminate the periphery without affecting the core. *The periphery contributes to the high productivity of the core.*

Social stratification [68–70] may arise in a research field as a consequence of the concentration–dispersion effect. A phenomenon similar to the concentration–dispersion effect may be observed also on the level of scientific fields (the few of them that are current attract many citations, and the other fields attract a much smaller number of citations).

A hypothesis called the *Ortega hypothesis* [71–80] is closely connected to the concentration–dispersion effect. Ortega suggests the following:

The work of the average scientists on unambiguous projects is very important for the advance of the science. The work of these scientists leads to minor contributions but without these minor discoveries by the mass of scientists the breakthroughs of the truly inspired scientists will be not possible [81].

4.7.4 The Continuous Limit: From the Law of Lotka to the Distribution of Pareto. Pareto II Distribution

If the number of researchers in the group is very large and the number of papers they have published is very large, too, then one can use a continuous approximation, whereby the number of publications $x(t)$ is a function of t (x is no longer necessarily a natural number).

The continuous version of the law of Lotka is the distribution of (the law of) Pareto.

The distribution of Pareto [82, 83] is

$$p(x) = \frac{\alpha}{x_0}\left(\frac{x_0}{x}\right)^{1+\alpha}, \tag{4.38}$$

where

- $p(x)$: density of distribution of researchers;
- x_0: the minimum number of papers of researchers from the studied large group of researchers ($x_0 \leq x \leq \infty$).
- $\alpha > 0$

The law of Pareto can be obtained on the basis of two assumptions:

1. The time the researchers work on problems in some research area differs among the researchers from the group and is given by the distribution $p(t) = \nu \exp(-\nu t)$; (ν: parameter).
2. The number of publications of the researchers grows proportionally to the number of already written publications (more experience means a shorter time for writing a new publication): $dx/dt = \lambda x \rightarrow x(t) = x_0 \exp(\lambda t)$ (λ is a parameter; x_0 is the number of publications at the initial time t_0).

From the second assumption, $t = \frac{1}{\lambda} \ln\left(\frac{x}{x_0}\right)$. The substitution of this in the relationship for $p(t)$ from the first assumption leads to (4.38) with $\mu = \frac{\lambda}{1-\lambda x_0}$ and $\alpha = \frac{\mu}{\lambda} - 1$.

The Pareto distribution has a shortcoming that can be eliminated by the use of the Pareto II distribution. Let us discuss this in detail.

In the general Pareto distribution (4.38) above, x_0 is a scaling parameter. Let us define the standard Pareto distribution as

$$p_s = \frac{\alpha}{x^{\alpha+1}}. \tag{4.39}$$

Then if the random variable Y has standard Pareto distribution (4.39), the random variable $x_0 Y$ ($x_0 > 0$) has the general Pareto distribution (4.38).

The tail distribution function of the standard Pareto distribution and of the general Pareto distribution is (we assume $x > 1$)

$$P(Y > x) = \int_x^\infty dz \frac{\alpha}{z^{\alpha+1}} = \frac{1}{x^\alpha}. \tag{4.40}$$

Equation (4.40) shows very clearly a drawback of the standard Pareto distribution: *the smallest allowed value of* x *is* 1. In many distributions connected to science dynamics, however, values smaller than 1 are possible (one example is the value 0). In order to solve this problem, one may use the Pareto II distribution, which is obtained as follows [84, 85]: if Y is a random variable that has a standard Pareto distribution (4.39), then the random variable $X = \beta(Y - 1)$ has the Pareto II distribution

$$f_X(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}, \quad x \geq 0. \tag{4.41}$$

The tail distribution of the Pareto II distribution is

$$\Psi_X(x) = \left(\frac{\beta}{x+\beta}\right)^\alpha, \quad x \geq 0. \tag{4.42}$$

As one can see, the Pareto II distribution and its tail distribution are heavy-tailed: for large x, we have $f_X \propto 1/x^{\alpha+1}$; $\Psi_X \propto 1/x^\alpha$.

The Pareto II distribution can be adapted for variables in the interval $(1, \infty)$ by a simple shift $W = X + 1$. If the random variable X has Pareto II distribution, then the random variable W has the shifted Pareto II distribution

$$f_W(x) = \frac{\alpha\beta^\alpha}{(x+\beta-1)^{1+\alpha}} \tag{4.43}$$

and tail function

$$P(W > x) = P(X > x - 1) = \left(\frac{\beta}{x+\beta-1}\right)^\alpha. \tag{4.44}$$

Finally, the moments of the Pareto II distribution exist up to order $n < \alpha$, and the expected value of X^n is

$$E[X^n] = \beta^n n! \frac{\Gamma(\alpha - n)}{\Gamma(\alpha)}, \tag{4.45}$$

where $\Gamma(x)$ is the gamma function.

4.8 Rank Approach

4.8.1 Law of Zipf

The law of Zipf can be obtained from the law of Lotka as follows. The number of researchers who have at least i publications is

$$r_i = \sum_{k=i}^{i_{max}} N_k. \tag{4.46}$$

From the law of Lotka (4.23), we have $N_k = N_1 \frac{1}{k^{1+\alpha}}$. The substitution of the last relationship in (4.46) and letting $i_{max} \to \infty$ leads to

$$r_i = N_i \sum_{k=i}^{\infty} \frac{1}{k^{1+\alpha}} \approx \frac{N_1}{\alpha} \frac{1}{i^{\alpha}}. \tag{4.47}$$

Characteristics of r_i:
The number r_i *is called the rank. According to* (4.46), r_i *is the characteristic of the number (in an ordered list) of researchers that have i publications.*

Let us assume for simplicity that in the studied group we have researchers with different numbers of publications. Then the rank of the sole most productive researcher will be 1. If we take the number of publications of the second most productive researcher, then the number of researchers that have publications greater than or equal to the number of publications of the second most productive researcher will be 2 (and these are the most productive researcher and the second most productive researcher). Thus the rank of the second most productive researcher will be 2. The third most productive researcher will have rank 3, etc.

From (4.47), one obtains

$$i_r = \frac{B}{r^{\beta}}, \tag{4.48}$$

where

$$B = \left(\frac{N_1}{\alpha}\right)^{1/\alpha}; \quad \beta = \frac{1}{\alpha}.$$

Equation (4.48) *with* $\alpha = 1$ *is called the law of Zipf* [24, 86, 87].

4.8.2 Zipf–Mandelbrot Law

The Zipf–Mandelbrot law is obtained when we drop the assumption of infinite productivity of the most productive researcher, $i_{max} = \infty$, and instead of this assume finite productivity i_{max}. Then the result analogous to (4.47) is

$$r_k \approx \frac{N_1}{\alpha}\left(\frac{1}{i^\alpha} - \frac{1}{i_{max}^\alpha}\right). \tag{4.49}$$

From (4.49), one obtains the following rank distribution:

$$i_r = \frac{A}{(r+B)^\gamma}, \tag{4.50}$$

where

$$A = (N_1/\alpha)^{1/\alpha}; \quad B = [N_1/(i_{max}^\alpha \alpha)]; \quad \gamma = 1/\alpha,$$

which is called the Zipf–Mandelbrot law [88, 89].

If we set in (4.50) the value of α from the law of Lotka ($\alpha = 1$) and if we let $i_{max} \to \infty$, we shall obtain the law of Zipf (4.48).[2]

Equation (4.50) gives the differential rank form of the Zipf–Mandelbrot distribution. The integral rank form of the Zipf–Mandelbrot distribution is

$$R(r) = A \ln\left(\frac{r+B}{1+B}\right), \quad \gamma = 1 \tag{4.51}$$

and

$$R(r) = \frac{A}{\gamma - 1}\left[\frac{1}{(1+B)^{\gamma-1}} - \frac{1}{(r+B)^{\gamma-1}}\right], \quad \gamma \neq 1. \tag{4.52}$$

The Zipf–Mandelbrot law is much used not only in scientometrics but also in physics, applied mathematics, etc. [90–94]. Because of this, the practical aspects of fitting and testing this law are of great interest to researchers. A discussion of these aspects is provided in [95].

[2]Note that the Zipf law and the Zipf–Mandelbrot law give the rank of the researchers (or in general of some source of information) only approximately. The reason for this is that for example, there can be several researchers with the same production. If there are no researchers with the same production, then the laws are again approximate, because we silently changed the sums to integrals in order to obtain the approximate relationships for the final form of the laws.

4.8.3 Law of Bradford for Scientific Journals

The following simple classification of the scientific journals (with respect to the articles devoted to some scientific area) can be made:

1. **Core journals**: these are specialized in the corresponding area and contain many articles discussing different research questions from the area.
2. **Intermediate group of journals**: usually, these are journals devoted to closely related scientific areas and containing a certain number of articles on the scientific area being studied.
3. **Periphery journals**: Journals containing articles from other scientific areas and some articles from the studied scientific area.

One possible explanation for the appearance of such groups of journals is as follows. Every researcher tries to publish his/her manuscripts in the best journals. The number of pages of these journals is limited, however. Thus researchers have to publish in other journals as well. These other journals can be close to the research area, but some of them can be quite far from the area of research in the scientific field of interest.

Bradford applied the following approach to the ranked sources of information (journals) [96]. He separated them into groups containing sources of the same production: the journals were separated into a group of journals containing one paper on the studied research topic, then a group of journals containing two papers, etc. In doing so, Bradford obtained empirically the following law.

Law of Bradford:
The journals ranked with respect to the number of articles on a scientific topic can be separated within groups of journals, each group containing the same total number of articles. Then the relationship between the numbers of journals in each group is

$$N_1 : N_2 : N_3 : \ldots = 1 : q : q^2 : \ldots, \tag{4.53}$$

where $q > 1$ is some parameter (can be different for different research topics).

The law of Bradford can be written as follows:

$$R(n) = k \ln\left(\frac{n}{n_0} + 1\right) \to k \ln n \quad \text{for large } \frac{\mathrm{n}}{\mathrm{n_0}}, \tag{4.54}$$

where

- $R(n)$: the total number of papers in the first n journals from the highest ranked groups of journals.
- k: parameter depending on the number of papers in each group of journals and on the number q (for more information, see below).

- n_0: parameter depending on the number of journals in the group of highest ranking and on the number q (for more information, see below).

Equation (4.54) is obtained as follows. The number of journals in the L highest ranked groups of journals is $n = \sum_{i=1}^{L} n_i = n_1 \frac{q^L - 1}{q-1}$ (this comes from the geometric progression in (4.53)). Let n^* be the number of journals in each of the L groups. Then the total number of journals is $R = Ln^*$, and $L = R/n^*$. Substituting this relationship for L into the above relationship for n, we obtain the law of Bradford, where

- $k = \frac{n^*}{\ln q}$,
- $n_0 = \frac{n_0}{q-1}$.

In other words, the number of articles on a topic from a particular research area in the highest ranked n journals (if n is large) increases as the logarithm of the number of such journals.

With some algebra, one can obtain the following form of the law of Bradford:

$$R(n) \approx N_1 \ln n, \tag{4.55}$$

where N_1 is the number of researchers who have written one or more articles on topics from the studied research area. This interesting relationship connects the number of researchers working on the research area of interest, the number of highest ranked journals in this area, and the number of published papers in these journals.

Let us note the following: *the law of Bradford is correct if the value of the exponent α in the law of Lotka is close to* 1. Such values are often present in practical situations, but there can be cases in which α can be significantly different from 1. Thus the Bradford law (as well as the other laws discussed above) has to be applied very carefully [97–99].

Let us stress the following.

The strongly positive feature of the laws discussed in this chapter is that they give us an orientation in the complex world of scientific structures, systems, and processes. For example, the frequent occurrence of $\alpha \approx 1$ is evidence of some kind of structure of the organization of science.

Bradford's law may be used for obtaining information about the degree of inequality in scientific and technology between developed and developing nations and makes it possible to group them into three classes (core, middle, and periphery class) with respect to their science and technology self-reliance [100]. Bradford's law is observed also in the area of expending on research and development of firms and in processes of concentration of research and development [101].

Bradford's law describes the distribution of articles in a single discipline over the various journals. As some of the journals have become more and more interdisciplinary, this has complicated the conditions for the validity of the Bradford law

[102]. Bradford's law can depend on the stage of development of the corresponding scientific field. Thus the Bradford law can change over time [103].

The Bradford distribution can be connected to the Leimkuhler curve and to the index of Gini [104, 105]. In order to show the relationship between the Lorenz curve (much discussed in the previous chapter) and the Leimkuhler curve, we shall consider a population of N journals. For each journal, we consider a number of references (these references can be papers from a research area of interest). We assume that the numbers of references form a random variable X. Let us define:

- $F(j) = P(X \leq j)$;
- $\overline{F}(j) = \frac{r(j)}{N}$, where $r(j)$ is the rank of the journal carrying j references (i.e., the number of journals carrying at least j references).
- $\mu = \frac{M}{N}$, where M is the total number of references carried by the set of N journals being studied.

Now let

$$\Psi(j) = \sum_{i \geq j} \frac{iP(x = i)}{\mu}, \quad i = 1, 2, \ldots, \tag{4.56}$$

and

$$\Phi(j) = \sum_{i \leq j} \frac{iP(x = i)}{\mu}, \quad i = 1, 2, \ldots. \tag{4.57}$$

On the basis of the above definitions, we can define the Lorenz curve and the Leimkuhler curve as follows:

- **Lorenz curve**: the set of points $(F(j), \Phi(j))$;
- **Leimkuhler curve**: the set of points $(\overline{F}(j), \Psi(j))$.

The connection between the two curves becomes clear when one realizes that

$$\overline{F}(j) = 1 - F(j); \quad \Psi(j) = 1 - \Phi(j). \tag{4.58}$$

Hence if one can construct the Lorenz curve, then the construction of the Leimkuhler curve is an easy task.

4.9 Matthew Effect in Science

For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath.
The Gospel of Matthew, Matthew 25:29

Matthew effect: *a term for the phenomenon "the rich get richer and the poor get poorer" or "success breeds success."*

The term "Matthew effect" was introduced by Merton [106, 107] in order to give a name to a mechanism that increases the visibility of contributions to science by eminent researchers and reduces the visibility of comparable contributions by less well known authors. Merton assumed that a contribution would probably enjoy greater visibility when it was made by a scientist of higher eminence.

The Matthew effect helps the rapid diffusion of publications of eminent scientists [108] and especially of their publications that are not of top quality. Such papers written by high-ranking scientists are more likely to be widely diffused early than are papers of the same quality written by low-ranking authors.

The Matthew effect is observed at different scales of science dynamics. For example, there is a Matthew effect for countries: papers with authors who are from some countries get more citations than expected at the cost of others [109–113]. There exists a Matthew effect for journals (papers from more prominent journals are more frequently cited at the expense of papers from other journals) and even a Matthew effect for papers in one journal (the papers of authors from some nations are more cited than the papers by authors from other nations) [114–117]. The Matthew effect exists even with respect to the scientific centers that produce winners of scientific degrees and awards [118] as well as in the peer review process [119].

A measure of the characteristics of scientific systems connected to the Matthew effect is the Matthew index. It is defined as follows:

$$M = \frac{O - E}{E}, \tag{4.59}$$

where

- O: observed number of items (say citations);
- E: expected number of items;

The Matthew index can be made more complicated in order to account for geographic areas [118] (e.g., in order to study the Matthew effect for academicians elected by the Chinese Academy of Sciences):

$$M_{ij} = \frac{O_{ij} - E_j}{E_j}, \tag{4.60}$$

where

- O_{ij}: number of items from region i for year j;
- E_j: expected average number of items per region for year j.

Other characteristics of the Matthew effect can be studied by indexes of concentration discussed in Chap. 4 as well as by means of power law tests [120].

Another effect can be connected to Saint Matthew (second Matthew effect or invitation paradox [121]). This effect is connected to the fact that publishing in journals with a high impact factor does not imply a high number of citations, but offers a chance only: "For many are called, but few are chosen" (Matthew 22:14). Many papers published in journals of relatively high impact factor will be cited less frequently than the average, and relatively few papers obtain a high number of citations. This is not unexpected: if some papers are cited more than the expectation on the basis of the impact factor (which is an averaged quantity), then many papers will be cited less frequently than the expectation on the basis of the impact factor.

4.10 Additional Remarks on the Relationships Among Statistical Laws

In this chapter we have discussed the most famous statistical laws connected to bibliometrics and scientometrics. Bookstein [122] discusses the possibility that in spite of marked differences in their appearance, almost all statistical laws discussed in this chapter are variants of a single distribution. In several more words, Bookstein considers these statistical laws as differing manifestations of a single regularity, which he calls the *informetric law*. The basis for such a point of view is that the regularities (statistical laws) describe a population of discrete entities: researchers, journals, words, businessmen, etc., and each of these entities is producing something over a timelike variable (have some yield): researchers publish articles, articles occur in journals covering some scientific discipline, businessmen earn money, etc. Thus many of the statistical laws describe, in different ways, the same type of data: yields as distributed over a population of items.

Let us consider the above from the point of view of mathematics. Let us first write the classical statistical laws discussed above by means of a unified notation. Thus the law of Bradford may be written as

$$N_n = k^n N_0, \tag{4.61}$$

where k is a constant (equal to the constant q above in the text); N_0 and N_n are connected to the construction of Bradford: he formed a core of journals of central interest to the discipline, and then he formed rings of successively less productive journals, so that each ring contained the same number of relevant articles as the core. The number of journals in a ring divided by the number of journals in the preceding ring was approximately a constant k. Then N_0 is the number of journals in the core, and N_n is the number of journals from the nth ring. The Leimkuhler version of the Bradford law can be written as

$$Y = A \ln(1 + BN), \tag{4.62}$$

where the journals are ranked in decreasing order with respect to the productivity for the studied research discipline, and N is the number of journals required to yield Y articles. Here A and B are appropriate constants. Equation (4.62) (known also as a Leimkuhler distribution) can be written also as

$$N = A^*[\exp(B^*Y) - 1], \tag{4.63}$$

where A^* and B^* are constants.

Lotka's law was for the number f of researchers (chemists in the original version of the law) producing y articles,

$$f = \frac{A}{y^\alpha}, \tag{4.64}$$

where A is an appropriate constant and α is a constant approximately equal to 2. Zipf's law for the frequency y of word occurrence in natural text when the words are ranked according to the number of occurrences in the text is

$$ry = A, \tag{4.65}$$

where r is the rank of the word, y is the frequency (yield) of the word, and A is an appropriate constant. The Zipf–Mandelbrot law is written as

$$y = \frac{A}{(1 + Br)^\alpha}, \tag{4.66}$$

where A and B are appropriate constants.

Let us now briefly discuss the relations between the statistical laws. In the previous section, we have shown the equivalence between the Bradford law (4.61) and the logarithmic (Leimkuhler) form of this law (4.62) (for more mathematical detail about this equivalence, see [122]). Above we have shown that the laws of Zipf and Zipf–Mandelbrot can be obtained from the law of Lotka. Let us now show that there is a relationship between the law of Lotka (4.64) and the Leimkuhler form of the law of Bradford (4.62). Let us denote the expected maximum yield of an item by y_0 (we use the general terminology described at the beginning of this section). Let us rank the items with respect to their yield. Then the cumulative yield Y up to the items of rank r (the items of rank r are assumed to have yield y) is

$$Y = \sum_{n=y}^{y_0} nf_n, \tag{4.67}$$

where f_n is the number of items having a yield of n (e.g., the number of researchers who are authors of n articles). Assuming that the relationship for f_n is given by the law of Lotka (4.64), $f_n = (y_0/n)^\alpha$, we obtain

$$Y = y_0^\alpha \sum n^{1-\alpha}. \tag{4.68}$$

The integral approximation of (4.68), $Y \approx y_0^{\alpha} \int\limits_{y-1/2}^{y_0+1/2} dx\, x^{1-\alpha}$, is as follows:

1. Case $\alpha = 2$:

$$Y \propto y_0^2 \ln\left(\frac{y_0 + 1/2}{y - 1/2}\right). \tag{4.69}$$

2. Case $\alpha \neq 2$:

$$Y \propto \frac{y_0^{\alpha}}{2-\alpha}\left[\frac{1}{(y_0 - 1/2)^{\alpha-2}} - \frac{1}{(y - 1/2)^{\alpha-2}}\right]. \tag{4.70}$$

The rank r of the items of yield y in the presence of the law of Lotka is $\sum\limits_{x=y}^{y_0}(y_0/x)^{\alpha}$, which can be approximated as $\int_{y-1/2}^{y_0+1/2} dx\,(y_0/x)^{\alpha}$. Then (note that $\alpha \neq 1$)

$$r = \frac{y_0}{\alpha - 1}\left[\left(\frac{y_0}{y - 1/2}\right)^{\alpha-1} - \left(\frac{y_0}{y_0 + 1/2}\right)^{\alpha-1}\right]. \tag{4.71}$$

From (4.71), one obtains

$$\frac{1}{y - 1/2} = \left[\frac{(\alpha-1)r}{y_0^{\alpha}} + \left(\frac{1}{y_0 + 1/2}\right)^{\alpha-1}\right]^{1/(\alpha-1)}, \tag{4.72}$$

and the substitution of this in (4.71) leads to

1. $\alpha = 2$:

$$Y \approx A[(B + Cr)^{\alpha} + 1], \tag{4.73}$$

2. $\alpha \neq 2$:

$$Y \approx A \ln(1 + Br), \tag{4.74}$$

where A and B are appropriate constants that can be easily calculated by the interested reader. In a similar way, one can obtain the Zipf law from the Leimkuhler version of the law of Bradford as well as the law of Pareto from the law of Lotka [122].

4.11 On Power Laws as Informetric Distributions

As we have noted at the beginning of the chapter, the laws connected to the processes studied by scientometrics, bibliometrics, and informetrics should be understood not literally, but as statements about probability distributions, or as statements about the corresponding expected values. Let us consider a population of objects and let each object of this population have integer yield y that can be measured. We can associate

another yield to each of the objects: the expected yield x. This expected yield may be not be an integer, and it may not be measurable. Let the number of objects (e.g., researchers) $f(x)$ having an expected yield x (e.g., publications) be proportional to a function $h(x)$ of x [123]:

$$f(x) = Ah(x), \tag{4.75}$$

where A is a constant (which may be set if we assume $h(1) = 1$). The relation between the expected yield x and the actually measured value is as follows. Let $p(n \mid x)$ be the probability that an object (researcher) with expected yield of x units (articles) actually has n units. Then the number of objects with n units will be proportional to

$$g(n) = \int dx\, p(n \mid x)h(x). \tag{4.76}$$

If $p(n \mid x)$ is sharply peaked at n near x, then $g(n) \approx h(n)$. Under the condition of sharp-peaked conditional probability, we have $f(n) \approx Ah(n)$ if the density of expectations is proportional to $h(x)$ (even x is a noninteger). *Thus instead of discrete values n for the units, one may work with continuous values x of the yield variable consistent with the expected value interpretation*. Bookstein [124] gives an example of the usefulness of this approach: if $h(x) = 1/x^2$ and $p(n \mid x)$ is a Poisson distribution, then the expected number of objects yielding n events (units) is $A/[n(n-1)]$, which for large values of n is approximately A/n^2.

After validation of the possibility of working with a continuous variable x instead of with the discrete variable n (and to obtain correct results), let us discuss the question of the form of the distribution $h(x)$ from (4.75) if we change the time interval from an interval in which every object produces the expected value of x units to an interval in which every object produces an expected value of sx units. We impose the following conditions:

1. Our law $h(x)$ has to be stable over such kinds of changes, i.e., the form of our distribution for the case of sx units will again be of the form h: $h(sx)$.
2. *The members of the population of objects produce units at a constant rate*.
3. The population of objects is stable (there are no entries and no exits of objects).

The above conditions lead to statistical laws in the form of power-law distributions. Let us show this.

The population of considered objects (scientists) produced x units (articles) in the first period and $x' = sx$ units in the second period. The number of objects having expected value $x' + \Delta$ in the second period will have expected values between $x'/s + \Delta/s = x + \Delta/s$ in the first period. We know the number of these objects for the first period: on the basis of (4.75), this number is $f(x)(\Delta/s) = Ah(x)(\Delta/x) = Q$. And then the number of objects that have produced Δ units in the second period is $Ah(x)/s$ (i.e., Q/Δ).

The form of the distribution should be stable, i.e., it should remain a constant multiplied by the function h of the expected number of produced units. Then

$$A'h(sx) = A\frac{h(x)}{s}. \tag{4.77}$$

From the condition $h(1) = 1$ and setting $x = 1$, we obtain

$$A'h(s) = \frac{A}{s}. \tag{4.78}$$

From (4.77) and (4.78), we obtain

$$h(sx) = h(s)h(x). \tag{4.79}$$

Equation (4.79) determines the form of the function $h(x)$. Taking into account that $x + \Delta = x(1 + \Delta/x)$, we obtain from (4.79)

$$h(x + \Delta) = h(x)h(1 + \Delta/x), \tag{4.80}$$

and this leads us to the relationship

$$\frac{h(x + \Delta) - h(x)}{\Delta} = \frac{h(x)}{x}\frac{h(1 + \Delta/x) - h(1)}{\Delta/x}. \tag{4.81}$$

Assuming that $(d/dx)[h(1)]$ exists, we obtain by letting $\Delta \to 0$,

$$\frac{dh(x)}{dx} = \frac{h(x)}{x}\frac{dh(1)}{dx}, \tag{4.82}$$

where dh/dx evaluated at $h = 1$ is a constant that we denote by A. Then

$$\frac{dh(x)}{dx} = A\frac{h(x)}{x}, \tag{4.83}$$

which has a power-law function as general solution, i.e.,

$$h(x) = Ax^{-\alpha}, \tag{4.84}$$

where α may be an arbitrary constant (but in practice, $\alpha > 0$, since $h(x)$ is connected to a statistical distribution).

It is remarkable that the relationship (4.84) is present even if one relaxes the requirement for stability of the population of objects, i.e., when objects (researchers) may enter and leave the population [125]. In more detail, the form of $h(x)$, namely $h(x) = Ax^{-\alpha}$, is maintained if the objects enter and leave the population at arbitrary rates, provided that *the distribution in yield production of the items entering and*

leaving the population is the same as that of those initially in the population. In addition, the above form for $h(x)$ is the only form for which this is true. The same form occurs if the rates of production are varying, i.e., if the objects don't generate items at a fixed rate [123]. In more detail, the rate of change may be arbitrary, and the condition is that a change in the rate affects all objects in the same way.

> Thus the occurrence of a power-law relationship may be considered a sign of the inertia of productivity patterns. The stability of the power laws in bibliometrics (e.g., of the law of Lotka) is consistent with the recognition that the studied research discipline will experience slow periods and periods of acceleration. If these variations over time tend to influence all the members of the discipline in the same way, then the corresponding power law will be preserved.

Let us now discuss the relation between the power laws and the multiple authorship (several objects are coauthors of a unit). Lotka derived his law by giving full credit to the senior author and to him alone (i.e., nothing for the other coauthors). It can be shown [123] that if the power law $h(x) \propto 1/x^{\alpha}$ is valid for one accounting system for authorship, it will be valid for any other if certain regularities exist. In addition, this law will be unique in being invariant under changes in counting method for x the expected yield in published articles. *Then if one finds a* $1/x^{\alpha}$ *relation to describe productivity for the case in which a full publication credit is assigned to every author whose name appears on a paper, this will also be the case if we had assigned fractional authorship instead.*

Let us now add some mathematics to the statements above. The number of objects (researchers) that are expected to produce between x and $x + dx$ units (articles) in some time interval are (as above) $Ah(x)dx$, where A is a constant defined by the constraint $h(1) = 1$ and $h(x)$ describes the studied population of objects for *some basis of accounting*. The question is whether the form of $h(x)$ changes if we change the accounting system, e.g., to the accounting system we are currently using. We consider an object (researcher) that has produced N units (articles): n_1 as lone author, n_2 with 1 coauthor, n_3 with two coauthors, etc. Let us in general use an accounting system that assigns credit v_i for the ith unit(paper) of the object (author). Then the total number of units that will be assigned to the object of interest is

$$x = \sum_{i=1}^{N} = rN, \tag{4.85}$$

where r is defined by (4.85). On the basis of this system of accounting, we have $Ah(x)dx$ objects who are expected to yield between x and $x + dx$ units.

Now let us consider a different accounting system. From the point of view of this system, the object yields x' units. This can be written as

$$x' = \frac{r'}{r}x = \theta x; \quad \theta = r'/r, \tag{4.86}$$

where θ depends on the object and on the accounting system. Now we shall obtain an equation for θ starting from the question, How many objects (authors) will yield x' units (articles) in the new accounting system? The number of objects that yield x' units with respect to the new accounting system is equal to the number of objects that yield x'/θ units from the point of view of the old system of accounting. Taking into account that the number of objects that have values of θ between θ_0 and $\theta_0 + d\theta$ is $F(\theta)d\theta$ (where $F(\theta)$ is the probability density function of θ), we obtain that the function $A'h'(x')$ connected to the objects yielding x' units is

$$A'h'(x') = A \int d\theta \; F(\theta) \left[\frac{1}{\theta} h\left(\frac{x'}{\theta} \right) \right], \tag{4.87}$$

where the factor $1/\theta$ compensates for the change of size of dx before and after the transformation.

Now suppose that a change in the accounting system does not change the form of h, i.e.,

$$h'(x) = h(x). \tag{4.88}$$

The substitution of (4.88) in (4.87) leads to

$$A'h(x) = A \int d\theta \; F(\theta) \left[\frac{1}{\theta} h\left(\frac{x}{\theta} \right) \right]. \tag{4.89}$$

Taking into an account that $h(1) = 1$, we obtain

$$A' = A \int d\theta \; F(\theta) \left[\frac{1}{\theta} h\left(\frac{1}{\theta} \right) \right]. \tag{4.90}$$

The substitution of (4.90) in (4.89) leads to

$$h(x) \int d\theta \; F(\theta) \left[\frac{1}{\theta} h\left(\frac{1}{\theta} \right) \right] = \int d\theta \; F(\theta) \left[\frac{1}{\theta} h\left(\frac{x}{\theta} \right) \right]. \tag{4.91}$$

Given $F(\theta)$, (4.91) is an equation for $h(x)$. It is straightforward to check that $h(x) = 1/x^{\alpha}$ is a solution of this equation. Moreover, the power law is the only solution that satisfies all constraints imposed to the problem. *Then, we can conclude that if the above power law holds for one accounting method, it will hold for every other one in which the change in the typical amount of credit given to authors per paper may*

vary from author to author but does not depend strongly on how much the author published. Then if the objects are authors and units are articles [123]:

> the investigator is free to adopt any reasonable system of assigning credit, and can be confident that if power law isn't observed, it is not because he chose the wrong means of attributing articles to authors.

Finally let us note that if we have several classes of objects that yield units and the distribution of those units follows some power law, then if we for some reason do not distinguish between the classes of objects (e.g., between chemists and biologists), then the distribution of units connected to the class of all objects will be approximately a power law [123].

There are many models that lead to relationships connected to different aspects of research production. A large number of such models will be discussed in the next chapter.

References

1. D. de Solla Price, A general theory of bibliometric and other cumulative advantage processes. J. Am. Soc. Inf. Sci. **27**, 292–306 (1976)
2. L. Egghe, The exact place of Zipf's and Pareto's law amongst the classical informetric laws. Scientometrics **20**, 93–106 (1991)
3. M.G. Kendall, Natural law in the social sciences. J. R. Stat. Soc. **124**, 4–16 (1961)
4. B.C. Brookes, The derivation and application of the Bradford-Zipf Distribution. J. Documentation **24**, 247–265 (1968)
5. B.C. Brookes, Bradford's law and the bibliography of science. Nature **224**, 953–956 (1969)
6. R.A. Fairthorne, Empirical hyperbolic distribution (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. J. Documentation **25**, 319–349 (1969)
7. G. Scarrot, Will Zipf join Gauss? New Sci. **62**, 402–404 (1974)
8. J. Vlachy, Frequency distribution of scientific performance. A bibliography of Lotka's law and related phenomena. Scientometrics **1**, 109–131 (1978)
9. S.D. Haitun, Stationary scientometric distributions. Part I. Different approximations. Scientometrics **4**, 5–25 (1982)
10. L. Egghe, Zipfian and Lotkaian continuous concentration theory. J. Am. Soc. Inf. Sci. Technol. **56**, 935–945 (2005)
11. L. Egghe, Consequences of Lotka's law for the law of Bradford. J. Documentation **41**, 173–189 (1985)
12. L. Egghe, The power of power laws and an interpretation of Lotkaian infometric systems as self-similar fractals. J. Am. Soc. Inf. Sci. Technol. **56**, 669–675 (2005)
13. L. Egghe, Consequences of the Lotka's law in the case of fractional counting of authorship and first author counts. Math. Comput. Model. **18**, 63–77 (1993)
14. S.D. Haitun, Stationary scientometric distributions. Part II. Non-gaussian nature of scientific activities. Scientometrics **4**, 89–104 (1982)
15. S.D. Haitun, Stationary scientometric distributions. Part III. The role of the Zipf distribution. Scientometrics **4**, 181–194 (1982)
16. L. Egghe, R. Rousseau, Theory and practice of the shifted Lotka function. Scientometrics **91**, 295–301 (2012)

17. A.I. Yablonsky, On fundamental regularities of the distribution of scientific productivity. Scientometrics **2**, 3–34 (1980)
18. K. Prpic, The publication activity of young scientists: an empirical study. Scientometrics **49**, 453–490 (2000)
19. K. Prpic, Gender and productivity differentials in science. Scientometrics **55**, 27–58 (2002)
20. R.K. Toutkoushian, S.R. Porter, C. Danielson, P.R. Hollis, Using publications counts to measure an institutions;s research productivity. Res. High. Educ. **44**, 121–148 (2003)
21. S. Kyvik, Productivity differences, fields of learning, and Lotka's law. Scientometric **15**, 205–214 (1989)
22. A.J. Lotka, The frequency distribution of scientific productivity. J. Washington Acad. Sci. **16**, 317–323 (1926)
23. S.C. Bradford, Sources of information on specific subjects. Engineering **137**, 85–86 (1934)
24. G.K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Weasley, Cambridge, MA, 1949)
25. D. de Solla Price, Little Science, Big Science (Columbia University Press, New York, 1963)
26. J.M. Russel, R. Rousseau, Bibliometrics and institutional evaluation, in Encyclopedia of Life Support Systems (EOLSS), ed. by R. Arvantis. Part 19.3: Science and Technology Policy (EOLSS Publishers, Oxford, UK, 2002), pp. 1–20
27. E. Garfield, *Citation indexing—its theory and application in science, technology, and humanities* (Wiley, New York, 1979)
28. E. Garfield, Journal inpact factor: a brief review. Can. Med. Assoc. J. **161**, 979–980 (1999)
29. B.K. Sen, Normalised impact factor. J. Documentation **48**, 318–325 (1992)
30. E. Garfield, The impact factor and using it correctly. Unfallchirurg **101**, 413–414 (1998)
31. S.J. Bensman, Garfield and the impact factor. Ann. Rev. Inf. Sci. Technol. **41**, 93–155 (2007)
32. A. Smith, M. Thelwall, Web impact factors for Australasian universities. Scientometrics **54**, 363–380 (2002)
33. M. Thelwall, Journal impact evaluation: a webometric perspective. Scientometrics **92**, 429–441 (2012)
34. P. Ingwersen, The calculation of web impact factors. J. Documentation **54**, 236–243 (1998)
35. L. Leydesdorff, T. Opthof, Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. J. Am. Soc. Inf. Sci. Technol. **61**, 2365 – 2369 (2010)
36. S.D. Haitun, *Scientometrics: State and Perspectives* (Nauka, Moscow, 1983). (in Russian)
37. B.V. Gnedenko, On the theory of limit theorems for sums of independent random variables. Bull. Acad. Sci. URSS. Ser. Math. [Izvestia Akad. Nauk SSSR] **1939**, 643–647 (1939). (in Russian)
38. W. Doeblin, Sur l'ensemble de puissonces d'une loi de probabilitè. Studia Math. **9**, 71–96 (1940)
39. S.D. Haitun, *Quantitative Analysis of Social Phenomena* (URSS, Moscow, 2005). (in Russian)
40. N.H. Bingham, Regular variation and probability: the early years. J. Comput. Appl. Math. **200**, 357–363 (2007)
41. B.B. Mandelbrot, *Fractals and Scaling in Finance* (Springer, New York, 1997)
42. M. Falk, J. Hüsler, R. Reiss, *Laws of Small Numbers: Extremes and Rare Events* (Birkhäuser, Basel, 2011)
43. J.W. Lamperti, *Probability. A Survey of the Mathematical Theory* (Wiley, New York, 1996)
44. A.I. Yablonsky, Stable non-Gaussian distributions in scientometrics. Scientometrics **7**, 459–470 (1985)
45. N.L. Johnson, S. Kotz, N. Balakrishnan, *Continuous Univariate Distributions*, vol. I (Wiley, New York, 1994)
46. A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1984)
47. S.D. Haitun, Criteria of Gaussian/non-Gaussian nature of distributions and populations, in Informetrics 89/90, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1990), pp. 149–161
48. A. Lotka, *Elements of Physical Biology* (Williams and Wilkins, Baltimore, 1925)

49. A.J. Lotka, *Analytical Theory of Biological Populations* (Plenum Press, New York, 1998)
50. H. Voos, Lotka nad information science. J. Am. Soc. Inf. Sci. **25**, 270–272 (1974)
51. W. Glänzel, A. Schubert, Price distribution. An exact formukation of Price's "square root law". Scientometrics **7**, 211–219 (1985)
52. L. Egghe, Relations between the continuous and the discrete Lotka power function. J. Am. Soc. Inf. Sci. Technol. **56**, 664–668 (2005)
53. R.C. Coile, Lotka's frequency distribution of scientific productivity. J. Am. Soc. Inf. Sci. **28**, 366–370 (1977)
54. L.J. Murphy, Lotka's law in the humanities? J. Am. Soc. Inf. Sci. **24**, 461–462 (1973)
55. T. Radhakrishnan, R. Kennzan, Lotka's law and computer science literature. J. Am. Soc. Inf. Sci. **30**, 51–54 (1979)
56. K.H. Chung, R.A.K. Cox, Patterns of productivity in the finance literature: a study of the bibliometric distributions. J. Finance **45**, 301–309 (1990)
57. N. Kumar, Applicability to Lotka's law to research productivity of council of scientific and industrial research (CSIR), India. Ann. Libr. Inf. Sci. **57**, 7–11 (2010)
58. M.I.M. Sobrino, A.I.P. Caldes, A.P. Guerrero, Lotka law applied to the scientific production of information science area. Braz. J. Inf. Sci. **2**, 16–30 (2008)
59. B.K. Sen, Lotka's Law: a viewpoint. Ann. Libr. Inf. Stud. **57**, 166–167 (2010)
60. M.L. Pao, Lotka's law: a testing procedure. Inf. Process. Manage. **21**, 305–320 (1985)
61. M.T. Kinnucan, D. Wolfram, Direct comparison of bibliometric models. Inf. Process. Manage. **26**, 777–790 (1990)
62. P.D. Allison, D. de Sola, Price, B.G. Griffith, J.M. Moravcsik, J.A. Stewart, Lotka's law: a problem in its interpretation and application. Soc. Stud. Sci. **6**, 269–276 (1976)
63. V. Lariviere, B. Macaluso, E. Archambault, Y. Gingras, Which scientific elites? On the concentration of research funds, publications and citations. Res. Eval. **19**, 45–53 (2010)
64. A.H. Goodall, Highly cited leaders and the performance of research universities. Res. Policy **38**, 1079–1092 (2009)
65. D.W. Asknes, G. Sivertsen, The effect of highly cited papers on national citation indicators. Scientometrics **59**, 213–224 (2004)
66. I. Podllubny, K. Kassayova, Towards a better list of citation superstars: compiling a multidisciplinary list of highly cited researchers. Res. Eval. **15**, 154–162 (2006)
67. C. Cao, *China's Scientific Elite* (RouthledgeCurzon, London, 2004)
68. J.R. Cole, S. Cole, *Social Stratification in Science* (The University of Chicago Press, Chicago, 1973)
69. M. Albrow, E. King (eds.), *Globalization, Knowledge and Society* (Sage, London, 1990)
70. X. Zhou, A nonparametric index of stratification. Sociol. Methodol. **42**, 365–389 (2012)
71. J. Ortega y Gasset, *The Revolt of the Masses* (Norton, New York, 1932)
72. M.H. Macroberts, B.R. Macroberts, Testing the Ortega hypothesis: facts and artifacts. Scientometrics **12**, 293–295 (1987)
73. A. Meadows, Ortega hypothesis. Scientometrics **12**, 315–316 (1987)
74. A. Nederhof, A. van Raan, Citation theory and the Ortega hypothesis. Scientometrics **12**, 325–328 (1987)
75. M. Oromaner, The Ortega hypothesis and influential articles in American sociology. Scientometrics **7**, 3–10 (1985)
76. S. Cole, J. Cole, Testing the Ortega hypothesis: milestone or millstone? Scientometrics **12**, 345–353 (1987)
77. H. Kretschmer, Measurement of social stratification. A contribution to the dispute on the Ortega hypothesis. Scientometrics **26**, 97–113 (1993)
78. H. Kretschmer, R. Müller, A contribution to the dispute on the Ortega hypothesis: connection between publication rate and stratification of scientists, tested by various methods. Scientometrics **18**, 43–56 (1990)
79. W. Snizek, A re-examination of the Ortega hypothesis: the Dutch case. Scientometrics **9**, 3–11 (1986)

80. L. Bornmann, F. de Moya Anegon, L. Leydesdorff, Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. PLOS One **5**, e.11344 (2010)
81. J.R. Cole, S. Cole, The Ortega hypothesis. Science **178**(4059), 368–375 (1972)
82. B.C. Arnold, Pareto and generalized Pareto distributions, in *Modeling Income Distributions and Lorenz Curves*, ed. by D. Chotikapanich (Springer, New York, 2008), pp. 119–145
83. M.E.J. Newman, Power laws, Pareto dostributions and Zipf's law. Contemp. Phys. **46**, 323–351 (2005)
84. Q.L. Burrell, Extending Lotkaian infometrics. Inf. Process. Manage. **44**, 1794–1807 (2008)
85. B.C. Arnold, *Pareto Distributions* (International Co-operative Publishing House, The University of Michigan, MI, 1983)
86. X. Gabaix, Zipf's law for cities: an explanation. Q. J. Econ. **114**, 739–767 (1999)
87. W. Li, Information theory. IEEE Trans. Inf. Theory **38**, 1842–1845 (1992)
88. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1982)
89. Z.K. Silagadze, Citations and the Zipf-Mandelbrot law. Complex Syst. **11**, 487–499 (1997)
90. M.A. Montemurro, Beyond the Zipf-Mandelbrot law in quantitative linguistic. Phys. A **300**, 567–578 (2001)
91. V.P. Maslov, The Zipf-Mandelbrot law: quantization and an application to the stock market. Russ. J. Math. Phys. **12**, 483–488 (2005)
92. S. Abe, N. Suzuki, Scale-free statistics of time interval between successive earthquakes. Phys. A **350**, 588–596 (2005)
93. P. Louridas, D. Spinellis, V. Vlachos, Power laws in software. ACM Trans. Softw. Eng. Methodol. **18**(1)(2) (2008)
94. B. Manaris, D. Vaughan, C. Wagner, J. Romero, R.B. Davis, Evolutionary music and the Zipf-Mandelbrot law: Developing fitness functions for pleasant music. Lect. Notes Comput. Sci. **2611**, 522–534 (2003)
95. J. Izsak, Some practical aspects of fitting and testing the Zipf-Mandelbrot model. A short essay. Scientometrics **67**, 107–120 (2006)
96. S.C. Bradford, Sources of information on specific subjects. J. Inf. Sci. **10**, 176–180 (1985)
97. S. von Ungern-Sternberg, Bradford's law in the context of information provision. Scientometrics **49**, 161–186 (2000)
98. K.C. Garg, P. Sharma, L. Sharma, Bradford's law in relation to the evolution of a field. A case study of solar power research. Scientometrics **27**, 145–156 (1993)
99. W. Chongde, W. Zhe, Evaluation of the models for Bradford's law. Scientometrics **42**, 89–95 (1998)
100. H. Eto, P.M. Candelaria, Applicability of the Bradford distribution to international science and technology indicators. Scientometrics **11**, 27–42 (1987)
101. H. Eto, Bradford law in R&D expending of firms and R&D concentration. Scientometrics **6**, 183–188 (1984)
102. A. Bookstein, Towards multi-disciplinary Bradford law. Scientometrics **30**, 353–361 (1994)
103. R. Wagner-Döbler, Time dependencies of Bradford distributions: structures of journal output in 20th-century logic and 19-th century mathematics. Scientometrics **39**, 231–252 (1997)
104. Q.L. Burrell, The Bradford distribution and the Gini index. Scientometrics **21**, 181–194 (1991)
105. F.F. Leimkuhler, The Bradford distribution. J. Documentation **23**, 197–207 (1967)
106. R.C. Merton, The Matthew effect in science. Science **159**(3810), 56–63 (1968)
107. R.K. Merton, The Matthew effect in science II: cumulative advantage and symbolism of intelectual property. ISIS: J. Hist. Sci. Soc. **79**, 606–623 (1988)
108. J.R. Cole, S. Cole, *Social Stratification in Science* (The University of Chicago Press, Chicago, 1973)
109. M. Bonitz, E. Bruckner, A. Scharnhorst, Characteristics and impact of the Matthew effect for countries. Scientometrics **40**, 407–422 (1997)
110. M. Bonitz, Ten years Matthew effect for countries. Scientometrics **64**, 375–379 (2005)
111. M. Bonitz, E. Bruckner, A. Scharnhorst, The micro-structure of the Matthew effect for countries, in *Proceedings of the Seventh Conference of the International Society for Scientometrics and Infometrics* (Universidad de Colima, Colima, Mexico, 1999), pp. 50–64

112. V. Lariviere, Y. Gingras, Brief communication: the impact factor's Matthew effect: a natural experiment in bibliometrics. J. Am. Soc. Inf. Sci. Technol. **61**, 424–427 (2010)
113. J. Wang, Unpacking the Matthew effect in citations. J. Infometrics **8**, 329–339 (2014)
114. M.H. Biglu, The influence of references per paper in the SCI to Impact Factors and the Matthew effect. Scientometrics **74**, 453–470 (2008)
115. M. Bonitz, E. Bruckner, A. Scharnhorst, The Matthew index—concentration patterns and Matthew core journals. Scientometrics **44**, 361–378 (1999)
116. M. Bonitz, A. Scharnhorst, Competition in science and the Matthew core journals. Scientometrics **51**, 37–54 (2001)
117. V. Pislyakov, E. Dyachenko, Citation expectations: are they realized? Study of the Matthew effect for Russian papers published abroad. Scientometrics **83**, 739–749 (2010)
118. X. Yang, X. Gu, Y. Wang, G. Hu, L. Tang, The Matthew effect in China's science: evidence from academicians of Chinese Academy of Sciences. Scientometrics **102**, 2089–2105 (2015)
119. F. Squazzoni, C. Candelli, Saint Matthew strikes again: an agent based model of the scientific community structure. J. Infometrics **6**, 265–275 (2012)
120. B. Birkmaier, K. Wohlrabe, The Matthew effects in economics reconsidered. J. Infometrics **8**, 880–889 (2014)
121. P. Vinkler, Introducing the contemporary contribution index for characterizing the recent relevant impact of journals, in *Proceedings of the International Conference of the International Society for Scientometrics and Informetrics*, ed. by D. Torres-Salinas, H.F. Moed (CINDOC-CSIC, Madrid, 2007), pp. 753–760
122. A. Bookstein, Informetric distributions, Part I: unified overview. J. Am. Soc. Inf. Sci. **41**, 368–375 (1990)
123. A. Bookstein, Informetric distributions, Part II: resilience to ambiguity. J. Am. Soc. Inf. Sci. **41**, 376–386 (1990)
124. A. Bookstein, The bibliometric distributions. Libr. Q. **46**, 416–423 (1976)
125. A. Bookstein, Patterns of scientific productivity and social change: a discussion of Lotka's law and bibliometric symmetry. J. Am. Soc. Inf. Sci. **28**, 206–210 (1977)

Chapter 5
Selected Models for Dynamics of Research Organizations and Research Production

Dedicated to the memory of Anatoly Yablonsky. His studies on mathematical models of science contributed much to my interest in mathematical modeling of social and economic systems.

Abstract The understanding of dynamics of research organizations and research production is very important for their successful management. In the text below, selected deterministic and probability models of research dynamics are discussed. The idea of the selection is to cover mainly the areas of publications dynamics, citations dynamics, and aging of scientific information. From the class of deterministic models we discuss models connected to research publications (SI-model, Goffmann–Newill model, model of Price for growth of knowledge), deterministic model connected to dynamics of citations (nucleation model of growth dynamics of citations), deterministic models connected to research dynamics (logistic curve models, model of competition between systems of ideas, reproduction–transport equation model of evolution of scientific subfields), and a model of science as a component of the economic growth of a country. From the class of probability models we discuss a probability model connected to research publications (based on the Yule process), probability models connected to dynamics of citations (Poisson and mixed Poisson models, models of aging of scientific information (death stochastic process model and birth stochastic process model connected to Waring distribution)). The truncated Waring distribution and the multivariate Waring distribution are described, and a variational approach to scientific production is discussed. Several probability models of production/citation process (Paretian and Poisson distribution models of the h-index) as well as GIGP model distribution of bibliometric data are presented. A stochastic model of scientific productivity based on a master equation is described, and a probability model for the importance of the human factor in science is discussed. The chapter ends by providing information about some models and distributions connected to informetrics: limited dependent variable models for data analysis and the generalized Zipf distribution and its connection to the Waring distribution and Yule distribution.

© Springer International Publishing Switzerland 2016

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1_5

5.1 Introductory Remarks

The interest in models of research dynamics and research production has increased greatly since the publication of the book *Little Science, Big Science* [1] by Derek de Solla Price in 1963, in which the first systematic approach to the structure of modern science was presented. One began to construct models for the growth of the scientific literature, and this growth was assumed to be exponential (for all of science) but could be also logistic or even linear for some scientific disciplines. In addition, models of aging and obsolescence of scientific information appeared [2–4]. At approximately the same time as Price, Goffman and Newill [5] developed an intellectual epidemics model of scientific communication. From the point of view of this model, the diffusion of ideas in a population of scientists could be compared to the spreading of a virus in some population, causing an epidemic. The model of Goffman and Newill was followed by other models that connected science dynamics to dynamics of populations. Several such models will be discussed below.

The number of models in the area of research dynamics grows continuously. There are many mathematical models connected to the dynamics of research organizations that may supply useful information for support of assessment of research production. The focus of this book is mainly on science dynamics and on results obtained by research on publications and citations. This focus limits the set of models for discussion and determines the selection of the models presented below. In principle, two kinds of models may be developed: deterministic models and probability models. The discussion below begins with models for dynamics of research publications. First of all, several forms of growth function are described. Then two deterministic models of a kind epidemic (SI model and the Goffman–Newill model) are presented. As an example of a deterministic nonepidemiological model, the Price model of knowledge growth is discussed. The nucleation model of Sangwal for citations dynamics follows, and this is the only deterministic model connected to citation dynamics. The reason for this limited coverage is as follows. A citation may be considered a unit of importance of scientific information. But this unit is small, and in addition, citations may arise more frequently than the larger units of scientific information (research publications). Finally, citations may arise quite irregularly. Thus more attention to citation dynamics is given from the point of view of probability models. The presentation of deterministic models continues with a model of competition of ideas, which is important for the evolution of research structures and systems. Further, the reproduction transport equation model of dynamics of scientific fields is discussed. The part devoted to deterministic models ends with a model of science as a component of the economic growth of a country.

The greater part of the chapter is devoted to probability models. This part begins with several general remarks on Poisson processes and their connection to the distributions of Yule and Waring and to the GIGP distribution. Then a probability model of research publications based on the Yule stochastic process is described. After that, attention is focused on models connected to citations of research publications. These models are for citation dynamics of a set of simultaneously appearing research pub-

lications and citation behavior of sets containing subsets of publications published at the same time. The discussion is based on the Poisson distribution and on the mixed Poisson distribution, which will be related to the Yule distribution. Models for aging of scientific information follow (the aging of information is an important topic connected to the dynamics of citations of research publications). Two probability models of the aging of scientific information are considered: a model based on a death stochastic process and a model based on a nonstationary birth process. The last model leads to the Waring distribution and to the negative binomial distribution. The Waring distribution is discussed in greater detail: the truncated Waring distribution and multivariate Waring distribution are described. On the basis of the truncated Waring distribution, a model of brain drain in the case of massive migration through migration channels is mentioned. A description of a variational approach to research production and two models of a production–citation process follows. The GIGP model distribution for bibliometric data is discussed. A master equation model of scientific productivity follows. The chapter ends with a probability model for the importance of the human factor in science.

5.2 Deterministic Models Connected to Research Publications

5.2.1 Simple Models. Logistic Curve and Other Models of Growth

One may consider simple exponential or logistic models of the growth of a number of items. For the case of the exponential model, the assumption is that the growth is proportional to the number of existing items,

$$\frac{dN}{dt} = kN, \tag{5.1}$$

where k is a parameter. The solution of (5.1) is $N(t) = N_0 \exp(kt)$, where N_0 is the number of available items at $t = 0$. It is of interest to know in many cases when the initial number of items N_0 will double. This time is $t^* = \ln(2)/k$ for the case of the exponential model. The exponential model, e.g., may be considered an approximation of the initial increase in the number of research publications in a newly established research field (more details follow below).

If we consider a longer time interval, then the initial exponential increase of the number of items may cease. In this case, one may consider another model, the logistic model of growth:

$$\frac{dN}{dt} = kN(a - N), \tag{5.2}$$

where k and a are (positive) parameters. The solution of the logistic equation (5.2) is

$$N = \frac{a}{1 + \left(\frac{a}{N_0} - 1\right) \exp(-kat)}. \tag{5.3}$$

This solutions has regions of almost exponential growth (when $N \ll a$, a region of almost linear growth around $N = a/2$, and a region of saturation (almost negative exponential growth) around $N \approx a$.

Logistic curves are frequently applied for modeling a variety of processes, e.g., the growth of scientific publications [6–10]. In order to describe trajectories of growth or decline in socio-technical systems, one generally uses the following three-parameter logistic curve [11]:

$$x(t) = \frac{K}{1 + \exp[-\alpha t - \beta]}, \tag{5.4}$$

where the quantities are as follows:

- $x(t)$: number of units in the species or growing variable to study,
- K: the asymptotic limit of growth,
- α: growth rate, which specifies the "width" of the curve for $x(t)$,
- β: specifies the time t_m when the curve reaches the midpoint of the growth trajectory such that $x(t_m) = 0.5\ K$.

The parameters K, α, and β are usually obtained after fitting the available data. It is well known that many cases of epidemic growth can be described by parts of an appropriate logistic curve. But not every interaction scheme leads to logistic growth [12]. The evolution of systems in such regimes may be described by more complex curves such as a combination of two or more simple three-parameter functions [11, 13].

Let us consider in more detail the logistic growth of knowledge and aging of scientific information. The appearance of the logistic curve in this case is a consequence of two processes: an increase in the amount of scientific information and the aging of scientific information. If only increasing of scientific information exists, then the increase may be proportional to the amount of the available information,

$$\frac{dx}{dt} = \alpha x \rightarrow x = x_0 \exp(\alpha t), \tag{5.5}$$

where α is a coefficient (the assumption is that each element produces a new element with a constant intensity α). This leads to exponential growth of scientific information. Such a situation can be observed for new areas of research in which the information is relatively new (and not aged). For more mature research areas, the coefficient α depends on the amount of information $x : \alpha = f(x)$ and decreases with

the aging of the scientific information. A simple assumption is that the decrease in α is proportional to x. Then

$$\frac{dx}{dt} = (a - bx)x. \tag{5.6}$$

Equation (5.6) is the logistic equation. Its solution is

$$x(t) = \frac{a}{b[1 + \sigma \exp(-at)]}, \tag{5.7}$$

where σ is a coefficient that can be determined from the initial conditions. From (5.7), it follows that the speed of the increase of scientific information is

$$\text{Eff} = \frac{dx}{dt} = \frac{\sigma a^2}{b} \frac{\exp(-at)}{\{1 + \exp[\sigma \exp(-at)]\}}. \tag{5.8}$$

The quantity Eff can be considered a measure of the effectiveness of the scientific field. This effectiveness (i) increases when the scientific field is new; (ii) passes through a maximum at $t = \ln(\sigma/a)$ (the maximum "expectation" of the scientific field; (iii) tends to 0 as $t \to \infty$ (the scientific field is exhausted).

In general, the growth can be described by the relationship

$$\frac{dx}{dt} = \alpha(x)x. \tag{5.9}$$

If we are interested in the growth around some value $x = x_0$, then we can represent $\alpha(x)$ by a Taylor series,

$$\alpha(x) = \alpha(x_0) + \frac{1}{1!}\frac{d\alpha}{dx}\Big|_{x=x_0}(x - x_0) + \frac{1}{2!}\frac{d^2\alpha}{dx^2}\Big|_{x=x_0}(x - x_0)^2 + \frac{1}{3!}\frac{d^3\alpha}{dx^3}\Big|_{x=x_0}(x - x_0)^3 \ldots. \tag{5.10}$$

If we use only the first term from (5.10), then the local growth around $x = x_0$ is exponential. If we have to use the first two terms in (5.10), then the local growth can be logistic. If we have to use the first three or more terms from (5.10), then the local growth is more complicated.

Logistic growth is not the only possible growth connected to the evolution of scientific information. The study of Menard [14] revealed three types of research fields with respect to the type of growth of the total number of publications in a given research field: stable fields (linear or exponential growth at small rates); exponentially growing fields (rapidly growing fields); cyclic fields: cyclic change of periods of stable and fast growth [15, 16]. Let us note the mathematical relationships for several kinds of growth functions that may be of interest to readers who encounter growth phenomena in their research:

1. *Gompertz growth function* [10]

$$x(t) = DA^{B^t}, \tag{5.11}$$

where $D > 0$ and $\log(A)\log(B) > 0$.

2. *Ware growth function* [17]

$$x(t) = \delta(1 - \varphi^{-t}), \tag{5.12}$$

where $\delta > 0$ and the constant φ is greater than 1.

3. *Power law growth function* [16]

$$x(t) = a + bt^{\gamma}, \tag{5.13}$$

where $a > 0$ and $b > 0$. For $0 < \gamma < 1$, the growth is concave and without an upper limit; for $\gamma = 1$, the growth is linear: for $\gamma > 1$, the growth is convex.

5.2.2 Epidemic Models

Below, we discuss two epidemic models of diffusion of knowledge by research publications. Epidemic models were used originally in population dynamics [18–24]. And for many years, most models of population dynamics were of interest only to biologists [25–30]. Today, these models are applied in many more areas of science [26–40]. For the area of research on scientific systems, the epidemic models are of great interest, too. This is so because some stages of processes by which ideas spread within a population, e.g., of scientists, has features that are like those of the spread of epidemics [41–43].

Epidemic models are a subclass of the more general class of Lotka–Volterra models [44–49] that are used in research on systems in the fields of biological population dynamics, social dynamics, economics, as well as for modeling processes connected to the spread of knowledge, ideas, and innovations [50–53].

The central concept of the epidemic models is the concept that scientific results spread to scientific communities by an epidemic diffusion process whereby more and more members of the scientific community are "infected" by the new scientific ideas and results. An important channel for spreading of this "infection" is research publications.

5.2.3 *Change in the Number of Publications in a Research Field. SI (Susceptibles–Infectives) Model of Change in The Number of Researchers Working in a Field*

Three basic classes of populations are important in epidemic research: [54]:

- **The susceptibles** S, who can become infectives on coming in contact with infectious material (the infectious material in our case is the scientific ideas).
- **The infectives** I who host the infectious material.
- **The recovered** R who are removed from the epidemic.

Because of this, the name of a class of epidemic models is the SIR-model (susceptibles–infectives–recovered (removed)). Nowakowska [55] discussed several discrete epidemic models for predicting changes in the number of publications in a given scientific field. The main assumption of the models is that the number of publications in the next period of time (say one year) will depend on the number of publications that have recently appeared and on the degree to which the subject has been exhausted. The behavior of the number of publications is considered to be as follows. The numbers of publications appearing in successive periods of time should first increase, then reach a maximum, and as the problem becomes more and more exhausted, the number of publications should decrease. A mathematical relationship that reflects such behavior was proposed by Daley [56]:

$$p_{t+1} = c_t p_t \left(N - \sum_{i=1}^{t} p_i \right), \tag{5.14}$$

where

- p_t: number of publications written in the period t;
- N: number of publications that have to appear in order to exhaust the problems in the research field.
- c_t: coefficient that can be connected to the number of researchers x_t working in the field: $c_t = 1 - (1 - d)^{x_t}$, where d is a parameter.

The epidemic part of the model is connected to the researchers who produce publications in the corresponding research field. There are researchers who produce publications in the field, and the number of these researchers may change. Some factors contribute to a decrease in the number of researchers (they retire or are no longer interested in the corresponding research problems). And there is a factor that contributes to an increase in the number of authors in the research field: new authors may begin to write publications (young researchers that begin their research career or researchers who became interested in the problems from the corresponding research field). We shall treat the last increase in the number of authors as infection and the entire process as an epidemic.

Let us assume that at a certain moment t, the epidemic's state is (x_t, y_t), where

- x_t is the number of infectives: authors who write publications in the corresponding scientific field;
- y_t is the number of susceptibles.

Then:

1. for a sufficiently short time interval Δt, one may expect that the number of infectives $x_{t+\Delta t}$ will be equal to $x_t - ax_t\Delta t + bx_t y_t \Delta t$,
2. while the number of susceptibles $y_{t+\Delta t}$ will be equal to $y_t - bx_t y_t \Delta t$ (a and b are suitable constants).

Let the expected number of individuals who either "die" or "recover" during the interval $(t, t + \Delta t)$, be $ax_t\Delta t$, and let $bx_t y_t \Delta t$ be the expected number of new infections. The equations of this model are

$$\begin{aligned} x_{t+\Delta t} &= x_t - ax_t\Delta t + bx_t y_t \Delta t, \\ y_{t+\Delta t} &= y_t - bx_t y_t \Delta t. \end{aligned} \tag{5.15}$$

The coefficients a and b may depend on the attractiveness of research field, on its being exhausted, etc. After setting appropriate relationships for a and b, one may investigate numerically the dynamics of the infectives x and susceptibles y, i.e., the dynamics of researchers producing publications in the corresponding research field.

5.2.4 *Goffman–Newill Continuous Model for the Dynamics of Populations of Scientists and Publications*

The model discussed above is an example of a discrete model. Now let us consider a continuous epidemic model connected with the dynamics of researchers and publications. Such a model is the Goffman–Newill model.

The Goffman–Newill model of intellectual epidemics is based on the Reed–Frost epidemic model [57–59], which was developed during the 1930s by Lowell Reed and Wade Frost, of the Johns Hopkins University. In the Reed–Frost model, one assumes a fixed population of size N. At each time, there is a certain number of cases of disease, C, and a certain number of susceptibles, S. One assumes that each case is infectious for a fixed length of time, and ignores the latent period: when individuals recover, one assumes that they are immune to further infection. During the infectious period of each case, one assumes that susceptibles may be infected and the disease may propagate further. The Goffman–Newill model [5, 60, 61] exploits the idea that the spreading of scientific ideas within a population of scientists can be studied on the basis of the publications of the members of that population. The main process in the model is the transfer of infectious materials (ideas) between humans by means of an intermediate host (a written article).

Let a scientific field be F and SF a subfield of F. We shall use the following notation: N_0, the number of scientists writing papers in the field F at t_0; I_0, the number of scientists writing papers in SF at t_0 (the number of infectives). Thus $S_0 = N_0 - I_0$ is the number of susceptibles; there is no removal (i.e., no scientists move out of the corresponding population) at t_0, but there is removal $R(t)$ at later times t. In addition, N_0' is the number of papers produced on F at t_0, and I_0' is the number of papers produced in SF at this time.

The process of intellectual infection takes place as follows:

1. A member of F is infected by a paper from I';
2. After some latency period, this infected member produces "infected" papers in N', i.e., the infected member produces a paper in the subfield SF citing a paper from I';
3. These'"infected" papers may infect other scientists from F and its subfields, such that the intellectual infection spreads from SF to the other subfields of F.

Let β be the rate at which the susceptibles from class S become "intellectually infected" from class I and let β' be the rate at which the papers in SF are cited by members of F who are producing papers in SF. As the infection process develops, some susceptibles and infectives are removed, i.e., some scientists are no longer active, and some papers are no longer cited. In addition, let γ and γ' be the rates of removal of infectives from the populations I and I' respectively, and let δ and δ' be the rates of removal from the populations of susceptibles S and S'. Moreover, there can be a supply of infectives and susceptibles in F and SF. Let the rates of introduction of new susceptibles be μ and μ' (these are the rates at which new authors and new papers are introduced in F) and let the rates of introduction of new infectives be υ and υ' (these are the rates at which new authors and new papers are introduced in SF). In addition, within a short interval of time, a susceptible can remain susceptible or can become an infective or be removed; the infective can remain an infective or can be removed; the removed remains removed; the immunes remain immune and do not return to the population of susceptibles.

Let us impose also the condition that the populations are homogeneously mixed. Then the system of model equations is

$$\frac{dS}{dt} = -\beta SI' - \delta S + \mu; \qquad \frac{dI}{dt} = \beta SI' - \gamma I + \upsilon \tag{5.16}$$

$$\frac{dR}{dt} = \gamma I + \delta S; \qquad \frac{dS'}{dt} = -\beta' S'I - \delta S' + \mu' \tag{5.17}$$

$$\frac{dI'}{dt} = \beta' S'I - \gamma' I' + \upsilon'; \qquad \frac{dR'}{dt} = \gamma' I' + \delta' S'. \tag{5.18}$$

The conditions for development of an epidemic are as follows:

1. If as an initial condition at t_0, a single infective is introduced into the populations N_0 and N_0', then for an epidemic to develop, the change in the number of infectives must be positive in both populations.

2. Thus for $\rho = \frac{\gamma - \upsilon}{\beta}$ and $\rho' = \frac{\gamma' - \upsilon'}{\beta'}$, the threshold for the epidemic arises from the conditions $\beta SI' > \gamma I - \upsilon$ and $\beta' S' I' > \gamma' I' - \upsilon'$, so that the threshold is

$$S_0 S_0' > \rho \rho'. \tag{5.19}$$

3. The development of an epidemic is given by the equation for $\frac{dI}{dt}$.
4. The peaks of the epidemics occur at time points where $\frac{d^2 I}{dt^2} = 0$, while the epidemic's size is given by $I(t \to \infty)$.

The Goffman–Newill model stimulated much research in the area of modeling of processes in science by models from population dynamics and epidemiology. Let us mention here just the models of the growth of mathematics specialties [62] and of the growth of papers in a specialty [63–67]. One can add additional categories of researchers to the SIR type of models. One example of this is the adding of the class of researchers exposed to the corresponding scientific ideas. In such a way, one obtains a class of epidemic SEIR models of research production [68, 69].

5.2.5 *Price Model of Knowledge Growth. Cycles of Growth of Knowledge*

An example of nonepidemic model of knowledge growth is the model of Price [70, 71]. The model is based on the following assumptions:

1. The growth is measured by the number of important publications appearing at a given time.
2. The growth has a continuous character, and a finite time period $T = \text{const}$ is needed to build up a result of fundamental character.
3. The interactions between various scientific fields are neglected.

Let in addition the number of scientists publishing results in this field be constant. Then the rate of scientific growth (of the publications x) is proportional to the number of important publications at time t minus the time period T required to build up a fundamental result. The model equation is

$$\frac{dx}{dt} = \alpha x(t - T), \tag{5.20}$$

where α is a constant, and the initial condition $x(t) = \phi(t)$ is defined on the interval $[-T, 0]$.

Often, the population of researchers is varying. Then for consideration of the evolution of the average number of papers per researcher instead of the linear right-hand side (5.20), the following nonlinear model is used:

$$\frac{dx}{dt} = f(x(t - T), x(t)), \tag{5.21}$$

where f is a homogeneous function of degree one. The simplest form of such a function is a linear function. Let us assume that the population of researchers L grows at the constant rate $n = \frac{1}{L}\frac{dL}{dt}$ and let $z = x/L$ be the mean number of papers written by a researcher. Then the evolution of the number of papers written by a researcher has the form

$$\frac{dz}{dt} = \alpha z(t - T) - nz(t). \tag{5.22}$$

We note the following:

1. If $n = 0$ and $T = 0$, the Price model of exponential growth is recovered.
2. Equation (5.22) is linear, but cyclic behavior may appear because of the feedback between the delayed and nondelayed terms.

The Price model was criticized along the following points: the quality of research is omitted, and many scientific products that seem to be new are not really new; creativity and innovation are confused, and creative papers with new ideas and results have the same importance as trivial duplications. Price answered by formulating the hypothesis that one may study only the growth of *important* discoveries, inventions, and scientific laws, rather than all important and trivial things. Then every growth will follow the same pattern as that mentioned above, but the growth will be much slower.

5.3 A Deterministic Model Connected to Dynamics of Citations

Sangwal [72–75] proposed a model of the growth of citations of a scientist based on the progressive nucleation mechanism known from chemistry [76]. In chemistry, this mechanism describes simultaneous nucleation and growth of a nucleus to crystallites of visible size. If the initial volume of the crystallizing phase is V and the crystallized volume is $V(t)$, then one has the following relationship for the ratio V_c/V:

$$\alpha(T) = \frac{V_c(t)}{V} = \left\{1 - \exp\left[-\left(\frac{t}{\Theta}\right)^q\right]\right\}, \tag{5.23}$$

where the relationships for the time constant Θ and for the exponent q are

$$q = 1 + \nu d; \quad \Theta = \left(\frac{q}{kG^{q-1}J_s}\right)^{1/q},$$

and the parameters are as follows:

- $\nu > 0$: a constant;
- d: dimension of the growing nucleus (can be 1, 2, 3);
- k: shape factor of the nucleus ($k = 4\pi/3$ for a spherical nucleus);

- $G = \frac{r^{1/v}}{t}$;
- r: radius of the growing nucleus;
- J_s: rate of stationary nucleation.

When $kJ_s = G$, then $\Theta = \frac{q^{1/q}}{kJ_s}$, which will be the case of interest for us. In this case, the nuclear radius grows in time as $r(t) \propto t^v$.

The process of nucleation can also be used to describe the growth of citations of a paper written by scientist. In this case,

$$\alpha(t) = \alpha(t) = \frac{C(t)}{C_{max}} = \left\{1 - \exp\left[-\left(\frac{t}{\Theta}\right)^q\right]\right\}, \tag{5.24}$$

where C is the maximum number of citations that a paper can receive, and $C(t)$ is the cumulative number of citations of the paper in the time t. The other parameters are defined as above (we recall that ($\Theta = \frac{q^{1/q}}{kJ_s}$). The nucleation model can be transferred to a description of the accumulation of citations of a paper if several conditions are met:

- Citations received by a paper and the paper earning these citations compose a closed system in which the process of occurrence of citations is stationary.
- Occurrence of citations of a paper continues in time and finally approaches a constant value C_{max}, which is the maximum number of citations received by the paper at time T.
- The dependence of the cumulative number of citations $C(t)$ of the paper at time t is determined by the maximum number of citations C_{max}, a time constant Θ, and an exponent q. The citation pattern of different papers of an author is characterized by different values of $C(t)$, Θ, and q for each paper.

If a researcher has authored n papers, then the cumulative fraction $\alpha_s(t)$ of the citations of these papers is

$$\alpha_s(t) = \sum_{i=0}^{n} \alpha_i(t). \tag{5.25}$$

If we assume that the researcher publishes papers at equal time intervals ΔT, then

$$\alpha_s(t) = \sum_{i=0}^{n} \alpha_i[t - (i-1)\Delta T] = \sum_{i=1}^{n} \left\{1 - \exp\left[-\left(\frac{t - (i-1)\Delta T}{\Theta_i}\right)\right]\right\}. \tag{5.26}$$

One can fit the model parameters for the data of the researcher whose production is evaluated. In most cases, the fit describes very well the process of accumulation of citations [75].

5.4 Deterministic Models Connected to Research Dynamics

5.4.1 *Continuous Model of Competition Between Systems of Ideas*

Ideas can diffuse not only among scientists in one organization but also in space (e.g., from scientists from one country to scientists from other countries). Thus one may include spatial variables in the models describing the diffusion of ideas. Such models can be of great interest during periods of globalization of economies, knowledge, and technology [77–82]. Below, we describe a model closely connected to the space–time models of migration of populations [83, 84].

The diffusion of ideas is often accompanied by competition between systems of ideas. Let a population of N individuals occupy a two-dimensional plane. We assume that:

- there exists a set of ideas $P = \{P_0, P_1, \ldots, P_n\}$;
- N_i members of the population are followers of the set P_i of ideas;
- members N_0 of the class P_0 are not supporters of any set of ideas.

In such a way, the population is divided into $n + 1$ subpopulations of followers of different sets of ideas, and $N = N_0 + N_1 + \cdots + N_n$. Let a small region $\Delta S = \Delta x \Delta y$ be selected in the plane. In this region, there are ΔN_i individuals holding the ith set of ideas, $i = 0, 1, \ldots, n$. If ΔS is sufficiently small, the density of the ith population can be defined as $\rho_i(x, y, t) = \frac{\Delta N_i}{\Delta S}$. Further, we assume that members of the ith population are capable of moving through the borders of the area ΔS. Let $\mathbf{j}_i(x, y, t)$ be the current of this movement. The total change in the number of members of the ith population is

$$\frac{\partial \rho_i}{\partial t} + \mathrm{div}\mathbf{j}_i = C_i, \tag{5.27}$$

where the changes are summarized by the function $C_i(x, y, t)$.

The first term in (5.27) describes the net rate of increase of the density of the ith population. The second term describes the net rate of immigration into the area. The right-hand side of (5.27) describes the net rate of increase exclusive of immigration. The quantities $\mathbf{j}_i$ and C_i are as follows: $\mathbf{j}_i$ is assumed to have two parts, a nondiffusion part $\mathbf{j}_i^{(1)}$ and a diffusion part $\mathbf{j}_i^{(2)}$ that is assumed to have the general form of a linear multicomponent diffusion [77] (D_{ik} is the coefficient of diffusion):

$$\mathbf{j}_i = \mathbf{j}_i^{(1)} + \mathbf{j}_2^{(2)} = \mathbf{j}_i^{(1)} - \sum_{k=0}^{n} D_{ik}(\rho_i, \rho_k, x, y, t)\nabla\rho_k. \tag{5.28}$$

A further assumption is that some of the followers of the set of ideas P_i are capable of changing to another set of ideas, e.g., they can change P_i for P_j. It can be assumed that the following processes can occur with respect to the members of the subpopulations:

- **Deaths**: described by a term $r_i\rho_i$. We assume that the number of deaths in the ith population is proportional to its population density. In general, $r_i = r_i(\rho_\nu, x, y, t; p_\mu)$, where ρ_ν stands for $(\rho_0, \rho_1, \ldots, \rho_N)$ and p_μ stands for $(p_1, \ldots, p_M)$, containing parameters of the environment.
- **Noncontact conversion**: in this class are included kinds of changes between P_i and P_j exclusive of changes after interpersonal contact between the members of populations. A reason for noncontact conversion can be the existence of different kinds of mass communication media (scientific books, influence of mass media, etc.). For the ith population, the change in the number of members by this kind of conversion is $\sum_{j=0}^{n} f_{ij}\rho_j, f_{ii} = 0$. In general, $f_{ij} = f_{ij}(\rho_\nu, x, y, t; p_\mu)$.
- **Contact conversion**: this happens by interpersonal contacts among the members of the population. Such contacts can happen between members in groups consisting of two members (binary contacts), three members (ternary contacts), four members, etc. As a result of the contacts, members of each population can change their sets of ideas. For binary contacts, let it be assumed that the probability of change for a member of the jth population is proportional to the probability of, for instance, the number of contacts, i.e., proportional to the density of the ith population. Then the total number of "conversions" from P_j to P_i is $a_{ij}\rho_i\rho_j$, where a_{ij} is a parameter.
  Next, a change in the set of ideas can take place by ternary contact. For this, one must have a group of three members. We assume that such a group exists with a probability proportional to the corresponding densities of the concerned populations. In a ternary contact between members of the ith, jth, and kth populations, members of the jth and kth populations can change their sets of ideas to $P_i = b_{ijk}\rho_i\rho_j\rho_k$, where b_{ijk} is a parameter. In general, $a_{ij} = a_{ij}(\rho_\nu, x, y, t; p_\mu)$; $b_{ijk} = b_{ijk}(\rho_\nu, x, y, t; p_\mu)$; etc.

On the basis of all of the above the C_i term can be written as

$$C_i = r_i\rho_i + \sum_{j=0}^{n} f_{ij}\rho_j + \sum_{j=0}^{n} a_{ij}\rho_i\rho_j + \sum_{j,k=0}^{n} b_{ijk}\rho_i\rho_j\rho_k + \ldots. \tag{5.29}$$

Hence the model system of equations is

$$\frac{\partial\rho_i}{\partial t} + \operatorname{div}\mathbf{j}_i^{(1)} - \sum_{j=0}^{n} \operatorname{div}(D_{ij}\nabla\rho_j) = r_i\rho_i + \sum_{j=0}^{n} f_{ij}\rho_j + \sum_{j=0}^{n} a_{ij}\rho_i\rho_j + \sum_{j,k=0}^{n} b_{ijk}\rho_i\rho_j\rho_k + \ldots. \tag{5.30}$$

The density of the entire population is $\rho = \sum_{i=0}^{n} \rho_i$. This density can change over time. One possible assumption is that ρ changes over time according to the Verhulst law

$$\frac{\partial\rho}{\partial t} = r\rho\left(1 - \frac{\rho}{C}\right), \tag{5.31}$$

where $C(\rho_v, x, y, t; p_\mu)$ is the carrying capacity of the environment and $r(\rho_v, x, y, t; p_\mu)$ is a positive or negative growth rate.

Now let us consider the case in which the current $\mathbf{j}_i^{(1)}$ is negligible, i.e., $\mathbf{j}_i^{(1)} \approx 0$. In addition, we consider only the case in which all parameters are constants. The model system of equations becomes

$$\frac{\partial \rho_i}{\partial t} - D_{ij} \sum_{j=0}^{n} \Delta \rho_j = r_i \rho_i + \sum_{j=0}^{n} f_{ij} \rho_j + \sum_{j=0}^{n} a_{ij} \rho_i \rho_j + \sum_{j,k=0}^{n} b_{ijk} \rho_i \rho_j \rho_k + \ldots, \tag{5.32}$$

where

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \quad i = 0, 1, 2, \ldots, n. \tag{5.33}$$

Next we shall separate the dynamics of averaged quantities from the dynamics of fluctuations. If $q(x, y, t)$ is a quantity defined in an area S, then the corresponding plane averaged quantity is

$$\overline{q} = \frac{1}{S} \int \int_S dxdy\, q(x, y, t). \tag{5.34}$$

The fluctuations are denoted by $Q(x, y, t)$:

$$q(x, y, t) = \overline{q}(t) + Q(x, y, t). \tag{5.35}$$

We assume that territory S is large enough; every plane averaged combination of fluctuations vanishes; $\int \int_S dxdy \Delta Q_k$ is finite. Then $\overline{\Delta Q_k} = \frac{1}{S} \int \int_S dxdy \Delta Q_k \to 0$. On the basis of these assumptions, the dynamics of the averaged quantities are separated from the dynamics of fluctuations by means of a plane averaging of (5.32). The result is

$$\overline{\rho}_0 = \overline{\rho} - \sum_{i=1}^{n} \overline{\rho}_i; \quad \frac{d\overline{\rho}}{dt} = r\overline{\rho} \left(1 - \frac{\overline{\rho}}{C}\right) \tag{5.36}$$

$$\frac{d\overline{\rho}_i}{dt} = r_i \overline{\rho}_i + \sum_{j=0}^{n} f_{ij} \overline{\rho}_j + \sum_{j=0}^{n} a_{ij} \overline{\rho}_i \overline{\rho}_j + \sum_{j,k=0}^{n} b_{ijk} \overline{\rho}_i \overline{\rho}_j \overline{\rho}_k + \ldots. \tag{5.37}$$

Instead of (5.36), we can write an equation for $\overline{\rho}_0$ of the type of (5.37). Then the total population density $\overline{\rho}$ will not follow the Verhulst law.

Equations (5.36) and (5.37) represent the model of competition among sets of ideas proposed in [85]. There also exists a discrete version of this model [86], and it can be applied to competition between different sets of ideas (scientific, political, religious, technological, etc.).

5.4.2 *Reproduction–Transport Equation Model of the Evolution of Scientific Subfields*

By means of migration, people can move from one territory to another. The change of the field of research by a scientist may also be considered a migration process [82, 87]. In order to study this, let us map research problems by sequences of signal words or macro-terms $P_i = (m_i^1, m_i^2, \ldots, m_i^k, \ldots, m_i^n)$, which are registered according to the frequency of their occurrence in the texts. Then:

- Each point of the problem space, described by a vector $\mathbf{q}$, corresponds to a research problem, with the problem space containing all scientific problems (no matter whether they are under investigation or not).
- The scientists distribute themselves over the space of scientific problems with density $x(\mathbf{q}, t)$. Thus there is a number $x(\mathbf{q}, t)d\mathbf{q}$ of scientists working at time t in the element $d\mathbf{q}$.
- The field mobility processes correspond to a density change of scientists in the problem space, i.e., instead of working on problem $\mathbf{q}$, a scientist may begin to work on problem $\mathbf{q}'$.
- As a result, $x(\mathbf{q}, t)$ decreases and $x(\mathbf{q}', t)$ increases.

This movement of scientists can be described by means of a reproduction–transport equation:

$$\frac{\partial x(\mathbf{q}, t)}{\partial t} = x(\mathbf{q}, t)\, w(\mathbf{q} \mid t) + \frac{\partial}{\partial \mathbf{q}} \left(f(\mathbf{q}, x) + D(\mathbf{q}) \frac{\partial \mathbf{q}}{\partial x} \right). \tag{5.38}$$

In (5.38), self-reproduction and decline are represented by the term $w(\mathbf{q} \mid x)\, x(\mathbf{q}, t)$; for the reproduction rate function $w(\mathbf{q} \mid x)$, one can write the relationship

$$w(\mathbf{q} \mid x) = a(\mathbf{q}) + \int d\mathbf{q}' b(\mathbf{q}, \mathbf{q}' x(\mathbf{q}, t). \tag{5.39}$$

The local value of $a(\mathbf{q})$ is an expression of the rate at which the number of scientists in field $\mathbf{q}$ is growing through self-reproduction and decline. The function $b(\mathbf{q}, \mathbf{q}')$ describes the influence exerted on the field $\mathbf{q}$ by the neighboring field $\mathbf{q}'$. The field mobility is modeled by means of the term $\frac{\partial}{\partial \mathbf{q}} \left(f(\mathbf{q}, x) + D(\mathbf{q}) \frac{\partial}{\partial \mathbf{q}} x(\mathbf{q}, t) \right)$.

In order to use this equation, we need initial conditions and determination of the coefficients on the basis of statistical data for the distribution of the scientists with respect to the research problems.

5.4.3 Deterministic Model of Science as a Component of the Economic Growth of a Country

Below we discuss a component of the model of evolution of the GDP (gross domestic product) of a country. This component is connected to the role of technology for increasing GDP [88–90].

The GDP of a country may grow extensively by inflow of workforce or capital to the national economic structures and systems [91]. But the GDP of a country may grow also intensively by advancement in science and technology. Let us discuss a simple model in which the GDP Y has the form

$$Y(t) = Y(L(t), C(t), T(t)). \tag{5.40}$$

The quantities in (5.40) are as follows:

- $L(t)$: labor (human resources);
- $C(t)$: production resources;
- $T(t)$: technology level.

Note that the above quantities are not chosen arbitrarily. They represents important factors that may influence the GDP of a country.

The change in the GDP over time is given by

$$\frac{dY}{dt} = \frac{\partial Y}{\partial L}\frac{dL}{dt} + \frac{\partial Y}{\partial C}\frac{dC}{dt} + \frac{\partial Y}{\partial T}.\frac{dT}{dt}. \tag{5.41}$$

The term $(\partial Y/\partial T)(dT/dt)$ describes the change in the GDP because of the evolution of technology. This component of the change of the GDP will be of interest for us below. Let us note that if technology advances, $((dT/dt) > 0)$, this is a contribution to the growth of the GDP. If technology for some reason deteriorates, $((dT/dt) < 0)$, then it can contribute to a decrease in the GDP.

The change in the GDP due to technology may be assumed to be [92]

$$\frac{\partial Y}{\partial T} = \frac{Y}{T}. \tag{5.42}$$

Equation (5.42) means that the increase in the technology level leads to a proportional increase of the GDP. Then the studied term from (5.41) becomes

$$\frac{\partial Y}{\partial T}\frac{dT}{dt} = Y\left(\frac{1}{T}\frac{dT}{dt}\right). \tag{5.43}$$

Next we shall discuss how the term $(1/T)(dT/dt)$ depends on S_T: the growth in knowledge about technology. Then the growth in knowledge about technology will be connected to the growth in scientific knowledge, which will be denoted by S.

We adopt the following notation:

- I_T: the investment directed to applications of the results of new technologies (machines, processes, etc.);
- I_0: the investments in older technologies;
- γ: coefficient of proportionality between the growth of knowledge about technology S_T and growth of scientific knowledge S.

Then the relationship between T and S is

$$\frac{1}{T}\frac{dT}{dt} = \gamma \frac{I_T}{I_O}\frac{1}{S}\frac{dS}{dt}. \tag{5.44}$$

Equation (5.44) leads to the following conclusions:

1. **Importance of the fundamental research**: Research and especially fundamental research lead to an increase in scientific knowledge. If there is no growth in scientific knowledge, $((dS/dt) = 0)$, then there is no technological evolution, $((1/T)(dT/dt) = 0)$, and an important factor for the growth of the national GDP is lost.
2. **Importance of the transfer of scientific knowledge to knowledge about technology**: If $\gamma = 0$, i.e., there is no transfer, then $((1/T)(dT/dt) = 0)$ (no technology evolution) even if scientific knowledge grows. Thus what is important for a country is to increase γ (by strengthening engineering sciences by creating new engineering institutes, for example). The value of γ for developed countries is about 0.5 (1 % growth in scientific knowledge results in 0.5 % growth in the number of patents).
3. **Importance of investment in new technologies**: If there is no such investment $(I_T = 0)$, then there is no evolution of technology, $((1/T)(dT/dt) = 0)$, even if there is growth of scientific knowledge and an intensive transfer of knowledge about technology.

The rate of growth of scientific knowledge $(1/S)(dS/dt)$ is assumed to depend on two main factors: the funding of (investment in) science I and the labor L ("human resources" or the number of qualified scientists). Let us set

$$\frac{1}{S}\frac{dS}{dT} = \phi(I, L). \tag{5.45}$$

Let us assume that $\phi(I, L)$ is a homogeneous function of degree α with respect to the funding I and a homogeneous function of the factor β with respect to the human resources L. Then we can obtain the relationship

$$\phi = aI^{\alpha}L^{\beta} = \frac{1}{S}\frac{dS}{dt}, \tag{5.46}$$

where a is a coefficient of integration. Hence a power-law relationship may exist between the rate of growth of scientific knowledge and investment and the number

of qualified scientists. We stress the words *power law*, since such laws arise frequently in studies of research systems (for examples, see Chap. 5).

Equation (5.46) leads to interesting conclusions.

1. **Exponential growth of knowledge in an established research area**. Let us consider an established research area with constant investment in science: $I =$ const and a constant number of qualified scientists $L =$ const. From (5.46), we obtain the relationship
$$S = S_0 \exp[aI^\alpha L^\beta t] \tag{5.47}$$
(S_0 is a constant of integration), which means that the scientific knowledge in this area is growing exponentially.
2. **Double-exponential growth of scientific knowledge in a new research area**. Let us now consider a new research area in which the number of scientists grows exponentially over time, $L = \exp(\mu t)$, and the funding is constant: $I =$ const and large enough. Then the growth of scientific knowledge in this area is double-exponential,
$$S = S_0 \exp\left[\frac{aI^\alpha}{\mu\beta}\exp(\mu\beta t)\right]. \tag{5.48}$$

The substitution of (5.44)–(5.46) in (5.43) leads to the following relationship for the influence of science on the change of GDP of a country:

$$\frac{\partial Y}{\partial T}\frac{dT}{dt} = \gamma a \frac{I_T}{I_O} I^\alpha L^\beta Y. \tag{5.49}$$

Equation (5.49) shows that countries that have a large GDP possess advantages (since $\frac{\partial Y}{\partial T}\frac{dT}{dt} \propto Y$), and in addition, the human factor and investment in science are very important. Thus every nation should try to build a community of qualified researchers and should invest sufficiently in the national research system. If this is not the case, then the process of global competition among the nations will lead inevitably to a brain drain.

The model above represents a global point of view of the importance of science as a component of economic growth of a country. There exists also a local point of view regarding this importance. A local point of view means that one considers the growth of the output of a worker with advancing technology. A mathematical model of this relationship may be based on the Cobb–Douglass production function and on the Solow model. The form of the Cobb–Douglass production function is [93, 94]

$$Y = AK^\alpha L^{1-\alpha}, \tag{5.50}$$

where

- Y: output per worker;
- K: physical capital per worker;
- L: human capital per worker (labor);

- A: productivity;
- α: output elasticity of the physical capital;
- $\beta = 1 - \alpha$: output elasticity of the human capital.

Looking at (5.50), we can conclude that technological advance allows (by increasing productivity) given quantities of physical and human capital to be combined to produce more output than was possible when older technology was used. Hence changes in technology directly affect economic growth. In addition, human capital L per worker cannot grow infinitely. Then in order to increase the output Y, one has to increase the physical capital K per worker (there are also limits to this increase), or one can increase productivity A by advancing technology. Thus even when K and L have reached their maximum values, *as long as A (productivity) continues to grow as a consequence of technological advance, income per capita will continue to grow too*.

The result of the mathematical theory is that the rate of growth of the total output $Y^* = (1/Y)(dY/dt)$ per worker (in the steady state of the production system) is connected to the growth of productivity A (which means that there is a strong connection between the growth of the total output and technological progress). Namely, if the rate of advance of technology is $A^* = (1/A)(dA/dt)$, then

$$Y^* = A^* \left(\frac{1}{1-\alpha} \right). \tag{5.51}$$

Equation (5.51) tells us that technological advance (by research and development) is extremely important for economic growth.

5.5 Several General Remarks About Probability Models and Corresponding Processes

In many cases, in the mathematical models of mechanisms of production of scientific information, one uses the concept of population of "sources" producing "items" observed over time [95]. *The observation of the items produced by a source is equivalent to the observation of a stochastic point process: a sequence of events occurring randomly in time*. The modeling of the corresponding process requires specification of the probabilistic mechanism producing the observed events.

> The simplest available point process is the Poisson process, which corresponds to the situation that events occur completely at random over time with the overall average rate of occurrence remaining constant, so that the expected number of events occurring increases linearly with time.

In order to model more realistic situations, the rate of the Poisson process may:

1. vary in time deterministically [96]. In this case, the number of occurring events may have nonlinear variation in time, and the process is called an inhomogeneous Poisson process;
2. vary in time stochastically [97, 98]. Such a process is called a doubly stochastic Poisson process or Cox process.

Each of the three Poisson processes described above has independent increments. The Poisson process and the doubly stochastic Poisson process have stationary increments. Thus they are able to model situations in which the probability distribution of the number of events in a period of time depends only on the length of the period and not on the time at which it begins.

When the entire population of sources is studied, it may happen that some variability in the rate of production between different items exists. The observed process is then a mixture of the individual processes, and it can be modeled mathematically by mixing the parameters determining the rates of production of the individual sources. The resulting mixed process may still have stationary increments, but because of the mixing, the increments are no longer independent.

We are going to describe briefly three kinds of Poisson processes that will arise in the models discussed below: the Greenwood–Yule process (gamma–Poisson process), GIGP (generalized inverse Gaussian–Poisson process), and Waring process (a negative binomial process) [95]. Let us consider a source that produces X_t ($t \geq 0$) items in the interval $[0, t]$. The process of production of items (the point process) is specified by a parameter θ, and we know the form of the process $\{X_t \mid \theta\}$ for a given value of θ. For given θ, the increments of the process are stationary but not independent, and

$$p(X_t = r) = E_\theta P(X_t = r \mid \theta) = \int dx f_\theta(x) p(X_t = r \mid \theta = x). \tag{5.52}$$

The above-mentioned three processes will be obtained by specifying the probability distribution function $f_\theta(x)$ and the form of the conditional process $\{X_t \mid \theta\}$. For example, in order to obtain the Greenwood–Yule process (called also gamma–Poisson process), we have to assume that each source produces items as a Poisson process and the probability distribution function is for the gamma distribution. In detail,

$$p(X_t = r \mid \lambda) = \exp(-\lambda t) \frac{(\lambda t)^r}{r!}; \quad r = 0, 1, \ldots, \tag{5.53}$$

where λ is the rate of the Poisson process; λ has a gamma distribution with scale parameter β and index ν:

$$f_\lambda(x) = \frac{\beta^{-\nu} x^{\nu-1}}{\Gamma(\nu)} \exp(-\frac{x}{\beta}); \quad x > 0. \tag{5.54}$$

As a result of substituting (5.53) and (5.54) in (5.52), one obtains the negative binomial distribution of index ν and parameter $p_t = 1/(1+\beta t)$:

$$p(X_t = r) = \binom{r+\nu-1}{r}\left(\frac{1}{1+\beta t}\right)^{\nu}\left(\frac{\beta t}{1+\beta t}\right)^{r}; \quad r = 0, 1, \ldots. \tag{5.55}$$

The GIGP (generalized inverse Gaussian–Poisson process) is obtained when the probability distribution function for the rate λ of the Poisson process (5.53) is

$$f_{\lambda}(x) = c(\alpha, \gamma, \theta)x^{\gamma-1}\exp\left[-x\left(\frac{1}{\theta}-1\right)-\frac{\alpha^2\theta}{4x}\right], \tag{5.56}$$

where $x > 0$; $-\infty < \gamma < \infty$; $\alpha \geq 0$, and the constant ensuring the normalization is

$$c(\alpha, \gamma, \theta) = \frac{(1-\theta)^{\gamma/2}}{2(\alpha\theta/2)^{\gamma}}K_{\gamma}\{\alpha(1-\theta)^{1/2}\}, \tag{5.57}$$

where $K_{\gamma}\{\alpha(1-\theta)^{1/2}\}$ is the modified Bessel function of the second kind of order γ. The substitution of the density (5.56) in (5.52) leads to the distribution

$$p(X_t = r) = \frac{(1-\theta_t)^{\gamma/2}}{K_{\gamma}\{\alpha(1-\theta)^{1/2}\}}\frac{(\alpha_t\theta_t/2)^r}{r!}K_{r+\gamma}(\alpha_t); \quad r = 0, 1, \ldots, \tag{5.58}$$

where $\theta_t = (t\theta)/[1+\theta(t-1)]$ and $\alpha_t = \alpha[1+(t-1)\theta]^{1/2}$. This distribution is reduced to the GIGP distribution when $t = 1$ (then $\theta_t = \theta$ and $\alpha_t = \alpha$). Because of this, the process X_t described by (5.58) will be called a GIGP process and may be denoted by GIGP$(\alpha_t, \theta_t, \gamma)$. Sichel [99, 100] used $\gamma = -1/2$, i.e., the GIGP$(\alpha_t, \theta_t, -1/2)$ distribution

$$p(X_t = r) = \left(\frac{2\alpha_t}{\pi}\right)^{1/2}\exp[\alpha(1-\theta)^{1/2}]\frac{(\alpha_t\theta_t/2)^r}{r!}K_{r-1/2}(\alpha_t); \quad r = 0, 1, \ldots, \tag{5.59}$$

in many practical applications.

Finally, we consider the Waring process (which will be much discussed below in the text). For this process, each source produces items as a negative binomial process of parameter q and index ψ:

$$p(X_t = r \mid q) = \binom{r+\psi t-1}{r}q^{\psi t}(1-q)^r; \quad r = 0, 1, \ldots, \tag{5.60}$$

and the parameter q has a beta distribution with parameters a and b:

$$f_p(x) = \frac{1}{B(a,b)}\frac{\psi^a x^{b-1}}{(x+\psi)^{a+b}}. \tag{5.61}$$

The substitution of (5.60) and (5.61) in (5.52) leads to

$$p(X_t = r) = \frac{\Gamma(\psi t + a)}{B(a,b)\Gamma(\psi t)} \frac{\Gamma(r + \psi t)\Gamma(r + b)}{r!\Gamma(r + \psi t + a + b)}. \tag{5.62}$$

Equation (5.62) describes the generalized Waring distribution [101–103]; Γ is the gamma function, and B is the beta function.

Some remarks about the moments of the obtained distributions follow. Moments of all orders exist for the gamma–Poisson distribution and for the GIGP distribution. For the existence of moments of the generalized Waring distribution, one has to impose some requirements on the parameters of the distribution. For the gamma–Poisson distribution, the mean $E[X_t]$ and the variance $V[X_t]$ are

$$E[X_t] = \nu\beta t, \tag{5.63}$$

$$V[X_t] = \nu\beta t(1 + \beta t). \tag{5.64}$$

For the GIGP distribution with $\gamma = -1/2$,

$$E[X_t] = \frac{\alpha\theta t}{2(1-\theta)^{1/2}}, \tag{5.65}$$

$$V(X_t) = \frac{\alpha\theta t}{4(1-\theta)^{3/2}}[2(1-\theta) + t\theta]. \tag{5.66}$$

For the generalized Waring distribution,

$$E[X_t] = \frac{\psi bt}{a-1}; \quad a > 1, \tag{5.67}$$

$$V(X_t) = \frac{\psi b(a+b-1)}{(a-1)^2(a-2)}(a - 1 + \psi t); \quad a > 2. \tag{5.68}$$

5.6 Probability Model for Research Publications. Yule Process

Probability models are very interesting and powerful tools for the study of the dynamics of research systems and characteristics of research production. Let us demonstrate this with a discussion of a probability model of dynamics of research publications [104] that will lead us to the famous statistical distribution of Yule.

Let us now consider scientific publications from the following point of view. A researcher has x publications. Then he/she writes one more publication, and we shall consider this as a transition to another state characterized by $x + 1$ publications. The

occurrence of a new publication is a rare event, and because of this, we shall consider the process of the occurrence of a new publication to be a Poisson pure multiplicative random process where the probability of transition to a new state in the time interval $(t, t + \Delta t)$ depends on the state of the system at time t.

5.6.1 *Definition, Initial Conditions, and Differential Equations for the Process*

We begin our study at the point in time where a studied researcher has one publication. Let $p_x(t)$ be the probability that a researcher has x publications at time t. Then the initial condition is $p_x(0) = 1$ if $x = 1$ and $p_x(0) = 0$ if $x \neq 1$. The process evolves according to the following two rules:

1. The probability of a transition from state x to state $x + 1$ in the interval $(t, t + \Delta t)$ is proportional to the interval Δt. We denote this probability by $\lambda(x)\Delta t$.
2. The probability of two or more transitions for the interval Δt is negligibly small.

Because of the above rules, the probability of a lack of transition between the states x and $x + 1$ in the time interval $(t, t + \Delta t)$ is $1 - \lambda(x)\Delta t$.

The probability that our system (the researcher) is in the state x (has x publications) for the interval $(t, t + \Delta t)$ is the sum of the probability that the system jumped there from the state $x - 1$ within the time interval and the probability that the system has not jumped to the next state $x + 1$ within the time interval. In symbols, this reads

$$p_x(t + \Delta t) = [1 - \lambda(x)\Delta t]p_x(t) + \lambda(x - 1)p_{x-1}(t)\Delta t. \tag{5.69}$$

This can be written as the following system of differential equations for the probability:

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\lambda_0 p_0(t), \\ \frac{dp_x(t)}{dt} &= -\lambda(x)p_x(t) + \lambda(x - 1)p_{x-1}(t). \end{aligned} \tag{5.70}$$

5.6.2 *How a Yule Process Occurs*

In order to continue analysis of (5.70), we have to determine $\lambda(x)$. We shall use the linear hypothesis for the parameter $\lambda(x)$:

The probability of a transition increases proportionally to the number of publications:

$$\lambda(x) = \lambda x, \tag{5.71}$$

where λ *is a constant.*

In other words, there is a linear hypothesis of the following kind: *If an author has many publications, he/she doesn't need much time to produce another one. In this way, our stochastic process becomes a linear pure multiplicative process* (**Yule process**) [105–109].

Using (5.71), one obtains the following solution of the system of equations (5.70): $p_x(t) = 0$ when $x = 0$ and

$$p_x(t) = [1 - \exp(-\lambda t)]^{x-1} \exp(-\lambda t). \tag{5.72}$$

Let us recall that in the case under discussion, the distribution (5.72) gives the probability that a researcher will have x publications at time t if at time $t = 0$, he had one publication.

5.6.3 Properties of Research Production According to the Model

1. *Expected value.*
   The expected value is the mean number of publications that are expected to be written for time t. Then
   $$E[x(t)] = \exp(\lambda t), \tag{5.73}$$
   which is often observed in practice and is called the **law of exponential growth of science**.
2. λ: *a measure of the publication activity of the researchers.*
   After a "differentiation" of (5.73), one obtains
   $$\lambda = \frac{dx_t/dt}{x_t}, \tag{5.74}$$
   which means that λ is the rate of growth of the number of publications, i.e., a measure of the intensity of publication (and partially of the scientific) activity of a researcher.
3. *Research work in a research area for some finite time.*
   Usually, a researcher works for some (finite) time on problems from some research area and then changes the research area of work (or retires). This time depends

on the potential of the research area, on the talent of the researcher, on the age of the researcher, on the work conditions, etc. The finite time of work is different for different researchers and is a random variable whose distribution can be obtained from queuing theory. The distribution is

$$p(t) = \nu \exp(-\nu t), \tag{5.75}$$

where $\nu = 1/t^*$ and t^* is the average value of t. This random distribution of the time of activity in a research area can be incorporated in the Yule distribution as $p_x(t) = p(x/t)$. Then in order to obtain the probability distribution of the publications that are observed in a database, we have to calculate the following integral:

$$p(x) = \int_0^\infty dt\, p(x/t)p(t) =$$

$$\int_0^\infty dt\, [1 - \exp(-\lambda t)]^{x-1} \exp(-\lambda t)\nu \exp(-\nu t). \tag{5.76}$$

The integration of (5.76) leads to the *Yule distribution*

$$p(x) = \alpha B(x, \alpha + 1), \tag{5.77}$$

where:

- $B(x, \alpha + 1) = \frac{\Gamma(x)\Gamma(\alpha+1)}{\Gamma(x+\alpha+1)}$ is the beta function;
- $\Gamma(x) = (x - 1)!$ is the gamma function;
- $\alpha = \nu/\lambda$.

The Yule distribution obtained above leads to several interesting conclusions about research production.

1. **Asymptotic behavior**: For large x, one obtains $\frac{\Gamma(x)}{\Gamma(x+\alpha+1)} \approx \frac{1}{x^{\alpha+1}}$ (the Stirling approximation was used). Let us in addition assume that α has small values. Then $\Gamma(\alpha + 1) \approx 1$, and the Yule distribution is reduced to

$$p(x) \approx \alpha\Gamma(\alpha + 1)\frac{1}{x^{\alpha+1}} \approx \frac{\alpha}{x^{\alpha+1}}, \tag{5.78}$$

which is the law of Pareto for $x_0 = 1$ and small values of α. Thus on the basis of the hypothesis that the scientific activity is a random branching multiplicative process with linear increase of effectiveness of the researchers (Yule process), we have obtained one of the basic laws of research production.

2. **Evaluation of the parameter** α: This can be done on the basis of the Yule distribution for researchers who have just one publication. For these researchers,

$$p(1) = \frac{\alpha\Gamma(1)\Gamma(\alpha+1)}{\Gamma(\alpha+2)} = \frac{\alpha}{\alpha+1} \tag{5.79}$$

(we have used $\Gamma(1) = 1$ and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$). Then taking into account that $p(1) = N_1/N$ is the proportion of the number N_1 of researchers with one publication in a group of N researchers, we obtain

$$\alpha = \frac{p_1}{1-p_1} = \frac{N_1}{N-N_1}. \tag{5.80}$$

Thus we can evaluate α by taking N and N_1 from a large enough database.

5.7 Probability Models Connected to Dynamics of Citations

5.7.1 Poisson Model of Citations Dynamics of a Set of Articles Published at the Same Time

Citation analysis is one of the frequently used methods of assessment of research impact [110–114]. An important topic in the research on citations is the investigation of citation distributions. This research may follow two paths [115]:

1. *Path 1*: Take a particular source—book, article, journal issue, journal volume, etc.—and study the age distribution of the cited articles in the studied source [116].
2. *Path 2*: Take a collection of sources (articles published in a journal, or articles from some scientific field) at a given time and then follow up and note the times at which each source from the collection is cited [117, 118].

Below, we present a probability model obtained by following *Path 2* and assuming continuous time as well as the presence of aging of published material (in the course of time, the material becomes obsolescent (and less frequently cited)) and the existence of publications that are never cited. The model is as follows [115]. Let us consider a population of sources that produces items over time. The population (for the case of citation analysis) consists of a collection of articles published at the same time $t = 0$. The items produced by the papers are their citations. The assumption is that citations are received randomly over time. Since different articles are in different scientific areas (with different popularity) and have different relevance, etc., their citation rates are also different. We assume that these rates of a randomly chosen source are characterized by a random variable Λ that has probability distribution F_Λ over the population of sources. Let X_t be the number of citations to a *randomly chosen*

source (article) in the interval $[0, t]$. The probability that this number of citations will be equal to r is

$$p(X_t = r) = \int_0^\infty dF_\Lambda(\lambda^*)\, P(X_t = r \mid \Lambda = \lambda^*). \tag{5.81}$$

We can recognize the process $\{X_t, t \geq 0\}$ as a counting process, and the model (5.81) is a mixture of counting processes with mixing distribution F_Λ and mixing parameter λ. Next, one has to assume the nature of the process connected to the conditional term $P(X_t = r \mid \Lambda = \lambda^*)$. The initial assumption can be that this process is a Poisson process [119–122] with stationary and independent increments. This will lead us to the distribution

$$P(X_t = r \mid \Lambda = \lambda^*) = \exp(-\lambda^* t)\frac{(\lambda^* t)^r}{r!}; \quad r = 0, 1, 2, \ldots. \tag{5.82}$$

In (5.82), $\lambda^* = \text{const}$, and the mean of the Poisson distribution is $\lambda^* t$. We note here that numerous models of citation distribution have been proposed based on different probability distribution functions $f(\lambda^*)$, $(dF_\Lambda(\lambda^*) = f(\lambda^*)d\lambda^*)$ [123].

Let us now consider the case in which λ^* depends on time. Since λ^* can be associated with the citation rate of a given paper, it can vary with the time t. If $\lambda^* = \lambda^*(t)$, then (5.82) has to be substituted by the more complicated equation [124]

$$P(X_t = r \mid \Lambda = \lambda^*) = \exp[-M(\lambda^*, t)]\frac{M(\lambda^*, t)^r}{r!}; \quad r = 0, 1, 2, \ldots, \tag{5.83}$$

where

$$M(\lambda^*, t) = \int_0^t ds \lambda^*(s).$$

In the case of citations of articles, an almost universal citation pattern in time $c(t)$ can be observed. Then we can assume that the citation rate $\lambda^*(t)$ of a paper has the particular form

$$\lambda^*(t) = \lambda c(t), \tag{5.84}$$

where $\lambda = \text{const}$. Then

$$M(\lambda^*, t) = \int_0^t ds\, \lambda c(s) = \lambda C(t); \quad C(t) = \int_0^t ds\, c(s) \tag{5.85}$$

and

$$P(X_t = r \mid \Lambda = \lambda^*) = \exp[-\lambda C(t)]\frac{[(\lambda C(t)]^r}{r!}; \quad r = 0, 1, 2, \ldots. \tag{5.86}$$

The mean of the Poisson process is $\lambda C(t)$; $c(t)$ is called the *obsolescence density function*; and $C(t)$ is called the *obsolescence distribution function* ($t > 0$). We assume that $\lim\limits_{t\to\infty} C(t) < \infty$.

The substitution of (5.86) in (5.81) leads to the final relationship for the citation production distribution:

$$p(X_t = r) = \int\limits_0^\infty dF_\Lambda(\lambda)[\lambda C(t)]^r \left[\frac{\exp[-\lambda C(t)]}{r!}\right], \; r = 0, 1, 2, \ldots. \tag{5.87}$$

This can also be written as the expected value

$$p(X_t = r) = E_\Lambda[P(X_t = r \mid \Lambda)]. \tag{5.88}$$

From (5.87), one can obtain the first citation distribution. Let T be the time after publication of the first citation of a randomly chosen source (article). We can consider T a random variable. For times $t < T$, the number of citations of a paper is 0. Then let $F_T(t)$ be the cumulative distribution function of the first citation time: $F_T(t) = p(T \le t)$. Since $p(T \le t) = 1 - p(T > t)$ and $p(T > t)$ is the same as the probability $p(X_t = 0)$, we have

$$F_T(t) = 1 - p(X_t = 0) = 1 - \int\limits_0^\infty dF_\Lambda \exp[-\lambda C(t)]. \tag{5.89}$$

An interesting consequence obtained on the basis of the first citation distribution (5.89) is as follows.

There will be publications that will be never cited.

This feature follows from the relationship $\lim\limits_{t\to\infty} F_T(t) < 1$. Indeed, we can see that

$$\int\limits_0^\infty dF_\Lambda \exp[-\lambda C(t)] = L_\Lambda[C(t)]$$

is the Laplace transformation of Λ, which has the property $L_\Lambda(1) > 0$. Then

$$\lim_{t\to\infty} F_T(t) = 1 - \lim_{t\to\infty} p(X_t = 0) = 1 - \lim_{t\to\infty} L_\Lambda(C(t)) = 1 - L_\Lambda(1) < 1.$$

The model developed above can be used for obtaining the nth citation distribution [125]. The result for the nth citation distribution is

$$F_n(t) = p(T_n < t) = \int_0^{C(t)} ds \, \frac{s^{n-1}}{(n-1)!} E_\Lambda[\Lambda^n \exp(-\Lambda s)]; \, t < \infty, \tag{5.90}$$

$$p(T_n = \infty) = \int_1^{\infty} ds \, \frac{s^{n-1}}{(n-1)!} E_\Lambda[\Lambda^n \exp(-\Lambda s)]. \tag{5.91}$$

5.7.2 *Mixed Poisson Model of Papers Published in a Journal Volume*

The accumulation of citations has varying dynamic behavior over the lifetime of a paper, and among other things, this behavior is also influenced by the reputation of the journal in which the paper was published. In most cases, immediately after publication, the number of citations grows slowly, usually because it may take some time for citing papers to appear in print and to be entered in the citations databases. After this initial period, citations increase faster as citations lead to new readers who may also cite the publication. Finally, the material of the paper becomes outdated and/or obsolete. Then the number of citations per year decreases. This is the typical behavior, but there exist other patterns of behavior such as "sleeping beauties," "shooting stars," etc. [126, 127].

The investigation of citation behavior in journal volumes can be based on the mixed Poisson distribution [128–131] model of Burrell [115, 125]. A journal volume can be treated as a collections of paper, usually from the same years and with common characteristics. The main assumption is that each paper generates citations at a constant (latent) rate (λ) following the Poisson distribution but that these rates vary across the collection as a random variable Λ. Then the probability that a paper will generate r citations at time t is

$$p(Z_t = r \mid \Lambda = \lambda) = \exp(-\lambda t) \frac{(\lambda t)^r}{r!}. \tag{5.92}$$

The population distribution of randomly chosen papers of unknown λ will be a mixture of the Poisson distributions of the kind (5.92),

$$p(X_t = r \mid \Lambda) = \int_0^{\infty} dF(\lambda) \exp(-\lambda t) \frac{(\lambda t)^r}{r!}, \tag{5.93}$$

where $F_\Lambda(\lambda)$ is the cumulative distribution of λ (of the latent rate), also called the mixing distribution.

There are different possibilities for the form of mixing distribution [132–134], but the most widely used distribution is the gamma distribution of shape parameter ν and size α:

$$\frac{d}{d\lambda}F_\Lambda(\lambda) = \exp(-\alpha\lambda).\frac{\alpha^\nu \lambda^{\nu-1}}{\Gamma(\nu)} \tag{5.94}$$

The appearance of the gamma distribution above is not a coincidence. The gamma mixture of Poisson distributions follows a negative binomial distribution [135–137] (a fact proved by Greenwood and Yule [138]). Yule is the same scientist who first described the preferential attachment process (Yule process). This negative binomial distribution is

$$P(X_t = r) = \binom{r+\nu-1}{\nu-1}\left(\frac{\alpha}{\alpha+t}\right)^\nu\left(1-\frac{\alpha}{\alpha+t}\right)^r, \quad r = 0, 1, 2, \ldots. \tag{5.95}$$

In most cases, citations of a paper do not occur at constant intervals (evenly) in time. Thus in most cases, λ is not a constant. The rate $\lambda(t)$ will be different for different papers. It can be assumed [115] that $\lambda(t)$ may be written in the form

$$\lambda(t) = \lambda c(t), \tag{5.96}$$

where $c(t)$ describes some pattern of citation behavior that is the same for all articles from the discussed collection of articles (i.e., $c(t)$ describes a sort of obsolescence). The function $c(t)$ is the probability density function of obsolescence, and $C(t)$ is the cumulative distribution function of obsolescence.

With the obsolescence distribution, the model discussed above leads to the following negative binomial distribution for the probability that a paper in a collection of papers will have r citations [139]:

$$p(X_r = r) = \binom{r+\nu-1}{\nu-1}\left(\frac{\alpha}{\alpha+C(t)}\right)^\nu\left(1-\frac{\alpha}{\alpha+C(t)}\right)^r, \quad r = 0, 1, 2, \ldots. \tag{5.97}$$

Many assumptions can be made about the form of $C(t)$. Two possibilities are as follows:

- Logistic function: $C(t) = 1/(1 + a\exp(-bt))$;
- Weibull distribution: $C(t) = 1 - \exp[-(t/b)^2]$.

The values of $C(t)$ can be determined by fitting citation data. Additional information about the investigation of citations in several research disciplines can be found in [140], where a Poisson distribution and an exponential distribution are used for describing such data.

5.8 Aging of Scientific Information

As a consequence of the continuous research efforts of scientists, a continuous flow of new scientific information exists, and existing scientific information ages. As a consequence of these two processes, there is a continuous reorganization of the structure of scientific information. For example, suppose a scientist publishes an article. At first, interest in the article may be significant (a large number of citations, for example). Then interest decreases as the information in the article ages and the scientific potential of the obtained results decreases. If one studies closely the number of citations of a publication, three periods can usually be distinguished:

1. **First two years after publication**: with rare exceptions, articles are not cited much in this period (they are not very well known to the corresponding scientific community). The exceptions are extremely important, however: if an article is very much cited within this initial period, it is highly probable that it will become a very influential publication that may contribute much to the development of the corresponding scientific field.
2. **Next five years**: here the publication achieves most of its citations as it becomes well known. If there are no citations, the publication has been judged by the corresponding scientific community to be of little use. This judgment is valid in the general case, but there can be rare exceptions: "sleeping beauties" that suddenly become current many years after publication [141].
3. **More than seven years after publication**: the number of citations usually begins to decrease, and the publication slowly moves toward the scientific archives.

The above considerations show that by their continuous work in obtaining new knowledge, researchers continuously renew the structure of scientific information by opening a place for the new information and compressing the aged information (this information, compressed to citations, arises in some of the new publications). In this process, researchers mainly use the achievements of the previous generation of researchers.

5.8.1 Death Stochastic Process Model of Aging of Scientific Information

The main assumptions of the model are as follows [142]:

1. At the initial moment of the study, there is some portion of the scientific publications that are cited. The number of citations of these publications $x(t)$ decreases with advancing time. The number of citations at $t = 0$ is x_0.

2. The probability that in the interval $(t, t + \Delta t)$ there will be $x - 1$ citations if in the previous interval the number of citations was x is $\mu_x \Delta t$. Thus the probability that the number of citations will not decrease is $1 - \mu_x \Delta t$.

Then the probability that at the moment $t + \Delta t$ there will be x citations of the scientific publications is

$$p_x(t + \Delta t) = (1 - \mu_x \Delta t) p_x(t) + \mu_{x+1} p_{x+1}(t) \Delta t. \tag{5.98}$$

On the basis of the assumption that the intensity with which citations are decreasing is proportional to the number of citations, $\mu_x = \mu x$, and imposing the initial condition $p_x(0) = 1$ for $x = x_0 \geq 1$ and $p_x(0) = 0$ for $x \neq x_0$, one obtains the following solution of (5.98):

$$p_x(t) = \frac{x_0!}{x!(x_0 - x)!} \exp(-\mu x_0 t)[\exp(\mu t) - 1]^{x - x_0} \tag{5.99}$$

for $0 \leq x \leq x_0$. From (5.99), the average number of citations with advancing time is

$$x_t = x_0 \exp(-\mu t), \tag{5.100}$$

which means that

there occurs an aging of scientific information according to an exponential law. This is a rapid pace of aging, and significant scientific efforts are needed in order to compensate it by production of new scientific information.

5.8.2 *Inhomogeneous Birth Process Model of Aging of Scientific Information. Waring Distribution*

Another approach to the aging of scientific information was proposed by Schubert and Glänzel [143] and discussed by Schubert, Glänzel, and Schoepflin [4, 144]. As we shall see below, the model of Schubert and Glänzel is quite interesting, because it is a deterministic one, yet it is connected to the (not much known but very interesting) Waring distribution. We shall see in addition that this model (that can be connected to an inhomogeneous birth process) leads to the same results as the model discussed above that is based on a death process. And the Waring distribution will be of great interest to us, since it is a generalization of several important statistical distributions appearing in the area of research on science dynamics and research production. Below, we describe a simple model that leads to the Waring distribution. Then we

consider a particular case of a stochastic process connected to repetitive events, and finally, we shall consider a particular class of the process with repetitive events (such as publishing papers and obtaining citations), and we shall consider the aging of scientific information (scientific articles) from the point of view of obtained citations.

5.8.2.1 Waring Distribution

The Waring distribution is a distribution with a very long tail. Because of this property, the Waring distribution is quite suitable for describing characteristics of many systems from the areas connected to research on biology and society. We shall see below that the Waring distribution is connected to other interesting distributions that are presented in this book: the Yule and Zipf distributions.

The Waring distribution may be connected to publication activity, and publication activity may be considered a measure of research productivity. Within the context of the epidemic model of Goffman and Newill (discussed above), the susceptible and infected persons have to be continuously replaced by persons entering the system, i.e., the population of researchers should be considered an open population. As we shall see below, the model by Schubert and Glänzel [143] describes similar processes connected to publication activity. The model assumes three groups in the population: a group that is entering the system, a group that is in the system, and a group that is leaving the system. In more detail, we consider an infinite array of cells (boxes) indexed in succession by nonnegative integers. The amount x of some substance can move between the cells. Let x_i be the amount of the substance in the ith cell. Then

$$x = \sum_{i=0}^{\infty} x_i. \tag{5.101}$$

The fractions $y_i = x_i/x$ can be considered probability values of a distribution of a discrete random variable ζ:

$$y_i = p(\zeta = i),\ i = 0, 1, \ldots. \tag{5.102}$$

We assume that the expected value of the random variable ζ is finite and that the content x_i of any cell can change under any of the following three processes:

1. Some amount s of the substance x may enter the system of cells from the external environment through the 0th cell.
2. The rate f_i of the substance x can be transferred from the ith cell into the $(i + 1)$th cell;
3. The rate g_i from the substance x may leak out of the ith cell into the external environment.

The stochastic process connected to the movement of the substance between the cells is formed by a change in the content of the cells, e.g., by a change of papers

published by authors who have entered the system. In this case, $x(t)$ is the (random) number of published papers, and $p(x(t) = i) = y_i$ the probability that an author in the system has published i papers in the period t. The stochastic model is obtained if $x(t)$ is considered the publication activity process of an *arbitrary* author, and $p(x(t) = i) = y_i$ is the probability that this author has published i papers in the time interval between 0 and t.

The three processes mentioned above can be modeled mathematically by a system of ordinary differential equations:

$$\begin{aligned} \frac{dx_0}{dt} &= s - f_0 - g_0; \\ \frac{dx_i}{dt} &= f_{i-1} - f_i - g_i. \end{aligned} \tag{5.103}$$

The following forms of the relationships for the amount of the moving substances are assumed in [143] ($\alpha, \beta, \gamma, \sigma$ are constants):

$$\begin{aligned} &s = \sigma x; \quad \sigma > 0 \rightarrow \text{self-reproducing property}, \\ &f_i = (\alpha + \beta i)x_i; \quad \alpha > 0, \ \beta \geq 0 \rightarrow \text{cumulative advantage of higher cells}, \\ &g_i = \gamma x_i; \quad \gamma \geq 0 \rightarrow \text{uniform leakage over the cells}. \end{aligned} \tag{5.104}$$

Substitution of (5.104) in (5.103) leads to the relationships

$$\begin{aligned} \frac{dx_0}{dt} &= \sigma x - \alpha x_0 - \gamma x_0; \\ \frac{dx_i}{dt} &= [\alpha + \beta(i-1)]x_{i-1} - (\alpha + \beta i + \gamma)x_i. \end{aligned} \tag{5.105}$$

Let us sum the equations from (5.105). The result of the summation is

$$\frac{dx}{dt} = (\sigma - \gamma)x, \tag{5.106}$$

and the solution for x is

$$x = x(0)\exp[(\sigma - \gamma)t], \tag{5.107}$$

where $x(0)$ is the amount of x at $t = 0$. Three regimes of change of $x(t)$ follow from (5.107):

1. Regime of exponential growth ($\sigma > \gamma$).
2. Stationary regime ($\sigma = \gamma$).
3. Regime of exponential decay ($\sigma < \gamma$).

The distribution of y_i will lead us to the Waring distribution. From (5.105) and with the help of (5.107) and the relationship $\frac{dy_i}{dt} = \frac{1}{x^2}\left[x\frac{dx_i}{dt} - x_i\frac{dx}{dt}\right]$, one obtains

$$\frac{dy_0}{dt} = \sigma - (\alpha + \sigma)y_0;$$
$$\frac{dy_i}{dt} = [\alpha + \beta(i-1)]y_{i-1} - (\alpha + \beta i + \sigma)y_i. \tag{5.108}$$

The solution of (5.108) is

$$y_i = y_i^* + \sum_{j=0}^{i} b_{ij} \exp[-(\alpha + \beta j + \sigma)t], \tag{5.109}$$

where y_i^* is the stationary solution of (5.109) given by the relationships

$$y_0^* = \frac{\sigma}{\sigma + a},$$
$$y_i^* = \frac{\alpha + \beta(i-1)}{\alpha + \beta i + \sigma} y_{i-1}^*, \ i = 1, 2, \ldots. \tag{5.110}$$

The coefficients b_{ij} are determined by the initial conditions. In the exponential function there are no negative coefficients, and because of this, when $t \to \infty$, the sum in (5.109) vanishes and the system comes to the stationary distribution from (5.110). Thus the distribution of y_i tends to be stationary despite the fact that the system is in a stationary state only when $\sigma = \gamma$.

Thus starting from any initial distribution, after some time, the system reaches the steady state, where the content of each cell decays exponentially with (the same) characteristic time $\frac{1}{\sigma - \gamma}$ and the distribution of the substance among the cells is given by (5.110).

This distribution is called the Waring distribution.

The form of the Waring distribution is

$$P(\zeta = i) = \frac{ak^{[i]}}{(a+k)^{[i+1]}}; \quad k^{[i]} = \frac{(k+i)!}{k!}, \tag{5.111}$$

with parameters $k = \alpha/\beta$ and $a = \sigma/\beta$.

We note that the words "after some time" above mean that the Waring distribution can be considered a good approximation of the considered process for large enough finite times when the stationary state of distribution of substance among the cells has almost been reached.

5.8.2.2 Parameters and Particular Cases of the Waring Distribution

The Waring distribution is quite interesting, since it contains as particular cases the distributions of Yule and Zipf.

Let $a > 2$. The expected value of the Waring distribution is

$$E[\zeta] = \frac{k}{a-1}; \quad a > 1. \tag{5.112}$$

We note that $a > 1$ is a condition for a finite expected value (such a finite value was assumed above). Then from the definition of a, it follows that $\sigma > \beta$.

The variance of the Waring distribution is

$$D^2[\zeta] = \frac{ka(k+a-1)}{(a-1)^2(a-2)}; \quad a > 2. \tag{5.113}$$

Several special cases of the Waring distribution are

1. $\beta = 0$ (*geometric distribution*).
   In this case (called also the model of Frank and Coleman [145, 146] or case with absence of cumulative advantage because of $f_i = \alpha x_i$),

$$P(\zeta = i) = q(1-q)^i; \quad q = \frac{\sigma}{\sigma + a}. \tag{5.114}$$

2. $k = 0, \alpha = 0, \beta \neq 0$ (*Yule distribution*).
   Let then $k \to 0$. The Waring distribution reduces to the Yule distribution [147],

$$P(\zeta = i \mid \zeta > 0) = aB(a+1, i), \tag{5.115}$$

   where B is the beta function. Let us note that in this case, $f_i = \beta i x_i$, which is known also as Gibrath law, much used in economics for describing size distributions of business systems [148] or size distributions of cities [149].
3. $i \to \infty$ (*Zipf distribution*).
   As $i \to \infty$, the Waring distribution becomes

$$P(\zeta = i) \to \frac{c}{i^{(1+a)}}, \tag{5.116}$$

   which is the frequency form of the Zipf distribution (c is an appropriate constant depending on the parameters of the distribution).

5.8.2.3 Truncated Waring Distribution

For some applications, one may need a model with a finite number of cells. In this case, we consider an array of $N+1$ cells (boxes) indexed in succession by nonnegative integers, i.e., the first cell has index 0, and the last cell has index N. We assume that there exists an amount x of some substance that is distributed among the cells. Let x_i be the amount of the substance in the ith cell. Then

$$x = \sum_{i=0}^{N} x_i. \tag{5.117}$$

The fractions $y_i = x_i/x$ can be considered probability values of the distribution of a discrete random variable ζ,

$$y_i = p(\zeta = i),\ i = 0, 1, \ldots, N. \tag{5.118}$$

The process of transfer of substance between the cells can be modeled mathematically by a system of ordinary differential equations:

$$\begin{aligned}
\frac{dx_0}{dt} &= s - f_0 - g_0;\\
\frac{dx_i}{dt} &= f_{i-1} - f_i - g_i,\ i = 1, 2, \ldots, N-1;\\
\frac{dx_N}{dt} &= f_{N-1} - g_N.
\end{aligned} \tag{5.119}$$

The forms of the amounts of the moving substances are the same as in (5.104). The substitution of (5.104) in (5.119) leads to the relationships

$$\begin{aligned}
\frac{dx_0}{dt} &= \sigma x - \alpha x_0 - \gamma x_0;\\
\frac{dx_i}{dt} &= [\alpha + \beta(i-1)]x_{i-1} - (\alpha + \beta i + \gamma)x_i,\ i = 1, 2, \ldots, N-1,\\
\frac{dx_N}{dt} &= [\alpha + \beta(N-1)]x_{N-1} - \gamma x_N.
\end{aligned} \tag{5.120}$$

Let us now derive the distribution of y_i. From (5.120), we obtain

$$\begin{aligned}
\frac{dy_0}{dt} &= \sigma - (\alpha + \sigma)y_0;\\
\frac{dy_i}{dt} &= [\alpha + \beta(i-1)]y_{i-1} - (\alpha + \beta i + \sigma)y_i,\ i = 1, 2, \ldots, N-1;\\
\frac{dy_N}{dt} &= [\alpha + \beta(N-1)]y_{N-1} - \sigma y_N.
\end{aligned} \tag{5.121}$$

We search for a solution of (5.121) in the form

$$y_i = y_i^* + F_i(t), \tag{5.122}$$

where y_i^* is the stationary solution of (5.122) given by the relationships

$$\begin{aligned} y_0^* &= \frac{\sigma}{\sigma+\alpha}; \\ y_i^* &= \frac{\alpha+\beta(i-1)}{\alpha+\beta i+\sigma} y_{i-1}^*,\ i = 1, 2, \ldots, N-1; \\ y_N^* &= \frac{\alpha+\beta(N-1)}{\sigma} y_{N-1}^*. \end{aligned} \tag{5.123}$$

For the functions F_i, we obtain the system of equations

$$\begin{aligned} \frac{dF_0}{dt} &= -(\alpha+\sigma)F_0; \\ \frac{dF_i}{dt} &= [\alpha+\beta(i-1)]F_{i-1} - (\alpha+\beta i+\sigma)F_i,\ i = 1, 2, \ldots, N-1, \\ \frac{dF_N}{dt} &= [\alpha+\beta(N-1)]F_{N-1} - \sigma F_N. \end{aligned} \tag{5.124}$$

The solutions of these equations are

$$F_0(t) = b_{00} \exp[-(\alpha+\sigma)t], \tag{5.125}$$

$$F_1(t) = b_{10} \exp[-(\alpha+\sigma)t] + b_{11} \exp[-(\alpha+\beta+\sigma)t], \tag{5.126}$$

$$\ldots$$

$$F_i(t) = \sum_{j=0}^{i} b_{ij} \exp[-(\alpha+\beta j+\sigma)t];\ i = 1, 2, \ldots, N-1, \tag{5.127}$$

$$F_N(t) = \sum_{j=0}^{N} b_{Nj} \exp[-(\alpha+\beta j+\sigma)t], \tag{5.128}$$

where

$$\begin{aligned} b_{ij} &= \frac{\alpha+\beta(i-1)}{\beta(i-j)} b_{i-1,j};\ i = 1, \ldots, N-1;\ j = 0, \ldots, i-1; \\ b_{Nj} &= -\frac{\alpha+\beta(N-1)}{\alpha+j\beta} b_{N-1,j},\ j = 0, \ldots, N-1; \\ b_{NN} &= 0. \end{aligned} \tag{5.129}$$

The b_{ij} that are not determined by (5.129) may be determined by the initial conditions. In the exponential function in $F_i(t)$ there are no negative coefficients, and because of this, as $t \to \infty$, we have $F_i(t) \to 0$, and the system comes to the stationary distribution from (5.123). The form of this stationary distribution is

$$P(\zeta = i) = \frac{a}{a+k}\frac{(k-1)^{[i]}}{(a+k)^{[i]}};\ \ k^{[i]} = \frac{(k+i)!}{k!};\ \ i = 0, \ldots, N-1,$$
$$P(\zeta = N) = \frac{1}{a+k}\frac{(k-1)^{[N]}}{(a+k)^{[N-1]}}, \tag{5.130}$$

with parameters $k = \alpha/\beta$ and $a = \sigma/\beta$.

The obtained distribution is called the truncated Waring distribution. The distribution (5.130) has a concentration of substance in the last cell (i.e., in the Nth cell). For the case of the nontruncated Waring distribution, the same substance is distributed in the cells $N, N+1, \ldots$.

5.8.2.4 A Nonstationary Birth Process. Negative Binomial Distribution, Papers, and Citations

Let us consider the nontruncated version of the Waring distribution. In addition, let us assume that the system is completely isolated from external influences. This means that no substance enters or leaves the system. Thus the amounts of the moving substances are

$$\sigma = 0;\ g_i = 0;\ f_i = (\alpha + \beta i)x_i;\ \ \frac{\alpha(t)}{\beta(t)} = N > 0. \tag{5.131}$$

The last of the above relationships shows that the process is nonstationary (since the substance flow can depend on time). The governing equations become

$$\frac{dy_0}{dt} = -\beta(t)Ny_0;$$
$$\frac{dy_i}{dt} = \beta(t)[(N+i-1)y_{i-1} - (N+1)y_i]; \tag{5.132}$$

with initial conditions $y_i(0) = 1$ if $i = 0$ and $y_i(0) = 0$ otherwise. What one needs is to obtain the distribution $y_i = p(x(t) = i)$ connected to the process. We recall that $p(x(t) = i)$ is the probability that an author in a system has published i papers in the period t. This distribution can be obtained from (5.132), and its form is very similar to the form of the distribution obtained on the basis of the model of death process above [4]:

$$p(x(t) = k) = \binom{N+k-1}{k} \exp[-N\rho(t)]\{1 - \exp[-\rho(t)]\}^k, \tag{5.133}$$

where $\rho(t) = \int_0^t d\tau\ \beta(\tau)$. Equation (5.133) is the relationship for the *negative binomial distribution*. In addition to the probability $p(x(t))$, one can define also transition probabilities $p_{i,k}(s,t)$ for the probability that at time t, the substance is in the kth unit if at time $s < t$ it was in the ith unit. From the point of view of the case with scientists and articles, $p_{ik}(s,t)$ is the probability that an author will own k articles at time t if at time s he/she owns $i \leq k$ articles. In this case, the evolution of the transition probability [144] is given by

$$\frac{\partial p_{i,k}(s,t)}{\partial t} = \beta(t)[(N+k-1)p_{i,k-1}(s,t) - (N+k)p_{i,k}(s,t)], \tag{5.134}$$

with initial conditions $p_{i,k}(s,s) = 1$ if $k = i$ and $p_{i,k}(s,s) = 1$ otherwise.

Citations are repetitive events exactly like papers. Thus all discussions about the nonstationary birth process connected to papers are the same for the nonstationary birth process connected to citations. In the first case, we have a scientist who publishes papers. In the second case, we have a paper that receives citations. Then (5.133) gives the probability that a paper will have received k citations at time t, and (5.134) gives the transitional probability that a paper will have received k citation at time t if it has i citations at the time s. The distribution connected to the transitional probability $p_{i,k}$ is also a negative binomial distribution. In more detail, the number of received citations for the time $t - s$ when the number of received citations at until time s was i, $p_{i,j}(s,t) = p[x(t) - x(s) = j \mid x(s) = i]$, is

$$p_{i,j}(s,t) = \binom{N+i+j-1}{j} \exp\{-[\rho(t) - \rho(s)](N+i)\}(1 - \exp\{-[\rho(t) - \rho(s)]\})^j, \tag{5.135}$$

i.e., the substance flow during the time period $t - s$ has a negative binomial distribution with parameters $\exp[-r(t) + r(s)]$ and $N + j$, where j is the index of the unit that was reached by the substance at time s [143, 144, 150].

With respect to the aging of scientific information, it is important to study the mean value function $M_i(s,t)$. It will show us that a paper that has received some number of citations during the time s after its publication is expected to receive (during an arbitrary time period $t - s$ after the moment s) a linear expression in what it had received previously:

$$M_i(s,t) = E[x(t) - x(s) \mid x(s) = i] = (N+i)\{\exp[\rho(t) - \rho(s)] - 1\} = c_s(t)i + d_s(t). \tag{5.136}$$

We note that $\frac{d_s(t)}{c_s(t)} = N = \text{const}$ is independent of time, and $c_s(t)$ is a characteristic of the aging process. Large $c_s(t)$ characterizes slowly aging literature.

Let us define

$$M(s,t) = E[x(t) - x(s)] = N \exp[\rho(t) - \rho(s)] \tag{5.137}$$

and

$$q(s,t) = \frac{E[x(s) + N]}{E[x(t) + N]}. \tag{5.138}$$

Then (5.135) can be written as

$$p_{i,j}(s,t) = \binom{N+i+j-1}{j} q(s,t)^{N+i}[1 - q(s,t)]^j, \tag{5.139}$$

and the expected citation rate during the time period $t - s$ under the condition that the corresponding paper has received i citations during the time span s is

$$M_i(s,t) = (N+i)\frac{E[x(t) - x(s)]}{E[x(s)] + N}. \tag{5.140}$$

Finally, from (5.139), one obtains that the probability that an article that has received $i \geq 0$ citations will no longer be cited is

$$p_{i,0}(s,t) = p[x(t) - x(s) = 0 \mid x(s) = i] = q(s,t)^{N+i}. \tag{5.141}$$

The lifetime distribution of a process $\{X(t)\}$ is defined by

$$F(t) = \frac{M(0,t)}{M(0,\infty)}, \quad t \geq 0. \tag{5.142}$$

Let us choose the following particular form of f_i [151]:

$$f_i = (N+i)\alpha^*\beta^* \exp(-\alpha^* t)x_i = \beta^* N(1 + i/N)\alpha^* \exp(-\alpha^* t),\ N > 0,\ \alpha^* > 0, \beta^* > 0. \tag{5.143}$$

The time-invariant part of f_i is proportional to $1 = i/N$, and because of this, increases by transfer from the ith cell to the $(i+1)$th cell (which can be considered a local form reflection of the cumulative advantage principle). The time-dependent component of f_i reflects the local exponential aging of the process (aging of the content relative to an individual unit). Then

$$M(s,t) = N\{\exp[\beta^*(1 - \exp(-\alpha^* t)) - \exp[\beta^*(1 - \exp(\alpha^* s))]]\} \tag{5.144}$$

and

$$F(t) = \frac{\exp[\beta^*(1 - \exp[-\alpha^* t])] - 1}{\exp(\beta^*) - 1}. \tag{5.145}$$

Finally, let us discuss the particular cases in which the model describes articles that obtain citations. One can define the *obsolescence function* $H(s)$: the probability that a paper will not be cited beyond a given time s. The definition is

$$H(s) = p(x(\infty) - x(s) = 0). \tag{5.146}$$

The obsolescence function for our particular case is

$$H(s) = \{1 + \exp(\beta^*) - \exp[\beta^*(1 - \exp(-\alpha^* s))]\}^{-N}. \tag{5.147}$$

We note that $H(\infty) = 1$, i.e., at infinity, every publication is obsolete. We have $H(0) = \exp(-\beta^* N)$, i.e., the probability that a paper is already obsolete at the moment it is published equals the probability that it will never be cited.

5.8.2.5 A Case of Brain Drain: Migration Channel for Research Personnel

Let us now discuss one application of the truncated Waring distribution. We consider a sequence of $N + 1$ countries that form a channel. As a result of a large migration movement, a flow of researchers moves through this channel from the country of entrance to the final destination country that is attractive to them in terms of good conditions for life and work. We may assume a situation of war in some region and motion of a large group of researchers from that region to another (more attractive region). The motion starts from an entry country, and the researchers have to move through a sequence of countries in order to reach a (very attractive from the point of view of the researchers) final destination country. We may think about the sequence of countries as a sequence of boxes (cells). The entry country will be the box with label 0, and the final destination country will be the box with label N. Let us consider a number x of researchers that have entered the channel and are distributed among the countries. Let x_i be the number of researchers in the ith country. This number can change on the basis of the following three processes: (a) A number s of researchers enter the channel from the external environment through the country of entrance (0th cell); (b) A number f_i of researchers move from the ith country to the $(i + 1)$th country; (c) A number g_i of the researchers change their status (e.g., they do not move farther in the direction of the final destination country and they are no longer active in the field of research). Fot the case of a large number of migrating researchers, the values of x_i can be determined by (5.103). The relationships (5.104) mean that (a) the number of researchers s that enter the channel is proportional to the number of researchers in all countries that form the channel; (b) there may be a preference for some countries, e.g., migrants may prefer the countries that are around the end of the migration channel (and the final destination country may be the most preferred one); (c) it is assumed that the conditions along the channel are the same with respect to "leakage" of researchers, e.g., the same proportion γ of researchers move out of the area of research work in every country of the channel.

As can be seen from (5.107), the change in the number of researchers depends on the values of σ and γ. If $\sigma > \gamma$, the number of researchers in the channel increases exponentially. If $\sigma < \gamma$, the number of researchers in the channel decreases exponentially. The dynamics of the distribution of the researchers in the channel is modeled by (5.108). When the time since the beginning of the operation of the channel become large enough, the distribution of the researchers in the countries that form the migration channel becomes close to the stationary distribution described by (5.110). Let us stress that the stationary distribution described by (5.110) is very similar to the Waring distribution, but there is a significant difference between the two distributions due to the finite length of the migration channel: there may be a large concentration of researchers in the final destination country especially, if this country is very attractive for researchers.

The parameters that govern the distribution of researchers in the countries that form the channels are σ, α, β, and γ. The parameter σ is the "gate" parameter, since it regulates the number of researchers that enter the channel. If σ is large, then the number of researchers in the channel may increase very rapidly, and this can lead to problems in the corresponding countries. We note that σ participates in each term of the truncated Waring distribution. This means that the situation at the entrance of the migration channel influences significantly the distribution of researchers in the countries of the channel.

The parameter γ regulates the "absorption" of the channel, since it regulates the change of the status of some researchers. They may settle in the corresponding country and may accept a job that is out of the area connected to their research. A large value of γ may compensate for the value of σ and may even lead to a decrease in the number of researchers in the channel. The parameter α regulates the motion of the researchers from one country to the next country of the channel. A small value of α means that the researchers tend to concentrate in the entry country (and eventually in the second country of the channel). An increase in α leads to an increase in the proportion of researchers that reach the second half of the migration channel and especially the final destination country.

The parameter β regulates the attractiveness of the countries along the channel. Large values of β mean that the final destination country is very attractive to researchers (e.g., has excellent conditions for work and the salaries are large). This increases the attractiveness of the countries in the second half of the channel (researchers are more desirous of reaching these countries because the distance to the final destination country is thereby decreased). If for some reason β is kept at a high value, then almost all the researchers may settle in the final destination country.

5.8.2.6 Multivariate Waring Distribution

One can define the multivariate Waring distribution as follows [152]. Let a and b be positive real numbers. Let $a^{(k)} = \frac{\Gamma(a+b)}{\Gamma(a)}$, where $\Gamma(x) = \int\limits_0^\infty dt\ \exp(-t)t^{x-1}$ is the

gamma function [153]. Let $p(x_1 = k_1, \ldots, x_n = k_n; a, b_1, \ldots, b_n)$ be the probability that $x_1 = k_1, \ldots, x_n = k_n$ with parameters $a, b_1, \ldots, b_n$. The multivariate Waring distribution is given by the relationship

$$p(x_1 = k_1, \ldots, x_n = k_n; a, b_1, \ldots, b_n) = a \frac{\Gamma\left(\sum_{i=1}^{n} k_i - n + 1\right) \Gamma\left(\sum_{i=1}^{n} b_i + a\right)}{\Gamma\left(\sum_{i=1}^{n} k_i + \sum_{i=1}^{n} b_i - n + a + 1\right)} \prod_{i=1}^{n} \frac{\Gamma(k_i + b_i - 1)}{\Gamma(k_i)\Gamma(b_i)}, \tag{5.148}$$

where $k_i = 1, 2, \ldots$ and $i = 1, \ldots, n$, a and b_i are positive real numbers. For $n = 1$, the multivariate Waring distribution is reduced to the univariate Waring distribution

$$p(x = k; a, b) = a \frac{\Gamma(b + k + 1)\Gamma(a + b)}{\Gamma(b)\Gamma(a + b + k)}. \tag{5.149}$$

Let $a^{(b)} = \frac{\Gamma(a+b)}{\Gamma(a)}$. Then the univariate form of the Waring distribution can be written as

$$p(x = k; a, b) = a \frac{b^{(k-1)}}{(a + b)^{(k)}}. \tag{5.150}$$

Two interesting properties of the multivariate Waring distribution are as follows:

1. Let the multivariate random variable $(x_1, \ldots, x_n)$ follow the multivariate Waring distribution (5.148). Then the corresponding expected value is

$$E(x_1, \ldots, x_n) = a \int_0^1 dx\, (1 - x)^{a-n-1} \prod_{i=1}^{n} (1 - x + b_i x). \tag{5.151}$$

2. Every marginal distribution of the multivariate Waring distribution is also a Waring distribution

$$\sum_{k_s=1}^{\infty} \cdots \sum_{k_n=1}^{\infty} p(x_1 = k_1, \ldots, x_s = k_s, x_{s+1} = k_{s+1}, \ldots, x_n = k_n; a, b_1, \ldots, b_n) = p(x_1 = k_1, \ldots, x_s = k_s; a, b_1, \ldots, b_n). \tag{5.152}$$

The simplest case of the multivariate Waring distribution is the bivariate Waring distribution

$$p(x = k, y = j; a, b, c) = a \frac{(k + j - 2)! b^{(k-1)} c^{(j-1)}}{(a + b + c)^{(k+j-1)(k-1)!(j-1)!}}, \tag{5.153}$$

with expected value

$$E(x, y) = 1 + \frac{b+c}{a-1} + \frac{2bc}{(a-1)(a-2)} \tag{5.154}$$

and covariance

$$\mathrm{Cov}(x, y) = 1 + \frac{b+c}{a-1} + \frac{2bc}{(a-1)(a-2)} - \left(1 + \frac{b}{a-1}\right)\left(1 + \frac{c}{a-1}\right). \tag{5.155}$$

If (x, y) follows the bivariate Waring distribution, then the conditional probability $p(x = k \mid y = m)$ is

$$p(x = k \mid y = m) = \frac{1}{(k+1)!} \frac{(a+c)^{(b)}}{(a+c+m)^{(b)}} \frac{b^{(k-1)} m^{(k-1)}}{(a+b+c+m)^{(k-1)}}, \tag{5.156}$$

and the conditional expectation $E(x \mid y = m)$ is

$$E(x \mid y = m) = 1 + \frac{b}{a+c-1} m. \tag{5.157}$$

The multivariate Waring distribution was applied to the study of scientific productivity among authors in six main Chinese journals of information science during the three-year periods 1987–1989 and 1990–1992 [152].

5.8.3 *Quantities Connected to the Age of Citations*

After publication of an article, some tame elapses before the article is cited. Let T be the time between publication of the article and the publication of the citing source. In general, T is a random variable, and one can study distributions of the time to the first citation [115], or to the nth citation [125]. Here we mention several quantities connected to the time of first citation (these quantities can be applied also to the time of second citation, etc.) [154]. Let us assume that T is a continuous quantity, and let $f(t)$ be the probability density function of the distribution of T. Then one can define the age-specific citation rate

$$r(t) = -\frac{d}{dt}[\ln R(t)], \tag{5.158}$$

where

$$f(t) = \frac{dR}{dt},$$

and $R(t) = R_T(t) = p(T > t)$ is called the reliability function of T (here $p(T > t)$ means the probability that $T > t$). From (5.158), it follows that

$$R(t) = \exp\left(-\int_0^t dsr(s)\right). \tag{5.159}$$

Assuming different kinds of distributions for $f(t)$, we can obtain the corresponding relationship for the age-specific citation rate. Since citations (in most cases) can be considered rare events, we can use distributions connected to the theory of extreme events, such as the following:

- The exponential distribution $f(t) = \lambda \exp(-\lambda t)$. In this case, $R(t) = \exp(-\lambda t)$ and

$$r(t) = \lambda. \tag{5.160}$$

  Thus a constant age-specific citation rate implies an exponential distribution of the citation age.
- The Weibull distribution of citation age T with shape parameter $\beta > 0$ and scale parameter $\alpha > 0$. Here the reliability function is $R(t) = \exp[-(t/\alpha)^{\beta}]$, and the age-specific citation rate is

$$r(t) = \frac{\beta t^{\beta-1}}{\alpha^{\beta}}. \tag{5.161}$$

5.9 Probability Models Connected to Research Dynamics

5.9.1 Variation Approach to Scientific Production

The occurrence of laws in the form of hyperbolic relationships (such as the laws of Zipf and Pareto, for example) and the persistence of such laws may lead to the following assumption:

> *A research organization is in an equilibrium state with respect to scientific production if the statistical laws for the characteristic quantities of this productivity are given by hyperbolic relationships.*

We can even extend the above assumption by the additional assumption that the parameters of the statistical laws have selected values (for example, $\alpha = 1$) when the research organization is in an equilibrium state. And if the distributions of the quantities are not described by the appropriate hyperbolic relationships, then the research organization (and its structure and system of functioning) may not be in an equilibrium state.

Equilibrium states of various systems may be studied by variational methods [155]. A hint at the possible applicability of a variational approach in the social sciences is connected to George Zipf, who explained what is now known as Zipf's law in the field of linguistics [156] by means of the principle of least effort:

> *Human communication is based on two opposite tendencies: the one who speaks tries to use the minimum number of words, and this one who hears tries to understand the speaker by investing minimal effort.*

Let the effort $E(x)$ of a researcher to produce x publications be proportional to the time he or she invests for research: $E(x) \propto t$. There is a law for an exponentially growing science that states that scientific production growths exponentially with invested time: $x(t) = \exp(\lambda t)$, where λ is a parameter. From here, $t = \frac{1}{\lambda} \ln(x)$ and

$$E(x) \propto \frac{1}{\lambda} \ln(x) = \rho \ln(x). \tag{5.162}$$

This relationship will be introduced in the relationships for the variational principle of Boltzmann below [104, 157].

The principle of maximum entropy (variational principle of Boltzmann) is for systems whose states x are distributed with probability $p(x)$ ($\int dx\, p(x) = 1$). Then at an equilibrium state with energy

$$E = \int dx\, p(x) E(x), \tag{5.163}$$

the entropy

$$H = -\int dx\, p(x) \ln[p(x)] \tag{5.164}$$

has a maximum value.

The function $p(x)$ above is the probability that a researcher has produced x publications, and we shall treat $E(x)$ below as a measure of the mean effort (mean "energy") spent in the course of the scientific work. The solution of the above variational problem is

$$p(x) = (1/Z) \exp[-\lambda^* E(x)] = (1/Z)(1/x^{\rho\lambda^*}), \tag{5.165}$$

where Z is the statistical sum and λ^* is a parameter that can be determined from the normalization condition and the boundary condition.

Here we shall discuss as the least-value state the state $x_0 = 1$ (researchers must have at least one publication). Then

$$E = \int_1^\infty dx\, p(x)E(x) \tag{5.166}$$

and

$$p(x) = (\rho/E)1/(x^{1+\rho/E}) = \alpha/(x^{1+\alpha});\quad \alpha = (\rho/E). \tag{5.167}$$

This is the law of Pareto (called also the Zipf–Pareto law).

The entropy of a system that obeys the law (5.167) is

$$H = -\int_1^\infty dx\, p(x)\ln[p(x)] = 1 + \frac{1}{\alpha} - \ln(\alpha); \tag{5.168}$$

"Temperature": *The analogy with the thermodynamics may be continued: one may introduce a quantity called "temperature." This quantity is a measure of the external influence on the scientific system.*

"Temperature" can be introduced by comparing the results for Lagrange multipliers in statistical mechanics (where $\lambda^* \propto 1/T$) with the case of scientific production (where $\lambda^* = (1+\alpha)/\rho$). Thus the "temperature" is

$$T \propto \frac{\rho}{1+\alpha}. \tag{5.169}$$

Using (5.169), we can write the Zipf–Pareto law (5.167) as

$$p(x) = \frac{\alpha}{x^{k\rho/T}}, \tag{5.170}$$

where k is a coefficient of proportionality. From (5.169), $\alpha = 1 - \frac{k\rho}{T}$, and the final form of the Zipf–Pareto law (5.170) is

$$p(x) = \frac{1 - \frac{k\rho}{T}}{x^{k\rho/T}}. \tag{5.171}$$

There are two parameters in (5.171):

- k: characteristic of the efforts of the researcher in the publication process. These efforts can depend on the talent of the researcher but also on the conditions of work, salary, etc. Increasing research efforts lead to a decreasing value of k.
- T: characteristic of external influence on research organization. The parameter T can be connected to different flows toward the scientific structures (e.g., to money

flows). Then if the money flow increases, the system is "heated," and if the money flow decreases, the system is "frozen."

Let us analyze (5.171). We shall see the role of better work conditions and increased funding in increasing research production.

1. Let us fix the number of publications x. Thus we can study the influence of ρ and T. Let us fix also T (for example, a fixed quantity of money flows to the scientific organization, and other external conditions are fixed). Then a decrease in ρ will increase the numerator of (5.171) and will decrease its denominator. Hence p will increase. *This means that initiatives to decrease the necessary expenditures of effort by researchers in the publication production process (for example, an initiative for better work conditions or better social networking in the research organization) may increase the probability that researcher will have a larger number of publications*.
2. Let us now fix x and ρ and increase T (for example, by increasing the money flow toward the research organization). The numerator of (5.171) increases, and the denominator decreases. Thus p increases, which means that *one can expect that research production will increase with increased funding*.

Finally, let us note that thermodynamic models are also used in other areas of science such as technological forecasting and the theory of manpower systems [158, 159].

The variational approach can also be applied to the case of discrete distributions (e.g., for studying the circulation of documents) [160]. Let us consider a finite probability distribution $P = \{p_1, \ldots, p_n\}$, where $p_i \geq 0$ for $i = 1, \ldots, n$ and $\sum_i = 1^n p_i = 1$. The entropy attached to this probability distribution is

$$H_n(P) = -\sum_{i=1}^{n} p_i \ln(p_i). \tag{5.172}$$

The entropy is a measure of uncertainty. The uncertainty is maximal when the outcomes are equally likely. Since the uniform distribution maximizes the entropy, it contains the largest amount of uncertainty.

Let $X = \{1, \ldots, n\}$ be a random variable and p_i the probability of the occurrence of the value i. We have the constraint

$$\sum_{i=1}^{n} p_i = 1, \tag{5.173}$$

and we impose an additional constraint about the expected value of the distribution X:

$$E(X) = \sum_{i=1}^{n} i p_i = \mu. \tag{5.174}$$

According to the principle of maximum entropy, we have to find the distribution P that maximizes the entropy (5.172) subject to the constraints (5.173) and (5.174). Introducing two Lagrange multipliers α and β, we have to find a maximum for the functional

$$L = H_n(P) - \alpha\left(\sum_{i=1}^{n} p_i - 1\right) - \beta\left(\sum_{i=1}^{n} ip_i - E(X)\right). \tag{5.175}$$

The Euler equations for L from (5.175) are

$$\begin{aligned} \partial L/\partial p_i &= -\ln(p_i) - 1 - \alpha - \beta i; \quad i = 1, \ldots, n, \\ \partial L/\partial \alpha &= 1 - \sum_{i=1}^{n} p_i, \\ \partial L/\partial \beta &= E(X) - \sum_{i=1}^{n} ip_i. \end{aligned} \tag{5.176}$$

The solution of these equations is

$$p_i = \frac{\exp(-\beta_0 i)}{\sum_{i=1}^{n} \exp(-\beta_0 i)}, \tag{5.177}$$

where β_0 is the solution of the equation

$$\sum_{i=1}^{n} [i - E(X)] \exp[-(i - E(X))] = 0. \tag{5.178}$$

A similar calculation can also be made for the case of more than two constraints.

5.9.2 Modeling Production/Citation Process

Joint modeling of production and citation processes in science attracted considerable attention after the introduction of the h-index of Hirsch. Below, we shall consider two models of the processes connected to the h-index.

5.9.2.1 Model of h-Index Based on Paretian Distributions

Discrete Paretian distributions and the Price distribution are distributions that are widely used for modeling publication activity and citation processes [161]. The properties of these distributions needed for investigation of the Hirsch index are

represented by means of Gumbel's characteristic extreme values [162]. The reason for this is that the Hirsch index can be defined on the basis of Gumbel's rth characteristic values.

Gumbel's rth characteristic values are defined as follows. Let us consider a random variable X that gives the citation rate of a paper. We define

- $p_k = P(X = k)$: probability distribution of X $(k \geq 0)$;
- $F(k) = P(X < k)$: cumulative distribution function of X.

Gumbel's rth characteristic extreme value is then defined as

$$u_r = \max\{k : G(k) \geq r/n\}, \tag{5.179}$$

where

- $G(k) = G_k = 1 - F(k) = P(X \geq k)$;
- n: given sample with distribution F.

The Hirsch index can be defined analogously to Gumbel's rth characteristic extreme value as follows:

$$h = u_h. \tag{5.180}$$

5.9.2.2 Case of Paretian Distribution of the Random Variable X

A distribution of a random variable (in our case, the distribution of citations X) is a Paretian distribution if it obeys asymptotically Zipf's law:

$$\lim_{k\to\infty} \frac{G_k}{k^\alpha} \approx \text{const.} \tag{5.181}$$

Below, we shall use a prominent member of the class of Paretian distributions, namely the Pareto distribution $p_k = P(X = k) \approx \frac{d}{(N+k)^{-(1+\alpha)}}$. This distribution is Paretian as $k \to \infty$. For the case $k \gg N$, we obtain

$$G_k = P(X \geq k) \approx \frac{d_1}{k^\alpha}, \tag{5.182}$$

where d_1 is a positive constant. Then

$$u_r \approx c_1 \left(\frac{n}{r}\right)^{1/\alpha}, \tag{5.183}$$

where c_1 is a positive constant. Equation (5.183) leads to the following equation for the Hirsch index (in the presence of the assumption $n \gg 1$):

$$h = u_h \approx c_1 (n/h)^{1/\alpha} . \tag{5.184}$$

From here, we obtain

$$h \approx c_2 n^{1/(1+\alpha)}, \tag{5.185}$$

where $c_2 = c_1^{\alpha/(1+\alpha)}$.

We can draw the following conclusions from (5.185) (note that we work with the assumption that the citation distribution is a discrete Paretian distribution (with finite expectation)).

1. If the number of underlying papers is large enough, then the Hirsch index h is proportional to the $(1 + \alpha)$th root of the number of publications. Usually α is close to 1. Then h is proportional to the square root of the number of publications.
2. The number of citations of the papers from the Hirsch core (which contains the h-papers: papers that received at least h citations each) is proportional to h^2 for $\alpha > 1$ and a large value of k [161].

5.9.2.3 Case of Price Distribution of the Random Variable X

We recall that in our case, the random variable X is the citation rate of a paper. The Price distribution is [163]

$$p_k = P(X = k) = N\left(\frac{1}{N+k} - \frac{1}{N+k+1}\right) = \frac{N}{(N+k)(N+k+1)}, \tag{5.186}$$

where $k \geq 0$ and N is a positive parameter.

Note that N is a positive parameter. Thus N may be a noninteger. In addition, the Price distribution contains the case $k = 0$ as well as the law of Lotka (for research publications) when $k \gg N$. Moreover, no positive moments of the Price distribution exist. The distribution (5.186) is called the Price distribution, since it contains as a limiting case the square root law of Price (*which states that half of the scientific papers are contributed by the top square root of the total number of scientific authors*) [163]. Let us stress that the Price distribution is a particular case (when $\alpha = 1$) of

the Waring distribution [101, 164]

$$p_k = P(X = k) = \frac{\alpha}{N+\alpha}\frac{N}{N+\alpha+1}\ldots,\frac{N+k-1}{N+\alpha+k} \tag{5.187}$$

where $k \geq 0$ and α and N are positive parameters.

For the case in which the distribution of the citation rate is described by the Price distribution, one obtains

$$G_k = \frac{N}{N+k}. \tag{5.188}$$

Thus the distribution is Paretian (but note that the expected value of X for this distribution is ∞, in contrast to the finite expectation connected to the Pareto distribution discussed above).

The Gumbel rth extreme value is

$$u_r = \left[\frac{N(n-r)}{r}\right], \quad r = 1, 2, \ldots, n, \tag{5.189}$$

where $[\ldots]$ denotes the integer part of the corresponding argument.

The corresponding h index is a solution of the equation

$$h = u_h \approx \frac{N(n-r)}{r}. \tag{5.190}$$

The solution (for $n \gg 1$) can be approximated as

$$h = \left(\frac{N^2}{4} + nN\right)^{1/2} - \frac{N}{2} \approx (nN)^{1/2}, \tag{5.191}$$

which means the following:

The h-index is proportional of the square root of the number of publications (if the citation rate is described by the Price distribution and all other assumptions are valid).

5.9.2.4 Model of *h*-Index Based on the Poisson Distribution

Another model of the h-index is based on the publication–citation model of Burrell [165, 166]. This model is for the publishing record of a scientist who publishes papers at certain times. These papers then attract citations, and both the publication and citation accumulation processes are random. The assumption is that the scientist

starts his/her publishing career at $t = 0$, and by the time $T > 0$, one observes the following:

1. *Poisson process of publishing*
   The author publishes papers according to a Poisson process at rate θ. The distribution of the number of publications Y_T at time T is

$$P(Y_T = r) = \exp(-\theta T)\frac{(\theta T)^r}{r!}, \quad r = 1, 2, \ldots, \tag{5.192}$$

   with expected value $E[Y_T] = \theta T$.
2. *Poisson process of citations receiving*
   Each of the publications receives citations according to a Poisson process of rate Λ, which can vary from paper to paper.
3. *Variation of the rate* Λ
   The citation rate Λ varies over the set of publications of the scientist according to a gamma distribution of index $\nu > 1$ and parameter $\alpha > 0$:

$$f_\Lambda(\lambda) = \frac{\alpha^\nu}{\Gamma(\nu)}\lambda^{\nu-1}\exp(-\alpha\lambda), \tag{5.193}$$

   where $0 < \lambda < \infty$.

The model leads to the following distribution of the citations of a randomly chosen paper of the scientist [166]:

$$P(X_T = r) = \frac{\alpha}{T(\nu-1)}B\left(\frac{T}{\alpha+T}; r+1, \nu-1\right), \quad r = 0, 1, 2, \ldots, \tag{5.194}$$

where

$$B(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x dy\, y^{a-1}(1-y)^{b-1}$$

is the cumulative distribution of the beta distribution of the first kind, and a and b are parameters.

What remains to be calculated is $N(n; T)$: the expected number of papers receiving at least n citations by the time T.

- **Case of $n = 0$ citations**

$$E[N(0; T)] = \theta T, \tag{5.195}$$

  i.e., the number of uncited papers of the scientist is expected to have linear increase over time.

- **Case of $n \neq 0$ citations** In this case [166],

$$E[N(n;T)] = \theta T \left[1 - \frac{\alpha}{T(\nu - 1)} \sum_{r=0}^{n-1} B\left(\frac{T}{\alpha + T}; r + 1, \nu - 1 \right) \right], n = 1, 2, \ldots . \tag{5.196}$$

Equation (5.196) has interesting consequences:

1. *Publish or perish!*: The expected number of papers with n citations is proportional of the publication rate θ.
2. *A long career in science is a good thing!*: The expected number of papers with n citations is increasing in T for every n.
3. *No one is a genius!*: The expected number of papers with n citations is decreasing in n for every T.

Finally, the h-index can be defined as

$$h(T) = \max\{n : n \leq E[N(n, T)]\}, \tag{5.197}$$

and as we have seen just above, the h-index depends on the intensity of publication, the length of the scientific career, and other parameters (such as the parameters α and ν of the beta distribution, which can vary from scientist to scientist).

5.9.3 *The GIGP (Generalized Inverse Gaussian–Poisson Distribution): Model Distribution for Bibliometric Data. Relation to Other Bibliometric Distributions*

Up to now, we have discussed several distributions that may be used to model different aspects of research dynamics and to fit bibliometric data. Sichel [167, 168] argues that there exists a distribution that is very suitable for modeling bibliometric data: the GIGP (generalized inverse Gaussian–Poisson) distribution. The GIGP distribution seems to be complicated, but its goodness of the fit with respect to bibliometric data is usually very good. The GIGP distribution may be obtained as follows. Let us consider a researcher who has an average rate of publishing λ_i papers in unit time. Then the expected number of papers published by this researcher for time t will be $\lambda_i t$. Let us assume that the statistical variability around the average $\lambda_i t$ follows a Poisson distribution. If we have a group of researchers, then within this group, the value of λ_i will vary, since some researchers are more productive than others. Let us assume that the values of λ_i are distributed according to a generalized

inverse Gaussian distribution law (called a GIG distribution).[1] Then we arrive at the compound Poisson distribution called GIGP [170]:

$$p(r,t) = \frac{(1-\theta_t)^{\gamma_t/2}}{K_{\gamma_t}[\alpha_t\sqrt{1-\theta_t}]}\frac{(\alpha_t\theta_t)^r}{2^r r!}K_{r+\gamma_t}(\alpha_t), \tag{5.198}$$

where $r = 0, 1, 2, \ldots$; $0 \le \theta_t \le 1$; $-\infty < \gamma_t < \infty$; $\alpha_t \ge 1$; $K_\nu(z)$ is the modified Bessel function of the second kind of order ν; and t is the length of the considered time period. The time-dependent parameters are as follows:

$$\alpha_t = \alpha\sqrt{1+\theta(t-1)};\quad \theta_t = \frac{\theta t}{1+\theta(t-1)};\quad \gamma_t = \gamma. \tag{5.199}$$

From (5.198), one can calculate the probabilities $p(r)$ by means of a recurrence relation as follows if one knows $p(0)$ and $p(1)$ for $r = 0, 1, 2, \ldots$:

$$p(r) = \left(\frac{r+\gamma-1}{r}\right)\theta_t p(r-1) + \frac{\alpha_t^2\theta_t^2}{4r(r-1)}p(r-2). \tag{5.200}$$

The GIGP is also able do describe the domain $r = 1, 2, , 3, \ldots$. For this purpose, one has to perform zero truncation of the distribution from (5.198). The result is

$$p(r,t) = \frac{(\alpha_t\theta_t)^r K_{\gamma+r}(\alpha_t)}{2^r r!\{(1-\theta_t)^{-\gamma/2}K_\gamma[\alpha_t(1-\theta_t)^{1/2}] - K_\gamma(\alpha_t)\}}. \tag{5.201}$$

The GIGP distribution has been used to describe bibliometric data such as the number of articles published in the area of operations research, the scattering of literature in applied geophysics, the literature on mast cells, publications of a group of chemists several years after receiving their doctoral degrees, in-house journal use in libraries, etc. [167].

The GIGP distribution (5.198) has three parameters. If some of these parameters are known a priori, then the GIGP distribution can be reduced to several different distributions. Some examples of such reduction are as follows:

1. *Negative binomial distribution*: $\alpha = 0$; $\gamma > 0$.
2. *Zero-truncated negative binomial distribution*: $\alpha = 0$; $-1 < \gamma < 1$.
3. *Fisher logarithmic series distribution*: $\alpha =$; $\gamma = 0$.
4. *Inverse Gaussian–Poisson (IGP) distribution*: $\gamma = -1/2$; $r = 0, 1, 2, \ldots$.

[1] The form of this distribution may be written as

$$f(x) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}x^{(p-1)}\exp[-(ax+b/x)/2],$$

where $x > 0$, K_p is the modified Bessel function of the second kind, $a > 0$, $b > 0$, and p is a real parameter [169].

The upper tail (i.e., for large values of r) of the GIGP distribution is given by the following relationship [168]:

$$p(r) \sim \frac{c\theta^r}{r^{1-\gamma}}, \tag{5.202}$$

where c is a normalizing constant, $0 < \theta \leq 1$, and $-\infty < \gamma < \infty$. Taking the logarithm of both sides of (5.202), one can write

$$Y = A - (1 - \gamma)X - B\exp(X), \tag{5.203}$$

where $Y = \ln p(r); X = \ln r; A = \ln C; B = -\ln\theta$. Thus the tail of the GIGP distribution for $\gamma < 1$ is first linear, and then with increasing value of r, it becomes convex. Let $\theta = 1$. Then the tail of the GIGP distribution described by (5.203) becomes linear, and thus the GIGP distribution for this case corresponds to the distributions of Lotka and Zipf discussed in a previous chapter of this book.

5.9.4 Master Equation Model of Scientific Productivity

We know already that productivity is an important element in the evolution of a research community. It is possible to derive an equation that accounts for the stochastic fluctuations in the productivity of the members of a scientific organization [171]. In order to obtain this model equation, we assume that the main processes of evolution of scientific community are these:

1. the self-reproduction of scientists,
2. aging and death of scientists,
3. departure of scientists from the scientific field due to mobility or abandoning research activities.

Let a be the scientific age (number of years devoted to scientific research) of a researcher, and let a scientific productivity index ξ be incorporated into the researcher state space (ξ and a are assumed to be continuous variables with values in $[0, \infty]$). The dynamics of the research community are described by a number density function $n(a, \xi, t)$, which specifies the age and productivity structure of the scientific community at time t. For example, the number of researchers with age in $[a_1, a_2]$ and scientific productivity in $[\xi_1, \xi_2]$ at time t is given by the integral $\int_{a_1}^{a_2}\int_{\xi_1}^{\xi_2} da\, d\xi\, n(a, \xi, t)$.

The following master equation for this function $n(a, \xi, t)$ can be derived [171]:

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n(a, \xi, t) = -[J(a, \xi, t) + w(a, \xi, t)]n(a, \xi, t) + \int_{-\infty}^{\xi} d\xi' \chi(a, \xi - \xi', \xi', t) n(a, \xi - \xi', t), \tag{5.204}$$

where $w(a, \xi, t)$ denotes the departure rate of community members. If $x(t)$ is a random process describing the scientific productivity variation and $p_a(x, t \mid y, \tau)$ $(\tau < t)$ is the transition probability density corresponding to such a process, then

$$\chi(a, \xi, \xi', t) = \lim_{\Delta t \to 0} \frac{p(\xi + \xi', t + \delta t \mid \xi, t)}{\Delta t}. \tag{5.205}$$

The transition rate $J(a, \xi, t)$ at time t from the productivity level ξ is by definition

$$J(a, \xi, t) = \int_{-\xi}^{\infty} d\xi' \, \chi(a, \xi, \xi', t).$$

The increment ξ' may be positive or negative. The equation for $n(a, \xi, t)$ can be obtained in the following way. First, for the increment we have

$$n(a + \Delta a, \xi, t + \Delta t) = n(a, \xi, t) - J(a, \xi, t) n(a, \xi, t) \Delta t + \int_{-\infty}^{\xi} \chi(a, \xi - \xi', \xi', t) n(a, \xi - \xi', t) d\xi' \Delta t - w(a, \xi, t) n(a, \xi, t) \Delta t, \tag{5.206}$$

where:

- the term on the right-hand side, $[1 - J(a, \xi, t)\Delta t]n(a, \xi, t)$, describes the proportion of individuals whose scientific productivity does not change in $(t, t + \Delta t)$;
- the integral term describes the individuals whose scientific productivity becomes equal to ξ because of increase or decrease in $(t, t + \Delta t)$;
- the last term corresponds to the departure of individuals through stopping research activities or death.

After expanding $n(a + \Delta t, \xi, t + \Delta t)$ around a and t and retaining terms up to the first order in Δt, one obtains the master equation (5.204).

The above master equation is difficult for analysis, and because of this, it is often reduced to an approximation similar to the well-known Fokker–Planck equation. Let

$$\mu_k(a, \xi, t) = \int_{-\xi}^{\infty} d\xi' (\xi')^k \chi(a, \xi, \xi', t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} < (\xi')^k >; \quad k = 1, 2, \ldots, \tag{5.207}$$

where the brackets denote the average with respect to the conditional probability density $p_a(\xi + \xi', t + \Delta t \mid \xi, t)$. In addition, we make the following assumptions:

- $\mu_1, \mu_2 < \infty$;
- $\mu_k = 0$ for $k > 3$;
- $n(a, \xi, t)$ and $\chi(a, \xi, \xi', t)$ are analytic in ξ for all a, t, and ξ'.

The assumption $\mu_k = 0$ for $k > 3$ demands that productivity be continuous, i.e., when $\Delta t \to 0$, the probability of large fluctuations $\mid \xi' \mid$ must decrease so quickly that $<\mid \xi' \mid^3> \to 0$ more quickly than Δt.

When the above assumptions hold, the function n satisfies the equation

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n = -\frac{\partial(\mu_1 n)}{\partial \xi} + \frac{1}{2}\frac{\partial^2(\mu_2 n)}{\partial \xi^2} - wn. \tag{5.208}$$

The following notes are in order here.

1. If $w = 0$, (5.208) is reduced to the Fokker–Planck equation.
2. Equation (5.208) describes the evolution of the scientific community through a drift along the age component and a drift and diffusion with respect to the productivity component.
3. The diffusion term characterized by the diffusivity μ_2 takes into account the stochastic fluctuations of scientific productivity conditioned by internal factors (such as individual abilities, labor motivations, etc.) and external factors (such as labor organization, stimulation systems, etc.).
4. The initial and boundary conditions for (5.208) are:
   - $n(a, \xi, 0) = n^0(a, \xi)$, where $n^0(a, \xi)$ is a known function defining the community age and productivity distribution at time $t = 0$;
   - $n(0, \xi, t) = \nu(\xi, t)$, where the function $\nu(\xi, t)$ represents the intensity of input flow of new members at age $a = 0$ and $\nu(\xi, 0) = n^0(0, \xi)$.
5. In addition, $n(a, \xi, t) \to 0$ as $a \to \infty$.

The general solution of (5.208) with the above initial and boundary conditions is still a difficult task. But for many practical applications, knowledge of the first and second moments of the distribution function $n(a, \xi, t)$ is sufficient. Equation (5.208) can be solved numerically or can be reduced to a system of ordinary differential equations [171].

In a similar way, a model of personal movement can be obtained [172]. The model equation for this case is

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n(a, t) = -n(a, t)[w_1(a, t) + w_2(a, t)] + r(a, t)\nu(t), \tag{5.209}$$

where a is the age variable, t is the time, $n(a, t)$ is the density of researchers having age a at time t, w is the age intensity of researchers' departure, $\nu(t)$ is the intensity of the input flow of new researchers at the moment of time t, $r(a, t)$ is the density

of the input flow age distribution, $w_1(a,t)$ is the intensity of departure due to death, retirement, etc., $w_2(a,t)$ is the intensity of the regulated departure of researchers ($w(a,t) = w_1(a,t) + w_2(a,t)$). Also, a_0 denotes the minimum age of researchers and A denotes the maximum admissible age of researchers; a_0 and A participate in the initial condition

$$n(a,0) = n^0(a), \quad a_0 \leq a \leq A, \tag{5.210}$$

and the boundary condition is

$$n(a_0,t) = 0, \quad t \geq 0. \tag{5.211}$$

5.10 Probability Model for Importance of the Human Factor in Science

Below, we shall discuss a probability model connected to the importance of the human factor in science. One often hears that technological evolution is closely connected to the growth of science and that the growth of science depends heavily on the human factor (number and quality of scientists). Such statements are no surprise, since a connection has been observed between the values of scientometric indicators of the research production of a country's researchers and the corresponding GDP [173–177]. A research organization may have a perfect structure with respect to research positions and research equipment associated with those positions. The research positions may be connected by a perfect system of relations, and the processes in the organization may be carefully planned. But this is not enough. In order to put all the above into effective action, one needs researchers. Researchers of good quality have to fill the research positions. Researchers have to perform actions that contribute to a smooth flow of the processes in a research organization. Only then can the work of this organization be effective. In addition, a researcher does not work alone [178–183]. Teamwork and collaboration among scientists and scientific groups is becoming ever more for solving the scientific problems of today [184–189].

This shows that the human factor is of extreme importance for research organizations. Because of this, we shall discuss below (with the help of mathematics) the importance of the size of the research community.

5.10.1 The Effective Solutions of Research Problems Depend on the Size of the Corresponding Research Community

It is intuitive that larger research communities can solve more complex problems [92]. Let us consider some research problem and let β be the mean probability that a qualified researcher will solve the problem. Then:

- $1-\beta$ is the probability that the researcher will not solve the problem.
- $(1-\beta)^n$ is the probability that a group of n qualified researchers will not solve the problem.

Thus the probability that the same group of n qualified researchers will solve the complex problem (which is not likely to be solved by a single researcher, i.e., $\beta \ll 1$) is

$$p_n = 1-(1-\beta)^n = 1-\exp[n\ln(1-\beta)] \approx 1-\exp(-n\beta). \tag{5.212}$$

If the research group is small, i.e., $n\beta \ll 1$, then from (5.212), we obtain the linear relationship

$$p_n \approx \beta n. \tag{5.213}$$

Then an increase in the size of the group of qualified researchers increases the probability of solving the problem. When the group is small, the probability of solving the problem is proportional of the group size. When the size of the group increases, the nonlinear terms become significant, and the probability p_n increases faster than a linear function.

5.10.2 *Increasing Complexity of Problems Requires Increase of the Size of Group of Researchers that Has to Solve Them*

Scientific organizations evolve and usually become more complex [190, 191]. One factor for such a development is the need to solve research problems of increasing complexity. This increasing complexity leads to a decreasing probability β that a single researcher can solve such a problem. In order to compensate this decrease, one may increase the size of the research group that has to solve the problem.

Let us study the above situation with the help of mathematics. To compensate the decrease of probability means that one has to keep $(dp_n/dt) \geq 0$. Then from (5.212), one obtains

$$\frac{1}{n}\frac{dn}{dt} \geq -\frac{1}{\beta}\frac{d\beta}{dt}. \tag{5.214}$$

Taking into account that $(d\beta/dt) < 0$, the increase in the size of the research group with increasing complexity of the solved problem has to be

$$\frac{dn}{dt} \geq \frac{n}{\beta}\left(-\frac{d\beta}{dt}\right). \tag{5.215}$$

The above simple model leads to the following conclusions. *As the complexity of scientific problems increases with time, one needs larger research collectives in*

order to support a large probability of solving the problems. Thus if a government wants an effective solution of national scientific and technological problems, it has to support a large enough national research community. A decrease in the number of researchers diminishes the national scientific capacity: the probability of solving problems important to the society decreases at least proportionally to the decrease in the size of the corresponding research community.

Note that the value of the parameter β plays an important role in the above model. This value must be kept as large as possible. In other words, an effective scientific community consists of qualified scientists. In addition, let us note that research groups in most cases consist not only of researchers. There are also supporting staff. In connection with this, certain scaling properties may exist for research units [192]. For example, a power law relationship may exist between the number of supporting staff N_s and the number of academic staff N_A of a research institution: $N_s = CN_A^{\beta}$, where C is a constant and β is the exponent of the power law. For the case of the UK National Health System, $C \approx 0.07$ and $\beta \approx 1.3$. The last relationship is an example of a quantitative power law relationship connected to the parts of research (and other) organizations. Such power laws have been discussed in Chap. 4.

5.11 Concluding Remarks

In this chapter, selected classes of deterministic and probability models connected to science dynamics and research production have been discussed. The focus was on the models connected to dynamics of research systems and especially on models for deterministic and statistical properties of the process of publication and the process of citation of research publications. Some of the models have been described very briefly, while for some (probability) models, more discussion has been provided (for the case in which one can obtain interesting conclusions without having to perform long mathematical calculations). This manner of presentation permitted a description of more that twenty models in relatively few pages. We hope that the selected set of models has provided a good impression to the reader about the mathematical tools and methods used for modeling of complex processes and the nonlinear dynamics connected to research systems.

There exist also other deterministic and probability models. For example, there exists a model of science as a part of a global model of a social system. In this model, the scientific system can be treated as a system that has entrances and exits [92]. The input (different flows) comes from the other parts of the social system to the entrances of the science subsystem. At the exit, there are scientific output flows to other parts of the social system. The input flows can be flows of funding or human resources, for example. The main output flow is scientific knowledge. Part of this flow is the flow of publications.

Finally, let us make several remarks on the limited dependent variable models and on the generalized Zipf distribution, since these topics are of significant interest for research in the area of informetrics.

Limited dependent variable models (e.g., binary, ordinal, and count data regression models) may be used for analysis of all kinds of categorical and count data in bibliometrics and scientometrics (such as assessment scores, citation counts, career transitions, editorial decisions, or funding decisions) [193]. The main advantage of limited dependent variable models is that in using them, one may identify the main explanatory variables in a multivariate framework, and in addition, one may estimate the size of the (marginal) effects of these variables.

Let us consider the group of regression models. Limited dependent variable models are a subgroup of this group with a limited range of possible values of the variable of interest. This variable may have a binary outcome (e.g., whether a journal article was cited over a certain period). The variable may take multiple discrete values (e.g., for the case of assessment of research or for the case of peer reviews).

In the case of a binary regression model, we have a variable y_i that can take only the values 0 and 1. We may model the probability that this variable will take value 1 depending on the values of other variables $x_{1i}, \ldots, x_{ki}$ as follows:

$$p(y_i = 1 \mid, x_{1i}, \ldots, x_{ki}) = L(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}), \tag{5.216}$$

where $L(x) = \frac{\exp(x)}{1+\exp(x)}$ is the logistic function (whose range is between 0 and 1). The model (5.216) is called the logit model. The coefficients β_i of the logit model may be estimated by maximizing the likelihood of the data with respect to the coefficients.

The binary logistic model may be used for analyzing or predicting (or for analyzing and predicting) whether articles will be cited [194], for analysis of funding and editorial decisions [195], for analysis of winning scientific awards [196], etc. [197, 198]. One illustration of the application of the model can be seen in [193], in which the dependent variable measures whether an article was cited in another published article during the calendar year following its publication.

For the case of the ordinal regression model, the variable of interest y_i is an ordinal variable that can take only the values $j = 1, 2, \ldots, J$. In this model, the cumulative probability is the probability that an observation i is in the jth category or lower: $p(y_i) \leq j = \delta_{ij}$ can be modeled by the logit relationship

$$\text{logit}(\delta_{ij}) = \alpha_j - \beta_1 x_{1i} - \cdots - \beta_k x_{ki}, \tag{5.217}$$

where $\text{logit}(p) = \log(\frac{p}{1-p}) = \log(p) - \log(1-p)$. Ordinal regression models are applied when we are interested in additional characteristics of the investigated variables with respect to a characteristic modeled by the binary regression model. For example, in the case of binary regression analysis of citations, it was of interest to know whether an article has been cited. If an article has been cited, it may not be of interest how many citations of this article exist. If we are interested in the number of citations, we may use the ordinal regression model above. Such models are used in peer assessment of research groups [199] and for predicting the impact of international coauthorship on citation impact [200].

Finally, one may use count data models if the modeled variable represents the frequency of an event. The count data models can be Poisson models, negative

binomial models, etc. The Poisson model is for a count variable y_i that can take only nonnegative integer values: $0, 1, \ldots$. It is assumed that y_i conditional on the independent variables has a Poisson distribution ($y = 1, 2, \ldots$)

$$p(y_i = y \mid x_{1i}, \ldots, x_{ki}) = \frac{\mu_i^y \exp(-\mu_i)}{y!}, \tag{5.218}$$

where μ_i is the expected value of the distribution that is modeled by

$$\mu_i = \exp[\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}]. \tag{5.219}$$

A limitation of the Poisson regression model is that the Poisson distribution is completely determined by its mean and that the variance is assumed to equal the mean. This restriction may be violated in many applications, since the variance is often greater than the mean. Then there is overdispersion: the variance is greater than the variance implied by assuming a Poisson distribution. One possibility for dealing with overdispersion is to use a negative binomial regression model. This model allows the conditional mean μ_i of y_i to differ from its variance $\mu_i + a\mu_i^2$ by estimating an additional dispersion parameter a.

A Poisson model may be used to identify the effects of coauthorship networks on performance of scholars [201]. Negative binomial regression models can be applied to study citation counts for the purpose of determining the relative importance of authors and journals [202], for comparing sets of papers [203], and for modeling the number of papers [204].

There is a generalization of the Zipf distribution (called the generalized Zipf distribution) that contain as particular cases a family of skew distributions found to describe a wide range of phenomena both within and outside the information sciences and referred to as being of Zipf type. The generalized Zipf distribution is defined as follows [205]. Let

$$d(k \mid f) = \frac{\log[f(k+1)] - \log[f(k)]}{\log(k+1) - \log(k)}, \tag{5.220}$$

where $f(k) > 0$ and the integer k is greater than 1. Let N be the set of natural numbers $1, 2, \ldots$ and Z a random variable defined on N. Let $P(k) = P(X = k)$ and $F(k) = P(X \geq k) = \sum_{i \geq k} P(i)$ be the corresponding distributions connected to Z. A distribution F defined on N is a generalized Zipf distribution with exponent $\alpha > 0$ if and only if $d(k \mid f) \to -\alpha$ as $k \to \infty$, i.e.,

$$\lim_{k \to \infty} d(k \mid f) = \frac{\log[F(k+1)] - \log[F(k)]}{\log(k+1) - \log(k)}. \tag{5.221}$$

It is easily to check that the Waring distribution with

$$F(k) = \frac{\beta^{(k-1)}}{(\alpha + \beta)^{(k-1)}}, \ \beta^{(k)} = \beta(\beta - 1) \ldots (\beta + k - 1) \tag{5.222}$$

is a particular case (belongs to the class) of the generalized Zipf distribution. But the geometric distribution ($P(k) = \theta(1-\theta)^{k-1}$ and $F(k) = (1-\theta)^{k-1}$) does not belong to the class of generalized Zipf distributions.

The class of generalized Zipf distributions has several properties. In order to define the first property, we need to know when a function $\varphi(k)$ varies gradually. Let $\varphi(k)$ be a positive function defined on N. Then $\varphi(k)$ varies gradually if and only if

$$\lim_{k\to\infty} d(k\varphi) = \lim_{k\to\infty} \frac{\log\varphi(k+1) - \log\varphi(k)}{\log(k+1) - \log(k)} = 0; \tag{5.223}$$

$F(k)$ is a generalized Zipf distribution of exponent $\alpha > 0$ if and only if [205]

$$F(k) = \frac{\varphi(k)}{k^{\alpha}}, \tag{5.224}$$

where $\varphi(k)$ is a gradually varying function. An example of a distribution that belongs to the class of generalized Zipf distributions is the Yule distribution, with

$$F(k) = \frac{(k-1)!}{(\alpha+1)^{(k-1)}}. \tag{5.225}$$

We can write this distribution in the form (5.224), where

$$\varphi(k) = \frac{(k-1)!}{(\alpha+1)^{(k-1)}} k^{\alpha}. \tag{5.226}$$

One can define the quantities proportional hazard rate

$$r(k) = \frac{kP(k)}{F(k)}, \tag{5.227}$$

and the conditional expectation

$$e(m) = E[X \mid X \geq m] = \sum_{k\geq m} k \frac{P(x=k)}{P(X \geq m)}. \tag{5.228}$$

Then the following two statements can be proved [205]. First of all, $F(k)$ is a generalized Zipf distribution with exponent $\alpha > 0$ if and only if

$$\lim_{k\to\infty} r(k) \to \alpha. \tag{5.229}$$

Next, $F(k)$ is a generalized Zipf distribution with exponent $\alpha > 1$ if and only if

$$\lim_{k\to\infty} \frac{e(k)}{k} = \lim_{k\to\infty} [e(k+1) - e(k)] = \frac{\alpha}{\alpha - 1}. \tag{5.230}$$

References

1. D. de Solla Price. *Little Science, Big Science*. (Columbia University Press, New York, 1963)
2. D.P. Wallace, The relationship between journal productivity and obsolescence. J. Am. Soc. Inf. Sci. **37**, 136–145 (1986)
3. L. Egghe, On the influence of growth on obsolescence. Scientometrics **27**, 195–214 (1993)
4. W. Glänzel, U. Schoepflin, A bibliometric study on ageing and reception process of scientific literature. J. Inf. Sci. **21**, 37–53 (1995)
5. W. Goffman, V.A. Newill, Generalization of epidemic theory. An application to the transmission of ideas. Nature **204**(4955), 225–228 (1964)
6. P. Nyhius, Logistic curves, in *CIPR encyclopedia of production engineering*, ed. by L. Laperriere, G. Reinhart (Springer, Berlin, 2014), pp. 759–762
7. A. Fernandez-Cano, M. Torralbo, M. Vallejo, Reconsidering Price's model of scientific growth: an overview. Scientometrics **61**, 301–321 (2004)
8. V. Volterra, Population growth, equilibria, and extinction under specified breeding conditions: a development and extension of the theory of the logistic curve, in *The Golden Age of Theoretical Ecology: 1923–1940*, ed. by F.M. Scudo, J.E. Ziegler (Springer, Berlin, 1978), pp. 18–27
9. C.-Y. Wong, L. Wang, Trajectories of science and technology and their co-evolution in BRICS: Insigths from publication and patent analysis. J. Inf. **9**, 90–101 (2015)
10. L. Egghe, I.K. Ravichandra, Rao. Classification of growth models based on growth rates and its applications. Scientometrics **25**, 5–46 (1992)
11. P.S. Meyer, Bi-logistic growth. Technol. Forecast. Soc. Chang. **47**, 89–102 (1994)
12. M. Ausloos, On religion and language evolutions seen through mathematical and agent based models, in *Proceedings of the First Interdisciplinary CHESS Interactions Conference*, ed. by C. Rangacharyulu, E. Haven (World Scientific, Singapore, 2010), pp. 157–182
13. P.S. Meyer, J.W. Yung, J.H. Ausubel, A primer on logistic growth and substitution: the mathematics of the Loglet Lab software. Technol. Forecast. Soc. Chang. **61**, 247–271 (1999)
14. H.W. Menard, *Science: Growth and Change* (Harvard University Press, Cambridge, MA, 1971)
15. G.N. Gilbert, Measuring the growth of science: a review of indicators of scientific growth. Scientometrics **1**, 9–34 (1978)
16. D. Wolfram, C.M. Chu, X. Lu, Growth of knowledge: bibliometric analysis using online database data, in *Informetrics 89/90*, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1990), pp. 355–372
17. G.O. Ware, A general statistical model for estimating future demand levels of data-base utilization within an information retrieval organization. J. Am. Soc. Inf. Sci. **24**, 261–264 (1973)
18. N. Bailey, Some stochastic models for small epidemics in large populations. Appl. Stat. **13**, 9–19 (1964)
19. M.S. Bartlet, *Stochastic Population Models in Ecology and Epidemiology* (Wiley, New York, 1960)
20. W. Goffman, An epidemic process in an open population. Nature **205**, 831–832 (1965)
21. D. Mollison, Dependence of epidemic and population velocities on basic parameters. Math. Biosci. **107**, 255–287 (1991)
22. F.C. Hoppensteadt, *Mathematical Theories of Populations: Demographics, Genetics and Epidemics* (SIAM, Philadelphia, PA, 1975)
23. K. Cooke, P. van den Driessche, X. Zou, Interaction of maturation delay and nonlinear birth in population and epidemic models. J. Math. Biol. **39**, 332–352 (1999)
24. H.W. Hethcote, The mathematics of infectious diseases. SIAM Rev. **42**, 599–653 (2000)
25. V. Colizza, A. Barnat, M. Barthelemy, A. Vespigniani, The modeling of global epidemics: stochastic dynamics and predictability. Bull. Math. Biol. **68**, 1893–1921 (2006)
26. R.M. May, Simple mathematical models with very complicated dynamics. Nature **261**(5560), 459–467 (1976)

27. H. Caswell, *Matrix Population Models* (Wiley, New York, 2001)
28. R.D. Holt, Population dynamics in two-patch environments: some anomalous consequences of an optimal habitat distribution. Theoretical Population Biology **28**, 181–208 (1985)
29. M.P. Hassell, H.N. Comins, R.M. May, Spatial structure and chaos in insect population dynamics. Nature **353**(6341), 255–258 (1991)
30. Z. Ma, J. Li, Basic knowledge and developing tendencies in epidemic dynamics, in *Mathematics for Life Sciences and Medicine*, ed. by Y. Takeuchi, Y. Iwasa, K. Sato (Springer, Berlin, 2007), pp. 5–49
31. N.K. Vitanov, M. Ausloos, Knowledge epidemic and population dynamics models for describing idea diffusion, in *Models for Science Dynamics*, ed. by A. Scharnhorst, K. Börner, P. van den Besselar (Springer, Berlin, 2012), pp. 69–125
32. C. Antonelli, *The Economics of Localized Technological Change and Industrial Dynamics* (Kluwer, Dordrecht, 1995)
33. P. Anderson, Perspective: complexity theory and organization science. Organ. Sci. **10**, 216–232 (1999)
34. M.A. Nowak, Five rules for the evolution of cooperation. Science **314**(5805), 1560–1563 (2006)
35. W. Weidlich, G. Haag, *Concepts and Models of a Quantitative Sociology: The Dynamics of Interacting Populations* (Springer, Berlin, 1983)
36. D. Strang, Adding social structure to diffusion models. Sociol. Methods Res. **19**, 324–353 (1991)
37. P.A. Geroski, Models of technology diffusion. Res. Policy **29**, 603–625 (2000)
38. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. Rev. Mod. Phys. **81**, 591–646 (2009)
39. N.K. Vitanov, Z.I. Dimitrova, Application of the method of simplest equation for obtaining exact traveling-wave solutions for two classes of model PDEs from ecology and population dynamics. Commun. Nonlinear Sci. Numer. Simul. **15**, 2836–2845 (2010)
40. R. Baptista, The diffusion of process innovations: a selective review. Int. J. Econ. Bus. **6**, 107–129 (1999)
41. I.Z. Kiss, M. Broom, P.G. Craze, I. Rafols, Can epidemic models describe the diffusion of topics across disciplines? J. Inf. **4**, 74–82 (2010)
42. H.G. Landau, A. Rapoport, Contribution to the mathematical theory of contagion and spread of information. I: spread through a thoroughly mixed population. Bull. Math. Biophys. **15**, 173–183 (1953)
43. W. Goffman, Mathematical approach to the spread of scientific ideas—the history of mast cell research. Nature **212**, 449–452 (1966)
44. A. Lotka, *Elements of Physical Biology* (Williams and Wilkins, Baltomore, 1925)
45. V. Volterra, Variations and fluctuations of the number of individuals in animal species living together. Journal du Conseil/Conseil Permanent International pour l'Exploration de la Mer **3**, 3–52 (1928)
46. F.J. Ayala, M.E. Gilpin, J.G. Ehrenfeld, Competition between species: theoretical models and experimental tests. Theor. Popul. Biol. **4**, 331–356 (1973)
47. M.E. Gilpin, F.J. Ayala, Global models of growth and competition. PNAS **70**, 3590–3593 (1973)
48. R.D. Holt, J. Pickering, Infectious disease and species coexistence: a model of Lotka-Volterra form. Am. Nat. **126**, 196–211 (1985)
49. Y. Takeuchi, *Global Dynamical Properties of Lotka-Volterra Systems* (World Scientific, Singapore, 1996)
50. A. Castiaux, Radical innovation in established organizations: being a knowledge predator. J. Eng. Technol. Manag. **24**, 36–52 (2007)
51. K. Dietz, Epidemics and rumors: a survey. J. R. Stat. Soc. A **130**, 505–528 (1967)
52. S. Solomon, P. Richmond, Power laws of wealth, market order volumes and market returns. Phys. A **299**, 188–197 (2001)

53. S. Solomon, P. Richmond, Stable power laws in variable economics. Lotka—Volterra implies Pareto—Zipf. Eur. Phys. J. B **27**, 257–261 (2002)
54. W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A **115**, 700–721 (1927)
55. M. Nowakowska, Epidemical spread of scientific objects: an attempt of empirical approach to some problems of meta—science. Theory Decis. **3**, 262–297 (1973)
56. D.J. Daley, Concerning the spread of news in a population of individuals who never forget. Bull. Math. Biophys. **29**, 373–376 (1967)
57. A.D. Barbour, S. Utev, Approximating the Reed-Frost epidemic process. Stoch. Process. Appl. **113**, 173–197 (2004)
58. H. Abbey, An examination of the Reed-Frost theory of epidemics. Hum. Biol. **24**, 201–233 (1952)
59. J.A. Jacquez, A note on chain-binomial models of epidemic spread: what is wrong with the Reed-Frost formulation? Math. Biosci. **87**, 73–82 (1987)
60. W. Goffman, V.A. Newill, Communication and epidemic process. Proc. R. Soc. Lond. Ser. A **298**, 316–334 (1967)
61. G. Harmon, Remembering William Goffman: mathematical information science pioneer. Inf. Process. Manag. **44**, 1634–1647 (2008)
62. M. Cohen, A. Blaivas, A model for the growth of mathematical specialties. Scientometrics **3**, 265–273 (1981)
63. B.M. Gupta, L. Sharma, C.R. Karisiddappa, Modelling the growth of papers in a scientific speciality. Scientometrics **33**, 187–201 (1995)
64. M. Kochen, Stability and growth of knowledge. Am. Doc. **20**, 186–197 (1969)
65. A.N. Tabah, Literature dynamics: studies on growth, diffusion and epidemics. Annu. Rev. Inf. Sci. Technol. **34**, 249–286 (1999)
66. B.M. Gupta, P. Sharma, C.R. Karisiddappa, Growth of research literature in scientific specialities. A modeling perspective. Scientometrics **40**, 507–528 (1997)
67. B.M. Gupta, S. Kumar, S.L. Sangam, C.R. Karisiddappa, Modeling the growth of world social science literature. Scientometrics **53**, 161–164 (2002)
68. L.M.A. Bettencourt, A. Cintron-Arias, D.I. Kaiser, C. Castillo-Chavez, The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. Phys. A **364**, 513–536 (2002)
69. L.M.A. Bettencourt, D.I. Kaiser, J. Kaur, C. Castillo-Chavez, D.E. Wojick, Population modeling of the emergence and development of scientific fields. Scientometrics **75**, 495–518 (2008)
70. M. Szydlowski, A. Krawiez, Growth cycles of knowledge. Scientometrics **78**, 99–111 (2009)
71. D.J. de Solla Price, The exponential curve of science. Discovery **17**, 240–243 (1956)
72. K. Sangwal, Progressive nucleation mechanism and its application to the growth of journals, articles and authors in scientific fields. J. Inf. **5**, 529–536 (2011)
73. K. Sangwal, On the growth of citations of publication output of individual authors. J. Inf. **5**, 554–564 (2011)
74. K. Sangwal, Progressive nucleation mechanism of the growth behavior of items and its application to cumulative papers and citations of individual authors. Scientometrics **92**, 643–655 (2012)
75. K. Sangwal, Growth dynamics of citations of cumulative papers of individual authors according to progressive nucleation mechanism: concept of citation acceleration. Inf. Process. Manag. **49**, 757–772 (2013)
76. D. Kashchiev, *Nucleation: Basic theory with applications* (Butterworth-Heinemann, Oxford, 2000)
77. E.H. Kerner, Further considerations on the statistical mechanics of biological associations. Bull. Math. Biophys. **21**, 217–253 (1959)
78. J.C. Allen, Mathematical model of species interactions in time and space. Am. Nat. **109**, 319–342 (1975)
79. A. Okubo, *Diffusion and Ecological Problems: Mathematical Models* (Springer, Berlin, 1980)

80. W.G. Willson, A.M. de Roos, Spatial instabilities within the diffusive Lotka—Volterra system: individual—based simulation results. Theor. Popul. Biol. **43**, 91–127 (1993)
81. Y.F. le Coadic, Information system and the spread of scientific ideas. R&D Manag. **4**, 97–111 (1974)
82. E. Bruckner, W. Ebeling, A. Scharnhorst, The application of evolution models in scientometrics. Scientometrics **18**, 21–41 (1990)
83. N.K. Vitanov, I.P. Jordanov, Z.I. Dimitrova, On nonlinear population waves. Appl. Math. Comput. **215**, 2950–2964 (2009)
84. N.K. Vitanov, I.P. Jordanov, Z.I. Dimitrova, On nonlinear dynamics of interacting populations: coupled kink waves in a system of two populations. Commun. Nonlinear Sci. Numer. Simul. **14**, 2379–2388 (2009)
85. N.K. Vitanov, Z.I. Dimitrova, M. Ausloos, Verhulst-Lotka-Volterra (VLV) model of ideological struggle. Phys. A **389**, 4970–4980 (2010)
86. N.K. Vitanov, M. Ausloos, G. Rotundo, Discrete model of ideological struggle accounting for migration. Adv. Complex Syst. **15**, Article No. 1250049 (2012)
87. W. Ebeling, A. Scharnhorst, Evolutionary models of innovation dynamics, in *Traffic and Granular Flow'99. Social, Traffic and Granular Dynamics*, ed. by D. Helbing, H.J. Herrman, M. Schekenberg, D.E. Wolf (Springer, Berlin, 2000), pp. 43–56
88. E. Borensztein, J. De Gregorio, J.-W. Lee, How does foreigh direct investment affect economic growth? J. Int. Econ. **45**, 115–135 (1998)
89. J. Dedrick, V. Gurbaxani, K.L. Kraemer, Information technology and economic performance: a critical review of the empirical evidence. ACM Comput. Surv. **35**, 1–28 (2003)
90. S.W. Popper, C. Wagner, New foundations of growth: The U.S. innovation system today and tomorrow. RAND MR-1338.0/1-OSTP (2001)
91. E. Mansfield, *Industrial Research and Technological Innovation: An Econometric Analysis* (Norton, New York, 1968)
92. A.I. Yablonskii, *Mathematical Methods in the Study of Science* (Nauka, Moscow, 1986). (in Russian)
93. C.W. Cobb, P.H. Douglas, A theory of production. Am. Econ. Rev. **18**(Supplement), 139–165 (1928)
94. A. Aulin, *The Impact of Science on Economic Growth and its Cycles* (Springer, Berlin, 1998)
95. Q.L. Burell, Predictive aspects of some bibliometric processes, in *Informetrics 87/88*, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1988), pp. 43–63
96. Q.L. Burrell, A note on ageing in a library circulation model. J. Doc. **41**, 100–115 (1985)
97. D.R. Cox, Some statistical methods connected with series of events (with discussion). J. R. Stat. Soc. B **17**, 129–164 (1955)
98. J. Grandell, *Doubly stochastic Poisson processes*, vol. 529, Lecture Notes in Mathematics (Springer, Berlin, 1976)
99. H.S. Sichel, On a distribution representing sentence-length in written prose. J. R. Stat. Soc. A **137**, 25–34 (1974)
100. H.S. Sichel, Repeat-buying and the generalized inverse Gaussian-Poisson distribution. Appl. Stat. **31**, 193–204 (1982)
101. J.O. Irvin, The generalized Waring distribution. Part I. J. R. Stat. Soc. A **138**, 18–21 (1975)
102. J.O. Irvin, The generalized Waring distribution. Part II. J. R. Stat. Soc. A **138**, 204–227 (1975)
103. J.O. Irvin, The generalized Waring distribution. Part III. J. R. Stat. Soc. A **138**, 374–384 (1975)
104. A.I. Yablonsky, *Mathematical Models in Science Studies* (Nauka, Moscow, 1986). (in Russian)
105. G.U. Yule, A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Philos. Trans. R. Soc. B **213**, 21–87 (1925)
106. H.A. Simon, C.P. Bonini, The size distribution of business firms. Am. Econ. Rev. **48**, 607–617 (1958)
107. M. Brown, S. Ross, R. Shorrock, Evacualtion of a Yule process with immigration. J. Appl. Probab. **12**, 807–811 (1975)

108. N. O'Connell, Yule process approximation for the skeleton of a branching process. J. Appl. Probab. **30**, 725–729 (1993)
109. D.J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. **16**, 23–34 (2001)
110. S. Redner, How popular is your paper? An empirical study of the citation distribution. Eur. Phys. J. B **4**, 131–134 (1998)
111. S. Redner, Citation statistics from 110 years of physical review. Phys. Today **58**, 49–54 (2005)
112. C.C. Sarli, E.K. Dubinsky, K.L. Holmes, Beyond citation analysis: a model for assessment of research impact. J. Med. Libr. Assoc. **98**, 17–23 (2010)
113. M.Y. Wang, G. Yu, D.R. Yu, Minimg typical features for highly cited papers. Scientometrics **87**, 695–706 (2011)
114. M. Wang, G. Yu, S. An, D. Yu, Discovery of factors influencing citation impact based on a soft fuzzy rough set model. Scientometrics **93**, 635–644 (2012)
115. Q.L. Burrell, Stochastic modeling of the first-citation distribution. Scientometrics **52**, 3–12 (2001)
116. L. Egghe, I.K. Ravichandra Rao, Citation age data and the obsolescence function: fits and explanations. Inf. Process. Manag. **28**, 201–217 (1992)
117. R. Rousseau, Double exponential models for first-citation processes. Scientometrics **30**, 213–227 (1994)
118. L. Egghe, A heuristic study of the first-citation distribution. Scientometrics **48**, 345–359 (2000)
119. D.R. Cox, V.I. Isham, *Point Processes* (Chapman & Hall, London, 1980)
120. J.F.C. Kingman, *Poisson processes* (Clarendon Press, Oxford, 1992)
121. T. Mikosch, *Non-life Insurance Mathematics. An Introduction with the Poisson Process* (Springer, Berlin, 2009)
122. H.C. Tijms, *A First Course in Stochastic Models* (Wiley, Chichester, 2003)
123. S. Nadarajan, S. Kotz, Models for citations behavior. Scientometrics **72**, 291–305 (2007)
124. S.M. Ross, *Stochastic Processes* (Wiley, New York, 1996)
125. Q.L. Burrell, The *n*-th citation distribution and obsolescence. Scientometrics **53**, 309–323 (2002)
126. A.F.J. van Raan, Sleeping beauties in science. Scientometrics **59**, 467–472 (2004)
127. Q.L. Burrell, Are "sleeping beauties" to be expected? Scientometrics **6**, 381–389 (2005)
128. J. Grandell, *Mixed Poisson processes* (Chapman & Hall, London, 1997)
129. S.A. Klugman, H.H. Panjer, G.E. Wilmot, *Loss Models. From Data to Decisions* (Wiley, Hoboken, NJ, 2008)
130. M. Bennet, *Stochastic Processes in Science, Engineering and Finance* (Chapman & Hall, Boca Raton, FL, 2006)
131. R.-D. Reiss, M. Thomas, *Statistical Analysis of Extreme Values* (Birkhäuser, Basel, 1997)
132. Q.L. Burrell, A simple stochastic model for library loans. J. Doc. **36**, 115–132 (1980)
133. Q.L. Burrell, Predictive aspects of some bibliometric processes, in *Infometrics 87/88*, ed. by L. Egghe, R. Rousseau (Amsterdam, Elsevier, 1988), pp. 43–63
134. Q.L. Burrell, Using the gamma-Poisson model to predict library circulation. J. Am. Soc. Inf. Sci. **41**, 164–170 (1990)
135. J.M. Hilbe, *Negative Binomial Regression* (Cambridge University Press, Cambridge, 2007)
136. N.L. Johnson, A.W. Kemp, S. Kotz, *Univariate Discrete Distributions* (Willey, Hoboken, NJ, 2005)
137. J.H. Pollard, *A Handbook of Numerical and Statistical Techniques: With Examples Mainly from the Life Sciences* (Cambridge University Press, Cambridge, 1977)
138. M. Greenwood, G.U. Yule, An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or repeated accidents. J. R. Stat. Soc. A **83**, 255–279 (1920)
139. J. Mingers, Q.L. Burrell, Modeling citation behavior in management science journals. Inf. Process. Manag. **42**, 1451–1464 (2006)

140. E.S. Vieira, J.A.N.F. Gomes, Citation to scientific articles: its distribution and dependence on the article features. J. Inf. **4**, 1–13 (2010)
141. C. Lachance, V. Lariviere, On the citation lyfecycle of papers with delayed recognition. J. Inf. **8**, 863–872 (2014)
142. A.I. Yablonskii, *Models and Methods of Mathematical Study of Science* (AN USSR, Moscow (in Russian), 1977)
143. A. Schubert, W. Glänzel, A dynamic look at a class of skew distributions. A model with scientometric application. Scientometrics **6**, 149–167 (1984)
144. W. Glänzel, A. Schubert, Predictive aspects of a stochastic model for citation processes. Inf. Process. Manag. **31**, 69–80 (1995)
145. R. Frank, Brand choice as a probability process. J. Bus. **35**, 43–56 (1962)
146. J.S. Coleman, *Introduction to Mathematical Sociology* (Collier-Macmillan, London, 1964)
147. H.A. Simon, On a class of skew distribution functions. Biometrica **42**, 425–440 (1955)
148. Y. Ijiri, H. Simon, *Skew Distributions and the Sizes of Business Firms* (North Holland, Amsterdam, 1977)
149. J. Eeckhout, Gibrath's law for (all) cities. Am. Econ. Rev. **94**, 1429–1451 (2004)
150. W. Glänzel, *Bibliometrics as a Research Field: A Course on Theory and Application of Bibliometric Indicators* (Ungarische Akademie der Wissenschaften, Budapest, 2003)
151. W. Glänzel, U. Schoepflin, A stochastic model for the ageing of scientific literature. Scientometrics **30**, 49–64 (1994)
152. S. Shan, G. Yang, L. Jiang, The multivariate Waring distribution and its application. Scientometrics **60**, 523–535 (2004)
153. M. Abramowitz, I.A. Stegun (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1972)
154. Q.L. Burrell, Age-specific citation rates and the Egghe-Rao function. Inf. Process. Manag. **39**, 761–770 (2003)
155. P. Fronczak, A. Fronczak, J.A. Holyst, Publish or perish: Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. Phys. Rev. E **75**, Art. No.026103 (2007)
156. K.G. Zipf, *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949)
157. A.I. Yablonsky, On fundamental regularities of the distribution of scientific productivity. Scientometrics **2**, 3–34 (1980)
158. L. Hartman, Technological forecasting, in *Multinational Corporate Planning*, ed. by G.A. Steiner, W. Cannon (Crowell-Collier Publishing Co., New York, 1966)
159. G.W. Tyler, A thermodynamic model of manpower system. J. Oper. Res. Soc. **40**, 137–139 (1989)
160. I.K. Ravichandra Rao, Probability distributions and inequality measures for analysis of circulation data, in *Informetrics*, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1988), pp. 231–248
161. W. Glänzel, On the *h*-index—A mathematical approach to a new measure of publication activity and citation impact. Scientometrics **67**, 315–321 (2006)
162. E.J. Gumbel, *Statistics of Extremes* (Dover, New York, 2004)
163. W. Glänzel, A. Schubert, Price distribution. An exact formulation of Price's "Square root law". Scientometrics **7**, 211–219 (1985)
164. H. Boxenbaum, F. Pivinski, S.J. Ruberg, Publication rates of pharmaceutical scientists: application of the Waring distribution. Drug Metab. Rev. **18**, 553–571 (1987)
165. Q.L. Burrell, A simple model for linked infometric processes. Inf. Process. Manag. **28**, 637–645 (1992)
166. Q.L. Burrell, Hirsch's *h*-index: a stochastic model. J. Inf. **1**, 16–25 (2007)
167. H.S. Sichel, A bibliometric distribution which really works. J. Am. Soc. Inf. Sci. **36**, 314–321 (1985)
168. H.S. Sichel, Anatomy of the generalized inverse Gaussian-Poisson distribution with special application to bibliometric studies. Inf. Process. Manag. **28**, 5–17 (1992)

169. L. Perreault, B. Bobee, R. Rasmussen, Halphen distribution system. Mathematical and statistical properties. J. Hydrol. Eng. **4**, 189–199 (1999)
170. H.S. Sichel, Repeat-bying and the generalized inverse Gaussian-Poisson distribution. Appl. Stat. **31**, 193–204 (1982)
171. A.K. Romanov, A.I. Terekhov, The mathematical model of productivity—and age-structured scientific community evolution. Scientometrics **39**, 3–17 (1997)
172. A.K. Romanov, A.I. Terekhov, The mathematical model of the scientific personnel movement taking into account the productivity factor. Scientometrics **33**, 221–231 (1995)
173. P. Vinkler, Correlation between the structure of scientific research, scientometric indicators and GDP in EU and non- EU countries. Scientometrics **74**, 237–254 (2008)
174. L.C. Lee, Y.W. Chuang, Y.Y. Lee, Research output and economic productivity: a Granger causality test. Scientometrics **89**, 465–478 (2011)
175. P.W. Hart, J.T. Sommerfeld, Relationship between growth in gross domestic product (GDP) and growth in the chemical engineering literature in five different countries. Scientometrics **42**, 299–311 (1998)
176. F. de Moya-Anegon, V. Herrero Solana, Science in America Latina: a comparison of bibliometric and scientific-technical indicators. Scientometrics **46**, 299–320 (1999)
177. F. Ye, A quantitative relationship between per capita GDP and scientometric criteria. Scientometrics **71**, 407–413 (2007)
178. J. Sylvan Katz, B.R. Martin, What is research collaboration? Res. Policy **26**, 1–18 (1997)
179. A.F.J. van Raan, Science as an international enterprise. Sci. Public Policy **24**, 290–300 (1997)
180. M. Pezzoni, V. Sterzi, F. Lissoni, Career progress in centralized academic systems: Social capital and institutions in France and Italy. Res. Policy **41**, 704–719 (2012)
181. D.B. de Beaver, R. Rosen, Studies in scientific collaboration: Part I-The professional origins of scientific co-authorship. Scientometrics **1**, 65–84 (1979)
182. D.B. de Beaver, R. Rosen, Studies in scientific collaboration: Part II—Scientific co-authorship, research productivity and visibility in the French scientific elite 1799–1830. Scientometrics **1**, 133–149 (1979)
183. D.B. de Beaver, R. Rosen, Studies in scientific collaboration: Part III—Professionalization and the natural history of modern scientific co-authorship. Scientometrics **1**, 231–245 (1979)
184. T. Luukkonen, O. Persson, G. Sivertsen, Understanding patterns of international scientific collaboration. Sci. Technol. Hum. Values **17**, 101–126 (1992)
185. M. Meyar, O. Persson, Nanotechnology—interdisciplinarity, patters of collaboration and differences in application. Scientometrics **42**, 195–205 (1998)
186. A.E. Andersson, O. Persson, Networking scientists. Ann. Reg. Sci. **27**, 11–21 (1993)
187. G. Melin, O. Persson, Hotel cosmopolitan: a bibliometric study of collaboration at some European universities. J. Am. Soc. Inf. Sci. **49**, 43–48 (1998)
188. P. Mählck, O. Persson, Socio-bibliometric mapping of intra-department networks. Scientometrics **49**, 81–91 (2000)
189. T. Lukkonen, R. Tijssen, O. Persson, G. Sivertsen, The measurement of international scientific collaboration. Scientometrics **28**, 15–36 (1993)
190. C.S. Wagner, L. Leydesdorff, Network structure, self-organization, and the growth of international collaboration in science. Res. Policy **34**, 1608–1618 (2005)
191. R. Stichweh, Science in the system of world society. Soc. Sci. Inf. **35**, 327–340 (1996)
192. B. Jamweit, E. Jettestuen, J. Mathiesen, Scaling properties in European research units. PNAS **106**, 13160–13163 (2009)
193. N. Deschacht, T.C.E. Engels, Limited dependent variable models and probabilistic prediction in informetrics, in *Measuring Scholarly Impact. Methods and Practice*, ed. by Y. Ding, R. Rousseau, D. Wolfram (Springer, Cham, 2014), pp. 193–214
194. H.P. Van Dalen, K. Henkens, Signals in science—the importance of signaling in gaining attention in science. Scientometrics **64**, 209–233 (2005)
195. J.W. Fedderke, The objectivity of national research foundation peer review in South Africa assessed against bibliometric indexes. Scientometrics **97**, 177–206 (2013)

196. L. Rokach, M. Kalech, I. Blank, R. Stern, Who is going to win the next Association for the Advancement of Artificial Intelligence fellowship award? Evaluating researchers by mining bibliographic data. J. Am. Soc. Inf. Sci. Technol. **62**, 2456–2470 (2011)
197. P. Jensen, J.-B. Rouquier, Y. Croissant, Testing bibliometric indicators by their prediction of scientists promotions. Scientometrics **78**, 467–47 (2009)
198. P. Vakkari, Internet use increases the odds of using the public library. J. Doc. **68**, 618–638 (2012)
199. T.C.E. Engels, P. Goos, N. Dexters, E.H.J. Spruyt, Group size, *h*-index and efficiency in publishing in top journals explain expert panel assessments of research group quality and productivity. Res. Eval. **22**, 224–236 (2013)
200. S.-C.J. Sin, International coauthorship and citation impact: a bibliometric study of six LIS journals, 1980–2008. J. Am. Soc. Inf. Sci. Technol. **62**, 1770–1783 (2011)
201. A. Abbasi, J. Altmann, L. Hossain, Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. J. Inf. **5**, 594–607 (2011)
202. G.D. Walters, Predicting subsequent citations to articles published in twelve crimepsychology journals: author impact versus journal impact. Scientometrics **69**, 499–510 (2006)
203. L. Bornmann, H.D. Daniel, Selecting scientific excellence through committee peer review—a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. Scientometrics **68**, 427–440 (2006)
204. F. Barjak, S. Robinson, International collaboration, mobility, and team diversity in the life sciences: impact on research performance. Soc. Geogr. **3**, 23–36 (2008)
205. S. Shan, On the generalized Zipf distribution. Part I. Inf. Process. Manag. **41**, 1369–1386 (2005)

Chapter 6
Concluding Remarks

Governments will always play a huge part in solving big problems. They set public policy and are uniquely able to provide the resources to make sure solutions reach everyone who needs them. They also fund basic research, which is a crucial component of the innovation that improves life for everyone.
Bill Gates

Up to a certain level of economic development the production of basic science information does not increase the wealth of an underdeveloped country but on an advanced economic and social level, further development will not be possible without increasing the level of maintenance of fundamental research.
Peter Vinkler [1]

Abstract In this chapter, several concluding remarks are provided about the importance of science for society and about general characteristics of research systems. The importance of statistical laws for research systems is emphasized, and we stress the usefulness of mathematical models and methods for the study and understanding of the dynamics of science and scientific production.

6.1 Science, Society, Public Funding, and Research

Interest in the methodology for assessment (and especially in the methodology for quantitative assessment) of research systems is growing. The reasons for this are the importance of science for society and economics and the wish for effective use of public funds for research. It has been emphasized in Chap. 1 of this book that science is a system of organized knowledge that is a driving force of positive social evolution. Advances in science lead to technological innovation, and because of this, science may be important component for the growth of a country's GDP.

Scientific systems are both social and economic systems. They require specific management and large public investment. The good shape of research facilities and institutions and the high status of national researchers are important conditions for

© Springer International Publishing Switzerland 2016

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1_6

increasing research production and the number of technological innovations. Such investments should ensure a sufficient size of the national research community. This size is very important. If a nation has a scientific or technological problem, then an adequate size of the group of corresponding qualified researchers increases the probability of solving the problem.

Kealey [2] formulated several hypotheses about the research funding. These hypotheses are as follows:

1. The percentage of national GDP spent for research and development increases with national GDP per capita.
2. Public and private funding displace each other.
3. Public and private displacements are not equal: public funds displace more than they themselves provide.

The hypotheses of Kealey are consequence observing the evolution of funding in developed countries where the private funding of research and development (R & D) activities is large. But even in this case, private funding cannot substitute the public funding. Without public funding, developed countries may lose their leading technological position with respect to emerging large economies (some of which use massive public funding of R & D). This displacement may strike the private sector in the corresponding country, and as a consequence, the ability of the private sector to fund R & D may decrease. As a consequence, further displacement of the private sector of the country from world markets may follow.

Public funding of R & D is also extremely important for developing economies, where the ability of the private sector to fund research activities is limited. There are threshold values of many indicators that must be exceeded for successful economic development. One such threshold value is the percentage of GDP spent for R & D. Without sufficient public funding and with very low private funding, this threshold value may not be reached, and the corresponding developing country will remain an economic laggard.

The first hypothesis of Kealey is of limited validity even for developed countries, since the percentage of GDP spent for R & D cannot grow indefinitely. Kealey recognizes this and sets an upper bound of 10 % of GDP. We are far away from this value today (twenty years after Kealey's book). Different factors have already begun to influence spending for R & D. The increase of R & D funding has slowed in many countries. In other countries, one observes cuts in R & D spending. Hence it is not surprising that the economic growth rates have decreased: an important engine of growth does not have enough fuel.

Kealey's hypothesis that government funding of civil R & D disproportionately displaces private funding is quite interesting. If one believes in this hypothesis, then a decrease in public funding should lead to an increase in private funding. This is certainly not the case in developing countries. And even in developed countries, if a private company remains without sufficient public R & D support (and without other kinds of support supplied by the state), then it may soon experience problems with competitors from other countries whose governments support public funding

of R & D. Such public funding of R & D may be very useful for increasing the competitiveness of a nation's private companies.

Research systems are open and dissipative. Thus in order to keep such a system far from equilibrium flows of energy, matter and information must be directed toward the system. These flows ensure the possibility of self-organization, i.e., a sequence of transitions toward states of greater organization. If the above-mentioned flows decrease below some threshold level, then the corresponding dissipative structures can no longer exist, and the system may end at a state of equilibrium (with a great deal of chaos and minimal organization). Thus such a decrease can lead to instabilities and the degradation of corresponding systems.

Instabilities (crises) have an important role in the evolution of science. They may lead to changes in the state of research systems. This change may be positive, but it may also lead to destruction of the corresponding systems. Because of this, one has to be very careful in the management of a research system in the critical regime of instability. Appropriate management requires analysis, forecasting, and finding solutions that can lead to ending the instability. Mathematical modeling and quantitative tools are very important for all of the above. For example, the evolution of research fields and systems may be followed very effectively by constructing knowledge maps and landscapes [3–9].

6.2 Assessment of Research Systems. Indicators and Indexes of Research Production

In addition to knowledge about (i) the importance of science and (ii) the importance of a sufficient amount of knowledge about specific features of research systems, one may need to know about assessment of research systems and about quantitative tools for such assessment. These important topics have been discussed in Chaps. 2 and 3 of the book. The quality of scientific production is important, since scientific information of high quality produced by researchers may be transformed into advanced technology for the production of high-quality goods and services. In order to manage quality, one introduces certain quality management systems (QMS), which are sets of tools for guiding and controlling an organization with respect to aspects of quality: human resources; working procedures, methodologies, and practices; and technology and know-how. In order to understand research systems, one needs to know about their specific statistical features. One such specific feature is that an important difference may exist between the statistical characteristics of processes in nature and those in society. The statistical characteristics of most natural processes are Gaussian, while those of many social processes are non-Gaussian. Because of this, objects and processes in the social sciences usually depend on many more factors than the objects and processes studied in the natural sciences. And research systems are social systems, too.

The need for multifactor analysis becomes obvious when one has the complex task of evaluating the research production of researchers or groups of researchers. The production of researchers has many quantitative and qualitative characteristics. Because of this, one has to use a combination of qualitative and quantitative methods for a successful evaluation of researchers and their production. One should select carefully the sets of indicators, indexes, and tools for evaluation of research production. The principle of Occam's razor is valid also in scientometrics. *The number of indices applied should be the lowest possible, yet it must still be sufficient*. Thus evaluators should apply only those indicators and indexes that are absolutely necessary for the process of evaluation of individual researchers or groups of researchers [1].

Research productivity is closely connected to the communication of the results of research activities. This communication is channelled nowadays in large part through the scientific journals, where the majority of results are published. And most indexes for evaluation have been developed for analysis of research publications (as units of scientific information) and their citations (as units of impact of scientific information). Thus the focus in Chaps. 2 and 3 was on these two groups of indexes and indicators. The characteristics of research productivity that are subject to evaluation usually are latent ones (described by latent variables that are not directly measurable). But by means of systems of indicators and indexes, one may evaluate these latent variables. Usually one needs more than one indicator or index for a good evaluation of a latent variable.

6.3 Frequency and Rank Approaches to Scientific Production. Importance of the Zipf Distribution

Frequency and rank approaches are appropriate for describing the research production of different classes of researchers. The rank approach is appropriate for describing the production of the class of highly productive researchers, in which there are rarely two researchers with the same number of publications/citations, and the ranking may be constructed effectively. The frequency approach is appropriate for a description of the production of less-productive researchers, many of whom have the same number of publications, and because of this, they cannot be effectively ranked. The areas of dominance of the above-mentioned two approaches are different. The frequency approach is dominant in the natural sciences, while the rank approach is more likely to be used in the social sciences. Because of the central limit theorem, the normal distribution plays a central role in the world of Gaussian distributions. Because of the Gnedenko–Doeblin theorem, the Zipf distribution plays an important role in the world of non-Gaussian distributions. Non-Gaussian power-law distributions occur frequently in the area of dynamics of research systems. A consequence of these laws is the concentration–dispersion effect, leading to the fact that *in a research organization, there is usually a small number of highly productive researchers and a large number of less-productive researchers.* Let me stress

again that the laws discussed in Chap. 4 of this book (and the laws of scientometrics in general) must not be regarded as strict rules (such as, e.g., the laws in physics). Instead of this, the above-mentioned laws should be treated as statistical laws (i.e., as laws representing probabilities). Nevertheless, the statistical laws discussed in the book and the corresponding indicators and indices can be used for evaluation and forecasting: it is likely that a researcher's paper with large values of his/her h- and g-indexes will be more frequently cited than a paper by a scientist from the same research field whose values of the h- and g-indices are much lower. It is probable that a paper published in a journal that has a large impact (Garfield) factor will be more frequently cited than a paper on the same subject published in a journal with smaller impact factor.

6.4 Deterministic and Probability Models of Science Dynamics and Research Production

The main focus of this book is on the mathematical tools for assessment of research production, on mathematical modeling of dynamics of research systems, and especially on mathematical models connected to the dynamics of research publications and their citations. Such mathematical models can be deterministic or probabilistic. These two classes of models are discussed in Chap. 5. The deterministic models (e.g., epidemic models, logistic curve models, models of competition between systems of ideas) may be more familiar to the reader. Because of this, Chap. 5 is more focused on probabilistic models. Probabilistic models lead to an explanation of many interesting characteristics connected to the dynamics of research publications and their citations. For example, one can prove the (intuitive) fact that there are publications that will never be cited. Many well-known heavy-tail and other statistical distributions such as the Yule distribution, Waring distribution, negative binomial distribution, and rare event distributions such as the Gumbel distribution, Weibull distribution, etc., are used in these models to describe production/citation dynamics, aging of scientific information, etc. In addition to the statistical laws, two kinds of (Matthew) effects connected to citation information are described. The first effect is that researchers (or journals) that have a relatively high standard may obtain more citations than deserved. This effect is accompanied by a second effect, known as the "invitation paradox": many papers published in journals with a high impact factor are cited less frequently than expected on the basis of the journal's impact factor. Thus "for many are called, but few are chosen" (second Matthew effect).

Let us note that there are many more models connected to dynamics of science and technology [10–12]. Some of these models are evolutionary models [13–16]. In general, the models of science dynamics and technology are some of the mathematical tools, and models connected to social dynamics (for several references, see [17–40]), which is a rapidly growing research area drawing the attention of an increasing number of researchers.

6.5 Remarks on Application of Mathematics

Mathematics is used for the quantification of research structures, processes, and systems [41–44]. A large field of research is concerned with the application of mathematical models and statistics to research and to quantify the process of written communication. This field of research is covered by bibliometrics [45, 46]. Bibliometrics is used not only in the area of research evaluation. Methods of bibliometrics are applied, for example, to the investigation of the emergence of new disciplines, the study of interactions between science and technology, and the development of indicators that can be used for planning and evaluation of different aspects of scientific activity [47].

One has to be careful in the use of methods of bibliometrics for research evaluation, since these methods are based on the assumption that carrying out research and communicating the results go hand in hand. This assumption is not true in all cases, e.g., research for military purposes. An additional assumption is that publications can be taken to represent the output of science. This assumption is not true in all cases, e.g., in the case of research for the needs of large corporations, since a significant part of such research is not published. But in the cases in which the assumption holds, the arrays of publications can be quantified and analyzed to study trends of development in science (national, global, etc.) as well as to study the production of scientific groups and institutions.

Mathematical tools are also used in citation analysis. The analysis of citations, however, is not connected only to mathematics. There exist also qualitative aspects such as quality, importance, and the impact of citations on research publications. The quality of a citation is an inherent property of the research work [48]. Judgment of quality can be made only by peers who can evaluate cognitive, technological, and other aspects connected to the scientific work and to the place of the citation in the work. The importance of a citation is based on external appraisal [49]. Importance refers to the potential influence on surrounding research activities. We note that self-citations do not have an external appraisal. Because of this, they are not as important as other citations and are usually excluded from the citation analysis of an evaluated scientist, research group, or organization. Finally, the impact of a citation is also based on external appraisal. The impact of citations reflects their actual influence. A citation reflects to some extent the influence of the cited source on the research community. We note here that review articles are generally more frequently cited than regular research articles. In addition, numbers of citations differ across different areas of scientific research. The impact of citations may be measured by different indicators. Such indicators are, for example, number of citations for the corresponding paper, average number of citations per paper (this measures the impact of the corresponding scientist), number of citations of a paper for the past few (three, four, five, or more) years, age distribution of the citations of the corresponding article, etc. Let us note that citation analysis has other interesting aspects [50, 51], e.g., cocitations [52–55] (which can be visualized by the Jaccard index or Salton's cosine [56]). Cocitation analysis may also be used for visualization of scientific disciplines [57], for detection

of research fronts [58], or even as a measure of intellectual structure in a group of researchers [59].

Another field of mathematics that has been much used in recent years in studies on research systems is graph theory and the associated theory of networks [60]. Methods such as mapping and clustering are used for processing citation and cocitation networks, coauthorship networks, and other bibliometric networks [61–63], and corresponding software such as Gephi, Pajec, Sci2 [64–68] is used for visualization of these networks. In more detail, one may study the organization of large research systems on the basis of the information contained in the nodes and links of the corresponding large networks. There are community-detection methods [69, 70], that reveal important structures (e.g., strongly interconnected modules that often correspond to important functional units) in networks. One such method is the map equation method [71]. Let us consider a network on which a network partition is performed (say the n nodes of the network are grouped into m modules). The map equation specifies the theoretical modular description length $L(M)$ of how concisely we can describe the trajectory of a random walker guided by the possibly weighted directed links of the network. Here M denotes a network partition of the n network nodes into m modules, with each node assigned to a module. The description length $L(M)$ given by the map equation is then minimized over possible network partitions M. The network partition that gives the shortest description length best captures the community structure of the network with respect to the dynamics on the network. The map equation framework is able to capture easily citation flow or flow of ideas, because it operates on the flow induced by the links of the network. Because of this, the map equation method is suitable for analysis of bibliometric networks.

Finally, let us note that an entire research area exists called computational and mathematical organization theory. Researchers working in this area focus on developing and testing organizational theory using formal models [72–74]. The models of this theory can be very useful for managers and evaluators of research organizations. Let us mention several areas that employ such models:

1. Innovation diffusion from the point of view of complex systems theory [75];
2. Public funding of nanotechnology [76];
3. Technology innovation alliances and knowledge transfer [77];
4. Attitude change in large organizations [78];
5. Complexity of project dynamics [79];
6. Corruption in education organizations [80];
7. Reputation and meeting techniques for support of collaboration [81];
8. Spreading of behavior in organizations [82];
9. Communication and organizational social networks [83];
10. Politics [84].

6.6 Several Very Final Remarks

Not everything that counts can be counted, and
not everything that can be counted counts.
Albert Einstein

It is time to end our journey through the huge area of evolution of research systems and assessment of research production. There were two competing concepts as this book was being planned: (i) the concept of a scientific monograph and (ii) the concept of an introductory book with elements of a handbook. The first variant would lead to a book twice as big as it is now. Mathematical theorems would be proved there, indexes and indicators would be discussed in much greater detail, and larger sets of topics would be described. Such a book would meet the expectations of the members of group 3 of potential readers mentioned in the preface. But I wanted to write a book for a much larger set of readers: these from the target groups 1 and 2 from the preface. Because of this, the second concept was realized. The introductory character of the book allowed me to concentrate the text around science dynamics and assessment of important elements of research production. The aspect of a handbook allowed me to describe many indexes and models in a small number of pages. Of course, the realization of the concept of introductory text with the aspect of a handbook led to the fact that many topics from the area of research on have been not discussed. I have not discussed important questions such as how researchers choose the list of references for their publications: What is the motivation to cite some publications and not others? Are there reference standards? Can scientific information be institutionalized? And so on. Instead of this, the focus was set on mathematical tools and models. In addition, some indexes and models have been presented very briefly. This is compensated by a sufficient number of warning messages about the proper use of indexes; by the large number of references, where the reader will find additional information; and by clear statements about the condition of validity of the models discussed. There are numerous examples of calculation of indexes, and many more examples could be (i) provided on the basis of the excellent databases available and (ii) found in the lists of references by the interested reader. My experience shows that the shortest way to become familiar with the indexes and with the conditions for their proper application is to calculate them oneself. So my advice to the reader is to perform many such calculations in order to gain experience about the proper and improper application of the indexes. Many years ago (when I was much younger), I needed about an year of practice before I could begin to apply the quantities and tools of nonlinear time series analysis in a proper way. So be patient, carry out a large enough number of exercises, and the results will come.

This is an introductory book, and the introduction has been made from the point of view of mathematics. Once Paul Dirac said, *If there is a God, he's a great mathematician*. The achievements of the mathematical theory of research systems are very useful, for science dynamics and research production have quantitative characteristics, and knowledge about those characteristics may help evaluators to perform appropriate assessment of researchers, research groups, research organizations, and

systems. One of Plato's ideas was that a good decision is based on knowledge (and not only on numbers). I hope that this book may help the reader to understand better the processes and structures connected to the dynamics of science and research production. This may lead to better assessment and management of research structures and systems as well as to increased productivity of researchers. If this book contributes to an increased understanding of complex science dynamics and to better assessment of research even in a single country and even in a small number of research groups in that country, I will be happy, and the goal of the book will have been achieved.

References

1. P. Vinkler, *The Evaluation of Research by Scientometric Indicators* (Chandos, Oxford, 2010)
2. T. Kealey, *The Economic Laws of Scientific Research* (Macmillan, Houndmills, 1996)
3. A. Scharnhorst. Constructing knowledge landscapes within the framework of geometrically oriented evolutionary theories, in I*ntegrative Systems Approaches to Natural and Social Dynamics* ed. by M. Matthies, H. Malchow, J. Kriz (Springer, Berlin, 2001) pp. 505–515
4. R. Klavans, K.W. Boyack, Using global mapping to create more accurate document-level maps of research fields. J. Am. Soc. Inform. Sci. Technol. **62**, 1–18 (2011)
5. K.W. Boyack, R. Klavans, K. Börner, Mapping the backbone of science. Scientometrics **64**, 351–374 (2005)
6. R.M. Shiffrin, K. Börner, Mapping knowledge domains. PNAS **101**, 5183–5185 (2004)
7. K. Börner, L. Dall'Asta, W. Ke, A. Vespignani, Studying the emerging global brain: analyzing and visualizing the impact of co-authorship teams. Complexity **10**, 57–67 (2005)
8. A. Skupin, A cartographic approach to visualizing conference abstracts. Comput. Graphics Appl. **22**, 50–58 (2002)
9. D. Hakken, *The Knowledge Landscapes of Cyberspace* (Routledge, London, 2004)
10. E. Bruckner, W. Ebeling, A. Scharnhorst, The application of evolution models in scientometrics. Scientometrics **18**, 21–41 (1990)
11. E. Bruckner, W. Ebeling, M.A. Jimenez Montano, A. Scharnhorst. Hyperselection and innovation described by a stochastic model of technological evolution, in *Evolutionary Economics and Chaos Theory. New directions in Technology Studies* ed. by L. Leydesdorff, P. van den Besselaar (St. Martin's Press, 1994) pp. 79–90
12. E. Bruckner, W. Ebeling, M.A. Jimenez, Montano, A. Scharnhorst. Nonlinear stochastic effects of substitution—an evolutionary approach. J. Evol. Econ. **6**, 1–30 (1996)
13. E. Bruckner, W. Ebeling, A. Scharnhorst, Stochastic dynamics of instabilities in evolutionary systems. Sys. Dyn. Rev. **5**, 176–191 (1989)
14. W. Ebeling, Karmeshu, A. Scharnhorst. Dynamics of economic and technological search processes in complex adaptive landscapes. Adv. Complex Syst. **4**, 71–88 (2001)
15. W. Ebeling, A. Scharnhorst, Selforganization models for field mobility of physicists. Czech J. Phys. **36**, 43–46 (1986)
16. W. Ebeling, A. Scharnhorst, M.A.J. Montano, Karmeshchu. Evolutions-und Innovationsdynamik als Suchprozeß in *Komplexe Systeme und Nichtlineare Dynamik in Natur und Gesellschaft: Komplexitätsforschung in Deutschland auf dem Weg ins nächste Jahrhundert* ed. by K. Mainzer (Springer, Berlin, 1999)
17. D. Helbing, *Quantitative Sociodynamics: Stochastic Methods and Models of Social Interaction Processes* (Springer, Berlin, 2010)
18. F. Schweitzer, *Brownian Agents and Active Particles: Collective Dynamics in the Natural and Social Sciences* (Springer, Berlin, 2003)
19. F. Schweitzer (ed.), *Self-Organization of Complex Structures: From Individual to Collective Dynamics* (Gordon and Breach, Australia, 1997)

20. B. Skyrms, *Social Dynamics* (Oxford University Press, Oxford, 2014)
21. M. Matthies, H. Malchow, J. Kriz (eds.), *Integrative Systems Approaches to Natural and Social Dynamics* (Springer, Berlin, 2001)
22. J. Klüver. The dynamics and evolution of social systems, in *New Foundations of a Mathematical Sociology* (Kluwer, Dordrecht, 2000)
23. N.B. Tuma, M.T. Hannan. Social Dynamics. Models and Methods (Academic Press, Orlando, FL, 1984)
24. A. Bejan, W. Merkx, *Constructal Theory of Social Dynamics* (Springer, New York, 2007)
25. G. Naldi, L. Pareshi, G. Toscani (eds.), *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences* (Springer, New York, 2010)
26. P. Doreian, F.N. Stokman (eds.), *Evolution of Social Networks* (Routledge, Amsterdam, 2013)
27. S. de Marchi, *Computational and Mathematical Modeling in the Social Sciences* (Cambridge Iniversity Press, Cambridge, 2005)
28. G.A. Marsan, N. Bellomo, A. Tosin. *Complex Systems and Society. Modeling and Simulation*. (Springer, Berlin, 2013)
29. N. Bellomo. *Modeling Complex Living Systems. A Kinetic Theory and Stochastic Game Approach*. (Birkhäuser, Boston, 2008)
30. J. Lorenz, H. Rauhut, F. Schweitzer, D. Helbing, How social influence can undermine the wisdom of crowd effect. PNAS **108**, 9020–9025 (2011)
31. L.M.A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G.B. West, Growth, innovation, scaling, and the pace of life in cities. PNAS **104**, 7301–7306 (2007)
32. J.A. Holyst, K. Kacperski, F. Schweitzer. Social impact models of opinion dynamics, in *Annual reviews of Computationsl Physics IX* ed. by D. Stauffer (World Scientific, Singapore, 2001)
33. D. Helbing, P. Molnar, Social force model for pedestrian dynamics. Phys. Rev. E **51**, 4282–4286 (1995)
34. D. Helbing, *Verkehrsdynamik: neue physikalische Modellierungskonzepte* (Springer, Berlin, 2013)
35. D. Helbing, J. Keltsch, P. Molnar, Modelling the evolution of human trail systems. Nature **388**(6637), 47–50 (1997)
36. F. Schweitzer, J. Steinbrink, Estimation of megacity growth: simple rules versus complex phenomena. Appl. Geogr. **18**, 69–81 (1998)
37. F. Schweitzer, W. Ebeling, H. Rose, O. Weiss, Optimization of road networks using evolutionary strategies. Evol. Comput. **5**, 419–438 (1997)
38. F. Schweitzer (ed.), *Modeling Complexity in Economic and Social Systems* (World Scientific, Singapore, 2002)
39. F. Schweitzer, R. Mach. The epidemics of donations: logistic growth and power-laws. PLos One **3**, e 1458 (2008)
40. F. Schweitzer, L. Behera, Nonlinear voter models: the transition from invasion to coexistence. Eur. J. Phys. B **67**, 301–318 (2009)
41. D. Lucio-Arias, A. Scharnhorst. Mathematical approaches to modeling science from an algorithmic-historiography perspective, in *Models of Science Dynamics* ed. by A. Scharnhorst, K.Börner, P. van den Besselaar (Springer, Berlin, 2012), pp. 23–66
42. D. Crouch, J. Irvine, B.R. Martin. Bibliometric analysis for science policy: an evaluation of the United Kingdom's research performance in ocean currents and protein crystallography. Scientometrics **9**, 239–267 (1986)
43. N. Payette. Agent-based models of science, in *Models of Science Dynamics* ed. by A. Scharnhorst, K. Börner, P. van den Besselaar (Springer, Berlin, 2012) pp. 127–157
44. M. Hanauske. Evolutionary game theory and complex network of scientific information. *Models of Science Dynamics* ed. by A. Scharnhorst, K. Börner, P. van den Besselaar (Springer, Berlin, 2012), pp. 159–191
45. J.M. Russel, R. Rousseau. Bibliometrics and institutional evaluation, in *Encyclopedia of Life Support Systems (EOLSS). Part 19.3: Science and Technology Policy* ed. by In R. Arvantis (EOLSS Publishers, Oxford, UK, 2002), pp. 1–20

46. D.J. de Solla, Price. A general theory of bibliometric and other cumulative advantage processes. J. Am. Soc. Inform. Sci. **27**, 292–306 (1976)
47. J. Enders, R. Whitley, J. Glser (eds.), *The Changing Governance of the Sciences. The Advent of Research Evaluation Systems. Sociology of the Sciences Yearbook* (Springer, Dordrecht, 2007)
48. L. Leydesdorff, L. Bornmann, R. Mutz, T. Opthof, Turning the tables on citation analysis one more time: principles for comparing sets of documents. J. Am. Soc. Inform. Sci. Technol. **62**, 1370–1381 (2011)
49. O. Amsterdamska, L. Leydesdorff, Citations: indicators of significance? Scientometrics **15**, 449–471 (1989)
50. E.C.M. Noyons, H.F. Moed, M. Luwel, Combining mapping and citation analysis for evaluative bibliometric purposes: a bibliometric study. J. Am. Soc. Inform. Sci. **50**, 115–131 (1999)
51. C. Oppenheim, S.P. Renn, Highly cited old papers and the reasons why they continue to be cited. J. Am. Soc. Inform. Sci. **29**, 225–231 (1978)
52. H. Small, E. Sweeney, Clustering the science citation index using co-citations i: a comparison of methods. Scientometrics **7**, 391–409 (1985)
53. H. Small, E. Sweeney, E. Greenlee, Clustering the science citation index using co-citations. ii: mapping science. Scientometrics **8**, 321–340 (1985)
54. R. Rousseau, A. Zuccala, A classification of author co-citations: definitions and search strategies. J. Am. Soc. Inform. Sci. Technol. **55**, 513–529 (2004)
55. H. Small, Macro-level changes in the structure of co-citation clusters: 1983–1989. Scientometrics **26**, 5–20 (1993)
56. L. Leydesdorff, On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. J. Am. Soc. Inform. Sci. Technol. **59**, 77–85 (2008)
57. H.D. White, K.W. McCain, Vizualizing a discipline: an author co-citation analysis of information science, 1972–1995. J. Am. Soc. Inform. Sci. **49**, 327–355 (1998)
58. M. Zitt, E. Bassecoulard, Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. Scientometrics **30**, 333–351 (1994)
59. H.D. White, B.C. Griffith, Author cocitation: a literature measure of intellectual structure. J. Am. Soc. Inform. Sci. **32**, 163–171 (1981)
60. Y. Ding, R. Rousseau, D. Wolfram (eds.), *Measuring Scholarly Impact Methods and Practice* (Springer, Chaim, 2014)
61. N.J. van Eck, L. Waltman, Vizualizing bibliometric networks, in *Measuring Scholarly Impact Methods and Practice* ed. by Y. Ding, R. Rousseau, D. Wolfram (Springer, Chaim, 2014), pp. 285–320
62. K. Börner, *Atlas of Science: Visualizing What We Know* (MIT Press, Cambridge, MA, 2010)
63. K. Börner, C. Chen, K.W. Boyack, Visualizing knowledge domains. Annu. Rev. Inform. Sci. Technol. **37**(1), 179–255 (2003)
64. W.D. Nooy, A. Mrvar, Y.V. Batageli, *Exploratory Social Network Analysis with Pajek*, 2nd edn. (Cambridge University Press, Cambridge, 2011)
65. M. Bastian, S. Heymann, M. Jacomy. Gephi: An open source software for exploring and manipulating networks. in *Proceedings of the Third International ICWSM Conference (2009)*, pp. 361–362
66. Sci2 Team. Science of Science (Sci2) Tool: Indiana University and SciTech Strategies (2009), http://sci2.cns.iu.edu
67. K. Börner, D.E. Polley. Replicable science of science, in *Measuring Scholarly Impact Methods and Practic*, ed. by Y. Ding, R. Rousseau, D. Wolfram (Springer, Chaim, 2014), pp. 321–341
68. N.J. van Eck, L. Waltman, Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**, 523–538 (2010)
69. M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure. Proc. Nat. Acad. Sci. **105**, 1118–1123 (2008)
70. M. Rosvall, C.T. Bergstrom, Mapping change in large networks. PLoS ONE **5**, e8694 (2010)
71. L. Bohlin, D. Edler, A. Lancichinetti, M. Rosval, Community detection and visualization of networks with the map equation framework, in *Measuring Scholarly Impact* ed. by Y. Ding, R. Rousseau, D. Wolfram (Springer, Chaim, 2014)

72. K.M. Carley, Computational and mathematical organization theory: Perspective and directions. Comput. Math. Organ. Theory **1**, 39–56 (1995)
73. K.J. Arrow, R. Radner, Allocation of resources in large teams. Econometrica **47**, 361–385 (1979)
74. A.W. Bausch, Evolving intergroup cooperation. Comput. Math. Organ. Theory **20**, 369–393 (2014)
75. N. Nan, R. Zmund, E. Yatgin, A complex adaptive systems perspective of innovation diffusion: an integrated theory and validated virtual laboratory. Comput. Math. Organ. Theory **20**, 52–88 (2014)
76. N. Hoser, Public funding in the academic field of nanotechnology: a multi-agent based model. Comput. Math. Organ. Theory **19**, 253–281 (2013)
77. Z.-S. Jiang, Y.-H. Hao, Game analysis of technology innovation alliance stability based on knowledge transfer. Comput. Math. Organ. Theory **19**, 403–421 (2013)
78. L.A. Costa, J.A. de Matos, Attitude change in arbitrary large organizations. Comput. Math. Organ. Theory **20**, 219–251 (2014)
79. C.M. Schlick, S. Duckwitz, S. Schneider, Project dynamics and emergent complexity. Comput. Math. Organ. Theory **19**, 480–515 (2013)
80. A.L. Osipian, Corrupt organizations: modeling educators' misconduct with cellular automata. Comput. Math. Organ. Theory **19**, 1–24 (2013)
81. K. Hansson, P. Karlström, A. Larsson, H. Verhagen, Reputation, inequality and meeting techniques: visualising user hierarchy to support collaboration. Comput. Math. Organ. Theory **20**, 155–175 (2014)
82. Y. Zhang, Y. Wu, How behaviors spread in dynamic social networks. Comput. Math. Organ. Theory **18**, 419–444 (2012)
83. L. Chen, G.G. Gable, H. Hu, Communication and organizational social networks: a simulation model. Comput. Math. Organ. Theory **19**, 460–479 (2013)
84. C. Cioffi-Revilla, Simplicity and reality in computational modeling of politics. Comput. Math. Organ. Theory **15**, 26–46 (2009)

Index

© Springer International Publishing Switzerland 2016

N.K. Vitanov, *Science Dynamics and Research Production*, Qualitative and Quantitative Analysis of Scientific and Scholarly Communication,
DOI 10.1007/978-3-319-41631-1