LNCS 9860

Journal Subline

Amin Anjomshoaa • Patrick C.K. Hung
Dominik Kalisch • Stanislav Sobolevsky
Guest Editors

Transactions on Large-Scale Data- and Knowledge-Centered Systems XXVII

Abdelkader Hameurlain • Josef Küng • Roland Wagner
Editors-in-Chief

Special Issue on Big Data for Complex Urban Systems

Springer

Lecture Notes in Computer Science 9860

More information about this series at http://www.springer.com/series/8637

Abdelkader Hameurlain · Josef Küng
Roland Wagner · Amin Anjomshoaa
Patrick C.K. Hung · Dominik Kalisch
Stanislav Sobolevsky (Eds.)

Transactions on Large-Scale Data- and Knowledge-Centered Systems XXVII

Special Issue on Big Data for Complex Urban Systems

Springer

Editors-in-Chief

Abdelkader Hameurlain
IRIT
Paul Sabatier University
Toulouse
France

Josef Küng
FAW
University of Linz
Linz
Austria

Roland Wagner
FAW
University of Linz
Linz
Austria

Guest Editors

Amin Anjomshoaa
MIT Senseable City Lab
Cambridge, MA
USA

Patrick C.K. Hung
Faculty of Business and Information Technology
University of Ontario Institute of Technology (UOIT)
Oshawa, ON
Canada

Dominik Kalisch
Trinity University
Plainview, TX
USA

Stanislav Sobolevsky
New York University
Brooklyn, NY
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-662-53415-1 ISBN 978-3-662-53416-8 (eBook)
DOI 10.1007/978-3-662-53416-8

Library of Congress Control Number: 2016950413

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Germany
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

Editorial Preface

Living in cities is becoming increasingly attractive for many people around the world. According to the United Nations, more than 3.8 billion or 53.6 % of the world's population were living in urban agglomerations in 2014. Especially from an ecological point of view, cities are a central issue for the future. Cities consume enormous amounts of energy, raw materials, and space, additionally producing tons of waste and hazardous materials, while many places suffer from congestion, traffic jams, crime, etc.

Today's cities are using systems and infrastructure that are partly based on outdated technologies, making them unsustainable, inflexible, inefficient, and difficult to change. In addition, the increasing pace of urbanization and transformation of the cities challenges traditional approaches for urban system forecasting, policy, and decision-making even further. In order to solve these challenges, we have to understand cities as hyper-complex interdependent systems that, with their interconnected layers and subsystems, cannot be efficiently understood separately from one another, but form a complex interdependent system of infrastructural, economic, and social components that require a holistic system model.

On the other hand, modern challenges in complex urban system studies come together with new unprecedented opportunities, such as digital sensing. The technological revolution resulted in the broad penetration of digital technologies in the everyday life of people and cities, creating big data records of human behavior. Also, recent advances in network science allow for deeper interactions between people, companies, and urban infrastructure from the new complex network perspective.

There is already a modern trend in urban planning to use the data that are available to improve quality of life, reduce costs, and objectify planning decisions. This is especially true for many cities — like Chicago or New York — which have begun to roll out urban sensor data for managing the city. Data, analytics, and technology are therefore the keys to making these data not only accessible, but to gain meaningful insights into urban systems to understand the city, allow evidence-based decisions, and create sustainable solutions and innovations improving the quality of urban life.

However, the high complexity of modern urban systems creates a challenge for the data and analytic methods used to study them, calling for newer approaches that are more unified, robust, and efficient.

The goal of this proposed special issue is to delineate important research milestones and challenges of big data-driven studies of the complex urban systems, discussing applicable data sources, methodology, and their current limitations.

This special issue contains 12 papers that contribute in-depth research of the subject. The results of these papers were presented at the symposium Big Data and Technology for Complex Urban Systems held during the 49th Hawaii International Conference in System Sciences on January 5, 2016.

The first contribution is "Brazilians Divided: Political Protests as Told by Twitter" by Souza Carvalho et al. This paper presents two learning algorithms to classify tweets in Twitter for an exploratory analysis so as to acquire insights of the inner divisions and their dynamics in the pro- and anti-government protests in the Brazilian presidential election campaign in 2014. The results show that there are slightly different behaviors from both sides, in which the pro-government users criticized the opposing arguments prior to the event, whereas the group against the government generated attacks during different times, as a response to supporters of the government.

Next, the second contribution “Sake Selection Support Application for Countryside Tourism” by Iijamai et al. discusses a study to investigate a way of attracting foreign tourists to participate in “Sake Brewery Tours” for the Tokyo Olympic Paralympic Games in 2020. This paper demonstrates a related application to engage foreign tourists who are not originally interested in sake.

The following contribution by Kalisch et al. is “A Holistic Approach to Understand Urban Complexity” and gives an introduction to the interdependent complexity of urban systems, addressing necessity for research in this field. Based on an industry-funded qualitative research project, the paper outlines a holistic approach to understanding urban complexity. The goal of this project was to understand the city in a holistic way, applying the approach of system engineering to the field of urban development, as well as to identify the key factors needed to redesign existing and newly emerging cities in a more sustainable way. The authors describe the approach and share a summary of a case study analysis of New York City.

The contribution entitled “Real-Time Data Collection and Processing of Utility Customer’s Power Usage for Improved Demand Response Control,” by Shawyun Sariri et al., investigates potential demand response solutions that provide cost-effective alternatives to high priced spinning reserves and energy storage. The context of the study focuses on the implementation of a pilot program, which aids in the understanding of large data collection in dense urban environments. Understanding the power consumption behavior of a consumer is key in implementing efficient demand response programs. Factors affecting large data collection such as infrastructure, data storage, and security are also explored.

The paper “Development of a Measurement Scale for User Satisfaction with E-Tax Systems in Australia” by A. Alghamdi and M. Rahim explores satisfaction of e-government systems in general and e-tax systems in particular. The paper develops a satisfaction construct of such e-tax systems and evaluates the approach in two steps. The conceptual model construct is being evaluated by an expert panel, and there is also a pilot evaluation of the survey instrument developed based on that model. The authors present the first overview of factors that are important for user satisfaction with e-tax systems.

The next two papers focus on the creation of open government data (OGD) resources. The first OGD contribution, entitled “Data-Driven Governments: Creating Value Through Open Government Data” by Judie Attard et al., explores existing processes of value creation on government data. The paper identifies the dimensions that impact, or are impacted by, value creation and distinguishes between the different value-creating roles and participating stakeholders. The authors propose the use of linked data as an approach to enhance the value creation process and provide a value creation assessment framework to analyze the resulting impact. They also implement the assessment framework to evaluate two government data portals.

The second OGD contribution, entitled "Collaborative Construction of an Open Official Gazette" by Gisele S. Craveiro et al., aims at describing the strategies adopted for preparing the implementation of an open official gazette at the municipal level. The proposed approach is a combination of bibliographical review, documentary research, and direct observation. The paper also describes the strategies and activities put into effect by a public body and an academic group in preparing the implementation of the open official gazette and analyzes the outcomes of these strategies and activities by examining the tool implemented, the traffic, and the reported uses of the open Gazette.

The next contribution, entitled "A Solution to Visualize Open Urban Data for Illegally Parked Bicycles" by Shusaku Egami et al., presents a crowd-powered open data solution for the illegal parking of bicycles in urban areas. This study proposes an ecosystem that generates open urban data in link data format by socially collecting the data, complementing the missing data, and then visualizing the data to facilitate and raise social awareness about the problem.

The contribution, entitled "An Intelligent Hot-Desking Model Based on Occupancy Sensor Data and Its Potential for Social Impact" by Konstantinos Maraslis et al., proposes a model that utilizes occupancy sensor data in a commercial hot-desking environments. The authors show that sensor data can be used to facilitate office resource management with results that outweigh the costs of occupancy detection. The paper shows that the desk utilization can be optimized based on quality occupancy data and also demonstrates the effectiveness of the model by comparing it with a theoretically ideal, but impractical real-life model.

The following contribution, "Characterization of Behavioral Patterns Exploiting Description of Geographical Areas" by Zolzaya Dashdorj et al., investigates relationships existing between human behavior measured through mobile phone data records on one hand, and location context, measured through the presence of points of interest of different categories, on the other. Advanced machine-learning techniques are used to predict a timeline type of communication activity in a given location based on the knowledge of its context, and it is demonstrated that the classification based on point-of-interest data has additional predictive power compared with the official data, such as the land use classification.

The contribution "Analysis of Customers' Spatial Distribution Through Transaction Datasets" by Yuji Yoshimura et al. studies people's consumption behavior and specifically customer mobility between retail stores, using a large-scale anonymized dataset of bank card transactions in Spain. Various spatial patterns of customer behavior are discovered, including spatial distributions of customer activity with respect to the distance from the considered store.

The last contribution, "Case Studies for Data-Driven Emergency Management/Planning in Complex Urban Systems" by Kun Xie et al., considers five related case studies within the New York/New Jersey metropolitan area in order to present a comprehensive overview on how to use big urban data (including traffic operations, incidents, geographical and socio economic characteristics, and evacuee behavior) to obtain innovative solutions for emergency management and planning, in the context of complex urban systems. Useful insights are obtained from the data for essential tasks of emergency management and planning such as evacuation demand estimation, determination of evacuation zones, evacuation planning, and resilience assessment.

July 2016

Amin Anjomshoaa
Patrick C.K. Hung
Dominik Kalisch
Stanislav Sobolevsky

Organization

Editorial Board

External Reviewers

Mohammed Al-Kateb	Teradata, USA

Contents

Brazilians Divided: Political Protests as Told by Twitter

Cássia de Souza Carvalho[1], Fabrício Olivetti de França[1,3](✉),
Denise Hideko Goya[1,3], and Claudio Luis de Camargo Penteado[2,3]

[1] Center of Mathematics, Computing and Cognition (CMCC),
Federal University of ABC (UFABC), Santo André, SP, Brazil
cassia.carvalho@aluno.ufabc.edu.br,
{folivetti,denise.goya}@ufabc.edu.br
[2] Center of Engineering, Modeling and Applied Social Sciences (CECS),
Federal University of ABC (UFABC), São Bernardo do Campo, Brazil
claudio.penteado@ufabc.edu.br
[3] Nuvem Research Strategic Unit, Santo André, Brazil

Abstract. After a fierce presidential election campaign in 2014, the reelected president Dilma Rousseff became a target of protests in 2015 asking for her impeachment. This sentiment of dissatisfaction was fomented by the tight results between the two favorite runners-up and the accusations of corruption in the media. Two main protests in March were organized and largely reported with the use of Social Networks like Twitter: one pro-government and other against it, separated by two days. In this work, we apply two supervised learning algorithms to automatically classify tweets during the protests and to perform an exploratory analysis to acquire insights of their inner divisions and their dynamics. Furthermore, we can identify a slightly different behavior from both parts: while the pro-government users criticized the opposing arguments prior the event, the group against the government generated attacked during different times, as a response to supporters of government.

1 Introduction

In democratic elections, whenever the results are tight, the competing sides tend to express a negative sentiment towards each other, inciting a polarization among people. When this sentiment is accompanied by doubts about the legitimacy of voting system, it may influence a wave of protests and calls for a change of rules.

This situation occurred in the Brazilian presidential election of 2014, in which the two main candidates, Dilma Rousseff, representing the Workers' Party, and Party, and Aécio Neves, representing the Brazilian Social Democracy Party, obtained a result of 51.64 % and 48.36 % of votes respectively. These results, together with the spread of news about internal corruption in one of the largest semi-public multinational corporation, influenced the people from the opposing side to organize a series of protests.

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 1–18, 2016.
DOI: 10.1007/978-3-662-53416-8_1

These protests occurred inside their homes, on the streets [21] and throughout the two main social networks: Facebook[1] and Twitter[2]. These Social Networks played an important role for the organization and discussions of such protests.

With the widespread use of the Social Networks, it is possible to extract different information about these events. For the government and opposition sides, it is important to know who are the main actors of these events, the overall sentiments, the demands and the different parts that gathered for a common goal.

In this paper, we apply two classification algorithms [2] to determine the overall sentiment of the protesters on the events that occurred during the period of 13th and 15th of March 2015. The first event (13th of March) was organized by pro-government groups, while the second (15th of March) was organized by groups against government. We explore what information we can infer from the classes by plotting the temporal relations. Despite the usual literature on Sentiment Mining [9], we will label the sentiments pro or against the government.

The paper is organized as follows: In Sect. 2 we contextualize these two political protests to better understand the overall sentiment of both sides. In Sect. 3 we explain the two classification algorithms used in this work: Naive Bayes [12] and Support Vector Machine [17], as well briefly summarize some works found in the literature of twitter sentiment analysis, particularly focusing on political context. In Sect. 4 we explain the methodology and apply these two algorithms in our collected dataset and to analyze the information that can be extracted from the results. Finally, in Sect. 5 we conclude this paper with some insights for future work.

2 Brazilian Political Protests

After a polarized campaign between the two candidates, the president Dilma Rousseff was re-elected as President of Brazil by a small margin of votes, 3,459,963 (roughly 3.28 % of the electors). The presidential campaign of 2014 was marked by intense debates between the candidates since the first round, motivating supporters and militants to produce favorable information for their candidates in the Internet Social Networks.

Disagreeing with the loss of the candidate Aécio Neves, their supporters and groups opposed to the Workers' Party manifested their unhappiness on the Internet, maintaining an intense online political mobilization. As a result from this articulation, groups against the government organized via digital media (Facebook, Twitter, WhatsApp[3]) a protest that was known as *Panelaço* (pan beating). During the initial statement of president Dilma Rousseff in national broadcast on 8th of March 2015, the protesters beat pans and swore the president and her party.

On 15th of March 2015 took place the first and largest manifestation against Dilma Rousseff, in several different cities, asking for her impeachment.

[1] https://www.facebook.com.

[2] https://www.twitter.com.

[3] https://web.whatsapp.com/.

These manifestations united on Brazilian streets millions of people, dissatisfied with the current management of the country, inflation of prices and corruption reports, chiefly in Petrobras.

On the other hand, supporters of the government decided for a counterattack. A mobilization was organized by union and social movements on 13th of March 2015. Besides occupying the streets, the political debate also occurred on the Internet.

The government supporters accused the traditional mass media of diminishing the importance of pro-government protests on news, while giving a wide coverage on protests of opposition, notably Rede Globo TV Channel, the most popular and influential media group in Brazil.

Virtual militants and connected citizen have continued the political debate in cyberspace. After the mobilization studied in this paper, there were two others great protests against the Workers' Party, on 12th of April 2015 and 17th of May 2015 (this last one with a smaller adhesion).

3 Supervised Learning

In Machine Learning, Supervised Learning [18] refers to the set of algorithms and methods that learns a function $y = f(x)$ where x is the object of study and y is a predicted value. This is performed by feeding the algorithm with a set X of object examples, associated with the expected output given by a set Y. The algorithm creates a mapping from the observed data, being capable of inferring any new object, already observed or not.

There are many algorithms created for this task, with different characteristics and capable of handling different types of variables. In this work, we will use two well-known techniques: Naive Bayes [12], a technique known for its good trade-off of performance and simplicity; and Support Vector Machine [17], a state-of-the-art algorithm for many classification problems and datasets, but with the need of more specific adjustments.

In the following sub-sections we will briefly explain these techniques.

3.1 Naive Bayes

Naive Bayes is a non-parametric probabilistic algorithm, often used for classification of categorical data [3] and text mining [6]. This algorithm assumes that the variables describing the objects of study are independent from each other regarding their classification, thus making use of the Bayes Theorem. With this strong assumption, we can use the Bayes Theorem described as:

$$p(c|X) = \frac{p(c)p(X|c)}{p(X)}, \tag{1}$$

where X is the feature set describing the object and c is the class to which it belongs.

From a training data, it is easy to estimate $p(c)$ as the proportion of objects classified as c. The estimation of $p(X|c)$ and $p(X)$ makes use of the independence assumption as:

$$p(X) = p(x_1) \cdot p(x_2) \cdots p(x_n), \tag{2}$$

and

$$p(X|c) = p(x_1|c) \cdot p(x_2|c) \cdots p(x_n|c). \tag{3}$$

After estimating all of these probabilities, a new object can be classified by finding the class c which gives the maximum probability given the features of the object.

3.2 Support Vector Machine

The Support Vector Machine (SVM) is a technique that extends the linear regression model to alleviate two problems: (i) the assumption that the data is linearly separable and; (ii) the over-fitting of the training data.

For the first problem, the first and simpler assumption during the classification task is that the objects are linearly separable, i.e., the objects of different classes can be separated with a simple line equation. But in practice, this assumption rarely holds, so a new set of features should be crafted or learned as a non-linear combination of the original features set. With this transformation, it is expected that the new features set resides on a linearly separable space, but this adds the cost of transforming to every new object to be classified. In SVM, the idea of a Kernel function was introduced to alleviate this problem [5,15].

A Kernel function $k(x, y)$ takes as input two objects described by their original features set and calculates the distance between them in a different space chosen by the function being used. This calculation is performed without explicitly transforming the feature space, thus having an efficient computational cost. The main Kernel functions used on the literature are Linear Kernel, Polynomial Kernel and RBF Kernel, the last two non-linear.

The second problem, regarding the over-fitting, is alleviated by changing the objective-function of the separation line. In Linear Regression, the objective is to find the separation line which gives the minimum error regarding the training data. In SVM, the objective-function is the maximization of the margin enveloping the separation line. In other words, the algorithm seeks a separation line that has a maximum distance from the closest points of each class.

By maximizing this margin, not only the classification error for the training data is minimized, but also it keeps some space for generalization of unseen data.

3.3 Related Work

It is well know the usage of SVM and Naive Bayes as text classifiers, and recently applied to Twitter corpora and other micro-blogging platforms [1,8,14]. In particular, we briefly summarize some studies that utilized tweets as a source of public opinion manifestations.

In the context of political sentiment mining on Social Networks, Spaiser et al. [16] applied statistical and machine learning techniques to almost $700,000$ tweets, being able to observe how they had contributed to weaken Russian protest movements.

Livne et al. [10] collected tweets from US House and Senate candidates, applied text mining using a bag-of-words model, conducted graph analysis to estimate co-alliances and divergence among candidates and generated a predictive model for a certain candidate win or lose the election.

Lotan et al. [11] analyzed the Tunisian and Egyptian Revolutions as told by Twitter, identifying the main actors of the online manifestations and flow of information.

Turkmen et al. [19] collected and labeled tweets during recent Turkey protests and used SVM and Random Forest classifier to predict political tendencies in the messages.

4 Experiments

In this section, we give a complete description of data acquisition, methodology and analysis of a real-life event on the Twitter Social Network.

4.1 Methodology

During the period of 12th to 16th of March 2015, we collected the tweets with hashtags related to both protests (see Table 1) by using the Twitter API[4] with the streaming interface that continuously collects tweets in real time. After the data collection, we ended up with 274,645 tweets from 101,452 different users.

We added the tweets published on 13th of March of 2015 in one dataset (PROGOV) and those published on 15th of March of 2015 in another dataset (CONGOV). From these two datasets we extracted the bag-of-words model, transforming the features by using tf-idf (frequency inverse document frequency) [4].

For the classification task, we randomly picked 100 tweets from each dataset, 50 for each sentiment [5], and fitted this data using both classification algorithms. After that, another 100 tweets were chosen at random and classified using these models. If the classification accuracy (percentage of correct classification) were below 70 %, these 100 tweets were added to the training data, and the process repeated until the accuracy levels reached 70 % or more on the random data. This threshold is a compromise of the reported accuracy of the literature [1,8,14] that range between as low as 60 % and as high as 85 %.

After that, we classified the entire dataset and performed some exploratory analysis to extract information about the protests dynamics. A summary of the datasets characteristics is depicted in Table 2.

[4] https://dev.twitter.com/.

[5] We are aware that this dataset is possibly unbalanced, but to know the exact balance would imply a large quantity of manual classification.

Table 1. Hashtags used during the data collecting stage.

Hashtag	Meaning
#13Marco	Date of the protest supporting the government
#AcordaBrasil	Wake-up Brazil
#DilmaNaoMeRepresenta	Dilma (elected president) does not represent me
#DilmaVaiada	Dilma booed
#ForaDilma	Go away Dilma
#ForaPT	Go away PT (Workers' Party)
#ImpeachmentDilma	Impeachment of president Dilma
#PetrobrasEhBrasil13	Petrobras (Brazilian oil company) belongs to Brazil (supporters of the gov.)
#PronunciamentoDaDilma	Speech of president Dilma
#SouPetrobras	I am Petrobras (supporters)
#TodosContraOGolpe	All against the coup d'état
#VamosVaiarDilmaNaTV	Let us shout down Dilma on TV
#VemPraRua15DeMarco	Let us go to the streets on March, 15th
#br45ilnocorrupt	No corruption in Brazil (with a pun with the code 45 of the opposition party)
#globogolpista	Coup-backer Globo (Globo is one of the largest TV Station in Brazil)
#protestos	protests

Table 2. Summary of studied datasets.

Dataset	# of tweets	Unique words
PROGOV	84, 821	36, 070
CONGOV	189, 824	60, 684

In the next subsections we will present just the main results in order to preserve clarity and brevity of this paper. The full set of results with the corresponding IPython Notebooks will be made available at https://github.com/folivetti/POLITICS.

4.2 Classification Results

After sampling 100 tweets from the datasets and manually labeling them as PRO or CON, as in pro-government and against it respectively, we trained the Naive Bayes and SVM algorithms with these sampled tweets, and applied the classification process for the entire data set. After this first step, we sampled another batch of 100 tweets from the classified results of each algorithms.

In order to use a diversified set, without a bias towards one class, we have used the Reservoir Sampling technique [20] that samples items with equal probability from a large set. The algorithm is briefly described in Algorithm 1.

Algorithm 1. Reservoir Sampling.

```
input  : Data stream $D$, number of samples $k$.
output: Sampled data $S$

$S \leftarrow \varnothing$
for $sample \in D$ do
    if $sample.index <= k$ then
        $S.append(sample)$
    else if $r\ U(0, k) < k$ then
        $S[r] \leftarrow sample$
```

The algorithm starts by inserting the first k samples into the sampled data set. After that point, every subsequent data can replace a given sample, chosen randomly by an uniform distribution ($r\ U(0, k)$), with probability $1/k$.

After the sampling process, we manually verified the classes of data to estimate the accuracy of both classifiers.

As we can see from the Truth Tables in Tables 3 and 4, both classifiers had similar results, with an accuracy around 90 %. Although this may not be statistically significant for the whole dataset, the intention of this work is to perform a practical analysis of the protests data with the minimal human effort.

Table 3. Truth table for the classification results of Naive Bayes.

		Actual values		
		PRO	**CON**	**Total**
Predicted values	**PRO**	45	4	49
	CON	6	45	51
	Total	51	49	100

Table 4. Truth table for the classification results of SVM.

		Actual values		
		PRO	**CON**	**Total**
Predicted values	**PRO**	45	9	54
	CON	2	44	46
	Total	47	53	100

4.3 Distribution of Classes

It is expected that classes are biased by the theme of the day, i.e., PRO tweets mainly occur in the PROGOV dataset, and CON tweets in the CONGOV dataset. However, our question is how imbalanced the datasets actually are, and if there is a difference on the distributions for each day.

To answer such questions, Figs. 1 and 2 show the distributions for each day and for each classifier. As we can see, regarding the classifiers, they agree on the distribution of topics on both datasets, having a very similar distribution of classes. Also, those Figures confirm that the distribution is biased towards the central theme of each protest, on March 13th the majority are supporting the government while on March 15th, the majority is against it.

We observe that on March 13th the opposing group was less active than on March 15th. This indicates that the people against the government concentrated their efforts on the protest of March 15th and did not pay attention to this pro-government manifestation. On the other hand, the group supporting the government was considerably active on both days of protests, trying to contest the claims of the other group.

Furthermore, the Figures show that the absolute number of tweets supporting the government is about constant throughout the days, with a number of around $80,000$ tweets, while the number of people against the government steps up from around $20,000$ to about $150,000$, almost 7 times more. This indicates a more consistent pattern of activists supporting the government.

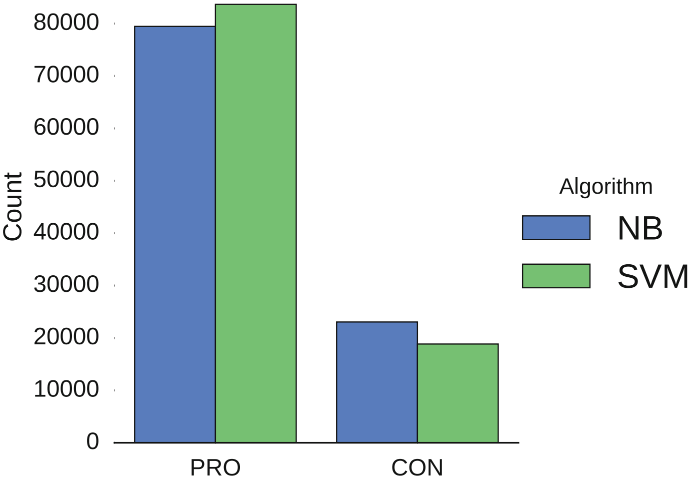


Fig. 1. Distribution of classes for March 13th.

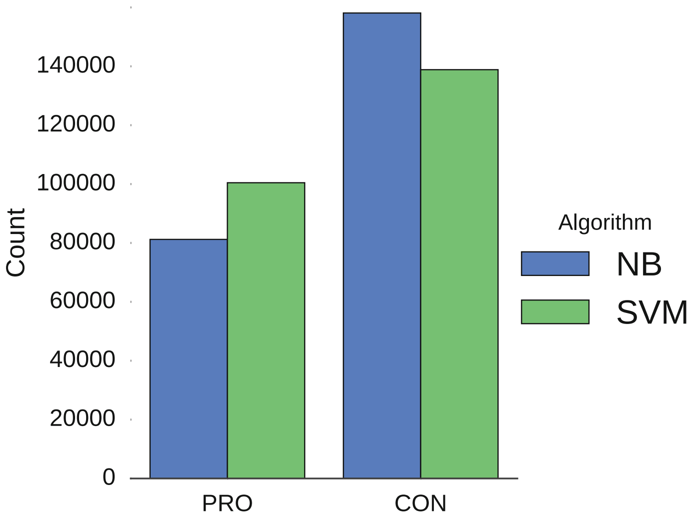


Fig. 2. Distribution of classes for March 15th.

4.4 Distribution of Words

After verifying the distribution of each class, it is also interesting to extract what people of each group are saying. For this matter we have extracted the Top 3 words used on the tweets for each class and on each type of protest.

The Figs. 3 and 4 show the results of these distributions. It is important to notice that both algorithms rendered the same set of words, so the results are grouped together on the bar plot depicted with the confidence intervals. The meaning of these words are explained on Table 5.

As we can see on March 13th, the majority of the tweets focused on the accusations against Globo TV Channel harming the democracy. In Brazilian history, Globo is often associated with the support of the military coup of 1964 [7] and the election of the only Brazilian president to suffer an impeachment [13]. The second and third more frequent words are associated with calling the people on the streets and stating they will not participate on the next protest against the government. The people against the govern limited themselves on calling people for the protests and asking the president to step out on her own.

On March 15th, the people supporting the government kept a similar behavior from the previous day, but additionally, they started a campaign claiming for democracy, stating that the people should accept the results from the past election as this is a democracy. The group against the govern intensified the use of the hashtag asking the president Dilma to step out together with the use of a similar hashtag related to her political party. The term *vemprarua* is perceived to have been used by both sides since this word is a more general term for calling people to the streets, without specifying the reason.

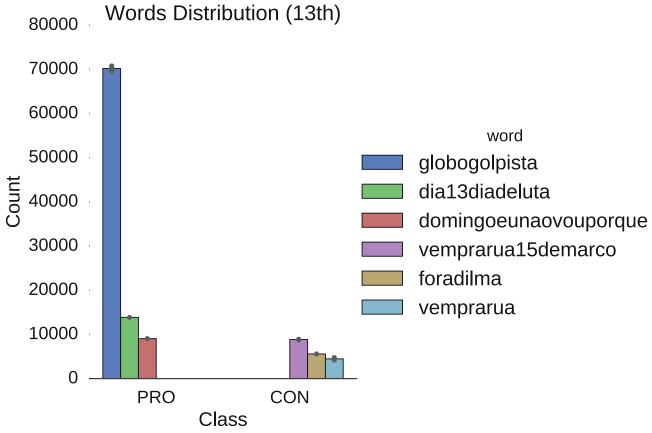


Fig. 3. Words distribution for March 13th.

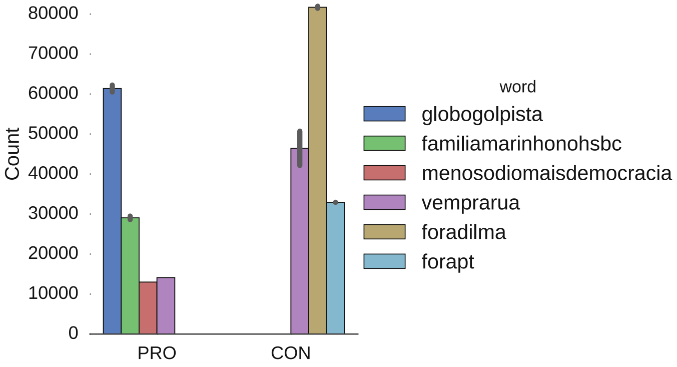


Fig. 4. Words distribution for March 15th.

4.5 Most Active Users

Another practical result of interest from these datasets is the identification of the most active users for each class. The identification of such actors may reveal the organizations and real motivation behind both manifestations. Even if they are not the *leaders* of such events, they represent a step towards finding such connections.

Initially, we analyzed the distribution of activity of all users in each day of protests. In Figs. 5 and 6 it is shown that the majority of users posted few tweets about the protests, while there were very few users responsible for about 800 tweets on March 13th and more than 1400 tweets on March 15th. This is similar to a power law distribution, indicating that few users are more active and possibly more influential than others. The next step was to identify those very active users and their role in the protests.

Table 5. Explanation of each hashtag.

Hashtag	Explanation
dia13diadeluta	Used to call the people for March 13th event
domingoeunaovouporque	Stating that they will not participate on March 15th
familiamarinhohsbc	Related to the accusations against Globo TV Station (accused of supporting the movement against the government) and HSBC bank
foradilma	Asking for Dilma Rousseff to step out of presidency
forapt	Asking for the Workers' Party to step out
globogolpista	Claiming Globo TV Station is trying a coup
menosodiomaisdemocracia	Asking for less hate and more democracy
vemprarua	Calling people to the streets, used for both events
vemprarua15demarco	Calling people to the streets on March 15th

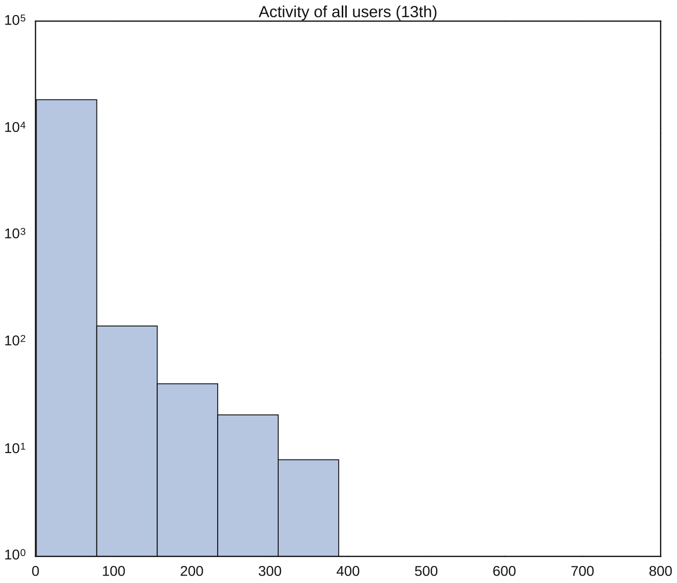


Fig. 5. Distribution of tweets from all users on March 13th, logarithmic scale for y axis.

In Figs. 7 and 8 we depict the distribution of the six most active users with confidence intervals. Regarding March 13th, the most active users for each group were *Larissa Alves* (/laripr), a twitter account of a person who actively tweets about the accomplishments of the current government, the suspicious and accusations of the opposing parties, and *Br45il No Corrupt* (/br45ilnocorrupt), an account with a pun on the number 45 corresponding to the opposing political party, replacing the letters 'A' and 'S' from *Brasil.* This account was specially

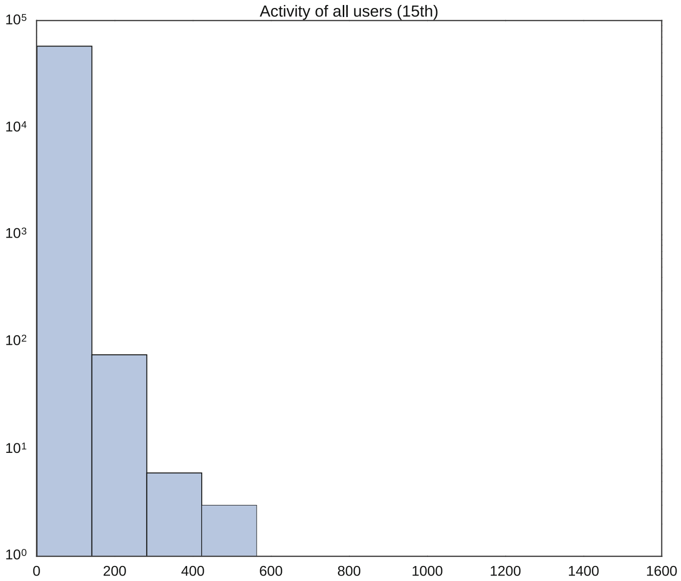


Fig. 6. Distribution of tweets from all users on March 15th, logarithmic scale for y axis.

created for accusing the Workers' Party of being corrupt and feed the discussions around the protests. This account was created by the non-profit organization of the same name that, while do not explicitly enlist a direct connection with the opposing party, it manifested support to them.

The account *#Dia13DiadeLuta* (/AdaByronKing) is an account related to a group of political activists against rumours, *#ForaDilma* (/jonhpaul11) was a common user that changed his name during the event to support the group against the government. There is no known connection with political parties but it is assumed that they have such support. The account *Revista Eletrônica* (/e_editora) refers to a self-claimed independent journalist media while *JoaoG* (/JGZZZO) seems to be a fake account created as a retweeting robot, also known as bot. These bots are computer programs created to share the messages of specific users, often used to fake the real impact of an opinion. The user is considered suspect of being a bot whenever they have more than 10 thousand tweets, consisting mostly of retweets, if they have many retweets in different languages, or have no tweet at all (i.e., retweet a message and delete some time later).

On March 15th, some of the tweets of the account *Br45sil No Corrupt* are probably incorrectly classified by one of the algorithms, generating a lower confidence. This misclassification occurred by a sequence of tweets without the common words used against the government. One example is the tweet literally translated to *Tomorrow we will be* 1 *million on the streets* that, without the

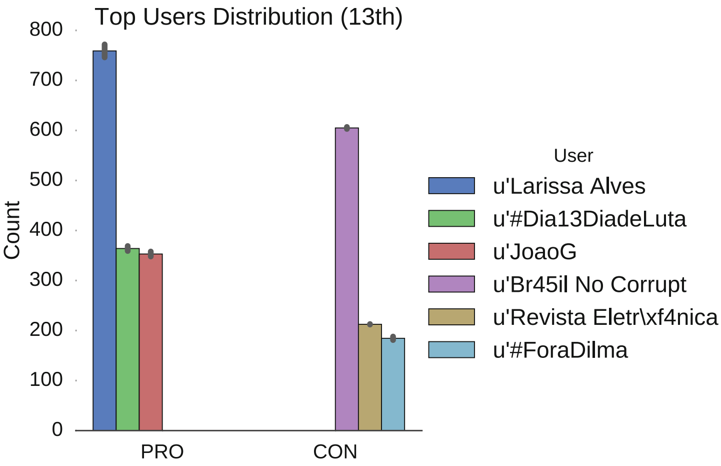


Fig. 7. Distribution of tweets from the six most active users on March 13th.

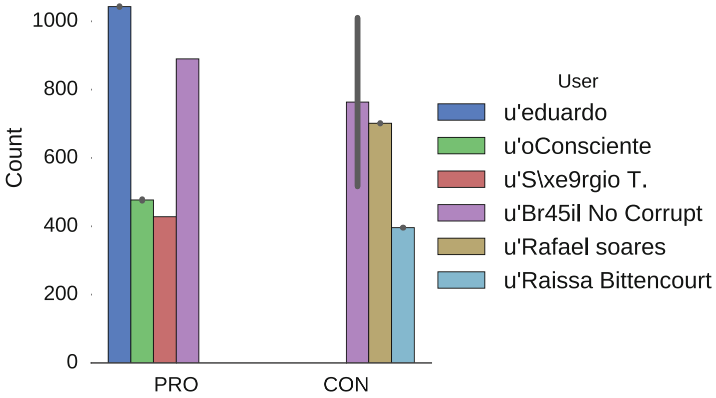


Fig. 8. Distribution of tweets from the six most active users on March 15th.

date of the tweet and the user that created the content, the correct classification cannot be inferred.

The user *Rafael Soares* (/KatycatBrasill), after manual inspection, seems to be an account created as a fan account for singer Katy Perry as a disguise for being another retweeting bot. This account has a long history of retweeting contents of different opinions in different languages. The user *Raissa Bittencourt* (/raissabittenco3) was a fake account and it is not active anymore, created probably with the purpose of retweeting opinions against the government. The user *eduardo* (/eduardonino) is a political activist supporting the government but aligned with more leftist parties. Finally, the user *oConsciente* (/oconsciente) is a political activist supporting the Workers' Party.

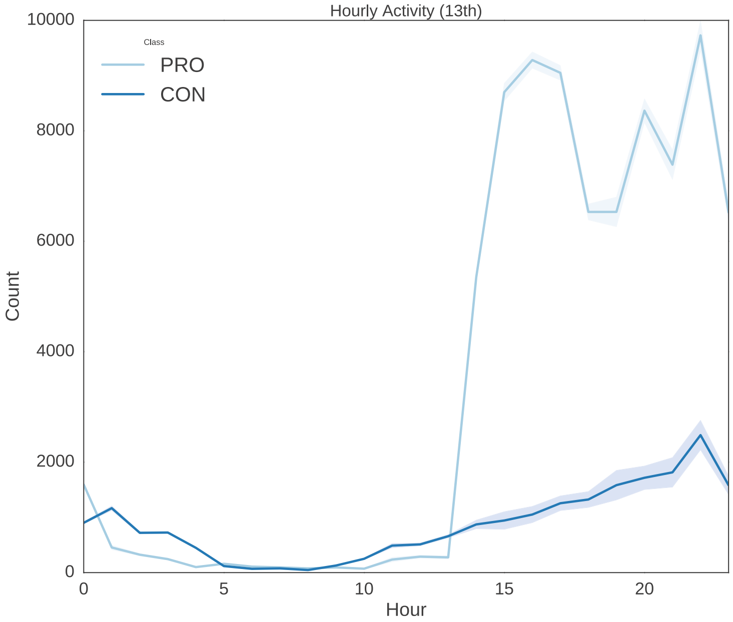


Fig. 9. Hourly distribution of classes on March 13th.

These results could find some interesting actors (i.e., *Br45il No Corrupt* and *oConsciente*) that are indicative of the organizations behind each group. But, also, it revealed the use of bots by both sides in order to inflate the importance of their claims.

4.6 Hourly Activity

Next we verify the hourly activity throughout both days of protests, first grouped by class and then by the top users. In Figs. 9 and 10 we can see the activities for each group on each day. We note that the protests took place during the afternoon of the corresponding days, thus the main activity was comprised from noon to midnight on both days. As it should be expected, the group supporting the government was more active than the group against it on March 13th, while on March 15th occurred the opposite.

However, the behaviors are different, as seen in these Figures. The first is regarding the behavior of the CON group during March 13th, as they kept a low profile in the morning but started raising their activity after 10 a.m., reaching its peak at around 11 p.m. of the Friday night. This pattern seems reasonable as a kind of attack against the supporters group, when their manifestation started. Since this is the day preceding the weekend, the working time might have prevented most of the users of tweeting before 6 p.m.

During the events of March 15th, we observe an intensified activity of the supporters group early in the morning. They seem to have organized themselves

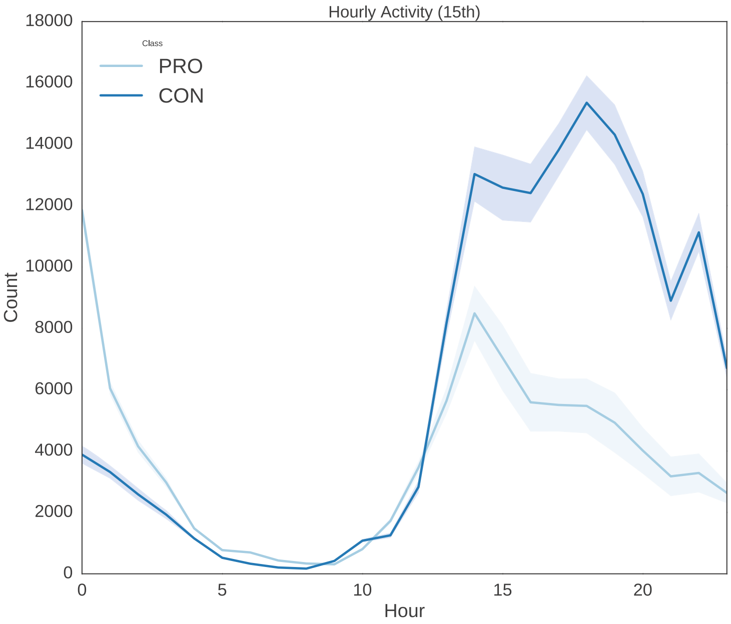


Fig. 10. Hourly distribution of classes on March 15th.

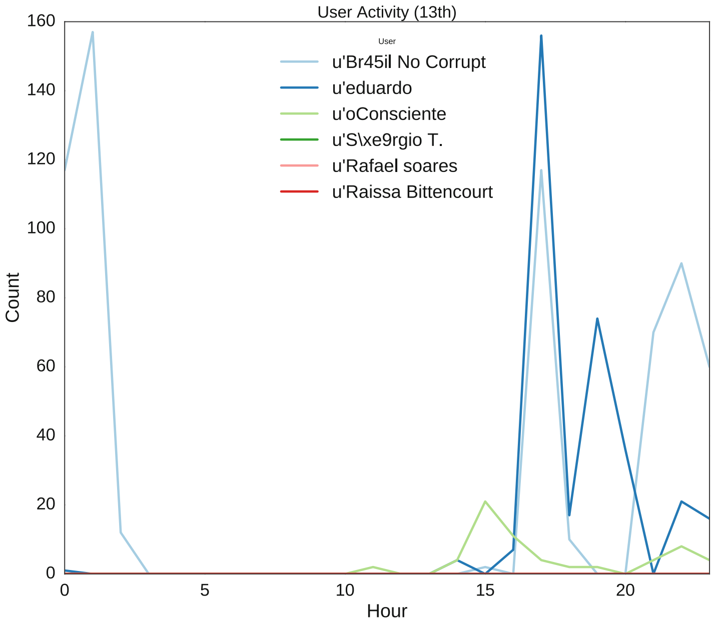


Fig. 11. Hourly user activity on March 13th.

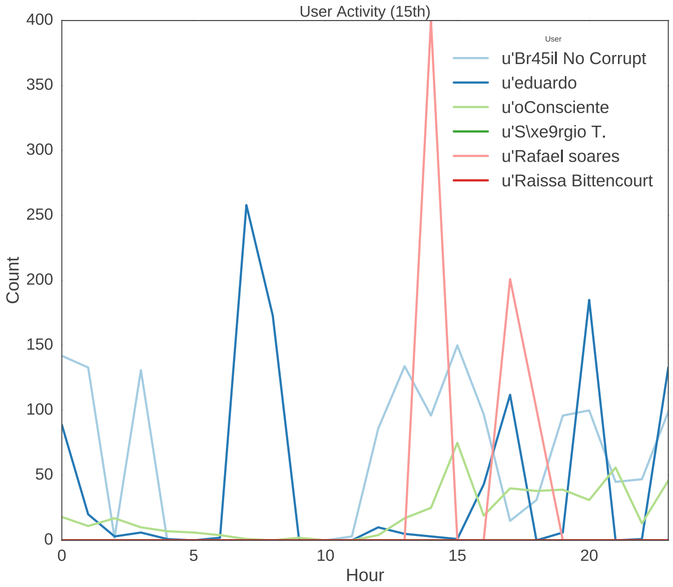


Fig. 12. Hourly user activity on March 15th.

to try attacking the protesters prior the event. Right after the start of the event, the supporters were also very active, trying to compensate for the rising of people tweeting against the government and, after that, followed the same trend of the protesters.

In Figs. 11 and 12, we depict the hourly activity of some of the top users from a previous analysis throughout each day. On March 13th, the users followed a similar behavior of the tweets by class, being more active during the afternoon. The users *eduardo* and *Br45il No Corrupt* were responsible for the most activities, having similar peaks at 5 p.m., at 7 p.m. and a final one at 9 p.m.. The events of 5 p.m. were about the presence of artists on the protest against the government, with a decay of such announcements on 7 p.m. and raising again at 9 p.m..

During March 15th, the activity of users did not match exactly the class hourly behavior, having several peaks throughout the day. The main peaks occurred on 6 a.m. by *eduardo* calling the people for a tweeting event against the protest. After that, at 2 p.m. the user *Rafael soares* chained a tweeting activity to raise the hashtags against the government on the trending topics. These users followed the same behavior later at 5 p.m. and, by 8 p.m., the user *eduardo* raised again a protest against the media trying to coup the government.

5 Conclusion

In this paper we show how we applied two algorithms for supervised learning, Naive Bayes and Support Vector Machine, in order to analyze the events of two opposing protests on the streets of Brazil, as told by Twitter users, as a consequence of the disputed presidential elections in 2014. These algorithms were trained using a very small sample of the data set in order to quickly estimate the numbers of both events.

The events were first separated in two datasets, being March 13th regarding the protest supporting the government and March 15th the protests from the group against the government. Both datasets were classified by the two algorithms on its entirety, and the distribution of the analyzed quantities were grouped together when convenient.

Ideally, to improve accuracy, a large set of labeled data should be available during the training process, so that the learning algorithms could face distinct examples that should pertain to the same class. But, in practice, we cannot always afford to manually separate a sufficient amount of data for this task, and not even verify the accuracy results. These experiments show that, even if you cannot guarantee high accuracy, some interesting information can still be extracted for using on a broader study.

The results showed that the activists supporting the government, although being a minor number, were more active throughout the weekend comprising both protests. They actively tried to reduce the importance of the protests against the government by accusing the organizations that supposedly were behind the event. On the other hand, the groups leading the protest against the government concentrated their efforts during the peak of the events, as an attempt of minimizing the importance of the other group and spread their goals.

Another interesting information found in these datasets was the use of retweeting robots from both groups to inflate the numbers of tweeters supporting each event. This not only may affect the perceivable intensity of the movements, but can also help to attract new people for both sides through the Twitter trending topics.

From this point, we have paths to follow for future research. On the Computer Science side, we will try to automatize the process of manual labeling for the training process or minimizing such efforts. We intend to do that by means of semi-supervised learning and the use of Topic Modeling. On the Data Science side, we will apply this procedure into a much larger data set containing all the events that happened during the presidential elections, and that motivated the current events.

Acknowledgment. This research was funded by FAPESP process number 2014/06331-1.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Proceedings of Workshop on Languages in Social Media, pp. 30–38. Association for Computational Linguistics (2011)
2. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 163–222. Springer, New York (2012)

3. Agresti, A., Kateri, M.: Categorical Data Analysis. Springer, Berlin (2011)
4. Aizawa, A.: An information-theoretic perspective of TF-IDF measures. Inf. Process. Manag. **39**(1), 45–65 (2003)
5. Amari, S.I., Wu, S.: Improving support vector machine classifiers by modifying kernel functions. Neural Netw. **12**(6), 783–789 (1999)
6. Berry, M.W., Castellanos, M.: Survey of text mining. Comput. Rev. **45**(9), 548 (2004)
7. Chong, A., Ferrara, E.L.: Television and divorce: evidence from Brazilian novelas. J. Eur. Econ. Assoc. **7**(2–3), 458–468 (2009)
8. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the OMG! In: ICWSM vol. 11, pp. 538–541 (2011)
9. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 415–463. Springer, New York (2012)
10. Livne, A., Simmons, M.P., Adar, E., Adamic, L.A.: The party is over here: structure and content in the 2010 election. In: ICWSM 2011 (2011)
11. Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., Boyd, D.: The Arab spring— the revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int. J. Commun. **5**, 31 (2011). http://ijoc.org/index.php/ijoc/article/view/1246
12. McCallum, A., Nigam, K., et al.: A comparison of event models for Naive Bayes text classification. In: AAAI-1998 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48. Citeseer (1998)
13. Miguel, L.F.: Mídia e eleições: a campanha de 1998 na rede globo. Dados [online] **42**(2) (1999). http://dx.doi.org/10.1590/S0011-52581999000200002
14. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC, vol. 10, pp. 1320–1326 (2010)
15. Roth, V., Steinhage, V.: Nonlinear discriminant analysis using kernel functions. In: Advances in Neural Information Processing Systems. Citeseer (1999)
16. Spaiser, V., Chadefaux, T., Donnay, K., Russmann, F., Helbing, D.: Social Media and Regime Change: The Strategic Use of Twitter in the 2011–2012 Russian Protests (2014). Available at SSRN
17. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
18. Thrun, S., Pratt, L.: Learning to Learn. Springer Science & Business Media, New York (2012)
19. Turkmen, A., Cemgil, A.: Political interest and tendency prediction from microblog data. In: 2014 22nd Signal Processing and Communications Applications Conference (SIU), pp. 1327–1330, April 2014
20. Vitter, J.S.: Random sampling with a reservoir. ACM Trans. Math. Softw. (TOMS) **11**(1), 37–57 (1985)
21. Watts, J.: Brazil: hundreds of thousands of protesters call for rousseff impeachment. The Guardian (2015). http://www.theguardian.com/world/2015/mar/15/brazil-protesters-rouseff-impeachment-petrobas

Sake Selection Support Application for Countryside Tourism

Teruyuki Iijima(✉), Takahiro Kawamura, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga

Graduate School of Information Systems,
University of Electro-Communications, Tokyo, Japan
{iijima.teruyuki,kawamura,sei,tahara,ohsuga}@ohsuga.is.uec.ac.jp

Abstract. For the upcoming Tokyo Olympic Paralympic Games in 2020, the number of foreign tourists coming to Japan is expected to rise. However, there has been a problem with tourists becoming less likely to visit places outside of the urban areas. In order to solve this issue, a commitment has been made by the government to use "Sake Brewery Tour" to draw tourists to less populated areas. The purpose of this study is to find a way to encourage foreign interest to sake and sake brewers, and participant in "Sake Brewery Tours". We developed an application for the foreign tourists who are not much interested in sake. The approach of the study involved the presentation of sake selection in connection with wines, which have surprising similarities to the sakes, and encourage the tourists access sake brewer sites. 20 test users used the application, and the average screen residence time was 55 (sec) including the sake brewer sites, which was longer than the application for comparison, which shows the sake information alone. Therefore, we confirmed that the users come to have an interest in sake and sake brewers by showing the surprising connections with wine.

1 Introduction

The Tokyo Olympic Paralympic Games are to be held in 2020 [1]. This will cause foreign tourists to gather in Japan's urban areas, and in turn, create a problem in that it will be difficult to prompt tourists to visit places outside of the urban areas. In recent years, there has been a problem that very few tourists have chosen to venture the outside of the urban areas. Therefore, various approaches have been taken as an attempt to activate the country areas. For example, one approach has been to conduct "Sake Brewery Tours" [2]. This is carried out in a similar way to wine tours in France and California, and utilizes sake brewer as the main tourist attractions, and thus invites foreign tourists to the country areas. In this study, we developed an application to encourage the foreign tourists to be interested in participating in sake brewing tourism. Thus, we have taken an approach of presenting the surprising connections between wine and sake to the user. The reason for choosing wine to make our connection is that wine has the same way of brewing as sake, and the foreign tourists are commonly known to enjoy wine. For example, in the case of "Seisyu Kitanohomare

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 19–30, 2016.
DOI: 10.1007/978-3-662-53416-8_2

Junmaigen-syu Samurai"(a sake name), the sake leads to "Kitanohomare Syuzou"(brewer) → "Otaru"(Location) → "Princess Mononoke"(Movie in the location) → "Hayao Miyazaki"(Director) → "Antoine de Saint-Exupery"(Writer who gave great influence on the director) → "CH.MALESCOT ST.EXUPERY" (Wine in winery that the writer's grandfather bought). We intended that the application invites the foreign tourists to the brewer and "Otaru" by showing connections such as the above.

The remainder of this paper is structured as follows. In Sects. 2 and 3, we present the proposed application and outline the background Linked Data to calculate the connections. Then, evaluations are reported in Sect. 4, before a discussion regarding related works in Sect. 5. In Sect. 6, we conclude this paper, and discuss the future works.

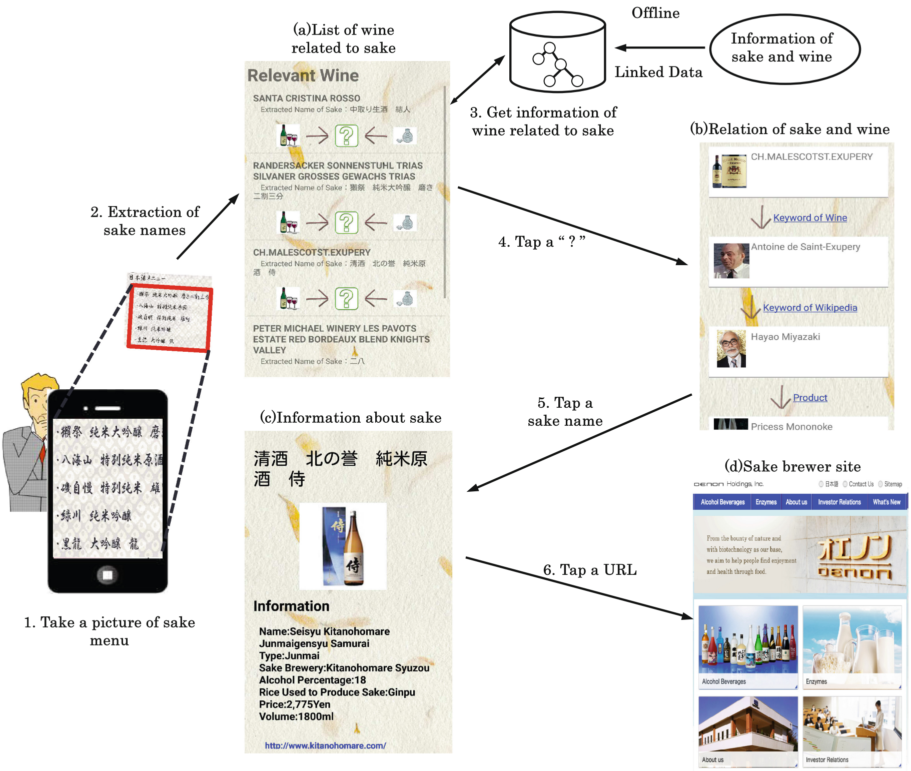
2 Proposed Application

We suggest use of our sake selection support application for people who are familiar with wine, but not much interested in sake. This application is able to use names of a sake list on a restaurant menu, to find a wine with surprising similarities to the particular variety of sake. Figure 1 shows a workflow of this application. The application is useful in the case that a user visits a Japanese restaurant, but is not familiar with the sake selection presented to him. There is already an application, that can provide the sake information such as brewers and flavors by reading labels on sake bottles [3]. However, there is no application, which provide sake's unique stock of knowledge related to wine. Figure 1(a) is of a screen that is displayed after taking a picture of a sake menu. A list of wines associated with the sake is displayed. However, due to the restriction of the screen size, the specific connections between the sake and the wine are displayed in the "?" mark at first. When the user recognizes a wine she/he is familiar with in the list, she/he taps the wine name. Then, the connection between the wine and the sake is indicated as shown in Fig. 1(b). If the user is interested in the sake, she/he can also tap the sake name. Then, Fig. 1(c) is displayed with the name of the sake and a picture of sake. Also, information such as the alcohol content of the sake and the URI of the brewer's website is listed at the bottom of the screen. If the user has become interested in obtaining more information at this point, she/he may access the brewer's website by tapping the URI.

2.1 System Architecture

Figure 2 indicates the system architecture to realize this application. The user starts the application, then takes a picture of the menu containing sake names. Then, the image that the user has taken is sent to a server. The server program analyzes the image, and sake names are extracted. Strings from the image are extracted using the Tesseract-OCR[1]. Tesseract-OCR is an OCR library. Also,

[1] https://github.com/tesseract-ocr.



Relevant Wine

SANTA CRISTINA ROSSO
Extracted Name of Sake：中取り生酒　結人

RANDERSACKER SONNENSTUHL TRIAS SILVANER GROSSES GEWACHS TRIAS
Extracted Name of Sake：獺祭　純米大吟醸　磨き二割三分

CH.MALESCOTST.EXUPERY
Extracted Name of Sake：清酒　北の誉　純米原酒　侍

PETER MICHAEL WINERY LES PAVOTS ESTATE RED BORDEAUX BLEND KNIGHTS VALLEY
Extracted Name of Sake：二八

CH.MALESCOTST.EXUPERY
Keyword of Wine
Antoine de Saint-Exupery
Keyword of Wikipedia
Hayao Miyazaki
Product
Pricess Mononoke

清酒　北の誉　純米原酒　侍

Information

Name:Seisyu Kitanohomare Junmaigensyu Samurai
Type:Junmai
Sake Brewery:Kitanohomare Syuzou
Alcohol Percentage:18
Rice Used to Produce Sake:Ginpu
Price:2,775Yen
Volume:1800ml

http://www.kitanohomare.com/

From the bounty of nature and with biotechnology as our base, we aim to help people find enjoyment and health through food.

Fig. 1. Application workflow

a SPARQL Protocol and RDF Query Language(SPARQL) query is performed on a Resource Description Framework(RDF) DB called Sesame[2], in order to get all the names of sake varieties. RDF is in the form of a <subject, property, object>, and a SPARQL is a query language for RDF. More details are described in Sect. 3. Then, by using the edit distance between the obtained sake name and strings of each line extracted from the image, a sake name with the smallest edit distance is retrieved. Finally, wines associated with the sake and connection information are acquired by performing a SPARQL search with the sake name. After obtaining all the associated wines by following the background Linked Data described in the next section, the connection information is sent to the client. The information includes the wine names associated with the sake and the connection information between the sake and the wines. The client side displays the information to the user. When the user taps a sake name, a SPARQL search is performed again in order to get the information about the sake, e.g. descriptions and brewer sites, from the client. Then, the obtained information is presented to the user.

[2] http://rdf4j.org.

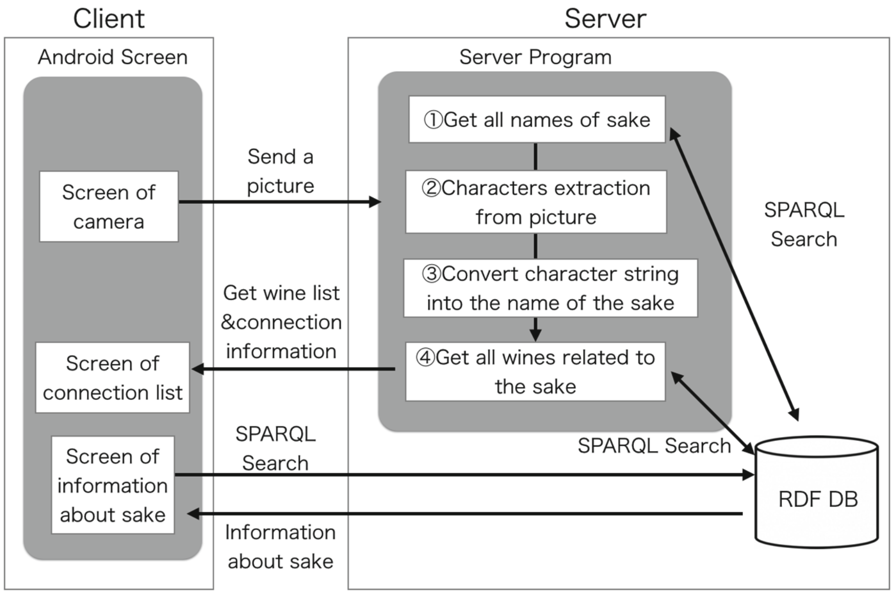

Client
Server
Android Screen
Server Program
Screen of camera
Send a picture
①Get all names of sake
②Characters extraction from picture
③Convert character string into the name of the sake
④Get all wines related to the sake
SPARQL Search
Get wine list &connection information
Screen of connection list
SPARQL Search
SPARQL Search
Screen of information about sake
RDF DB
Information about sake

Fig. 2. System architecture

2.2 Example of Search

For example, "Seisyu Kitanohomare Junmaigensyu Samurai" is a variety of sake. Figure 1(a) shows that there is a connection between the sake and the wine called "CH.MALESCOTST.EXUPERY". The sake is made in a brewer named "Kitanohomare Syuzou" found in Otaru City, Hokkaido in Japan. Otaru is known as a stage of a cartoon film called "Princess Mononoke", which is directed by "Hayao Miyazaki", whose favorite writer is "Antoine de Saint-Exupery". Also, there is a winery owned by his great-grandfather, and "CH.MALESCOT ST.EXUPERY" is one of wines produced by the winery. This is a wine that has been associated with the sake. By noticing such surprising connections between the sake and the wine, the users become interested in the sake, and hopefully the sake brewer, and its area outside of the urban district.

3 Background Linked Data

Linked Data is a graph data, which is used to publish and share data on the Web proposed by Tim Berners-Lee[3]. In this study, the background information related to wine, sake, and their brewers, etc. has been converted into Linked Data. We collected a large amount of data described about sake and wine in several websites, and converted them in the RDF format.

[3] http://www.w3.org/DesignIssues/LinkedData.html.

3.1 Conversion of Sake and Wine Data to Linked Data

As described above, we created a set of data related to sake and wine in Linked Data format. We collected the data from EC sites such as the Sake Brewer's official sites in Rakuten[4] and sites of sake tasting information. The converted data set consist of 186,000 triples <subject, predicate, object>, which corresponds to records in DB. For retrieving the wine data, we performed a morphological analysis on sentences in the wine comments, and also extracted the data from Wikipedia headwords. We used Mecab[5] as the morphological analysis engine. The extracted data are described with the DBpedia[6] resources. DBpedia is Linked Data, which contains Wikipedia infobox information. Then, properties are described in our own sake schema defined in our website[7]. Linking to the resources of the DBpedia made it easy to link the external data. We also used the data about the sister cities of the Council of Local Authorities for International Relations[8]. The data of the sister cities are used in order to make it easy to search the connection of brewers. In the previous example, a place was a stage of a cartoon file, and also a location of the sake brewer. However, less data would be used to make the connections in other places. Therefore, we used the Linked Data of the sister cities to facilitate the search. In addition, we used Linked Open Data called Location Site of Japanimation(LSJ)[9]. LSJ includes information about locations that have become stages of cartoon filmes. Figure 3 shows a sake called "DASSAI 23" in the RDF format. The resource is indicated as <`Sake:dassai23`>, and a property is a <`rdf:label`>, and an object is described as a literal "DASSAI 23". Although the sake brewer that made this sake is "Asahi Brewery" in Yamaguchi Prefecture, it is difficult to distinguish the same brewery in Oita Prefecture. Therefore, the URI is described as a representative URI, <`Sake_bre:Asahi_Yamaguchi`>. Information such as the polishing ratio of rice and amino acid level of the sake is also converted to the RDF format. Table 1 shows some of the properties that we have defined, where "Sake_pro:" is a prefix of <http://www.ohsuga.is.uec.ac.jp/sake/property/>.

3.2 Search Method

In this application, a server-side program is used to search wines related to the same based on the semantic relationship. A SPARQL query leads to the location of the sake brewer. Various contents are then associated with municipalities. For example, the content to be used in the search includes locally famous persons, stages of films, sister cities, and so on. The program searches for a wine through these contents.

[4] http://www.rakuten.co.jp.

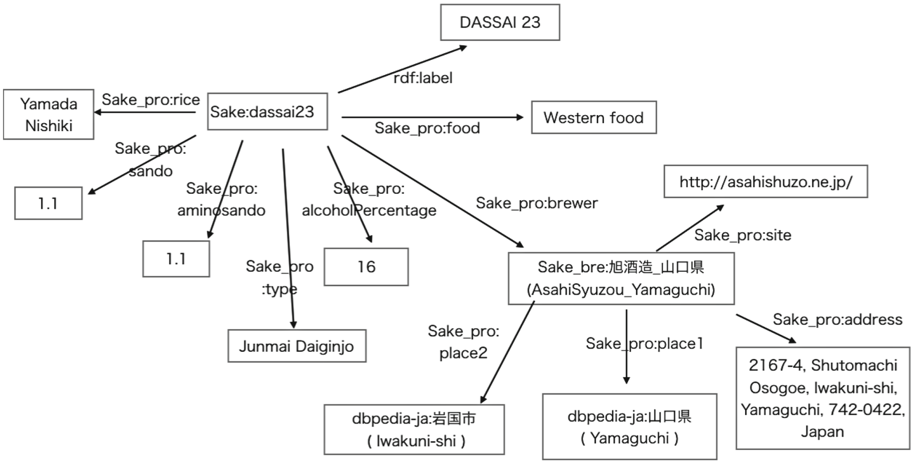
[5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

[6] http://wiki.dbpedia.org.

[7] http://www.ohsuga.is.uec.ac.jp/sake/property/wiki.

[8] http://www.clair.or.jp/j/exchange/shimai/data150831.xlsx.

[9] http://cheese-factory.info.



```
PREFIX Sake:<http://www.ohsuga.is.uec.ac.jp/sake/>
PREFIX Sake_bre:<http://www.ohsuga.is.uec.ac.jp/sake/brewer>
PREFIX Sake_gov:<http://www.ohsuga.is.uec.ac.jp/sake/gov>
PREFIX Sake_pro:<http://www.ohsuga.is.uec.ac.jp/sake/property/>
PREFIX dbpedia-ja:<http://ja.dbpedia.org/resource>
PREFIX rdf:<http://www.w3.org/2000/01/rdf-schema/>
```

Fig. 3. Example of RDF

Table 1. List of properties

Defined property	Description
Sake_pro:brewer	sake brewery
Sake_pro:type	Type
Sake_pro:volume	Volume
Sake_pro:alcoholPercentage	Alcohol percentage
Sake_pro:rice	Rice used in the brewing
Sake_pro:food	Food that matches well
Sake_pro:temperature	Temperature suitable to drink
Sake_pro:smellTaste	Smell and taste
Sake_pro:price	Price
Sake_pro:site	Sake brewery website
Sake_pro:address	Address
Sake_pro:place1	Address1
Sake_pro:place2	Address2
Sake_pro:wiki	Word of Wikipedia which has the relation

For another example, in the case of “Kagatobi Junmaidaiginjou”(a sake name), the sake leads to “Fukumitsuya”(Brewer) → “Kanazawa City”(Location) → “COIL A CIRCLE OF CHILDREN”(Cartoon based on the location) → “Wearable Computers”(Key items in the cartoon) → “The Expendables 3”(Movie that uses the same items) → “Arnold Alois Schwarzenegger”(Actor in the movie) → “California” (State that the actor has been inducted into the office of governor) → “RIESLING SONOMA COUNTY”(Wine of the state). Figure 4 shows the above relation. The server program executes several SPARQL queries. Then, if it obtains wines in the resulted connections, it sends the data of the wine and any related contents to the client side of the application.

4 Evaluation

The purpose of the evaluation is to measure effectiveness of this application by analyzing the user behavior. In addition, we accessed whether the user is interested in sake and wine, or not.

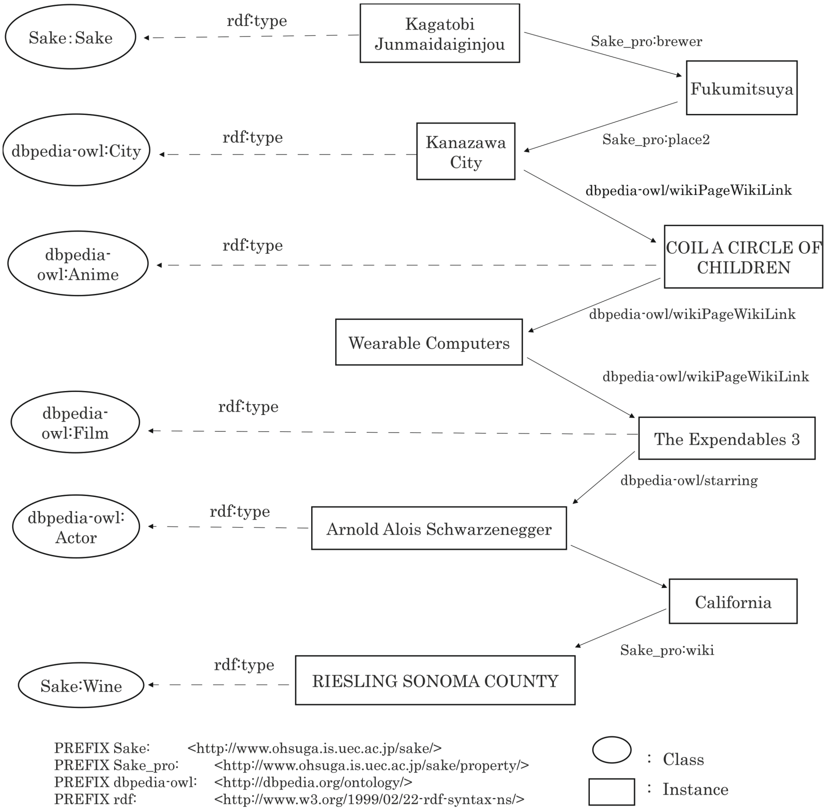
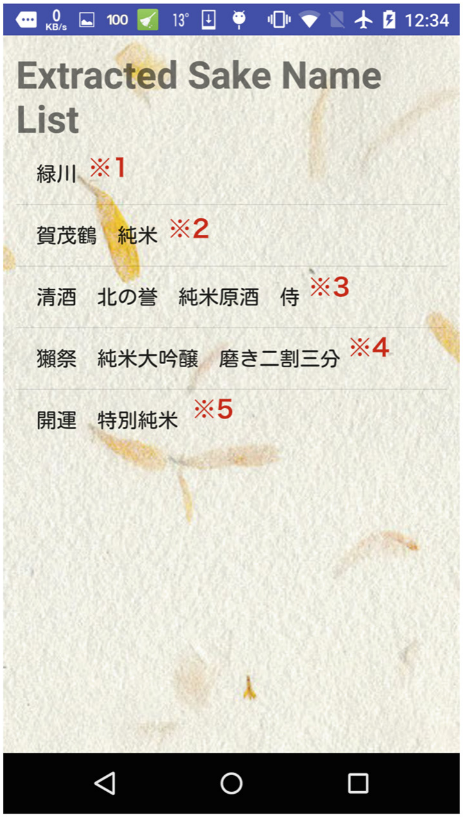

Sake:Sake
rdf:type
Kagatobi Junmaidaiginjou
Sake_pro:brewer
Fukumitsuya
dbpedia-owl:City
rdf:type
Kanazawa City
Sake_pro:place2
dbpedia-owl/wikiPageWikiLink
dbpedia-owl:Anime
rdf:type
COIL A CIRCLE OF CHILDREN
dbpedia-owl/wikiPageWikiLink
Wearable Computers
dbpedia-owl/wikiPageWikiLink
dbpedia-owl:Film
rdf:type
The Expendables 3
dbpedia-owl/starring
dbpedia-owl:Actor
rdf:type
Arnold Alois Schwarzenegger
California
Sake_pro:wiki
Sake:Wine
rdf:type
RIESLING SONOMA COUNTY
PREFIX Sake: <http://www.ohsuga.is.uec.ac.jp/sake/>
PREFIX Sake_pro: <http://www.ohsuga.is.uec.ac.jp/sake/property/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns/>
: Class
: Instance

Fig. 4. Example of search method



Extracted Sake Name List

緑川 ※1

賀茂鶴　純米 ※2

清酒　北の誉　純米原酒　侍 ※3

獺祭　純米大吟醸　磨き二割三分 ※4

開運　特別純米 ※5

※1 : Midorikawa
※2 : Kamotsuru
※3 : Kitanohomare
※4 : Dassai
※5 : Kaiun

Fig. 5. Screen of comparison application

4.1 Experiment on Effectiveness

In order to confirm the effectiveness of providing sake and wine connections, we utilized Google Analytics v4[10]. Google Analytics is a free tool that can analyze users' behavior in the application. The evaluation items are the average residence time of each screen in terms of screen view and view rate. We also created an application that displays only the information of sake for comparison, which corresponds to the conventional application described in Sect. 2. If a user takes a picture of a sake menu, a list of sake names is read and displayed as shown in Fig. 5. Then, if the user taps a sake name, only the information of the sake is displayed as shown in Fig. 1(c). As with the proposed application, the brewer site is also displayed as shown in Fig. 1(d) when the user taps the URI. We compared

[10] https://developers.google.com/analytics/devguides/collection/android/v4/.

Table 2. Results of evaluation

	Degree of interest of sake	Degree of interest of wine	Avg.time on screen(seconds)	Screen views	%View
Compared application	1	1	0:00	0	0.00 %
	1	2	0:13	6	100.00 %
	2	1	0:24	3	100.00 %
	2	2	0:22	6	57.15 %
	2	3	0:06	7	50.00 %
	3	2	0:22	5	60.00 %
Proposed application	1	1	0:00	0	0.00 %
	1	2	0:55	3	100.00 %
	2	1	0:09	2	100.00 %
	2	2	0:39	7	62.50 %
	2	3	0:19	5	100.00 %
	3	2	1:06	4	75.00 %

these two applications by their subject use. The evaluation items included the screen residence time, the number of screen views and the view rate in the screen of Fig. 1(d), which is the last screen. The user information is the degree (1: hate, 2: neither, 3: love) of preference for wine and sake. We invited 20 users to participate in the evaluation. However, the number of sake varieties used for the sake menu in the experiment was five for now.

4.2 Performance Comparison

Table 2 shows the result of the evaluation. In terms of the average screen residence time, the screen staying time of the proposed application was longer than the applications to compare. For people who answered that they are not much interested in sake, the average screen residence time in the application to compare was 13 (sec), but the proposed application achieved an average of 55 (sec). Although there was no change in the number of screen views, the proposed application has higher scores than the application to compare in terms of the view rate. The average view rate of the proposed application was 73.00 %. On the other hand, the average view rate of the application for comparison was 61.20 %. If the screen residence time and the view rate will increase, the possibility that the users see the sake brewer sites will also increase. Thus, we can confirm the effectiveness of the proposed application.

5 Related Work

Sakenomy[11] is an existing application, providing a service related to drinking sake. Sakenomy is a sake information retrieval application that uses the recorded information of sake. Information that is recorded in the application is about 800 bottles of sake that are exhibited in a sake competition called "SAKE COMPETITION"[12]. If the user takes a picture of the label of sake, they can view information about the taste of the sake. In addition, the user can record information about sake tasting results, and it is possible to compare the results of the professional tasting with their own tasting. Ministry of Economy, Trade and Industry in Japan also developed an application similar to the above in the Cool Japan Initiative [3].

This application offers recommendations for sake selection. However, the user's preference data for sake are used for the recommendation and thus the application is not suitable for users, who are not familiar with sake.

A study of Nasugawa includes natural language processing of murmurs in Twitter [4]. This study analyzed 373 tweets including 131 shops located in Tokyo, and as a result, information about 10 taverns was obtained. Although it was difficult to identify tweets for analysis due to excessive noise, evaluation of the tavern identified was high. This showed the effectiveness of the micro-blog as a knowledge source.

As the recommendation of the relevant studies using the Linked Data, there is research of Khrouf [5]. Meta-information such as the location of the event information site is converted to a set of Linked Data. The event information recommendation system is constructed by a content-based approach. The method uses the similarity of the data structure and calculation of the sentence degree of similarity, by applying the topic model method to sentence events. Elahi et al. studied recommendation of pictures using the data converted into RDF from the user information on Facebook and Flickr [6]. Passant et al. proposed a method called "Linked Data Semantic Distance" to calculate a semantic distance between Linked Data, and performs a music recommendation [7]. Moreover, Mian et al. proposed the technique of recommending music to be associated the location information of the user [8]. Mirizzi et al. proposed a method for recommending movies by using the vector space model as a source of information for the DBpedia [9]. However, the method for recommendation from the semantic structure has not yet been applied to the liquor to the best of our knowledge.

6 Conclusion

In order to lead foreign tourists to sightseeing in the sake brewers in the countryside as the Tokyo Olympic Paralympic Games held in 2020, we developed an application that prompts the tourists to have an interest in sake by showing the surprising connections between sake and wine. Then, we evaluated the

[11] http://www.sakenomy.jp.
[12] http://sakecompetition.com.

application in terms of the view rate, the screen residence time and the number of screen views to measure the degree of relevance and interest for the users. Although the main contribution of this paper is a novel application proposal for supporting countryside tourism, the experiment of 20 test users showed that the application has the possibility that exposing the connections between wine and sake may cause the interest in sake and sake brewers in the user.

However, in this work, we developed a non-personalized application in order to avoid troublesome operations that is necessary for input the user profile (metadata). Also, the extraction of the user's tasting profile will take longer time to analyze. However, since a mechanism to customize the target wines based on tourists' metadata or specific answered questions by the user could substantially enhance the performance of the application, we intend to include the user's profile in the future version. We plan to analyze the view rate, the screen residence time and the number of screen views to estimate the user's profile including the preferable relations between the user and the sake/wine. In addition, we will incorporate more information about sake and wine, and increase SPARQL queries to find more surprising connections.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Numbers 24300005, 26330081, 26870201.

References

1. Tokyo 2020 - The Tokyo Organising Committee of the Olympic and Paralympic Games. The Tokyo Organising Committee of the Olympic and Paralympic Games. https://tokyo2020.jp/en/. Accessed 15 Sept 2015
2. Sake Brewery Tours | Immerse yourself in brewing tradition, John Gaunter, Etsuko Nakamura and Michi Travel. http://saketours.com. Accessed 15 Sept 2015. (in Japanese)
3. Cool Japan Initiative, Ministry of Economy, Trade and Industry. http://www.meti.go.jp/policy/mono_info_service/mono/creative/file/1406CoolJapanInitiative.pdf. Accessed 15 Sept 2015
4. Nasukawa, T., Yoshida, I., Nishiyama, R., Yoshikawa, K., Ikawa, Y., Ohno, M., Kanayama, H., Suzuki, S., Murakami, A.: Attempt of micro blog utilization as the knowledge source which finds a good store of sake from a large amount of tweets. In: Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing, pp. 820–823 (2015, in Japanese)
5. Khrouf, H., Troncy, R.: Hybrid event recommendation using linked data and user diversity. In: Proceedings of the 7th ACM Conference on Recommender Systems, pp. 185–192 (2013)
6. Elahi, N., Karlsen, R., Holsb, E.J.: Personalized photo recommendation by leveraging user modeling on social network. In: Proceedings of International Conference on Information Integration and Web-based Applications, pp. 68–71 (2013)
7. Passant, A.: dbrec — music recommendations using DBpedia. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part II. LNCS, vol. 6497, pp. 209–224. Springer, Heidelberg (2010)

8. Wang, M., Kawamura, T., Sei, Y., Nakagawa, H., Tahara, Y., Ohsuga, A.: Music recommender adapting implicit context using 'renso' relation among linked data. J. Inf. Process. **22**(2), 279–288 (2014)
9. Mirizzi, R., Di Noia, T., Ragone, A., Ostuni, V.C., Di Sciascio, E.: Movie Recommendations with Linked Data, IIR. In: CEUR Workshop Proceedings, vol. 835, pp. 101–112. CEUR-WS.org (2012)

A Holistic Approach to Understand Urban Complexity

A Case Study Analysis of New York City

Dominik Kalisch[1](✉), Steffen Braun[2], and Alanus von Radecki[2]

[1] Saint Mary's College of California, Moraga, CA 94575, USA
dkalisch@stmarys-ca.edu

[2] Fraunhofer-Institute for Industrial Engineering IAO, 70569 Stuttgart, Germany
{steffen.braun,alanus.radecki}@iao.fraunhofer.de

Abstract. In 2012, the Fraunhofer Society, under the leadership of the Fraunhofer Institute for Industrial Engineering, started an ambitious innovation network project called Morgenstadt: CityInsights. For this system research initiative, 12 Fraunhofer institutes worked together to analyze innovative solutions in six different cities around the globe for a sustainable city. The goal of this project was to understand the city in a holistic way, applying the approach of system engineering to the field of urban development, as well as to identify the key factors to redesign existing and newly emerging cities in a more sustainable way. In this paper we will describe a systematic and holistic approach in city analysis and illustrate initial sector-related results of the on-site research in New York City in 2013. We will further analyze project and process structures of the studied projects and describe what other cities can learn from New York City. We complete the paper with an outlook on the second project phase that started earlier this year.

1 Introduction

According to the United Nations (United Nations 2012), 60 % of the world's population will live in urban areas by 2030. While many cities around the world are growing and expanding, at the same time a large number of cities in the northern hemisphere are facing reverse trends, e.g., caused by the demographic change. As a result of these trends and the comprehensive globalization, cities are competing within a global market for companies and well-educated inhabitants. As an additional challenge, the climate change revealed his powerful forces during the last decades as seen in hurricanes Katrina and Sandy in 2005 and 2012, respectively, or typhoon Haiyan in 2013. In this context, cities are facing an extremely difficult assignment: an innovative sustainable development of the city, including ecologic, economic and social dimensions. This task includes two central requirements, making the city livable on the one hand and resilient against external factors such as natural disasters or other crises on the other. This paper outlines innovative approaches in New York City in order to achieve

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 31–47, 2016.
DOI: 10.1007/978-3-662-53416-8_3

the goal of a sustainable city of tomorrow. The paper is based on an interdisciplinary long-term research project called "Morgenstadt: City Insights" (M:CI), which analyzed innovative and sustainable solutions and projects of the city sectors mobility, water infrastructure, production and logistics, governance, buildings, energy, security and ICT in six leading cities around the world in order to identify common characteristics and structures of success stories. Therefore, the paper first presents the research methodology of the M:CI project, followed by an overview of the examined sectors, projects and cities. Subsequently, the key findings regarding the examined sectors in New York City will be presented and the role of each sector for an innovative and sustainable city development will be outlined. Finally, the paper discusses the transferability of the identified approaches and tries to illustrate possible strategies to implement such innovative and sustainable solutions.

2 Morgenstadt: City Insights Project

The following section of the paper provides a brief introduction into the M:CI project. First the underlying idea for the project is outlined, followed by the developed and applied research methodology.

2.1 Idea

The urban knowledge economy is facing a tremendous transformation that will affect the society technologically, organizationally and systemically. Individual technological sectors, such as energy or mobility, will be affected. But since these sectors are highly cross linked, especially in cities and urban regions, the change in one sector will affect all others as well as the urban system itself. To understand the interdependent links between the urban sectors, the Fraunhofer Society launched the innovation network M:CI. For this system research initiative, 12 Fraunhofer institutes work together to investigate innovative solutions for a sustainable city. To achieve this goal a holistic research approach was developed in order to analyze the city system in its interdependent structure (Kalisch et al. 2013a). The main goal of the first period (2012–2013) of the M:CI project was to identify the status quo and establish a starting point for the research and development of innovations for urban systems. Based on the findings of the first period and the systemic understanding of urban areas, the second period (2014–2015) will focus on discovering and implementing systemic approaches that successfully respond to the increasing problems of the selected technology fields in leading cities. By detecting and analyzing innovative but already field-tested approaches, their feasibility for other complex environments and demands for an urban future will be evaluated. To verify this, expertise will be pooled to develop smart and individually customized strategies together with our network partners from industries and cities, aiming at the future requirements for further concepts' efficient implementation.

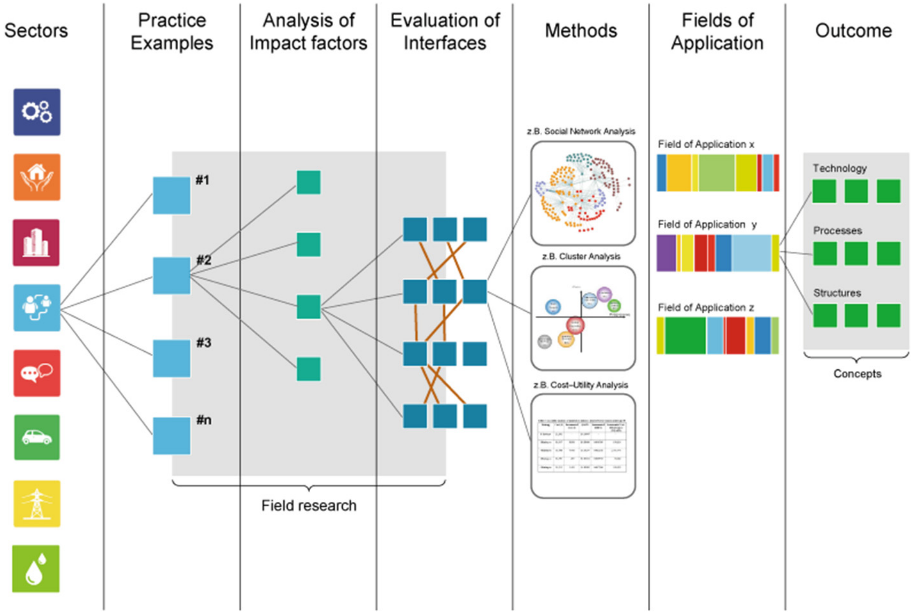


Fig. 1. Overview of the research process from sectors to areas of application (Kalisch et al. 2013a)

2.2 Methodology

The M:CI project follows a trans-disciplinary research approach; its first phase has been divided into seven phases (Fig. 1). At first more than 270 global good practices in more than 250 cities around the world that were applicable to bring the city forward towards a liveable, resilient, zero-waste and CO2 free city were studied. The examples were ranked by researchers from the corresponding field by innovative technologies, business models, forms of organization used, and the transferability to other cities. Based upon this assessment 80 solutions were defined as best practices. All 80 best practices were evaluated in a systemic way which included assessment of core sustainability indicators on social, economic and environmental impact and a cross-sectoral analysis of systemic interfaces with other sectors. The amount of identified best practices per city served as reference for the city ranking. Further, a meta-analysis of cities that appeared in different indices lists was conducted. Based on this list a meta-ranking of the cities was compiled that reflects their overall performance. The final ranking was realized by integrating the best practice-ranking (70 %) and the global meta-ranking (30 %) into one list of inspiring and leading global cities in the field of urban sustainability. The first 24 cities of the final ranking were taken as base items for defining the top 12 list. This was done by referring to the preferences of project partners, to a fair regional distribution and to a good distribution of sector-specific best practices. Based on the top 12 list, the project partners chose six cities (Berlin, Copenhagen, Freiburg im Breisgau, New York, Singapore and Tokyo) that were studied on-site (phase 1 and 2 in Fig. 1). Prior to the two week research visit, the mayor's offices were contacted and asked to support the fieldwork with a letter of recommendation and support. Additionally several other locally-based institutions such as universities, German associations, etc. were also contacted in advance to request support in lining up interviews with the persons that were responsible for the studied best practice examples.

The M:CI project team defined 15–65 indicators with the associated data for each sector in the given city and saved this information in a relational database that was developed for this project (Kalisch and Wetzel 2013). The same was done with information and data that were collected from each studied practice example in the city. Prepared with the results of this desktop research, a group of Fraunhofer researchers stayed in each of the six cities for two weeks and mainly conducted narrative interviews with relevant actors within each practice example. The interviews, typically 1.5 h in duration, were conducted on the basis of a part standardized questionnaire which was adapted to each interview. The interviews were recorded, when permitted, and later analyzed. The practice examples were, whenever possible, viewed and visited, in order to gain a personal impression.

Each night the involved researchers came together to share the insights they gained during the day. This step was not only done for a group dynamic reason, but to gain trans-disciplinary insights from the other researchers. By sharing and discussing the experiences, the researchers were challenged to view the studied example within their own sector from another perspective and also to rethink the projects of other sectors from one's own perspective (see Roe 2012; Mille Bojer et al. 2008). Additionally, all actors that were involved in the city's key projects were invited to an evening event during which the project, as well as the researcher's first impressions of the city, were presented. The city's sustainability initiatives were discussed during a panel discussion and a subsequent reception. The feedback of the participants was incorporated in the analysis and accounted in the following interviews. During the so-called "Morgenstadt: City Labs" several hypotheses relating the examined practice examples were developed following a defined methodology and discussed with the M:CI project partners. The discussions served to help the researchers recognize inherent patterns in the implementation of projects and solution approaches (phase 3 and 4 in Fig. 1). Based on the qualitative interviews and available quantitative data, impact factors for certain processes were identified. The analysis of impact factors uncovers why a certain progress happens in a particular way in a specific urban system. Accordingly, they describe general forces that push or hinder the process of sustainable development on many different levels. The identification of impact factors is complex and requires a trans-disciplinary reflection by the researchers. The researchers therefore reflected every day on the identified drivers and framework conditions. One important tool for this was collaborative mind-maps to structure the identified factors. Further, a mixed methods approach was applied, utilizing social network analysis and cluster analysis (phase 5 and 6 in Fig. 1).

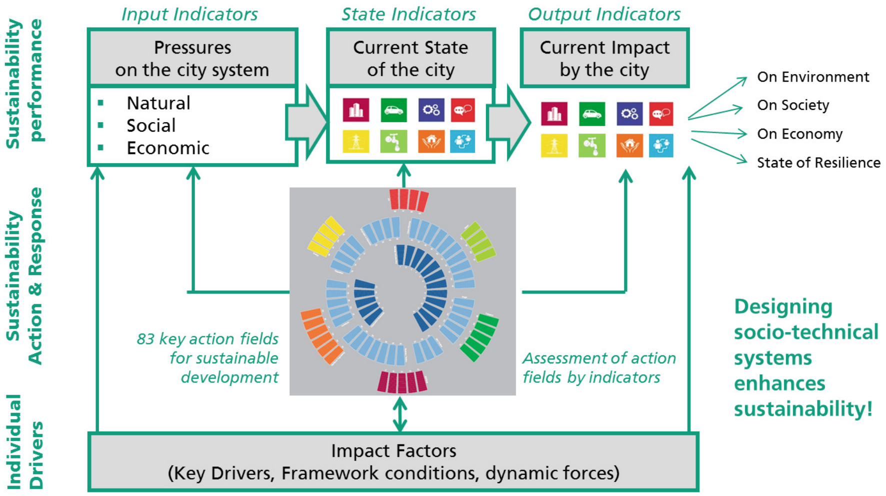

Sustainability performance
Input Indicators
Pressures on the city system
Natural
Social
Economic
State Indicators
Current State of the city
Output Indicators
Current Impact by the city
On Environment
On Society
On Economy
State of Resilience
Sustainability Action & Response
83 key action fields for sustainable development
Assessment of action fields by indicators
Designing socio-technical systems enhances sustainability!
Individual Drivers
Impact Factors
(Key Drivers, Framework conditions, dynamic forces)

Fig. 2. Morgenstadt model for sustainable urban development (Fraunhofer IAO et al., 2013a, p. 211)

Starting from a three-level-approach (indicators, impact factors and action fields) of urban systems analysis, the M:CI research network developed a first generic model for sustainable urban development (see Fig. 2). After the on-site research visits, all prior defined indicators had been evaluated. The assessment showed that most variables are only available in some cities and therefore not useful for general city comparisons. A revision of the M:CI indicators provided a set of less than 100 urban indicators that define the state of sustainability of a city. These indicators are listed in the final project report (Fraunhofer Institute for Industrial Engineering IAO 2013). The 83 defined key action fields for sustainable development represent the core of the Morgenstadt model (Wendt et al., 2014). These action fields describe the sustainable actions and responses of the cities. They can be related to indicators and allow the M:CI researchers to assess whether a response of a city is in line with existing pressures or state conditions and therefore helps optimize outputs for enhanced sustainability. The key action fields were further assessed by the participating researchers. They rated the impact of each key action field on each other based on their field of expertise. This so called cross-impact matrix of key action fields was subsequently evaluated by the sum of active and passive ratings. By plotting the sums of each key action field, three groups of action fields could be separated that have a significant relevance for sustainable development of a city (see Fig. 3).

- The “drivers” were key action fields that bring ideas and initiatives forward.
- The “enabler” enables the city to perform certain actions.
- The “levers” amplify given actions.

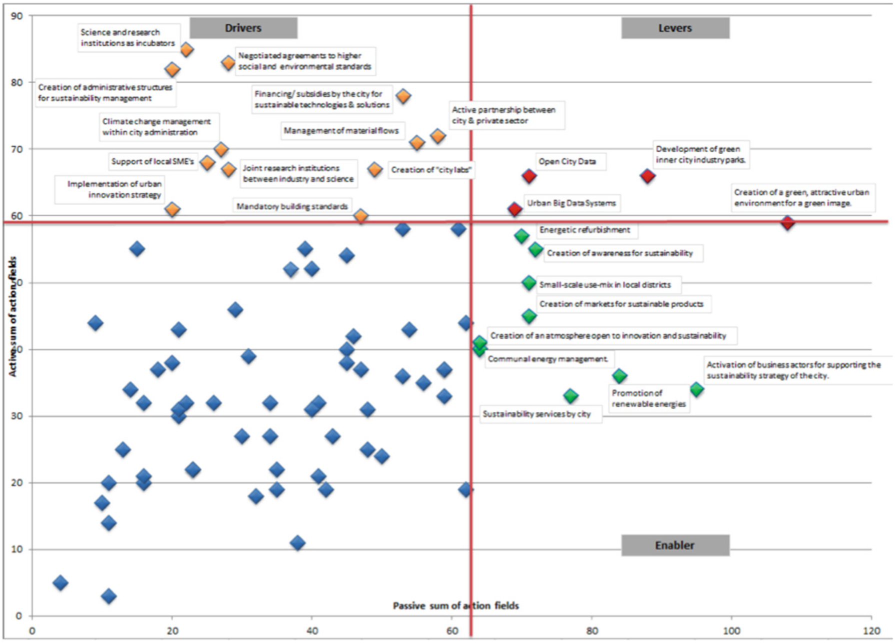


Fig. 3. Cross-impact analysis of key action fields (Wendt et al., 2014, p. 536)

The cross impact of each key action field to each other is also dynamically visualized and accessible for project members through the project website.

3 Sector Results

It has proven to be quite difficult to compare cities in terms of their sustainability and their projects designed to increase sustainability, as no uniform assessment criteria exist and because the framework conditions of each city are unique. This brings rise to the following: Is it even possible to learn from the experiences of individual cities?

The M:CI project argues that while every city with sustainability-oriented projects and approaches reacts to specific challenges, uses locally-available resources and implements its projects under local framework conditions, the main challenges addressed are, nevertheless, comparable to the challenges faced by many cities worldwide. The projects are planned and implemented according to similar patterns. As such, the objective of the M:CI project is to understand the activities within the individual cities, to identify the specific framework conditions present, and to recognize the patterns within these activities.

Thus, the M:CI research visits were conducted with the following objectives in mind:

- To analyze the selected practice examples in relation to their motivation, conception, planning, successful implementation and measurements of success;
- To identify the key drivers and framework conditions which have affected the projects and solution approaches either positively or negatively;
- To analyze the network of actors, their roles within the studied projects and their solution approaches;
- To discuss the transferability of projects and solution approaches to different cities.

For the Fraunhofer M:CI project six researchers visited New York City between April 8 and April 23, 2013 to conduct 50 interviews (Kalisch et al., 2013b) with experts, political leaders and scientists from the different sectors. The following results are a summary of the City Report for New York City (Kalisch et al., 2013b).

3.1 ICT

The cooperation between NYC's mayor and police chief has been a significant structural effect factor. The implementation of CompStat and the resulting revolutionized police work in NYC was possible thanks to former NYC mayor Rudolph W. Giuliani and former chief of police Bratton who jointly developed a strategy to improve safety in the city back in 1994. The mayor of a city has the ability to set comprehensive priorities and involve other relevant public authorities in the process; because of that, inter-dependencies with other sectors can be examined and modified if needed. Local differences in a city, and the corresponding adjustments required to adapt to individual circumstances and conditions in the various districts, pose another important factor for success. For example, in NYC local representatives are involved in the strategy formulation process for the city's police. An important part of the development of strategies and the implementation of locally adapted approaches in NYC are the CompStat meetings in which police chiefs meet with their key employees once a week to exchange knowledge on successful factors, identify existing barriers and discuss how to resolve these barriers in order to improve the city's overall anti-crime strategy. It must be ensured that such a strategy is continuously evolving and adapting in order to ensure that crucial exchange and learning is an ongoing process. Data analysis is central to the fight against crime in NYC. A continuous review of strategies and the results of procedures contribute to the ongoing evaluation of data. Information gathered on the location, time, and specifics of a crime, combined with details gathered on the offender(s), is evaluated to optimize the fight against crime. Timely evaluation is essential and effective evaluation can, for example, lead to more focused policing of certain identified areas and enhance adaptation to local conditions. Another important factor is to gain the support and involvement of the population in order to obtain information about crime in different neighborhoods. This has been achieved through community policing initiatives, which can also help to improve the relationship between the public and the police. NYC's outcome-oriented approach has been a central factor contributing to the city's continued and dramatic reduction in crime rates. The focus here has not been on predicting individual crimes but on uncovering general patterns. This approach was successfully implemented to reduce auto theft in NYC.

3.2 Security[1]

Overall, NYC is promoting three key strategic security missions: catastrophe and disaster management, big data, and infrastructure protection. In the wake of Hurricane Sandy, NYC has undergone vital measures to better prepare for and respond to natural disasters and the short and long-term consequences thereof. Based on the successful implementation of PlaNYC, A stronger and More Resilient New York, a nearly US $20 billion resiliency plan, was implemented. This plan is a comprehensive endeavor to unite and concentrate the city's core capabilities in the field of sustainability with the aim of incorporating infrastructure and activities related to the built environment, such as coastal protection, insurance, utility supply, healthcare, water and transportation with specific community rebuilding efforts and resilience planning. The plan foresees the participation of not only official and professional bodies, but also New Yorkers themselves and therefore works to keep residents thoroughly informed on the various initiatives and projects announced in the plan. Hurricane Sandy hit NYC and the surrounding urban areas with such unexpected intensity that experts agree that the city and its neighbors have begun to reconsider the city's close proximity to the ocean and the threats that may occur due to its specific location. Thus, the NYC Office of Emergency Management (OEM) is revising all flood and security-related maps to better prepare for both natural disasters and man-made catastrophes. Big data systems are at the forefront of NYC's security strategy. The city's surveillance system, known as the Domain Awareness System (DAS), which was launched by the NYPD, provides an example of the city's interconnected big data systems. The DAS combines CCTV camera footage, reports from over 3,000 radiation sensors, license plate detectors and public data streams for the identification of threats on the streets. NYC has made it a priority to support crime prevention as well as crisis management operations using existing as well as new sensor and data systems which are based on the sharing of extremely large amounts of data. Such interoperable information gathering systems have become crucial to the work of all security-related authorities. Systems such as NYPD's DAS are designed to be transferable to other metropolitan areas which are equally densely populated and have a similar urban infrastructure. However, the cultural context in which such systems are placed is crucial for their implementation since they may interfere with civil and privacy rights causing controversies and a lack of acceptance among citizens. As a third fundamental security mission, NYC is on the forefront of critical infrastructure and building protection. The city is still deeply stricken by the

[1] This paragraph is co-authored by Hanna Leisz.

very recent consequences of Hurricane Sandy and the events of September 11 have left the city deeply scarred. The reconstruction of the World Trade Center as a key business district is strongly grounded in developing technological and emergency response-related security measures. In particular, site access control systems, above all the Vehicle Security Center, show that preparation for a possible terrorist attack is a core motivator of the overall security planning and implementation measures taken for both individual building complexes as well as surrounding interconnected infrastructure complexes in the corresponding city districts.

3.3 Water[2]

Since 1842 New York City has received water from outside the city's boundaries. Nowadays, more than 9 million inhabitants and visitors of the city are relying almost completely on water sources up to 250 km away from the city. Consequentially Mayor Bloomberg asked, as he came into office, "What could literally close down this city?" A failure of the supply system, transporting water into the city would have done that (Flegenheimer 2013). While the water supply infrastructure was aging, several droughts in the 1980s made the limitation of the water resources obvious. At the same time the population was and still is steadily growing. Due to these conditions, the city successfully started several strategic plans and initiated measures to achieve water conservation, to modernize the existing supply infrastructure, and to guarantee that the water resources will be sufficient for serving the population even in future times. While the city set up rules for water conservation, in one prominent district, the Battery Park City (BPC), even higher standards were developed by the local authorities, that have to be achieved for new buildings, leading to most innovative solutions in terms of water reuse and efficiency, decentralized waste water treatment, and energy efficiency within buildings. The practice examples of BPC are impressive showcases, presenting the water reuse and efficiency potential in combination with a high level of living quality in modern buildings within densely populated areas of a city. Increased awareness of the city's attractiveness brought the value of the many surface water bodies of the city more and more into focus in recent years. At the same time, more frequent flooding of an ever broader range of communities occurred, leading amongst others to regular combined sewer overflows (CSO) into the city's waterways. To prevent flooding and to avoid the pollution of the water bodies by CSOs, several strategic issues, such as the Sustainable Stormwater Management Plan, were incorporated within the city's strategic master plan, PlaNYC. The different issues NYC is confronted with in the water sector occur all over the world more and more often. The solutions of the city, the strategic processes targeting many small and larger measures, and its consequent implementation with a documentation of its progress, can help cities everywhere to cope with their individual issues. However, the efforts

[2] This paragraph is co-authored by Felix Tettenborn.

New York City has undertaken depend to a large extent on the active engagement of the authorities, on the awareness of the population and last but not least on the technological progress, which still has not come to an end.

3.4 Buildings[3]

One of the strongest factors in NYC's recent development is the governmental support of building innovation, energy efficiency and sustainable city planning. A clear guideline for all decision makers and offices is manifested in PlaNYC. This helps provide transparency and facilitates faster processing and decision-making. The energy efficiency regulations have a strong influence on building development, both for new buildings under construction and old buildings required to undergo retro-commissioning. As part of the Greener Greater Buildings Plan (GGBP) local laws were implemented to insure energy audits of larger buildings. Such laws create new understanding and demonstrate that economic incentives for improvements and innovation pay off in the long term. It is important to remember that while sustainability is the goal, sustainable development is only achievable if it is proven financially viable. Therefore, investments into green building practices and retro-commissioning must be able to prove themselves economically beneficial in order to succeed and become widely adopted. Another way of creating better understanding of critical environmental issues is through education on sustainability. CUNY, a 'green university', provides an excellent case in point. The university is collaborating with the local government on a project that will, in time, help shape public opinion and make developers and residents aware of the need for sustainable buildings, thereby turning sustainability features into something people will value and want in a building. CUNY's green campuses set a positive example of green development and exemplify values of sustainability in a public space thus creating curiosity and admiration. The education and programs provided by the university produces future experts in sustainable technologies and trades. Additionally, program graduates have practical experience from contributing to their universities' green development initiatives. A green university is the ideal place to conduct research on developing new methods and concepts for sustainable buildings and cities. Another strong concept to create economic benefit from sustainable buildings is the public-private-partnership (PPP). By entrusting the project with valuable goals and clear guidelines to a private partner, to implement and treat it as a normal source of income, the government can reduce its financial investment. On the other hand, the private partner is provided with a profitable project that would not have been available to them without the incentives provided by the government. In this way, innovative projects can be realized much faster and with more security for both parties involved.

[3] This paragraph is co-authored by Elvira Ockel.

3.5 Mobility[4]

NYC ranks first in the nation in terms of passenger miles flown, transit passenger miles traveled and truck freight volume. In the year 2006, transit alone accounted for 1.8 billion passenger trips carrying 8 million passengers per day (almost 70 % in subways). New Yorkers are heavily dependent on public transportation and have a much lower car ownership rate (23 %) than any other major city in the country (78 % average). Moreover, NYC is the only city in the United States where more than half of the households do not own a car. Were the city to follow general car ownership patterns, the city would have an additional 4.5 million cars on its streets. The transport sector emitted 11.4 million tons of CO2 in 2010 (69 % from passenger cars) and is the second largest CO2 emitting sector after electricity generation. Due to low private car use, about 48 billion miles (approx. 77 billion km) of travel are avoided yearly, saving the city 23 million tons of transport-related CO2 emissions.

3.6 Governance[5]

In 2007 the master plan for New York City, the ‘PlaNYC 2030’ has been released and attracted attention as a global example of sustainable community and economic development. Three main challenges functioned as key drivers for the development of a comprehensive, strategic plan for NYC’s development: the expansion of population, the city’s aging infrastructure and the impacts of climate change on NYC. Moreover, the 9/11 events have raised awareness that a city must not only provide public services, but also create a safe space in which the future-oriented economic, social and environmental needs of a diverse and prosperous city can be met. Furthermore, projections for climate change impacts on the Big Apple highlighted the need for NYC to take action by preparing for inevitably negative impacts while striving to minimize its own impact on global warming. Thus, the concepts of sustainability and resilience became central guidelines for the future development of NYC. PlaNYC is an ambitious agenda aimed at creating a ‘greener, greater New York’ even as the city’s population continues to grow towards a projected nine million residents by 2030. The ten fields of action which are part of the city’s sustainability strategy include: Parks and Public Space, Energy, Brownfield, Air Quality, Waterways, Solid Waste, Climate Change, Water Supply. Additionally, PlaNYC presents seven topics, which are cross-sectoral: Public Health, Food, Natural Systems, Green Building, Waterfront, Economic Opportunity, and Public Engagement. The conception of PlaNYC and the implementation of its numerous initiatives is the result of a joint effort on part of the city, state and federal governments, citizens, neighborhood groups, non-profit organizations, community boards, private companies, as well as research institutions and universities. While McKinsey and Company assisted in writing the plan, the Mayor’s Office of Long-Term Planning and Sustainability (OLTPS) released the plan. Support from the mayor and

[4] This paragraph is co-authored by Martha Loleit.

[5] This paragraph is co-authored by Katrin Eisenbeiss.

top administration officials has been fundamental for the successful and efficient implementation of PlaNYC.

4 Analysis of Projects and Processes

The description of structures within a city must always be understood as a still-life, capturing a specific moment in time. The transformation of a city towards a sustainable state requires the transformation of these structures, which is why the analysis of projects and processes - taking into account their time-related dimensions - are of central importance in this research project. The key question is: What is required in order to shape these transformational processes successfully in each individual project? In order to identify the causes underlying the successful implementation of projects, it is helpful to divide the processes into project phases, as shown in Fig. 4. Each project phase depicts a different structure of actors involved. A project tends to be successful only when the implementation of all phases is successful. If, for example, a project's goals are not clearly enough defined, or, if at the end of the project the resources available are not sufficient or the responsibilities have not been laid out clearly enough, optimal project implementation will not be achievable. The approach of dividing the process into project phases can be applied to individual projects, long-term accompanying processes (such as, for example, the Sustainability Council) as well as the entire transformational process towards a more and more sustainable future as a whole.

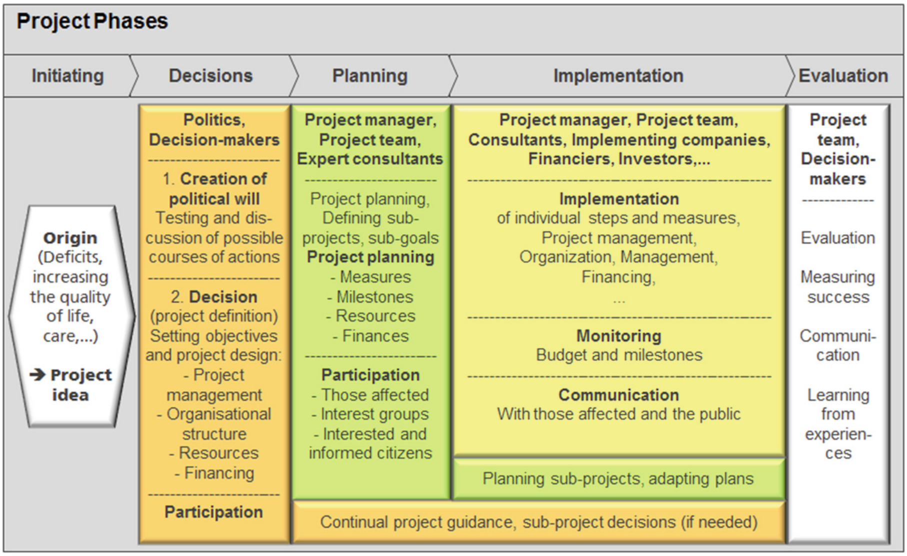

Project Phases
Initiating
Decisions
Planning
Implementation
Evaluation
Origin (Deficits, increasing the quality of life, care,...)
➔ Project idea
Politics, Decision-makers
1. Creation of political will
Testing and discussion of possible courses of actions
2. Decision
(project definition)
Setting objectives and project design:
- Project management
- Organisational structure
- Resources
- Financing
Participation
Project manager, Project team, Expert consultants
Project planning, Defining sub-projects, sub-goals
Project planning
- Measures
- Milestones
- Resources
- Finances
Participation
- Those affected
- Interest groups
- Interested and informed citizens
Project manager, Project team, Consultants, Implementing companies, Financiers, Investors,...
Implementation
of individual steps and measures, Project management, Organization, Management, Financing,
...
Monitoring
Budget and milestones
Communication
With those affected and the public
Planning sub-projects, adapting plans
Continual project guidance, sub-project decisions (if needed)
Project team, Decision-makers
Evaluation
Measuring success
Communication
Learning from experiences

Fig. 4. Typical project phases in a transformation process (Fraunhofer IAO et al., 2013b, p. 105)

4.1 Key Success Factors

Successful implementation of a project depends on solid planning. However, external drivers exert pressure on projects, which influences successful implementation. Some of these factors and their effects are known at the beginning of the project. These will exert influence throughout the duration of the project and are already taken into consideration during the planning phase. Other factors only become significant during the course of the project, and may require adaptation of the project. Both types of factors - and the boundary between the two is fluid - can prove to be either beneficial or damaging to the project. This research has the goal of identifying the most important drivers within a city, in order to understand the reasons behind the courses the projects take and to gain insight into the transferability of the practice examples analyzed. This is valuable information, since it can be assumed that transferability is a given, provided the most important factors (in this case success factors) within the city studied are also present in the city the project is being transferred to. In NYC's practice examples, 36 factors were identified with varying effects on the successful implementation of the practice examples. The factors were assigned to one of twelve categories, which led to an average of 3.61 factors per category.

4.2 Reciprocity of Factors

Figure 5 visualizes the reciprocity of the factors. The placement of the factors was selected using the Kamadakawai-algorithm, which chooses the position based on the centrality index of the corresponding node. We can see that even though Mayor Bloomberg has a higher number of nominations, the three factors, 'public available data', 'open mind' and 'evidence-based policing', have a more central

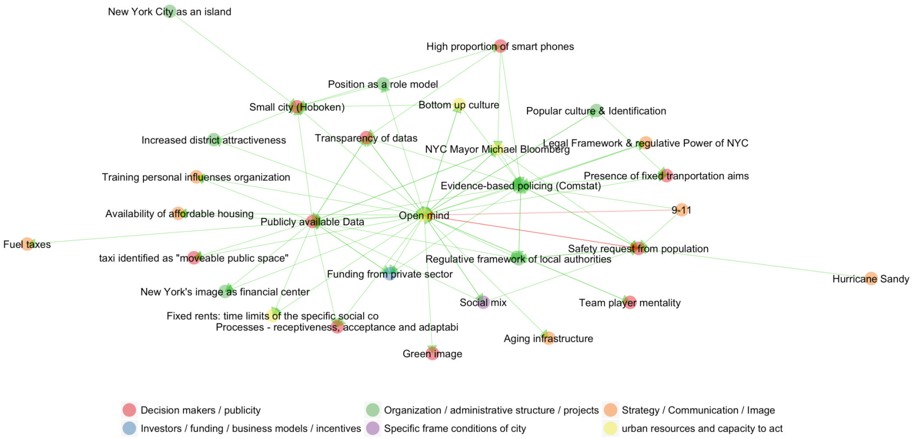

New York City as an island
High proportion of smart phones
Position as a role model
Small city (Hoboken)
Bottom up culture
Popular culture & Identification
Increased district attractiveness
Transparency of datas
NYC Mayor Michael Bloomberg
Legal Framework & regulative Power of NYC
Training personal influenses organization
Presence of fixed tranportation aims
Evidence-based policing (Comstat)
Availability of affordable housing
Open mind
9-11
Publicly available Data
Fuel taxes
taxi identified as "moveable public space"
Safety request from population
Regulative framework of local authorities
Hurricane Sandy
Funding from private sector
New York's image as financial center
Social mix
Team player mentality
Fixed rents: time limits of the specific social co
Processes - receptiveness, acceptance and adaptabi
Aging infrastructure
Green image
Decision makers / publicity
Organization / administrative structure / projects
Strategy / Communication / Image
Investors / funding / business models / incentives
Specific frame conditions of city
urban resources and capacity to act

Fig. 5. Representation of the reciprocity of the factors. Positive interactions are coded in green, negative interactions in red. (Color figure online)

position in the NYC urban system, at least in the investigated projects. Of these factors ‘open mind’ is in a prominent position. This becomes obvious when we take a look at the out degrees (Fraunhofer-Institute for Industrial Engineering IAO 2013). The open-minded population of NYC is a central factor in the success of the city’s project implementation and is one of the main cultural foundations of this city. Residents’ open-mindedness has allowed the city to forge new paths without meeting resistance. A good example of this is the availability of venture capital for start-ups. Where in Germany a start-up needs to prove a concept by referring to the successful implementation of other projects and processes, start-ups in the United States and especially in NYC have easier access to venture capital because, even if there is no proof of concept, the start-up can acquire capital if it can convince the stakeholders that their idea is innovative. This fundamental cultural characteristic opens the door to trying out new concepts that are unthinkable in German cities. However, this advantage comes with a price. On the one hand, actors in NYC can test innovations which elsewhere would be smothered in the early discussion stage. On the other hand, they run the risk that the project develops in a way that could negatively impact the population. An example is the data-driven society. The open data initiative has huge advantages in the blending of different entities or in a better understanding of social systems. The drawback, however, is that such systems can easily jeopardize citizens’ security and privacy.

4.3 Impact Factors

The most influential impact categories are the urban resources and political actors. The most influential political actor is, as already mentioned above, Mayor Bloomberg, who stepped down as Mayor in 2013 after 12 years in office. It is not possible to estimate what future impact his successor, Bill de Blasio, will have on NYC. Aside from the mayor there are also other political actors who are important for the described projects. For instance, in the case of the Open Data Initiative, Gale Arnot Brewer is of particular importance.

5 Learning from New York City

One of the central elements in NYC is the usage of data and IT. However, the usage of data and IT is not an end in itself. The process started with the citizens’ request for an overview of the city’s data in order to make the government accountable and to increase transparency. The citizens wanted to know what their tax money was being used for. United States residents, particularly New Yorkers, realized that economic market principles could also be used in governmental and political processes. Therefore, under the leadership of Mayor Bloomberg, the NYC administration implemented an assessment system that sets verifiable goals and measures their status with defined indicators, which were enshrined in PlaNYC, before applying policies as well as during the implementation process. Only if a policy is successful will the government continue the program without making adaptations. If a policy is not successful, the initiatives are either adjusted or stopped.

In NYC this evidence-based governance is highly IT and data driven. For this reason, Mayor Bloomberg created the ‘Office of Policy and Strategic Planning’, a group of civic-minded number crunchers, lead by Michael Flowers, who work directly with the mayors office. Flowers, while not connected to New York’s political system, was an external person with a good idea - using predictive informational techniques - that he presented to John Feinblatt, the Mayors chief policy adviser. Flowers, however, is not the only external person who has been brought on board by the city’s administration. The Bloomberg administration was known for seeking out expert knowledge when necessary to become more objective and evidence-based. As a result, the solution for a lot of things are not only based on ideology but more and more on the question of ‘does it work? Does it have a measurable benefit?’.

Applying this approach to the studied practice examples gives a diverse answer to questions about the projects’ benefits and adaptability. If we look at a project that has a comparatively low density, such as ‘Via Verde’, we need to conclude, according to Edward Glaser (Glaeser 2012), that from the perspective of sustainability this is not beneficial, however, it is from a community perspective. Based on this information, we now can decide which we consider more important. In other words, a decision must still be made, however, the decision is now based on a more objective analysis. To provide another example, we can also conclude that the ‘Electric Vehicle Pilot’ project works in NYC because of the city’s population density. We know that such a project can be adapted by cities with a similar density but should question whether it would also be successful in a low-density area. The IT and data approach, and the resulting increase in transparency, is not only useful for holding the government accountable but also for monitoring and assessing individual decisions and gives consumers a basis for their decisions so that they can make informed choices. The Solar Map initiative, for example, enables citizens to calculate the return on investment of the installation of a solar panel in any given location. Likewise, the LEED certificate provides information on building construction and retro-commissioning and provides estimates in regard to estimated costs. Overall, data and ICT plays a central role in NYC. We can say that NYC is the most ICT-based city of all cities studied in this project. It is important to note that the IT systems used enable the information usage and increase the accessibility to such information (i.e., publish data, analyze data, etc.). They are not sustainable by themselves, but can be used as a tool for sustainability. ICT is also used to automate a lot of processes like water treatment, quality measurement and security surveillance. The positive effects of this approach come at a cost. To get a benefit out of the data, one needs to be able to analyze it and understand the implications of the results found. This requires a high level of education, and computer science and statistics are becoming increasingly fundamental abilities, similar to reading and writing. Those who are unable to understand these cultural techniques are more likely to be over-proportionally disadvantaged. Knowing this, NYC tries to enhance the public school system and improve its universities as well as found new ones. Such initiatives are economically beneficial as they attract knowledge-based companies. Likewise, existing universities adjust their programs accordingly and offer more data-driven degrees and degree programs while also focusing more on sustainability aspects, like CUNY is doing. Overall, we can summarize the process as the transformation from an economic system to a knowledge-based system. We can see that Berlin is on a very similar path. It is approximately at the position that NYC was in about ten years ago. If Berlin continues down this path, similar approaches and results may be seen in Berlin in the future as were observed in NYC. In addition to being related to ICT, the success of NYC is also rooted in its cultural setting. The United States in general, and NYC in particular, has a very strong grass roots movement, which originates in strong community (not necessary neighborhood) relationships. This leads to a 'team player' mentality that is dominant in almost all studied projects. The citizens are also very open-minded and willing to try out new approaches and methods. The benefits of evidence-based policy (e.g., a tremendous reduction of crime within the city limits) strengthen this effect additionally because the policies can be seen to have a direct benefit. In addition to its cultural characteristics, it is interesting to see that New York City - under Bloomberg - had a very central style of planning. This was physically expressed in the arrangement of the mayor's office: his desk was in the middle of an open office surrounded by his employees. He was responsible for the data driven approach, the PlaNYC, OLTPS and other similar initiatives. Central support increases a project's weight and reputation. However, the city government, for the most part, functions as a framework that sets project boundaries while the actual implementation is often realized in a Public-Private-Partnership. The sustainability efforts must also be understood under this maxim. The government sets the goal for the city to become more sustainable, but the approaches need to have a positive measurable outcome for the city. Based on the culturally-founded subsidiarity principle, Mayor Bloomberg, like the intellectual urbanists Benjamin Barber (Barber 2013) or Edward Glaeser (Glaeser 2012), sees the city as being responsible the problems and able to provide the solutions for the challenges in sustainability.

6 Prospect

The recently started second phase of the Morgenstadt project will be a transformation of the project into an ongoing alliance of industry, cities, and research partners that will join forces for the purpose of accelerating innovation throughout the various research sectors and for creating showcases for transformative urban projects. The focus in this phase of the project will be the development of detailed, innovative cross-sectoral urban sustainability projects and their implementation within context-specific complex city systems. The primary mission of the City Insights Network will be to identify, conceive, initiate and implement pilot and demonstration projects for sustainable urban solutions in cities around the world. Projects will be developed in variable consortia made up of industry, city, and research partners. The City Insights Network is designed to address the challenges that were mentioned above with a new collaborative approach. The aim of the second phase of the Morgenstadt project is therefore to initiate and accelerate the long-term transitions of selected cities towards sustainable urban systems and to thereby create international reference projects on the level of entire cities. Morgenstadt aims to become the first global alliance for planning and implementing large-scale sustainable urban solutions in a range of cities around the world.

References

Werkzeuge des Wandels: Die 30 wirksamsten Tools des Change Managements. Stuttgart (2012)

Benjamin, R.B.: If Mayors Ruled the World Dysfunctional Nations, Rising Cities. Yale University Press, New Haven (2013)

Flegenheimer, M.: After Decades, A Water Tunnel Can Now Serve All of Manhattan. The New York Times, New York City (2013)

Fraunhofer IAO, Fraunhofer IFF, Fraunhofer IBP, Fraunhofer IGB, Fraunhofer ISE, Fraunhofer IML, Fraunhofer EMI, Fraunhofer IPA, Fraunhofer FOKUS, and Fraunhofer ISI: Innovation Network "Morgenstadt: City Insights". Technical report, Fraunhofer-Gesellschaft, Stuttgart (2013a)

Fraunhofer IAO, Fraunhofer ISE, and Fraunhofer IBP: Morgenstadt City Report. Technical report, Fraunhofer Gesellschaft, Stuttgart (2013b)

Fraunhofer-Institute for Industrial Engineering IAO. Morgenstadt: City Insights. Technical report, Stuttgart (2013)

Glaeser, E.: Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier, 2nd edn. Penguin Books, London (2012)

Kalisch, D.P.H., Wetzel, T.: Morgenstadt DB. Technical report (2013)

Kalisch, D.P.H., Schatzinger, S., Braun, S., von Radecki, A.: Morgenstadt: City Insights. A research approach for systems research in urban development. In: Proceedings RealCORP 2013 (2013a)

Kalisch, D.P.H., Tettenborn, F., Eisenbeiss, K., Loleit, M., Leisz, H., Ockel, E., von Radecki, A.: Morgenstadt City Report. Technical report, Fraunhofer Gesellschaft, Stuttgart (2013b)

Mille Bojer, M., Roehl, H., Knuth, M.: Mapping Dialogue Essential Tools for Social Change. The Taos Institute Publications, Chagrin Falls (2008)

United Nations: World Urbanization Prospects. The 2011 Revision. Technical report, New York (2012)

Wendt, W., Kalisch, D.P.H., Vandieken, T., Engelbach, W.: Smart cities and ICT-insights from the morgenstadt project. In: Proceedings of RealCORP 2014, pp. 533–541, May 2014

Real-Time Data Collection and Processing of Utility Customer's Power Usage for Improved Demand Response Control

Shawyun Sariri(✉), Volker Schwarzer, Dominik P.H. Kalisch,
Michael Angelo, and Reza Ghorbani

2540 Dole Street Homes Hall #302, Honolulu, HI, USA
{shawyun,volkers,mangelo,rezag}@hawaii.edu,
dkalisch@trinity.edu

Abstract. A large growth in energy demand has increased renewable energy penetration into existing power grid infrastructures, as well as spurring increased research into demand response programs. But before implementing an efficient demand response program, it is first necessary to understand the power usage behaviors of a consumer. This paper presents a real-time data acquisition system for the collection and storage of power data that will allow the study of demand response in an urban area. Demand response programs are an ideal alternative to costly energy storage and spinning reserves. Detailed power consumption data is necessary to study proper demand response programs and implement efficient control decisions. A pilot system has been implemented on the island of Oahu in Hawai'i to prove the feasibility of a data collection system in a dense urban environment. The pilot program has implemented a smart metering device that is collecting power data at a high resolution and transmitting it to a server for load forecasting analysis. The architecture of the system will be discussed as well as preliminary results and scalability of the pilot system as it relates to the implementation of the system into a large urban center.

Keywords: Demand response · Load forecasting · Power profile signature · Urban center

1 Introduction

A 2013 report by the American Society of Civil Engineers (ASCE) gave the American electrical grid a "D+" rating, on an A to F scale [1]. Operation failures were mentioned as a main source of outages across the country because of congestion in transmission lines. Utility companies are relying on a current grid infrastructure that still has components from the 19th century. Expanding US energy capacity after 2020 will be a main concern for the utilities, and one way to alleviate some of the pressure of capacity expansion will be to increase consumer side power generation using renewable resources, but the addition of more generation comes with logistical issues, especially when transmitting power generated from stochastic sources. Rather than invest large amounts of money replacing the current grid infrastructure, the ASCE suggests research

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 48–63, 2016.
DOI: 10.1007/978-3-662-53416-8_4

in smart grids and real-time forecasting as alternatives. Smart grids will be necessary in areas containing dense populations such as large urban centers.

With a majority of the world population living in urban areas by 2030 [2], cities themselves will need to become large power generators. This is because in times of peak power draw, factors such as lengthy transmission lines and lack of fuel supply can have crippling effects on a large urban populations, as was the case in 2014 during the US Polar Vortex [3, 4]. Cities have been turning to distributed generation (DG) as a way to become more self-sufficient in regards to power generation [5]. This is because many DG units now allow for more reliability, increased efficiency and cost effectiveness as well as an opportunity to use renewable generation sources [6]. Hybrid renewable systems being used as distributed generation (DG) provide a way for utility companies to move peak loads and deliver reliable power transmission [7].

Utilizing DG can allow a more reliable and cost-effective solution to consumers, and in cases where renewable generation sources are installed, a more maintainable and ecofriendly alternative to fossil fuels [6]. However, with more DG generation becoming interconnected into the current grid infrastructure, and DG sources potentially feeding power back into the current grid system, utilities will need to be able to better monitor different points within the grid to ensure grid stability. As renewable energy generation becomes more abundant and affordable, distributed generation use will only increase and become more interconnected with current grid infrastructure, necessitating a further need to collect large amounts of data to analyze and predict grid states in real-time. To do this, smart meter devices will be needed to collect large amounts of grid data to be analyzed. Thus contributing to the development and maintenance of demand response programs.

Utility companies today are needing to evolve from their historic position of producing energy, to managing energy production not only from the supply side, but from the consumer side as well. The topic of this paper revolves around the implementation of a pilot system that allows a power producer to collect large data and analyze it in order to create cost effective energy management strategies for urban centers.

1.1 A Smarter Grid

The transition to a "smarter" grid will grant utilities the ability to become more proactive in how they manage power supply in the transmission infrastructure. In the past, utility companies have needed to increase spinning reserves, and invest in generators with faster start up times to counter intermittent generation created by renewable energy sources [8–10]. Demand response is an option to alleviate the issues that come with renewable energy penetration, and are an alternative to costly large scale energy storage [10]. Even though there has been research into the feasibility of renewables into the current grid infrastructure, utilities and policymakers find themselves still requiring ways to understand the benefits and drawbacks of demand response programs [11, 12].

The North American Electric Reliability Corporation categorized demand response as a "subset" of Demand-Side Management (DSM), which looks to create efficient energy programs focused on the consumer end (node) of power consumption [13]. Many current grid infrastructures have a utility generating energy at a plant, and sending it through a network to the consumer [14]. In a demand response program, the consumer has a direct connection to the utility, whether it be through Direct Control Load Management (DCLM), or and Interruptible Demand. DCLM involves the utility having the ability to remotely turn on/off, or cycle devices within a home, or business, thereby reducing demand on the consumer side. Interruptible demand is an agreement between the consumer and the utility where the utility can request that a consumer curtail their energy use during peak hours, or have the ability to remotely trip devices within the consumers property as long as notice is given beforehand. In exchange, a consumer will receive discounts and/or credits towards their energy bills.

Because demand response is relatively new solution to controlling peak loads, large data collection with high sampling rates will be necessary to provide as much detailed data as possible. The necessity for large amounts of data comes from the fact that there is still a lack of experience with long term demand response programs [15].

Demand response for a large urban area is hard to model as it is complex and multilayered, so data is needed to properly simulate demand response in a densely populated area [15]. To better understand the factors that affect demand response programs, data relating to consumer behaviors, as well as external factors such as weather, price sensitivity, and the changing of seasons must be obtained, and researched. An outline of the demand response logic as it pertains to the pilot system is displayed in Fig. 1.

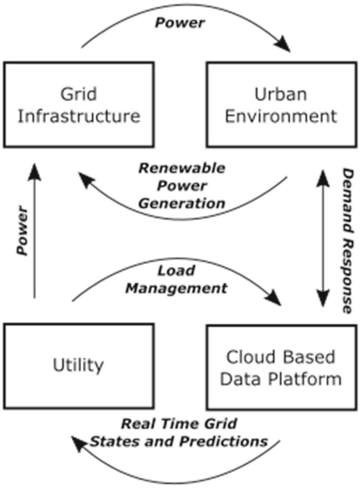

Power
Grid Infrastructure
Urban Environment
Renewable Power Generation
Power
Demand Response
Load Management
Utility
Cloud Based Data Platform
Real Time Grid States and Predictions

Fig. 1. The system demand structure for a data collection system is presented. A cloud based platform will store and analyze data collected from a home, or business, in real-time, allowing for quick control decisions in demand response programs.

Devices that measure power consumption have been used in research, however, most studies do not offer high frequency data with the resolution to detect small transient changes. Current research on the pilot system collects and analyzes data at higher resolutions. A 1 Hz resolution, or better, will provide a good sampling rate for large data collection and the ability to see transient patterns in power usage, such as the warming of a stove, or the brightness of a television. Results from the pilot system have shown that different devices such as a stove top, or a water heater, create a specific power profile signature when their power draw is monitored. This signature can be thought of as a "power fingerprint." Having the ability to determine device usage from power data allows cost efficiency in power monitoring because rather than installing a power monitoring meter on each device within a building, software can instead analyze and determine which devices on a property are in use based on the power signatures found within an aggregate power data set for an entire home, or business.

Power producers will be able to monitor a home, or business, and understand which devices can be cycled during peak loads to relieve grid pressure, especially in high energy consumption areas like urban centers where large percentages of a population tend to live. In order to accomplish this, a device is needed to record a consumer's power usage. A pilot program has been created at the University of Hawai'i that currently involves monitoring aggregate power usage from 20 homes on the island of Oahu using a smart power meter (SPM). The components, challenges and scalability of the pilot system will be discussed, as well as future work pertaining to demand response programs, which will be discussed in the following sections.

1.2 Related Research

The study and feasibility of demand response as it relates to power grids is ongoing, and the pilot program looks to contribute to that research in the areas of large data collection, storage and analysis [12, 13].

Demand response programs allow for increased peak load reduction as well as the ability to balance supply and demand of energy in power grids [12]. Stability and load shifting are two factors that are important in maintaining grid stability, which can be accomplished through demand response programs. Cost efficiency is another benefit of demand response because there is no need to maintain spinning reserves and large power storage infrastructure [8].

Similar research is being done on smart meters to collect and analyze data. A group from the University of Bath investigated the use of smart metering devices in combination with voltage control techniques. Their research focused on analyzing the consumer side of demand response as a way to create cost efficiency for a consumer as well as a tool to restore grid system faults and maintain transmission stability. The Lon Local Operating System (LonWorks) and ZigBee Wireless Network Standard were two suggestions for creating a system of communication between smart meters and controllers to handle real-time data [31].

A research group in Europe proposed the use of local area networks (LAN) and wireless local area networks (WLAN) in combination with KNX communication standards as an option to set up communication between smart metering devices. The use of ZigBee and KNX components were deemed feasible to monitor load consumption of devices in order to create a timetable of *shiftable loads*. The load shifts refer to the rescheduling of device usage from peak hours to times that do not provide large strains on the grid. Real-time analysis and visualization would allow consumers to make the proper choices in energy consumption that are related to cost efficiency. An algorithm based on tariffs was the basis for the load timetables [32].

Researchers in Canada proposed a smart metering system based on load disaggregation where a power signal is analyzed into the various device components that produce it. Their research focused on the factors that affect load disaggregation such as noisy signals, simultaneous loading, computational costs and privacy issues. They noticed that devices produced different power signals when cycled, for example, constant vs. periodic loads. To train algorithms in detecting a device, the research group suggested algorithm training based on probabilities and the clustering of individual devices. The research group deemed the definition of *deferrable actions* as necessary in their proposed system. *Deferrable actions* are those relating to devices whose utilization is not a priority and cycling can instead be scheduled at an alternative time, which would allow for load shedding. These devices include washer/dryers, ovens and dishwashers [33].

A UK-based power utility, National Grid, looked into the affect the power usage of certain devices had on the grid. They found that millions of kettles are cycled around 5 pm, knowledge such as this allows a utility to know when to cycle specific loads within home. National Grid uses the aforementioned knowledge to maintain grid frequency. Aggregating these cycling patterns with the loads of other houses in a neighborhood, or region, allow for the ability to maintain grid stability throughout sections of a power grid [30].

2 SPM Pilot System

Because of the island's geography and dense population, Oahu provides an ideal location to understand renewable energy penetration into an existing power grid, and how it relates to demand response programs. Several factors allow for Oahu to be the location to implement the pilot system, these factors include high solar radiation on the island, access to a dense urban populations, and Oahu being an isolated power grid. In 2015, the Hawaii state legislature voted to have 100 % energy generation from renewable sources by 2045 [16, 17]. Hawaii's commitment to alternative energy sources allows for a continued study of an urban area with high renewable energy generation, and the effects of this generation on demand response. Because most buildings have circuit breaker boxes, a common interface is already in place to install the SPMs. The device collects data at one-second intervals and sends it through a local WiFi network to a remote cloud server using a SSH tunnel. Data storage, analysis, forecasting and control can all occur within the cloud. The server will have the ability to send control signals based on analysis of the power data to the consumer, where an installed client can cycle devices in accordance with demand response programs to reduce peak loads. Figure 2 illustrates the overall pilot system.

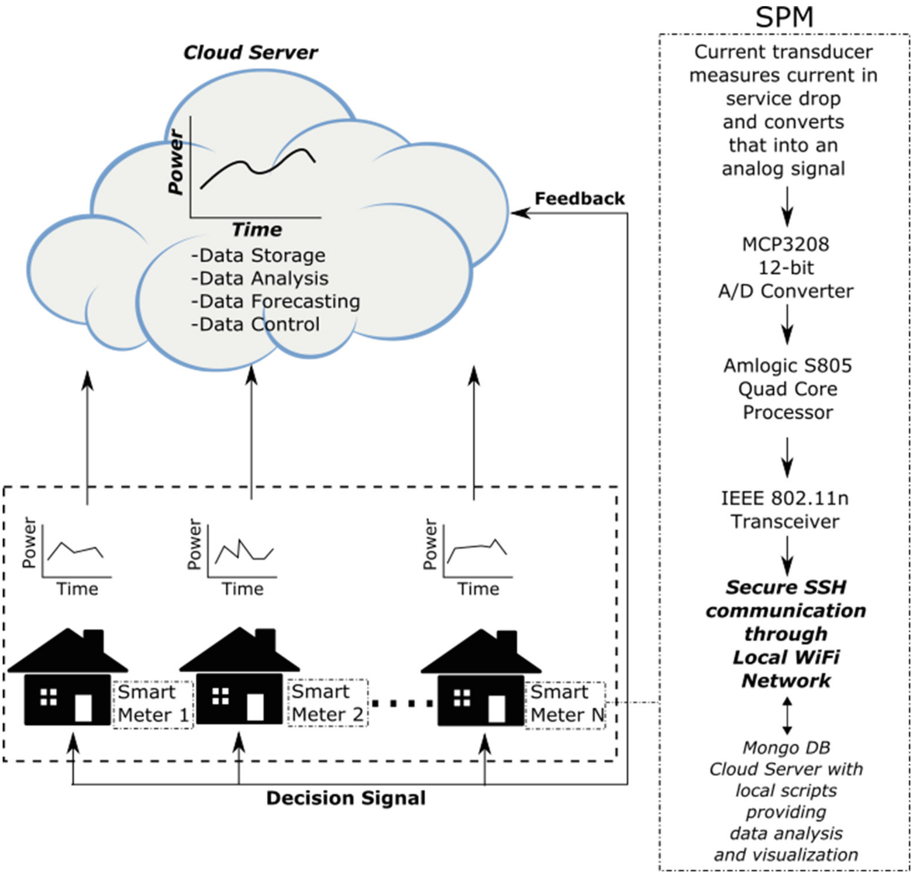

Cloud Server
Power
Time
-Data Storage
-Data Analysis
-Data Forecasting
-Data Control
Feedback
Power
Time
Power
Time
Power
Time
Smart Meter 1
Smart Meter 2
Smart Meter N
Decision Signal
SPM
Current transducer measures current in service drop and converts that into an analog signal
MCP3208 12-bit A/D Converter
Amlogic S805 Quad Core Processor
IEEE 802.11n Transceiver
Secure SSH communication through Local WiFi Network
Mongo DB Cloud Server with local scripts providing data analysis and visualization

Fig. 2. The setup of the proposed system implements a SPM to monitor and transmit data from a circuit box. Data is then transmitted to a server for analysis. The current server can be scaled to cloud storage, so that more nodes can participate in the pilot program and provide more data for load forecasting analysis.

2.1 Data Acquisition

The data acquisition is performed by a power metering device at the local consumer level. The device can fit within a circuit breaker box, is non-invasive, and allows for easy installation, setup and maintenance while delivering accurate power measurement, data preprocessing and server communication. The SPM is powered through the circuit breaker box. Two current transducers, one connected to each service drop wire within the circuit breaker box, measure current signals, which are transformed into analog voltage signals, and sent to a MCP3208 12 bit analog digital converter (ADC), which collects data at 80kSps. Images of an installed device are shown in Fig. 3.

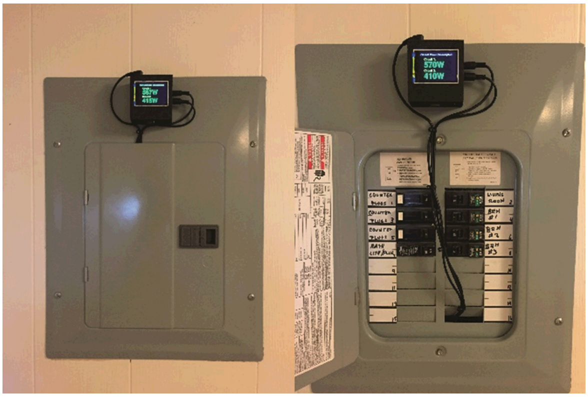


Fig. 3. A SPM meter is installed in the circuit breaker box of a home taking part in the pilot project.

An Amlogic Quad Core processor computes the power consumption for each phase. Power is calculated assuming a constant voltage. The median power pertaining to one second of collected data is obtained for each phase, and sent to the cloud server for storage and analyzing. Figure 4 describes data collection and transmission on the consumer level.

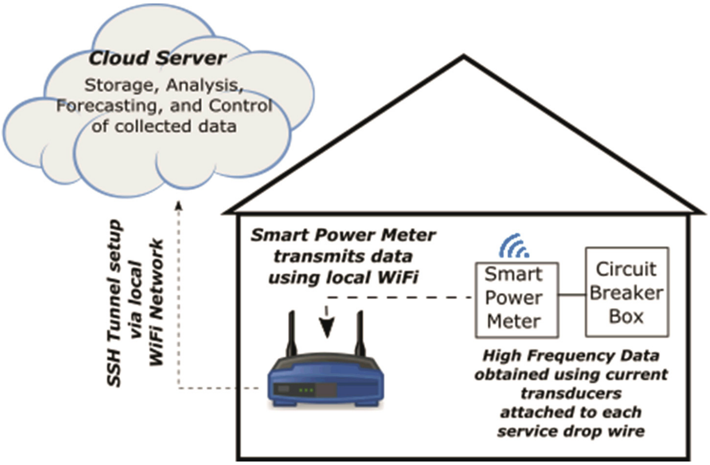

Cloud Server
Storage, Analysis, Forecasting, and Control of collected data
SSH Tunnel setup via local WiFi Network
Smart Power Meter transmits data using local WiFi
Smart Power Meter
Circuit Breaker Box
High Frequency Data obtained using current transducers attached to each service drop wire

Fig. 4. Utilizing preexisting WiFi connections within a home allow for a cost effective solution for data transmission. Circuit breaker boxes are usually located in a remote area of a building, so it is necessary to utilize a wireless connection to allow for a robust system to monitor and transmit data from a node. A secure SSH connection allows for safe and reliable transmission of data to a server in real-time.

2.2 Communication

After the power data is collected and preprocessed by the SPM, the data is then transmitted to a remote server using a secure SSH tunnel via a local WiFi network. The advantage of this communication setup is that the SSH tunnel provides an added layer of security for what is confidential information. While the utilization of a preexisting local WiFi connection takes advantage of an already existing network, thus eliminating the added cost of building a new communication infrastructure. Data is stored directly into a MongoDB database hosted on a cloud server. Because data is being sent from multiple locations, each data set needs to be identified by the node it originated from, this is accomplished when the SPM assigns a node identifier to each outgoing data set. When there is a disturbance in the WiFi connection, or a communication delay, the SPM will buffer until a connection is reestablished to minimize data-loss. Despite the 1 Hz transmission rate of the SPM, bandwidth and storage requirements are kept minimal. Each database query consists of just three integers, which total 24 bytes of data per second on a 64 bit system. Households are currently transmitting approximately 2 MB/d. The island of Oahu has a population of approximately 950,000, assuming 200,000 households, 400 GB of power data would be sent to the servers each day at a rate of 4.63 MB/s.

2.3 Data Storage/Analysis

The MongoDB database on the cloud server, is a document based open-source database. It is utilized as a multiuse agent that acts as a central node where large amounts of power data is collected, streamed and queried for data analysis of real-time system states and forecasting.

Document based databases yield high scalability and data storage flexibility, which is quintessential for power analysis of large complex urban centers. Streams of real-time and recent data, as well as data queries for historical data must be performed as efficiently as possible to create predictions that will analyze data in real-time, thus allowing for fast and efficient conclusions and decisions. These conclusions will be utilized in future work to create control decisions to be sent back to the consumer where devices within a property can be controlled using a client. Thus granting the ability to create forecasts that enable efficient demand response programs to be implemented, which will reduce peak loads and ensure reliable power transmission within the grid infrastructure.

2.4 Control

Future work revolves around enabling the cloud server to analyze real-time and historic data in order to determine, and send control decisions for demand response programs. Smart control decisions enable the ability to better ensure grid stability and power transmission reliability. These commands include, but are not limited to, ON/OFF commands, as well as time constraint commands. The control clients executing the commands will have the ability to send feedback data to the cloud.

The server itself can be utilized by the consumer as an interface to monitor power consumption, or override control decisions.

3 Data Analysis

Data collection is currently in progress using a total of 20 nodes and has been ongoing since August 2015. Participants volunteered (not compensated) to participate in the study and the household sizes range from two to six members. The backgrounds of the various participants are varied, however, specific details are kept confidential for privacy reasons. There was no criteria for selecting participants, the only requirement was that they had an accessible circuit breaker box within their home.

Each phase in the circuit breaker box is measured, and the power for each phase is plotted. Figure 5 gives an example of data from a node for one day. Phase one and two are plotted in red and black, respectively.

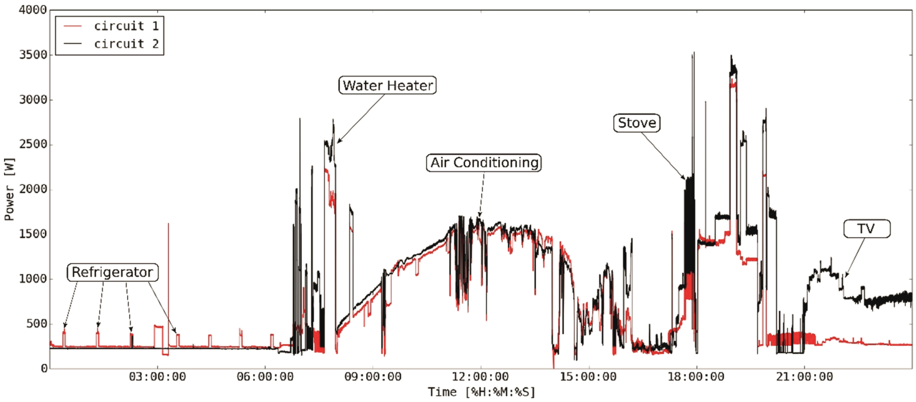


Fig. 5. Devices produce specific power signatures when in use. It can be observed when certain devices are cycled. The cycling of loads within a node displays the behavior and patterns of a consumer that can be used to predict and schedule power generation. (Color figure online)

It can be seen that there are unique device signatures throughout the day, which correspond to a combination of specific devices within the node. In the displayed example, from midnight to 7 am, the only signal that stands out is the refrigerator cycling, which is due to the fact no other major loads are present at the respective time interval. During the day air conditioning is the dominant load, which correlates to the heat in Oahu at midday. Evening loads are dominated by consumer electronics such as TV. Detailed power profiles over extended time periods grant an observer the ability to understand the energy needs of a consumer and predict when to schedule loads. Such is the case in Fig. 6 where a week of data has been plotted.

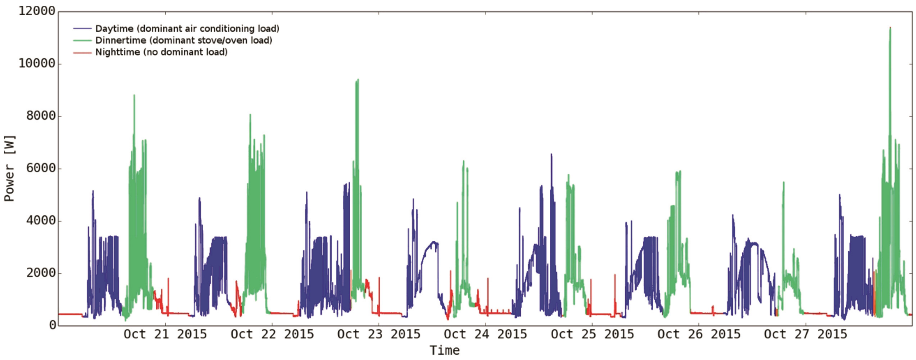


Fig. 6. One week of total power consumption is plotted for one family home. Consumer pattern behavior is evident from the increases in power consumption. (Color figure online)

The node displays a clear pattern of power consumption throughout a week. Dominant loads throughout the day are shown in blue and green, correlating to air conditioning and dinner-related activities, respectively. The family exhibits a fixed pattern of power consumption throughout the week that can be used for load prediction. Air conditioning loads dominate the day while cooking-related activities dominate evening loads. The two main load patterns stemming for air conditioning and cooking are repeated daily throughout the week. Nighttime loads are reduced to a bare minimum because of inactivity at night.

Aiding in the study of demand response it the fact that each device produces a specific power signature, or fingerprint, when spectral analysis is performed on the plotted power signal obtained by the SPM, as shown in Fig. 7.

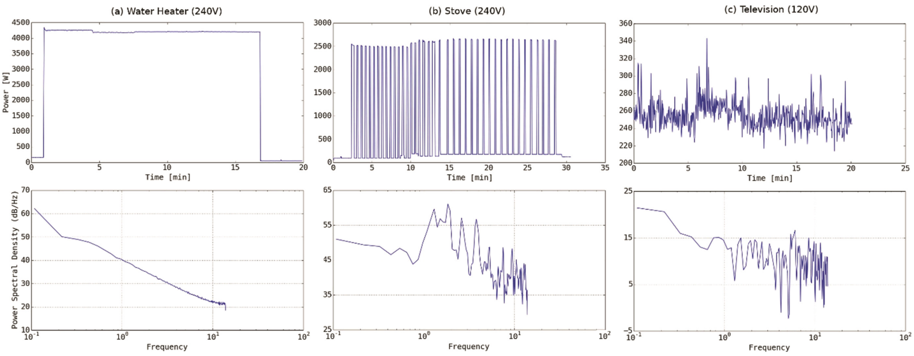


Fig. 7. The first row displays the power measured and transmitted by the SPM to the cloud server for a water heater, stove and television. Spectral analysis of the power signals correlating to each device are shown in the second row. A time based signal is converted to a frequency spectra that allows the ability to locate unique frequency signatures related to the time series signal.

Self-learning algorithms, such as ANNs, can be taught to detect power fingerprints in large data sets such as those shown in Figs. 5 and 6. Knowing which devices are in use, and when, will allow for scripts installed on a server to calculate optimal load schedules to cycle devices, such as water heaters and HVAC units within a node. Being able to distinguish when, and how often, a consumer uses a device will enable a power provider the ability to shed peak loads while not creating an interruption to a consumer's power usage. The capability to cycle a load can be automated, so that a client within a home can obtain decision signals from a cloud based server and implement the signals in real-time.

The results show that it is possible to determine which devices are consuming power at a given time. It is also clear that large quantities of data from a node permit the observation of consumer patterns as they relate to power usage. Combining the historical and real-time power data from multiple nodes within a section of the grid, allows a power producer to understand the needs of the consumer while providing efficient load management. However, it should be noted that the observed data can be sensitive as it displays patterns and behaviors of consumers, which must remain confidential to protect privacy.

4 Scalability

A large and flexible database is necessary for bulk amounts of data being collected from an urban center. MongoDB is a "NoSQL" cloud database where large data collection will be stored and analyzed when the pilot system is scaled.

A "NoSQL", or "non SQL" database is an alternative to the relational databases that use the Structured Query Language (SQL). There are alternative "NoSQL" databases such as Apache Cassandra and Couchbase, but recent studies have shown MongoDB to be more efficient in terms of reduced latencies when it came to read and update workloads [18, 34–36]. MongoDB contains a document database architecture, which provides the flexibility needed for scalability as the pilot system grows to include more nodes.

The use of a single server would lead to scalability issues as more data is collected and processed, MongoDB overcomes these issues with the potential to add more servers to accommodate large data as well as the utilization of automatic sharding, meaning that data is spread throughout multiple servers. Automatic sharding permits data to be accessed easier, and managed faster [19]. MongoDB utilizes a flexible data model, which allows the opportunity for easier development and scalability. This is because MongoDB does not use a rigid database schema, which determines how data is logically grouped. Rather, documents within MongoDB are assigned a primary key (id), which allows for a flexible schema where data can be easily queried. This is advantageous because factors that are not originally in prediction algorithms, but are later proven to be vital (as more data is collected) in load forecasting, can easily be added to the existing database (using key value pairs) and be used in prediction algorithms [20]. The "NoSQL" database MongoDB also takes advantage of bucket streaming URI (Uniform Resource Identifier), which is based on chunked transfer encoding. The streaming transfers permit data to be sent directly to the cloud server whenever data is available for transfer. This is because data is not buffered and saved to an isolated file, thus allowing for faster data transmission to the cloud servers [21].

There are drawbacks to using MongoDB, one being that the database performs poorly when it comes to aggregate functions, such as medians, modes, and sums. However, current research has not deemed this to be a problem when implementing algorithms into the cloud server. MongoDB also struggles with non-key values, but this too has not been deemed an issue in current research related to the pilot system. Because MongoDB is a "NoSQL" database, its implementation will require more effort than a SQL database due to the fact the schema in a "NoSQL" database is not as rigid. And because "NoSQL" databases have only recently gained in the popularity they have today, there is less support and literature as compared to a SQL database, which in many industries is considered a standard [20].

4.1 Data Security

Analyzing data will grant the ability to understand the behavior of a consumer, and as the pilot system is scaled up to include thousands of users within an urban environment, it will be necessary to protect sensitive information. The information is sensitive because it can reveal what a person, or persons, are doing at a specific time in the day. Many activities can be monitored, such as a person cooking, taking a shower, or working on the computer. It can also be determined when a person is home based on their air conditioning and heating usage. The monitoring of data can even analyze the power spectrum of a television, allowing for the TV power signal to be compared to the TV signatures of known channels, and from there determine what TV programs a person is watching. Unauthorized disclosure of this potentially sensitive information could allow an unauthorized agent to study the habits and routines of an end-user, thus creating potential threats to the privacy of the consumer.

Currently, the pilot system utilizes a single server, however, when scaling up the system to include consumers from a dense urban population, a cloud server will be used. Once the computational and storage limits of the single server are reached, the pilot system will be scaled to cloud computational storage. The use of cloud services has been increasing due to a number of factors, some of these factors include; the potential for scalability, geographic reach, cost savings and higher availability [22]. With the growth of cloud service and usage comes the need to address potential for security risks.

4.2 Addressing Threats

On the local WiFi network level, it will be important for the owner of the network to make sure their WiFi network has a strong network password as well as the most current firmware available for their router. In some cases, a user may disable the Wi-Fi Protected Setup (WPS) that is vulnerable to brute force attacks on the WPS PIN. It will also be important to apply "secure by design" principles when creating levels of security within a local WiFi network. One example will be determining who will have "root," or "admin," privileges in regards to the network, and whether these privileges will apply to the entire network, or in an "isolated environment" where only certain functions are available [23].

4.3 Unauthorized Internal Users and External Hackers

Two threats to the cloud server include unauthorized internal users, as well as external hackers. An unauthorized user, whether intentionally, or accidently, can access and manipulate data within the server. To prevent this, proper security protocols must be implemented that insure only authorized persons can access sensitive data within the servers. These protocols may include, but are not limited to, physically protecting servers, encryption tools, and anomalous behavior pattern detection [24].

External hackers may use techniques to prevent the proper function of a cloud server, or to obtain sensitive information. A common practice is the abuse of resources technique, which includes sending thousands of requests per second to disrupt and inhibit network computational resources. This technique is also known as a denial of service attack (DoS). The motivation for these types of attack are not to obtain data, but rather to overwhelm networks, and consumer computational resources. DoS attacks can be prevented by blocking an IP address where an attack (large number of requests within a certain amount of time) is found to originate, as well as implementing a security tool that can recognize when an attack takes place [25]. In addition to general hacking techniques, where someone will look to exploit perceived weaknesses in a system, external hackers will also employ password sniffing and man in the middle attacks (MITM). Password sniffing involves monitoring messages in a network with the goal of obtaining a password [26]. MITM attacks will involve a hacker impersonating two parties in hopes of obtaining confidential information, this technique can be applied to a server and user, which can lead to a compromised network [27]. Both password sniffing and MITM attacks can be prevented using strong passwords, ensuring a direct and authenticated connection to the server in use, as well as strong encryption techniques.

4.4 Vulnerabilities in Cloud Security

There are many vulnerabilities that are associated with cloud server use, a few will be mentioned to provide a foundation for future security protocols.

Data Interception. Data interception is a key concern due to the fact a large number of consumers will be sending sensitive data to a cloud server in the range of seconds. To remedy this, a secure shell (SSH) will implemented in the transfer of data from the consumer to the cloud. A SSH provides data encryption and the ability to implement a proxy for added security [28].

Data Leakage. There is potential for data within MongoDB to be leaked to unauthorized users, however, the developers of MongoDB look to actively recognize and address any issues relating to data leakage, which are usually related to versions of MongoDB that are outdated and unpatched. Other ways to prevent data leakage is using proper encryption methods, and recognizing when and where data is sent, so that it can be properly monitored. Physical protection of servers and personnel screening provide added security benefits [28].

Insecure or Ineffective Deletion of Data. When deleting data from a cloud server, there is always potential that data deletion may be incomplete, or insufficient. To counteract any potential issues from data deletion, it will be necessary to follow proper deletion protocols related to the cloud server platform, and in worst case scenarios, insure that a disk containing sensitive data is destroyed. Once again, proper encryption of data will decrease the risk related to ineffective data deletion [28].

Loss of Encryption Keys. The loss of encryption keys may be due to the accidental publication of a secret key, such as a secure socket layer (SSL), or a network password, which would create a vulnerability for potential threats. Several methods to mitigate encryption key loss are listed below [28]:

1. Storing encryption keys and the data in separate locations
2. Implementing audit trails to track who accesses data, and when the data is accessed
3. Backing up encryption keys onto a secured device
4. Encrypting the encryption keys themselves
5. Periodic changing of encryption keys [29]

Malicious Probes. In the case of a malicious probe, an unauthorized user may look to introduce a virus into the system. This can be prevented by creating database logs that record who and when someone attempts to access the database, so that any unauthorized user attempts may be blocked. Continually updating security patches will also insure that cloud security architecture is up to date. If a malicious probe does enter the system, historical data can be protected though periodical backup into a secured location [28].

5 Conclusion and Future Work

In order to provide the proper demand response program to a power grid, it is first necessary to collect large amounts of data in order to understand consumer behaviors and patterns. A pilot system was created with a smart metering device that can collect and transmit data at high frequencies (1 Hz or less) through a SSH tunnel to a server. A robust collection of data allows for the patterns and behaviors of a dense urban population to be analyzed. The small scale pilot system has proven the feasibility of data collection related to large-scale demand response. However, challenges will be present when scaling the pilot system to include more nodes. Topics to be addressed will include protecting sensitive consumer information, server infrastructure, security, and the management of big data.

Current research related to the pilot program is in the early stages of understanding consumer behavior. Human behavior is complex and is a study within itself, however we look to just understand device usage as it relates to demand response. Initial results are encouraging as patterns related to node power consumption can be detected. But because of the complex nature of human behavior, more data will need to be taken to see how external factors such as weather, holidays, and season affect consumer power consumption. However, the pilot system provides an initial foundation into the study of factors affecting consumer power usage. As more nodes are added to the current system and data collection continues a better understanding of consumer behavior as it relates to demand response programs will be achieved.

References

1. 2013 Report Card for America's Infrastructure. ASCE, Reston, VA (2013)
2. Global Trends 2030: Alternate Worlds. National Intelligence Council, Washington, DC, vol. 5 (2012)
3. Tweed, K.: Polar Vortex Cripples Power Generation, But Grid Survives, 9 January 2014. http://spectrum.ieee.org. Accessed 1 Nov 2015
4. Duke Energy: Causes of Power Outages (2015). https://www.duke-energy.com. Accessed 5 Nov 2015
5. Calvillo, C.F., et al.: Distributed energy generation in smart cities. In: Paper presented at the International Conference on Renewable Energy Research and Applications, Madrid, Spain, 20–23 October 2013
6. The Potential Benefits of Distributed Generation and Rate Related Issues that may impede their Expansion. US Department of Energy, Washington, DC (2007)
7. Salameh, Z.M., Davis, A.J.: Case study of a residential-scale hybrid renewable energy power system in an urban setting. In: Paper presented at the Power Engineering Society General Meeting, Toronto, Canada, 13–17 July 2003
8. Impacts of Solar Power on Operating Reserve Requirements. NREL, Golden, CO (2012)
9. Wesoff, E.: What are the impacts of high wind and solar penetration on the grid? 25 September 2013. http://www.greentechmedia.com. Accessed 2 Nov 2015
10. Smart Grids and Renewables: A Guide for Effective Deployment. IRENA, Abu Dhabi, UAE (2013)
11. Lew, D., et al.: The Western Wind and Solar Integration Study Phase 2. NREL, Golden (2013)
12. Demand Response and Advanced Metering. FERC, Washington, DC (2008)
13. Demand Response Discussion for the 2007 Long-Term Reliability Assessment. NAERC, Atlanta, GA (2007)
14. Smart Grid (2015). http://energy.gov. Accessed 20 Oct 2015
15. O'Connell, N., et al.: Benefits and challenges of electrical demand response: a critical review. Renew. Sustain. Energ. Rev. **39**, 686–699 (2014). Elsevier
16. Press Release: Governor Ige Signs Bill Setting 100 Percent Renewable Energy Goal in Power Sector. http://governor.hawaii.gov. Accessed 2 Nov 2015
17. Namata, B.: New Law requires 100-percent renewable energy in Hawaii by 2045, 8 June 2015. http://khon2.com. Accessed 21 Oct 2015
18. Scalability Benchmarking: MongoDB and NoSQL Systems. USA, Pleasanton, CA (2015)
19. Cattell, R.: Scalable SQL and NoSQL Data Stores. ACM Sigmod Rec. **39**(4), 12–27 (2010)
20. Parker, Z., et al.: Comparing NoSQL MongoDB to an SQL DB. In: Paper Presented at the ACMSE 2013 Proceedings of the 51st ACM Southeast Conference, Article no. 5, Savannah, GA, 4–6 April 2013
21. Google: Concepts and Techniques (2015). https://cloud.google.com/storage/docs/concepts-techniques?hl=en#streaming. Accessed 30 Oct 2015
22. Cloud Computing Trends: 2014 State of the Cloud Survey. RightScale, Santa Barbara (2014)
23. Wikipedia: Secure by design (2015). https://en.wikipedia.org. Accessed 25 Oct 2015
24. Chou, T.S.: Security threats on cloud computing vulnerabilities. IJCSIT **5**(3), 79–88 (2013)
25. Weiss, A.: How to Prevent DoS Attacks, 2 July 2012. http://www.esecurityplanet.com. Accessed 1 Nov 2015

26. Armstrong, D.: Password Sniffing, 25 October 1996. http://cng.seas.rochester.edu. Accessed 27 Oct 2015
27. Karapanos, N., Capkun, S.: On the effective prevention of TLS man-in-the-middle attacks in web applications. In: Paper Presented at the Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, 20–22 August 2014
28. Computing, Cloud: Benefits, Risks and Recommendations for Information Security. ENISA, Crete (2009)
29. Protect your most critical data and your access to it by following these tips for securing encryption keys. http://aspg.com. Accessed 25 Oct 2015
30. National Grid: Frequency Response Services (2015). http://www2.nationalgrid.com/uk/services/balancing-services/frequency-response. Accessed 22 Nov 2015
31. Gao, C., Redfern, M.A.: A review of voltage control in smart grid and smart metering technologies on distribution networks. In: Paper Presented at the 46th International Universities Power Engineering Conference, Soest, Germany, 5–8 September 2011
32. Kunold, I., et al.: A system concept of an energy information system in flats using wireless technologies and smart metering devices. In: Paper presented at the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Prague, Czech Republic, 15–17 September 2011
33. Makonin, S.: The cognitive power meter: looking beyond the smart meter. In: Paper Presented at the 26th IEEE Canadian Conference Of Electrical and Computer Engineering (CCECE), Regina, Canada, 5–8 May 2013
34. Olavsrud, T.: 9 MongoDB success stories, 24 November 2015. http://www.cio.com/article/3008114/open-source-tools/9-mongodb-success-stories.html. Accessed 10 Dec 2015
35. Bhattacharjee, A.: NoSQL vs SQL – Which is a Better Option? 8 May 2014. https://blog.udemy.com/nosql-vs-sql-2. Accessed 11 Dec 2015
36. McNulty, E.: SQL VS. NOSQL- WHAT YOU NEED TO KNOW, 1 July 2014. http://dataconomy.com/sql-vs-nosql-need-know/. Accessed 10 Dec 2015

Development of a Measurement Scale for User Satisfaction with E-tax Systems in Australia

Abdullah Alghamdi(✉) and Mahbubur Rahim

Caulfield School of Information Technology,
Monash University, Melbourne, Australia
abdullah.alghamdi@Monash.edu

Abstract. Governments worldwide have introduced various types of e-tax systems, as an important e-government agenda, to provide citizens and residents with a channel to lodge their tax claims at their convenience. An understanding of what constitutes taxpayer satisfaction with using e-tax systems is thus important for government agencies to further improve the quality of services delivered through these systems. However, to date limited research has been devoted to evaluate user satisfaction with e-tax systems. In this paper, we thus report on the development of a satisfaction construct which is rigorously evaluated using a three stage process. We find the emergence of several dimensions which require further investigation.

Keywords: E-tax system · E-government · Satisfaction · Measurement scale · Australia

1 Introduction

The proliferation of the Internet and Web 2.0-based technologies has encouraged government agencies worldwide to offer electronic government (e-government) initiatives [7]. These initiatives enable government agencies to disseminate important information and encourage the public to receive government services at their convenient time and location [18]. E-government initiatives can be of different types. However, Government-to-Citizen (G2C) initiatives have received considerable attention in the literature [26]. One interesting example of G2C initiative is e-tax systems [8]. According to Fu et al. [20], e-tax systems refer to the automation of all business processes and transactions relevant to taxation for improving the efficiency of lodging and collecting taxes.

Information Systems (IS) and e-government literature streams report studies on e-tax systems for developed and developing countries alike. Examples include those undertaken in such countries as Australia [4], Greece [18], India [22, 34], Japan [6], Malaysia [2, 16, 35], Nigeria [33], Philippines [8], and Taiwan [7]. The primary focus of these studies is however on the adoption and acceptance of e-tax systems, and relatively less attention has been given to post-implementation aspects. As a good proportion of citizens, particularly in the developed nations, are known to use e-tax

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 64–83, 2016.
DOI: 10.1007/978-3-662-53416-8_5

systems for a considerable time, we thus argue that research attention needs to be shifted to the post-implementation issues of these systems. Given the fact that the use of e-tax systems is not mandated in most countries, user satisfaction with these systems in particular needs to be evaluated because the continuous use of IT systems is known to be largely influenced by the level of user satisfaction with those systems [13].

Some studies have been reported on user satisfaction with e-government in general [1, 32]. A few studies [21, 25] also exist that concern with e-tax user satisfaction. Despite the existence of these studies, it is important to undertake further studies in this area because perceptions of citizens towards e-government services differ among countries [22]. Such differences in perceptions are attributed to the variations in legislative issues, public access to government information, and public access to government services [43]. Moreover, technological awareness and readiness of a country and its citizens vary widely across countries. We further note the existence of a disagreement in the e-tax literature about the dimensions included to measure e-tax user satisfaction. This is because e-tax satisfaction has been evaluated from two different perspectives: tax officers and tax payers. Hence, several scholars have called for more research on e-tax systems [33]. In response, we have thus undertaken an exploratory study with an aim to develop a measurement scale for the Australian e-tax users' (citizens) satisfaction by identifying its key dimensions. We acknowledge that in Australia a few studies have examined e-tax systems implementation success (e.g. [4, 5]) but they do not look at success from the taxpayer (user) satisfaction perspective. The aim of our research is addressed by developing a conceptual model which is then empirically evaluated using a rigorous three-stage process. We find the scale to be made of four integrated dimensions (information trustworthiness, e-tax usability, time related benefits, and accessibility) unlike others reported in the broader satisfaction literature. The implications of this finding are discussed and further explorations are recommended. Our paper makes a modest contribution to theory and practice. The integrated nature of most dimensions included in our e-tax satisfaction construct indicates the need for further exploration for the conceptual clarity of the dimensions of e-tax satisfaction. We believe the satisfaction construct would still encourage government agencies responsible for developing e-tax systems in Australia to further improve their online services by specifically focusing on the dimensions included in the construct.

2 E-tax System: An Introduction

2.1 Characteristics of E-tax Systems

Electronic taxation systems (e-tax) or online tax systems represent one type of electronic applications provided by government agencies. These systems are classified as revenue-collection applications and considered to be one of the most critical innovations offered by governments [11]. Therefore, in those countries in which paying taxes is mandatory, tax agencies have expressed interest in moving from manual, paper-based tax filing process to the use of IT applications [11]. The viewpoint regarding e-tax systems varies among researchers. For example, Fu et al. [20] introduce a broad definition of the electronic filing of personal income taxes as the automation of all business processes and transactions relevant to taxation to improve the efficiency of lodging and collecting taxes. In another study, Hu et al. [24] define an e-tax system as an online service that helps in improving service quality by reducing costs for taxpayers as well as enhancing the efficiency of the tax agency. According to Shao et al. [39], e-tax systems include such services as the provision of tax filing software, process of taxpayers' e-filing, and tax related online consulting. These systems are designed to unify tax preparation, tax filing, and tax payment by providing enhanced tax service for businesses and government alike [24]. It appears that Fu et al. [20] has defined e-tax systems from the perspective of efficiency improvement. On the other hand, Hu et al. [24] define e-tax systems from the service quality perspective. Neither of these viewpoints acknowledges the distinction between online tax Web sites and e-tax software. In this paper, the term "e-tax systems" is used to refer to both tax Web sites and e-tax software.

2.2 Benefits of E-tax Systems

A number of benefits can be experienced by the users and tax authorities as a result of acceptance and use of e-tax systems by taxpayers. Yusuf [50] claims increased taxpayers' compliance level and revenue generation of a country through wider adoption and use of e-tax systems. In addition, e-tax systems have the possibility to ease the process of tax filing for individuals, providing them with time saving and cost efficiency benefits [38]. These benefits can be achieved when e-tax systems are introduced to meet the expectation of individuals using these systems.

2.3 E-tax Systems in Australia

Australian Taxation office (ATO) E-tax is a government owned software developed to help lodging tax return. The software can be installed from ATO website. It is a stand alone application that can be installed and run in a desktop or a laptop computer. According to e-tax accountants' website (etax.com.au), individuals' using ATO E-tax are on their own with insufficient help and assistance to lodge their tax return. ATO E-tax might involve more than a hundred pages, which makes it a complex and difficult technology to use.

3 Related Background Literature

Literature on e-tax systems although limited but is gradually evolving. A review of the e-tax literature indicates the presence of three key themes that received much of the attention from the scholars. These include: e-tax adoption factor, usage of e-tax, and post-adoption issues of e-tax systems. For example, scholars like Connolly and Bannister [11] and Schaupp et al. [37] have looked at the adoption of e-tax systems. Likewise, Chu and Wu [9] have examined the factors contributing to usage of e-tax systems. Lai [28] and Lai and Choong [29] looked at the challenges faced by the taxpayers for using e-tax systems. Post-implementation impacts like benefits and satisfaction have also received some attention [36]. As this paper is concerned with satisfaction, a brief but critical analysis of satisfaction literature related to e-tax, e-government, and IS in general is provided in order to understand how various scholars have conceptualized 'satisfaction' construct in the IS and e-government literatures.

The notion of satisfaction is not new; however its application to e-government context represents a relatively new phenomenon. In general, satisfaction within e-government context is conceptualized by scholars in two broad ways. One group of scholars view satisfaction as an independent variable that has an effect on other human behaviors (e.g. sustained usage, word of mouth recommendation). Three key characteristics of e-government studies adopting this view include: (a) satisfaction is evaluated without focusing on a specific e-government application and/or service (e.g. e-tax and e-voting), (b) the primary focus is not on the assessment of various dimensions comprising satisfaction (e.g. [10, 47]), and (c) satisfaction is regarded as an independent factor that relates to either adoption or success of e-government initiatives (dependent factor) (e.g. [19]). According to these studies, only 3 to 4 indicators are used to operationalize the concept of "satisfaction". The works of Colesce and Dobrica [10] and Wang and Liao [47] represent examples of this stream of literature. The primary focus of Colesce and Dobrica [10] is to evaluate the adoption of electronic government services, while that of Wang and Liao [47] is to assess the success of e-government systems. By drawing on IS adoption theories and 481 responses received from Romanian respondents, Colesce and Dobrica [10] evaluate citizens' adoption of e-government services. While investigating adoption, they identify several factors (e.g. information quality and accuracy) that can be used to evaluate user satisfaction with online government services. Colesce and Dobrica [10] find correlations between the constructs identified to evaluate users' adoption of e-government, whereas perceived ease of use, perceived usefulness, and perceived quality were found to affect user satisfaction. In another study, [47] identify different dimensions of user satisfaction. The dimensions are basically drawn from the IS success model proposed by DeLone and McLean [13]. According to Wang and Liao [47], user satisfaction can be measured indirectly through information quality, service quality, and system quality.

In contrast, another group of e-government scholars considers satisfaction as a dependent variable. According to them, factors from different theoretical backgrounds are used to develop satisfaction construct. Typical works representing this view of satisfaction include those of Abhichandani et al. [1], Verdegem and Hauttekeete [44] and Verdegem and Verleye [45]. Abhichandani et al. [1] have proposed the EGOVSAT framework to measure user satisfaction with online transportation systems as an example of e-government services. The framework includes five factors (utility, reliability, efficiency, customization, and flexibility) to affect user satisfaction. In their work, Verdegem and Hauttekeete [44] focus on quality of access and quality of service indicators to formulate a conceptual model for measuring user satisfaction with e-government services in general. Based on their quantitative analysis, the following indicators are considered significant measures of user satisfaction with electronic government services in general: reduced administrative burden, reliability, security, usability, content readability, ease of use, content quality, cost effective, privacy/personal information protection, transparency, courtesy, responsiveness, accessibility, flexibility, and personal contact. Yet in another study, Verdegem and Verleye [45] have developed a model to explain how satisfaction with e-government services in general is influenced by the actual use of e-government services. Their results indicate that nine indicators are considered to be significant in measuring the level of user satisfaction with regard to e-government services. Those indicators are: cost, awareness, security/privacy, content, usability, technical aspects, customer friendliness, availability, and infrastructure.

The viewpoint of the second group of e-government scholars is in line with scholars from other relevant disciplines. For example, in taxation information systems satisfaction research, the focus is on identifying a set of factors to measure satisfaction with e-tax systems. These factors are generally identified from system and service quality perspectives (e.g. [7, 21]). In Business-to-Consumer (B2C) e-business satisfaction literature, quality (e.g. [27, 30]) and security and convenience perspectives (e.g. [41, 51]) are used to frame the factors affecting satisfaction with e-business applications. Self-Service Technology (SST) satisfaction literature is popularly represented by the work of Meuter et al. [31] who used critical incident technique whereas customers told the experiences they have had with technology-based self-services. Based on those incidents, a set of self-service technology characteristics (in other words, factors) were identified that contributes towards making users satisfied. The literature on End-User Computing satisfaction is fundamentally influenced by the pioneering work of Doll and Torkzadeh [15]. They develop End-User Computing Satisfaction model to evaluate user satisfaction with IT applications within organizational contexts. Their model is based on non-Internet IS/IT applications; but still has received considerable recognition from the IS/IT scholars.

IS Success Model takes the lead in developing the constructs for satisfaction with IT applications and services. The model is developed by DeLone and Mclean [13]. The model includes “satisfaction” along with “use” as factors affecting IS success in organization. Both satisfaction and use can be measured indirectly through information quality, systems quality and service quality. The model has its influence on IT satisfaction research in which quality dimensions are widely used as a guide to develop satisfaction models. For example, Chen [7] and Gotoh [21] acknowledge that satisfaction is influenced by factors related to quality (e.g. system, services, information, preparation, process, and result).

4 Research Model

From a review of literature on satisfaction (from such areas as e-government, B2C e-commerce, SST and EUC), a total of 85 dimensions were identified that could potentially constitute taxpayer satisfaction with e-tax systems. It would be difficult to operationalize and empirically evaluate a model based on the inclusion of so many dimensions. As such, a two-phase filtering process was followed to shortlist these dimensions relevant for e-tax context. Phase 1 identifies the dimensions that have overlapping meanings. A total of 46 dimensions were identified after removing all redundant dimensions. Phase 2 identifies those dimensions that are supported in the literature from both the theoretical and empirical perspectives. By applying these criteria, the number of dimensions was further reduced from 46 to 15 (Fig. 1).

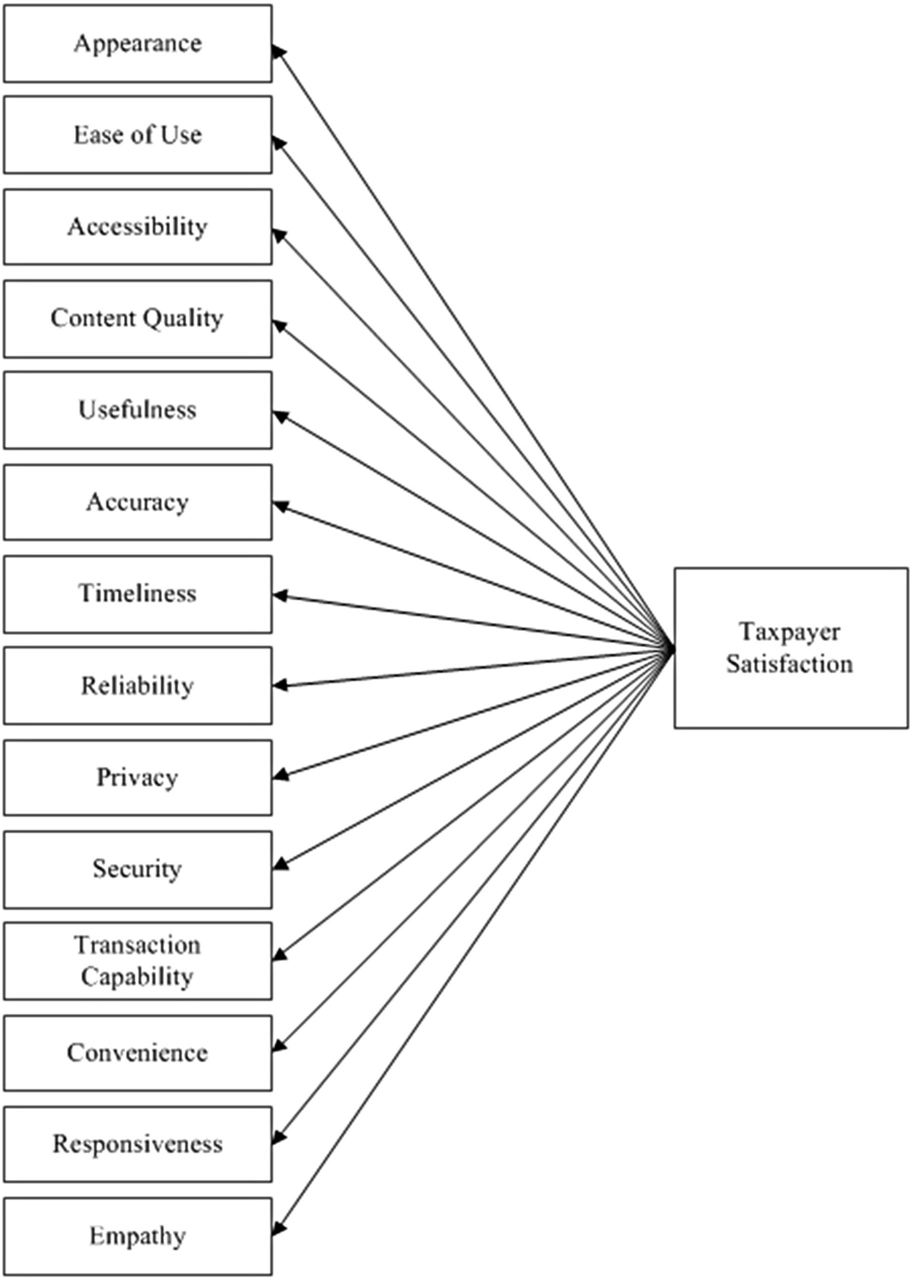


Fig. 1. Research model

Appearance: According to Kim and Stoel [27], appearance emphasizes how well a system guides its users and how easy it is to follow. They examine the effect of systems' appearance and design on users' perception of quality and satisfaction, and report that appearance is one of the most critical factors that influence user satisfaction with online systems. For e-tax context, we thus believe that appearance would influence taxpayer satisfaction with e-tax systems.

Ease of use: It refers to the ability of users to operate electronic systems with minimal difficulties [7]. Ease of use is an important dimension in measuring user satisfaction in the context of End-User Computing [15], e-government services [32] and e-tax systems [7, 19, 25].

Interactivity: It refers to "*the extent to which the communicator and the audience respond to, or are willing to facilitate, each other communication needs*" [23]. The definition can be conceptualized in terms of electronic services as the ability of electronic systems to intelligently respond to user needs. Interactivity has been found to be a significant factor to measure user satisfaction with online and electronic services. Interactivity is one of the significant dimensions that constitute Web customer satisfaction [30]. In the electronic services literature, interactivity is considered to be a significant dimension that can be used to measure taxpayer satisfaction with e-tax systems as well [7].

Accessibility: It is defined as the ability to access the system at all times [30]. Web site accessibility is an important dimension of measuring user satisfaction with online services [49]. In terms of e-government satisfaction, Verdegem and Hauttekeete [44] use accessibility to measure citizens' satisfaction with e-government services. In addition, accessibility was found to be related to taxpayer satisfaction with e-tax systems [7].

Content Quality: For the context of e-tax satisfaction, it is defined as the adequacy and clarity of information provided by a system so that it meets users' needs [32]. The content quality of information is considered to be an important indicator in measuring citizen satisfaction with e-government services [44]. For the context of e-tax satisfaction, content quality is a significant construct of taxpayer satisfaction with e-tax systems [7].

Usefulness: It refers to the degree to which a user can believe that a system will enhance performance [12]. In terms of the recipients' perspective, usefulness refers to the degree to which a person believes that using an e-tax Web site will enhance his or her efficiency and provide benefits. Devaraj et al. [14] examine the relationship between usefulness and satisfaction and find that it is a key determinant of user satisfaction with e-commerce.

Accuracy: It is defined in terms of information as being free from errors [17]. Accuracy is one of the most significant factors that affect End-Users Computing Satisfaction [15]. In terms of measuring satisfaction with e-tax systems, accuracy is a vital dimension to measure taxpayer satisfaction [7].

Timeliness: This indicates that a system can provide up-to-date information for a required task [42]. Timeliness has been widely used in satisfaction literature. In e-commerce literature, timeliness is found to positively affect user satisfaction [14]. In terms of e-tax satisfaction, Hwang [25] and Fu et al. [19] find that timeliness is significantly relevant to citizens' satisfaction with e-tax systems.

Reliability: It is the ability of a system to provide information and service dependably [48]. Service and system reliability have been proven to impact user satisfaction in the context of e-services. Reliability is an important antecedent of online service quality that affects user satisfaction with online services [49]. In the e-tax satisfaction literature, Chen [7] finds reliability to be an important factor to measure taxpayer satisfaction with e-tax systems.

Privacy: It refers to users' perception that their personal information is protected and is not disclosed to a third party [44]. Privacy is a very important determinant for citizens and should be used to measure satisfaction [44, 45].

Security: It is defined as "*freedom from risk or doubt during the service process*" [51]. In e-services literature, security is found to be more important than the appearance of Web sites as well as information provided by these Web sites. Various satisfaction studies have considered security to be one significant factor affecting user satisfaction [41, 44].

Transaction Capability: It refers to the extent to which a system can support its business functions [27]. Transaction capability can significantly affect online user satisfaction. This argument is supported by Kim and Stoel's [27] findings, as they find that transaction capability is a significant factor that affects user satisfaction with e-retailing.

Convenience: It refers to simplifying business processes by the adoption of information technology [41]. For the e-retailing context, online convenience is known to influence user satisfaction with online retailing services [30].

Responsiveness: It refers the quality of services offered by employees who are willing to help electronic system users [7].

Empathy: It refers to the ability of employees to pay attention to electronic system customers' needs [7]. Responsiveness and empathy are significant constructs for satisfaction with e-services [14].

5 Research Approach

A qualitative approach involving two techniques was used to refine the model: expert panel evaluation of the conceptual model and a pilot evaluation of the survey instrument developed based on that model. This was followed by an exploratory survey. These are now briefly described below.

Domain panel evaluation: A group of domain experts involving four academics (whose areas of research include e-government) and three senior tax agents working in professional tax agencies evaluated the model. A brief profile of these experts is shown in Table 1. The academics were chosen by reviewing their profiles appearing in the university websites. The tax agents were selected from the yellow pages. An email was sent inviting them to participate in our research project as a domain expert. The email contained an explanatory statement and a consent form. Upon receiving their consents (via email replies), a document outlining 15 satisfaction dimensions included in the model was sent to these experts. A short interview was later organized with each domain expert after one week of sending the evaluation document. During each interview, the domain expert was requested to: a) evaluate the importance of each dimension on a scale of 1 to 5, in which 1 means “extremely unimportant” and 5 means “extremely important”, and b) identify any new dimensions not mentioned in our document.

Table 1. A brief profile of the participating domain experts

Domain expert	Type of domain expert	Gender	Experience	Highest qualification
A	Academic	Male	10 years (Teaching)	PhD
B	Academic	Male	20 years (Teaching)	PhD
C	Academic	Male	10 years (Teaching)	PhD
D	Academic	Female	3–5 years (Working in tax agencies)	PhD
E	Industry	Male	3–5 years (Working in tax agencies)	Bachelor
F	Industry	Male	3–5 years (Working in tax agencies)	Bachelor
F	Industry	Male	3–5 years (Working in tax agencies)	Bachelor

Pilot evaluation of the survey instrument: To improve the clarity of the survey instrument drawn from the dimensions shortlisted through the expert panel evaluation process, feedback from several experienced e-tax users was obtained. Various sessional tutors from a large Melbourne-based university were contacted via email. They were invited to participate in our research project. Among those who agreed to participate, four tutors were chosen because they met the following criteria: (a) they have at least 3 years’ experience of using the e-tax system, (b) they have used the e-tax system within the past five years, and (c) they are interested in the findings of our research project. An email was sent to these tutors including a document for evaluating the survey instrument and its items. The document consists of three sections. In Section A, they were advised to evaluate each item based on its relevance to the dimension it is associated with on a scale of 1 to 5, where 1 means “strongly irrelevant”, 2 means “somewhat irrelevant”, 3 means “neutral”, 4 means “somewhat relevant”, and 5 means “strongly relevant”. In section B, they were required to provide any suggestions regarding any changes to these items (e.g. revision, deletion, addition). In section C, the survey instrument was attached for general comments about the layout and the design of the questionnaire.

Administration of Survey: Participants involved at this stage of our research project include staff and students from a large Australian university. They were chosen randomly and survey questionnaires were distributed at such public places as campus centers, recreation facilities, and cafes where staff and students generally spend their free times on campuses. Participants were personally contacted to fill out the survey questionnaire. The purpose of our research was explained to the participants and the questionnaire was given to them. A total of 300 survey questionnaires were distributed among staff and students. However, only 162 responses were received. Of these, 100 (representing a response rate of 33 %) participants have acknowledged using the e-tax system. The survey data analysis was performed using SPSS. Those 62 non-e-tax users indicated the following four reasons for not using the e-tax system: (a) lack of time, (b) lack of confidence, and (c) availability of easily tax agents to prepare tax lodgment, among others.

6 Initial Analysis

Qualitative Evaluation of Domain Experts Feedback: The responses given by the experts for each dimension are summarized in Table 2. Based on the feedback collected from the domain experts, the following observations were made and actions were undertaken.

Table 2. Evaluation of dimensions by the domain experts

No	Dimensions	Domain experts									Overall Average	Retention status
		Academic experts					Industry experts					
		A	B	C	D	Avg	E	F	G	Avg		
D1	Appearance	5	5	4	X	4.6	5	3	4	4	4.3	Yes
D2	Ease of use	5	5	5	X	5	4	4	5	4.3	4.6	Yes
D3	Interactivity	5	5	5	X	5	5	3	5	4.3	4.6	Yes
D4	Accessibility	5	3	5	X	4.3	5	2	5	4	4.1	Yes
D5	Content quality	5	5	5	X	5	5	3	5	4.3	4.6	Yes
D6	Usefulness	4	5	5	X	4.6	4	2	5	3.6	4.1	Yes
D7	Accuracy	5	5	5	X	5	5	4	5	4.6	4.8	Yes
D8	Timeliness	5	1	5	X	3.6	5	3	5	4.3	4	Yes
D9	Reliability	5	4	5	X	4.6	4	4	5	4.3	4.5	Yes
D10	Privacy	5	5	5	X	5	5	5	5	5	5	Yes
D11	Security	5	5	5	X	5	5	4	5	4.6	4.8	Yes
D12	Transaction capability	5	5	5	X	5	5	3	5	4.3	4.6	Yes
D13	Perceived convenience	4	4	5	X	4.3	4	2	5	3.6	4	Yes
D14	Responsiveness	4	5	5	X	4.6	3	4	5	4	4.3	Yes
D15	Empathy	4	2	5	X	3.6	3	4	5	4	3.8	No

First, all but one domain experts evaluated the dimensions. This expert, however, provided many useful insights about the relevance and redundancy of the dimensions included in our model. Despite this, there is a broad agreement among the experts regarding the importance of the dimensions. We however decided to remove ‘empathy (D15)’ as a dimension because it received an overall average score 3.8 out of 5.

Second, two domain experts distinguished between the Australian Taxation office (ATO) Web site and e-tax software downloaded from that Web site and argued that responsiveness (D14) is more relevant for measuring satisfaction with online queries and interactions between citizens and ATO staff. In contrast, as the e-tax software downloaded from the Web site does not allow online communication through “live chatting” with ATO employees, responsiveness (D14) is of little relevance in measuring taxpayer satisfaction with the e-tax system. Thus, we decided to exclude ‘responsiveness (D14)’ from our model.

Third, two domain experts considered content quality (D5), transaction capability (D12), and usefulness (D6) to have overlapping in their meanings and advised for combining them into a single dimension (usefulness). In addition, two other domain experts found security (D11) and privacy (D10) dimensions to be interrelated and recommended combining them into another single dimension (security and privacy). Based on these recommendations, we decided to merge content quality, transaction capability, and usefulness into one dimension (usefulness), while security (D11) and privacy (D10) dimensions are to be combined in one dimension (security and privacy).

Fourth, a number of issues regarding various e-tax aspects were suggested by the domain experts for consideration of possible inclusion in the research model. These include: Online help/support, Ease of download, List of items and transactions that are taxable or tax-deductable, Is the e-tax user friendly for the first time user, Is there any problem in lodging first tax return using e-tax?, Does it create any problem before lodgment?, Does it preserve data accurately?, and How efficient is the identification process? These suggestions were compared against the definitions of the existing dimensions already identified in this research. We find that each aspect can be addressed by the existing dimensions. Hence, no new dimensions are included in the model.

After making amendments, the revised taxpayer satisfaction construct now includes 10 dimensions: appearance, ease of use, interactivity, accessibility, usefulness, accuracy, timeliness, reliability, security & privacy, perceived convenience.

Pilot Evaluation: An initial survey questionnaire consisting of 38 items was then developed from those ten dimensions shortlisted by the domain experts. The items were chosen from various scholarly sources and adapted for e-tax context. An operationalization of these dimensions is shown in Appendix A. As discussed earlier, a group of four experienced e-tax users evaluated the initial questionnaire. Based on their feedback, we note a broad agreement among these users about the relevance of these items. However, one item measuring ease of use dimension was removed as it received a median score of 3.5 out of 5 (See Appendix B).

The participating experienced e-tax users also provided insightful suggestions regarding removing items, rephrasing items, and improving the overall clarity of the survey questionnaire. Drawing on their suggestions, eleven items were identified for removal. Of these, five items were removed because they had exactly similar meanings to other items, two items were removed because they were considered to be vague and did not have specific meanings, and four items were removed because they did not actually measure the intended dimensions. Additional suggestions were offered to improve the clarity of the items. According to them, these items require rephrasing to improve clarity. Thus, further revisions were made and an improved version of the survey instrument was developed. The number of the items included in this refined instrument was reduced from 38 to 28 items which still measured those ten dimensions.

7 Findings and Discussion

The demographic characteristics of the survey participants who have used the e-tax system are summarized in Table 3. The following observations can be deduced: (a) a majority of e-tax users are male (70 %), (b) except users over 40 years, each age group is well-represented, (c) dominance of the participants with income in the range of A$35,001 – A$80,000 is observed, and very few participants (2 %) have income exceeding AU $180,000. This makes sense, as the context in which the survey was conducted represents a tertiary educational institution where the number of people from very high income group is very limited, and (d) a majority of the participants (62 %) have a postgraduate degree. This also makes sense for the tertiary educational institution.

Table 3. Demographic characteristics of survey participants

Variables	No.	(%)
Gender		
Male	70	70
Female	30	30
Age		
18-23	16	16
24-29	27	27
30-39	51	51
Over 40	6	6
Income		
0–6000	19	19
6001–35,000	25	25
35,001–80,000	36	36
80 000–180 000	18	18
> 180,000	2	2
Education		
Secondary college	11	11
Undergraduate	26	26
Postgraduate	61	61
Others	2	2

An exploratory iterative factor analysis was performed on data collected from 100 e-tax users. Factor analysis is a well-known statistical method which is generally used to investigate to what extent a group of variables (items) are associated with their underlying factors.A total of 15 items loaded on four distinct single constructs (Table 2), indicating that only four dimensions constitute taxpayer satisfaction with e-tax systems. This factor solution was obtained after applying a multiple iterative process of factor analysis and item deletion. Items were deleted when either of the following conditions was met: (a) an item had a factor loading of less than 0.40, and (b) an item loaded on more than a single dimension. The retained four factors together account for 66.63 % of the variation in satisfaction. The Eigenvalues and the percentage variance explained by these factors are also shown in Table 2. This factor solution is also statistically significant. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy and Bartlett's Test of Sphericity were examined to evaluate the reliability of responses received from participants. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was found to be 0.821, and Bartlett's Test of Sphericity was 0.000, which was significant at $p < 0.001$ (Table 4).

Table 4. Results of factor analysis

Item	Dimensions				Corrected Item-total Correlation
	D1	D2	D3	D4	
A1				.773	.489
A2				.828	.498
EU1		.771			.589
EU2		.821			.578
AC1	.769				.605
AC2	.785				.634
AC3	.774				.645
I2		.803			.482
I3		.816			.672
U4			.675		.609
T2			.610		.429
SP2	.783				.545
SP3	.604				.532
PC1			.728		.566
PC2			.642		.397
Eigenvalue	5.94	1.77	1.23	1.04	
Variance by individual dimension	39.6	11.87	8.20	6.93	
Cumulative variance	39.6	51.50	59.70	66.63	
Cronbach Alpha	.856	.872	.7	.71	

Drawing on the factor analysis, we now observe that only four dimensions appeared to be relevant for the e-tax system satisfaction context. We now review the meanings of these four dimensions. In our conceptual model (Fig. 1), 'accuracy' and ‘security and privacy’ were considered as two separate dimensions. However, the factor analysis demonstrated that these two dimensions could be grouped together into a single dimension (D1: Information trustworthiness). Likewise, ease of use and interactivity were found to be grouped into a single dimension and is called as ‘E-tax Usability” (D2). Another dimension (D3) grouped some of the items belonging to three such dimensions as usefulness, timeliness, and perceived convenience. This new dimension is now renamed as “Time Related Benefits” (D3). The last dimension (D4) represents a single dimension (i.e. accessibility) identified in our model. The reliability of each of these new dimensions is calculated (last row of Table 2) and is found to be satisfactory [23].

In our research model, ‘accuracy’ and ‘security and privacy’ were considered two separate dimensions. However, the factor analysis demonstrated that these two dimensions are to be grouped together into a single dimension (D1). In the e-tax literature, accuracy is considered as one antecedent of information quality [7]. Accuracy constitutes an important construct concerning data for measuring End-User Computing Satisfaction [15]. Thus, ‘security and privacy’ and ‘accuracy’ are clearly about the information received and sent via e-tax systems. In other words, together they measure how trustworthy is information. It is important that such systems as e-tax must provide sufficient security and privacy to users’ information (e.g. income sources) and maintain accuracy of income related information required to submit an application. The two factors together have thus be renamed as “Information trustworthiness”.

Ease of use and interactivity were found to be grouped into one dimension (D2). A possible justification for that is that ease of use and interactivity items are related to usability. Usability is defined as the individuals’ ability to interact with a website with no required training due to the ease of use [3]. Although usability is generally conceptualized as a multi-dimensional concept, It is reported that some studies about IS consider usability as one single dimension [46]. Ease of use and interactivity could thus be renamed as “E-tax usability”.

Another dimension (D3) is renamed as “Time Related Benefits” because it includes some of the items belonging to 3 dimensions: usefulness, timeliness, and perceived convenience. Upon close inspection, we find that these items have one common characteristic – which is receiving a benefit involving time. For example, one item is about flexibility of usage from time perspective, another item is about completion of a task on time, and yet another item is about auto closure of the application after certain time.

These findings bear the following observations. First, the integrated nature of dimensions discussed above for the e-tax systems context indicate the construct of user satisfaction is more complex than previously identified by the researchers. Second, e-tax systems involve dealing with income and expenses related data for which users expect the government to provide a secured platform that is capable of handling sensitive data. Satisfaction will suffer when users perceive an inability of the tax authorities to deliver such a secured platform. Third, no matter how secured an e-tax platform is delivered by the tax authorities, user satisfaction will decline when such a system is perceived to be unusable and unable to deliver benefits that relate to time (e.g. flexibility, on time completion of task, and auto closure after a certain time).

8 Conclusion

In this paper, we have reported the development of a scale for measuring taxpayer satisfaction with e-tax systems for the Australian context. Drawing on a three-stage process, our measurement scale is developed which eventually contains four dimensions. Three of these dimensions (e.g. Information trustworthiness, E-tax usability, and time related benefits) are however found to be integrated in nature which according to other scholars, exist as an independent dimension. This finding was not expected but they still raise an interesting question. Do these three dimensions really reflect a higher level aspects of satisfaction as we have discovered in this paper or are they artificially created due to the small sample used in this study (n = 100)? Further studies are thus required to answer this question. Nevertheless, our study is still useful to theory and practice.

For theory, developing a reliable instrument for measuring user satisfaction with e-tax systems represents a contribution to the IT/e-business literature. In particular, e-government researchers can adopt this instrument as a template to measure user satisfaction with other innovative online government service delivery systems for citizens. To practice, the government officials, responsible for promoting customer relations between government agencies and citizens, are advised to concentrate to those dimensions that can help design an improved version of e-tax systems. This could help in creating more satisfied taxpayers.

Finally, we caution about some of the limitations of our work reported in this paper. The survey response rate was relatively low (33 %) which constrains the generalizability of the research findings. One reason for the low responses rate is that many students were found to be non-users of e-tax systems. Hence, future research should involve a large sample involving staff and students from all campuses and faculties. In particular, it would be interesting to examine whether a large sample has any impact on the relationship between users' demographic characteristics and their level of satisfaction. Another limitation is that this study was conducted for a tertiary educational institution context. Future studies should involve participants from a wide range of professions including doctors, accountants, IT specialists, businessmen, and employees from a wide variety of organizations. It would be interesting to find out how the relevance of satisfaction dimensions can change over time. Hence, longitudinal studies should be conducted to identify the importance of dimensions comprising satisfaction with e-tax systems. In this study, some of the dimensions were merged into a single integrated one (e.g. information trustworthiness). We however acknowledge that the rationale used in proposing such an integrated dimension is not without questions. Hence, the indicators used to operationalize these dimensions need further theoretical scrutiny and a large survey needs to be undertaken to empirically confirm the existence of such integrated dimensions.

A Appendix A List of Items Used to Operationalize the Dimensions

Dimensions	Items	Literature source
Accessibility	The e-tax system is always accessible	[49]
	The e-tax system quickly loads all the contents	[30]
	I can get all relevant information from the e-tax system in time	[7]
	All the content of the e-tax system are accessible	Developed
Ease of use	It is easy for me to learn how to use the e-tax system	[7]
	It is easy for me to navigate through the e-tax system	[7]
	The e-tax system is user-friendly	[15]
	Using the e-tax system is easy for me	[14]
Accuracy	The e-tax system is an accurate source of information for me	[7]
	The information content is consistent with my previous experience	[7]
	The content of the e-tax system helps me to understand the system	Developed
	The e-tax system provides information that I can trust	Developed
	The e-tax system is accurate	[15]
Interactivity	I believe that my interaction with the e-tax system does not require much attention	[7]
	The e-tax system has natural and predictable screen changes	[7]
	My interaction with the e-tax system is clear and understandable	[7]
Reliability	The e-tax system meets all my needs	Developed
	Any problems resulting from using the e-tax system can be quickly solved	[7]
	The e-tax system is credible	[30]
	The e-tax system is trustworthy	[30]
Usefulness	Information on the e-tax system is informative	[30]
	I find the e-tax system to be quite useful	[14]
	The e-tax system provides precise information I need	[15]
	The content of the e-tax system is readable	Developed
	The content of the e-tax system is understandable	Developed
	The contents of the e-tax system provide sufficient information	[15]
Timeliness	The e-tax system provides me with up-to-date information.	[15]
	The e-tax system accomplishes tasks very quickly	[14]
	When my account is logged off due to time-out, it does not bother me	Developed

(*Continued*)

(*Continued*)

Dimensions	Items	Literature source
Security and privacy	I feel that my personal information is safe	[44]
	I feel that lodging my tax application using the e-tax system is secure	Developed
	ATO guarantees that my personal information will not be shared or disclosed	Developed
	Transactions using the e-tax system are safe	[44]
Perceived convenience	I spend less time on lodging my taxes online than doing my taxes manually	[41]
	I can use the e-tax system whenever and wherever I am	[41]
Appearance	The e-tax system provides an easy-to-follow interface	[41]
	The e-tax system displays a visually pleasing design	[27]
	The e-tax system is visually appealing	[27]

B Appendix Item Evaluation by Experienced E-tax Users

Item code	Expert users				Median	Retained item?
	A	B	C	D		
A1	5	5	5	5	5	Yes
A2	4	4	3	5	4	Yes
A3	4	4	5	3	4	Yes
A4	5	4	5	5	5	Yes
EU1	5	2	5	5	5	Yes
EU2	5	5	5	5	5	Yes
EU3	4	5	5	5	5	Yes
EU4	4	3	5	1	3.5	No
AC1	5	3	5	4	4.5	Yes
AC2	4	4	5	3	4	Yes
AC3	3	4	5	4	4	Yes
AC4	5	5	5	5	5	Yes
AC5	5	3	5	5	5	Yes
I1	5	3	5	3	4	Yes
I2	4	4	5	1	4	Yes
I3	5	5	5	1	5	Yes
RE1	4	3	5	5	4.5	Yes
RE2	4	3	5	5	4.5	Yes
RE3	4	5	5	5	5	Yes
RE4	5	5	5	5	5	Yes

(*Continued*)

(*Continued*)

Item code	Expert users				Median	Retained item?
	A	B	C	D		
U1	5	5	5	4	5	Yes
U2	5	5	5	5	5	Yes
U3	3	5	5	5	5	Yes
U4	5	3	5	3	4	Yes
U5	4	5	5	3	4.5	Yes
U6	5	5	5	3	5	Yes
T1	5	5	5	5	5	Yes
T2	5	3	5	5	5	Yes
T3	5	5	5	5	5	Yes
SP1	5	5	5	5	5	Yes
SP2	5	5	5	5	5	Yes
SP3	5	5	5	5	5	Yes
SP4	5	5	3	4	4.5	Yes
PC1	5	5	5	5	5	Yes
PC2	5	5	5	5	5	Yes
AP1	4	5	5	5	5	Yes
AP2	5	3	5	1	4	Yes
AP3	5	3	5	5	5	Yes
W1	5	5	5	1	5	Yes
W2	5	5	5	5	5	Yes
CI1	5	5	5	3	5	Yes
CI2	5	5	5	5	5	Yes

References

1. Abhichandani, T., Horan, T.A., Rayalu, R.: EGOVSAT: Toward a robust measure of e-government service satisfaction in transportation. In: International Conference on Electronic Government, pp. 1–12, Ottawa (2005)
2. Azmi, A.A.C., Bee, N.L.: The acceptance of the e-filing system by Malaysian taxpayers: a simplified model. Electron. J. e-Gov. **8**, 13–22 (2010)
3. Benbunan-Fich, R.: Using protocol analysis to evaluate the usability of a commercial web site. Inf. Manag. **39**, 151–163 (2001)
4. Chamberlain J., Castleman, T.: Transaction with citizens: Australian government policy, strategy, and implementation of online tax lodgement. In: 11th European Conference on Information Systems, Naples, Italy (2003)
5. Chamberlain, J., Castleman, T.: Moving personal tax online: the Australian taxation office's E-Tax initiative. Int. J. Cases Electron. Commer. **1**, 54–70 (2005)
6. Chatfield, A.T.: Public service reform through e-government: a case study of 'E-tax' in Japan. Electron. J. E-Gov. **7**, 135–146 (2009)
7. Chen, C.: Impact of quality antecedents on taxpayer satisfaction with online tax-filing systems- an empirical study. Inf. Manage. **47**, 308–315 (2010)

8. Chen, J.V., Jubilado, R.J.M., Capistrano, E.P.S., Yen, D.C.: Factors affecting online tax filing–an application of the is success model and trust theory. Comput. Hum. Behav. **43**, 251–262 (2015)
9. Chu, P.Y., Wu, T.Z.: Factors influencing tax-payer information usage behavior: test of an integrated model. In: PACIS 2004, Shanghai (2004)
10. Colesce, S.E., Dobrica, L.: Adoption and use of e-government services: the case of Romania. J. Appl. Res. Technol. **6**, 204–217 (2008)
11. Connolly, R., Bannister, F.: eTax filing & service quality: the case of the revenue online service. Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng. **2**, 56–60 (2008)
12. Davis, G.B., Olson, M.H.: Management Information Systems: Conceptual Foundations, Structure, and Development. McGraw-Hill Inc, New York (1984)
13. DeLone, W.H., McLean, E.R.: The DeLone and McLean model of information systems success: a ten-year update. J. Manage. Inf. Syst. **19**, 9–30 (2003)
14. Devaraj, S., Fan, M., Kohli, R.: Antecedents of B2C channel satisfaction and preference: validating e-commerce metrics. Inf. Syst. Res. **13**, 316–333 (2002)
15. Doll, W.J., Torkzadeh, G.: The measurement of end user computing satisfaction. MIS Q. **12**, 259–274 (1988)
16. Dorasamy, M., Marimuthu, M., Raman, M., Kaliannan, M.: E-Government services online: an exploratory study on tax E-filing in Malaysia. Int. J. Electron. Gov. Res. (IJEGR) **6**, 12–24 (2010)
17. Fisher, C.W., Kingma, B.R.: Criticality of data quality as exemplified in two disasters. Inf. Manage. **39**, 109–116 (2001)
18. Floropoulos, J., Spathis, C., Halvatzis, D., Tsipouridou, M.: Measuring the success of the Greek taxation information system. Int. J. Inf. Manage. **30**, 47–56 (2010)
19. Fu, J.R., Chao, W.P., Farn, C.K.: Determinants of taxpayers' adoption of electronic filing methods in Taiwan: an exploratory study. J. Gov. Inf. **30**, 658–683 (2004)
20. Fu, J.R., Farn, C.K., Chao, W.P.: Acceptance of electronic tax filing: a study of taxpayer intentions. Inf. Manage. **43**, 109–126 (2006)
21. Gotoh, R.: Critical factors increasing user satisfaction with e-government services. Electron. Gov. Int. J. **6**, 252–264 (2009)
22. Gupta, G., Zaidi, S., Udo, G., Bagchi, K.: The effect of espoused culture on acceptance of online tax filing services in an emerging economy. Adv. Bus. Res. **6**(1), 14–31 (2015)
23. Ha, L., James, E.L.: Interactivity reexamined: a baseline analysis of early business web sites. J. Broadcast. Electron. Media. **42**, 457–474 (1998)
24. Hu, P.J.H., Brown, S.A., Thong, J.Y., Chan, F.K., Tam, K.Y.: Determinants of service quality and continuance intention of online services: the case of eTax. J. Am. Soc. Inf. Sci. Technol. **60**, 292–306 (2009)
25. Hwang, C.S.: A comparative study of tax-filing methods: manual, internet, and two-dimensional bar code. J. Gov. Inf. **27**, 113–127 (2000)
26. Jaeger, P.T.: The endless wire: e-government as global phenomenon. Gov. Inf. Q. **20**, 323–331 (2003)
27. Kim, S., Stoel, L.: Apparel retailers: website quality dimensions and satisfaction. J. Retail. Consum. Serv. **11**(2), 109–117 (2004)
28. Lai, M.L.: Electronic tax filing system: benefits and barriers to adoption of system. In: The Chartered Secretaries Malaysia, pp. 14–16 (2006)
29. Lai, M.L., Choong, K.F.: Motivators, barriers and concerns in adoption of electronic filing system: survey evidence from Malaysian professional accountants. Am. J. Appl. Sci. **7**, 562–567 (2010)
30. McKinney, V., Yoon, K., Zahedi, F.M.: The measurement of web-customer satisfaction: an expectation and disconfirmation approach. Inf. Syst. Res. **13**, 296–315 (2002)

31. Meuter, M.L., Bitner, M.J., Ostrom, A.L., Brown, S.W.: Choosing among alternative service delivery modes: an investigation of customer trial of self-service technologies. J. Mark. **69**, 61–83 (2005)
32. Mohamed, N., Hussin, H., Hussein, R.: Measuring users' satisfaction with Malaysia's electronic government systems. Electron. J. e-Gov. **7**, 283–294 (2009)
33. Musptapha, B.: Evaluation of E-tax quality implementation criteria: the case of self-employed taxpayers in Nigeria. IJCER **4**, 39–45 (2015)
34. Ojha, A., Sahu, G.P., Gupta, M.P.: Antecedents of paperless income tax filing by young professionals in India: an exploratory study. Transforming Gov. People Process Policy **3**, 65–90 (2009)
35. Ramayah, T., Ramoo, V., Ibrahim, A.: Profiling online and manual tax filers: results from an exploratory study in Penang, Malaysia. Labuan e-J. Muamalat Soc. **2**, 1–8 (2008)
36. Saha, P.: Government E-service delivery identification of success factors from citizens' perspective. Ph.D. thesis. University of Technology, Luleå (2009)
37. Schaupp, L.C., Carter, L., ME, M.: E-file adoption: a study of US taxpayers' intentions. Comput. Hum. Behav. **26**, 636–644 (2010)
38. Schaupp, L.C., Carter, L.: The impact of trust, risk and optimism bias on E-file adoption. Inf. Syst. Frontiers **12**, 299–309 (2010)
39. Shao, B., Luo, X., Liao, Q.: Factors influencing E-tax filing adoption intention by business users in China" electronic government. Int. J. **11**, 283–305 (2015)
40. Straub, D., Boudreau, M., Gefen, D.: Validation guidelines for IS positivist research. Commun. Assoc. Inf. Syst. **13**, 380–427 (2004)
41. Szymanski, D.M., Hise, R.T.: E-satisfaction: an initial examination. J. Retail. **76**, 309–322 (2000)
42. Tee, S.W., Bowen, P.L., Doyle, P., Rohde, F.H.: Factors influencing organizations to improve data quality in their information systems. Acc. Financ. **47**, 335–355 (2007)
43. United Nations: From e-government to Connected Governance. Department of Economic and Social Affairs (2008). http://unpan3.un.org/egovkb/portals/egovkb/Documents/un/2008-Survey/unpan028607.pdf
44. Verdegem, P., Hauttekeete, L.: User centered e-government: measuring user satisfaction of online public services. In: IADIS International Conference e-Society (2007)
45. Verdegem, P., Verleye, G.: User-centered e-government in practice: a comprehensive model for measuring user satisfaction. Gov. Inf. Q. **26**, 487–497 (2009)
46. Wang, J., Senecal, S.: Measuring perceived website usability. J. Internet Commer. **6**, 97–112 (2007)
47. Wang, Y.S., Liao, Y.W.: Assessing e-government systems success: a validation of the DeLone and McLean model of information systems success. Gov. Inf. Q. **25**, 717–733 (2008)
48. Wixon, B.H., Todd, P.A.: A theoretical integration of user satisfaction and technology acceptance. Inf. Syst. Res. **16**, 85–102 (2005)
49. Yang, Z., Fang, X.: Online service quality dimensions and their relationships with satisfaction. Int. J. Serv. Ind. Manage. **15**, 302–326 (2004)
50. Yusuf, S.M.: The influence of taxpayers consciousness, tax services and taxpayers compliance on tax revenue performance. (Survey on the Individual Taxpayer in South Tangerang). Undergraduate thesis, Syarif Hidayatullah State Islamic University Jakarta (2013)
51. Zhang, X., Prybutok, V., Huang, A.: An empirical study of factors affecting e-service satisfaction. Hum. Syst. Manage. **25**, 279 (2006)

Data Driven Governments: Creating Value Through Open Government Data

Judie Attard(✉), Fabrizio Orlandi, and Sören Auer

University of Bonn, Bonn, Germany
{attard,orlandi}@iai.uni-bonn.de, auer@cs.uni-bonn.de

Abstract. Governments are one of the largest producers and collectors of data in many different domains and one major aim of open government data initiatives is the release of social and commercial value. Hence, we here explore existing processes of value creation on government data. We identify the dimensions that impact, or are impacted by value creation, and distinguish between the different value creating roles and participating stakeholders. We propose the use of Linked Data as an approach to enhance the value creation process, and provide a Value Creation Assessment Framework to analyse the resulting impact. We also implement the assessment framework to evaluate two government data portals.

Keywords: Government data · Value creation · Smart city · Data value network · Assessment framework

1 Introduction

Especially in recent years, open government initiatives have gone way beyond the simple publishing of data. In fact, the end aims of open data movements such as the Public Sector Information (PSI) Directive[1], U.S. President's Obama open data initiative[2], the Open Government Partnership[3], and the G8 Open Data Charter[4] focus on achieving *transparency*, *participatory governance*, and *releasing social and commercial value*. In order to have a well-functioning, democratic society, citizens and other stakeholders need to be able to monitor government initiatives and their legitimacy. Transparency means that stakeholders not only can access the data, but they also should be enabled to use, re-use and distribute it. The success to achieve transparency results in a considerable increase in citizen social control. Furthermore, through the publishing of government data, citizens are given the opportunity to actively participate in governance processes, such as decision-taking and policy-making, rather than sporadically voting in an

[1] http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information.

[2] http://www.whitehouse.gov/open/documents/open-government-directive.

[3] http://www.opengovpartnership.org/.

[4] https://www.gov.uk/government/publications/open-data-charter.

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 84–110, 2016.
DOI: 10.1007/978-3-662-53416-8_6

election every number of years. Hence, through open government data initiatives such as portals, stakeholders can also be more informed and be able to make better decisions [32]. This opportunity can have a major impact within so many dimensions, including, but definitely not limited to; urban management, marketing, service improvement, and citizens' quality of life.

All data, whether addresses of schools, geospatial data, environmental data, weather data, transport and planning data, or budget data, has social and commercial value, and can be used for a number of purposes that could be different than the ones originally envisaged. Governments are one of the largest producers and collectors of data in many different domains [15]. Considering its volume (huge amount of data produced), velocity (frequent gathering of data, especially sensor data), variety (different domains), and veracity (uncertainty of data), government data can be considered to be Big Data. By publishing such data the government encourages stakeholders to innovate upon it, and create new services. The main challenge in releasing social and commercial value is that open data has no value in itself, yet it becomes valuable when it is used [16], and there are many factors within an open government initiative that influence its success.

In this paper we attempt to answer the following research question: *What are existing processes of value creation on open government data?* With this research question we aim to address the niche in existing literature with regards to the creation of value on open government data. In this paper we hence identify the various processes in a government data value chain, as well as dimensions that, in some way or another, have an impact on value creation upon government data. We also distinguish between the different value creating roles of participating stakeholders within the government data value chain, and identify the resulting impacts of value creation and of exploiting data as a product. While we focus on government data, it is important to note that most of what we discuss is also valid for generic open data initiatives.

2 Methodology

In order to analyse existing approaches undertaken for creating value based on open government data, we review existing literature on open government data initiatives. We implement a systematic approach, where we define a number of search terms and perform a search on a number of digital libraries. Thereafter, we select which literature to include in our study by applying inclusion and exclusion criteria.

The search terms we defined are a combination of the following keywords: *government, data, portal, open, publishing, consuming*, and *public*. The latter were selected with the aim of obtaining results relevant to the research question defined in Sect. 1, or more specifically; any initiative that exploits government data on order to create value. We here stick to the definition of *Government Data* to entail any data that is government-related. It may or may not be produced or published by a governmental entity, and it may or may not be made openly available (it can have varying degrees of openness).

In order to cover the largest spectrum of relevant publications possible, we identified and used the most extensively used electronic libraries, namely: ACM Digital Library, Science Direct, Springer Link, IEEE Xplore Digital Library, and ISI Web of Knowledge. To achieve relevant results that are sufficiently comprehensive and precise, we apply these search terms on both the title and the abstract search fields.

After the systematic search was completed, we led a manual study selection based on exclusion and inclusion criteria. Basically, we only considered literature to be relevant if it regarded the actual exploitation of government data. This resulted in 74 publications that form our set of primary studies[5].

With the research question in mind, we analysed the 74 publications with the aim of identifying current practices of creating value using government data. We hence provide our observations, comments, guidelines, and conclusions in the rest of this paper. Of course, apart from the above-mentioned literature, we also conduct further lookup and exploratory searches [22] to identify related literature on which to base the contributions within this paper.

3 Background Literature

Data is increasingly becoming a commodity in our information society. It is steadily becoming the basis for many products and services, such as open data, Linked Data, or Big Data applications. Using open data, specifically open government data, has the potential of not only resulting in economic benefits, but also has good social and governmental impacts. Releasing government data will impact transparency and accountability factors, while the release of specific datasets can encourage stakeholders to create innovative services and boost economic growth. The release of information will also aid stakeholders in making informed decisions based on relevant data.

In order to reflect such a data-centric society, the concept of *value chains* [29] was coined to identify how value is created in order to achieve a product. The value chain model describes value-adding activities that connect an industry's supply side, such as raw materials and production processes, to its demand side, such as sales and marketing. The value chain model has been used to analyse and assess the linked activities carried out within traditional industries in order to identify where, within these activities, value is created. This was done with the aim to identify what activities are the source of competitive advantage within these industries.

As successful as the value chain concept was to achieve this aim, during these last years products and services are becoming increasingly digital, and exist in a more non-tangible dimension [27]. In addition, the traditional value chain model does not consider when information is used as a source of value in itself [30]. Thus, the original concept of value chain is becoming an inappropriate method with which to identify value sources in today's industries that produce non-tangible

[5] All primary studies can be accessed here: http://mnd.ly/1LFgFQJ.

products [27]. Newer definitions of the concept, such as in [9,19,20,24,27], cater for the digital dimensions; taking into account factors and activities which set this dimension apart from the more physical one.

Lee and Yang [20] define a value chain for knowledge, including the knowledge infrastructure, the process of knowledge management, and the interaction between the required components. Knowledge, a step further than information, is data organised in meaningful patterns. The process of reading, understanding, interpreting, and applying information to a specific purpose, transforms information into knowledge. This means that for an entity that is unable to understand knowledge, the knowledge is in fact still only information. This is the *data literacy* problem, where any effort invested in knowledge generation is lost if the target consumer is unable to actually understand the provided knowledge [37]. Similar to Porter, Lee and Yang classify the activities within the knowledge value chain in five categories, namely knowledge acquisition, knowledge innovation, knowledge protection, knowledge integration, and knowledge dissemination.

In [9], Crié and Micheaux provide us with a more generic value chain than Lee and Yang, including raw data in their definition. Within their paper, the authors aim to highlight any issues within the value chain, to provide an overview of the current progress, and also to encourage entities to view the benefits of participating within the data value chain. They focus on four aspects of the *Data Value Chain*, namely:

- *Obtaining the right data* – Capturing the right data is the first step to forming an information chain that aims to provide the best customer service and result in profits;
- *Data quality management* – Ensuring the data is of good quality increases the potential towards maximising returns from the data for both the entity and its customers;
- *Deriving information and knowledge from raw data* – The act of extracting information from data, and interpreting knowledge from information;
- *Using information and knowledge to satisfy customers and generate profits* – The use of good data increases the chance of making better decisions.

Peppard and Rylander [27] also discuss a value chain that is more suited where the product in question is digitised, and thus non-tangible. The authors introduce the concept of *Network Value*, where value is created by a combination of actors within the network. In contrast to the earlier definition of a value chain, network value does not necessarily follow a linear model, and accounts for the various interconnected actors that work together to *co-produce* value. While these actors or entities should be able to function independently, they operate together in a framework of common principles. This means that an action by a single entity can influence other entities within the network, or otherwise require further actions from them in order to achieve the final product. Morgan et al. [26] provide a similar discussion on the co-production of value through open-source software.

In line with more recent popular themes, Miller and Mork [24] and Latif et al. [19] focus on big data and Linked Data respectively. Similar to

Crié and Micheaux [9], Miller and Mork discuss the data value chain concerning all required actions in aggregating heterogeneous data in an organised manner and creating value (information/knowledge) that can influence decision-making. The authors divide their data value chain in three main categories, namely data discovery, integration, and exploitation. In contrast, Latif et al. propose the *Linked Data Value Chain*. Motivated by the still limited commercial adoption of the Semantic Web, the authors aim to drive the Semantic Web and the use of Linked Data closer to commercial entities. The authors discuss the entities participating in the Linked Data value chain, their assigned Linked Data roles, as well as the types of data processed within the chain. An interesting aspect that distinguishes the proposed Linked Data value chain from the ones previously mentioned is that actors within the chain are not necessarily bound to one specific role. Rather the assignment of roles to entities is more flexible where, in extreme cases, an entity can even occupy all roles at once.

Whichever model it follows, the data value chain is at the centre of a knowledge economy [1], where data products provide digital developments to more traditional sectors, such as transport, health, manufacturing, and retail. Being one of the largest producers and collectors of data in so many domains, governments play a vital role in data value chains. Essentially, the strategy beneath a data value chain is to extract the maximum value from data by building on the intelligent use of data sources [1]. The authors of [28] add that value-adding is one of the most important properties for information, where the objective of adding value to information is to develop information products and provide information service with social and economic value.

Based on the discussed literature, we provide our definition of a *Data Value Network*, as shown in Fig. 1. This network differs from the classic definition of a data value chain in that the *activities* within the network do not follow a sequential structure; rather, activities can be executed in tandem, and other activities can be skipped or repeated. Furthermore, each activity can be further broken down into more specialised *value creation techniques* (the thin arrows in

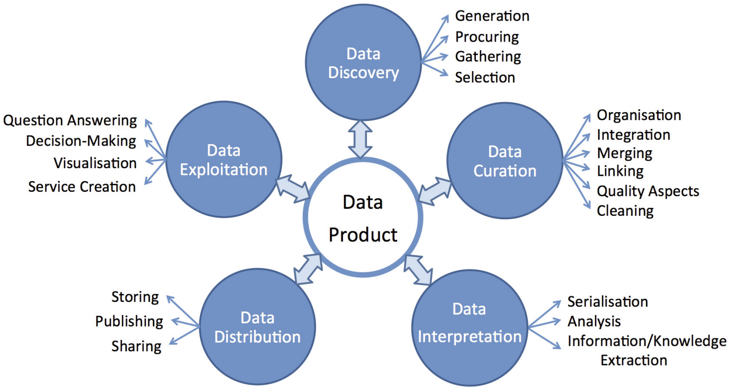


Fig. 1. The Data Value Network

the Fig. 1). While not exhaustive, the listed techniques are the most common and generic processes that can be executed on a data product. The Data Value Network also caters for multiple actors, where one or more actors can participate to co-produce value within an activity. We hence define the Data Value Network to be:

> *A set of independent activities having the aim of adding value to data in order to exploit it as a product*

where different **actors** can participate by executing one or more **activities**, and each activity can consist of a number of **value creation techniques**.

Data activities in a Data Value Network all have the purpose of adding value to data, which may or may not result in a new data product. We can consider 'adding value' to be equivalent to 'making the data more usable, or making it more fit for use in a specific use case'. So, for example, while data in PDF format is easily human-readable, it's conversion to RDF would make it more usable where the use case requires data to be machine readable. The opposite can also stand true. We here provide a brief description of the activities that add value to data. The value creation techniques will be described in detail in Sect. 4.

- **Data Discovery**: Data discovery is the process of obtaining data. Sources of data can be as varied as sensor data, the Internet, private companies, governmental entities, and social media, amongst others.
- **Data Curation**: This is a very generic activity that can encompass a large number of different value creation techniques, all of which modify the data in some way or another. This activity can recur numerous times, until the required data product is obtained.
- **Data Interpretation**: This activity involves presenting the data in a different manner, in order for it to be more understandable.
- **Data Distribution**: Data distribution is the activity involving making the data available as a product. This means that other entities can search for and discover this data.
- **Data Exploitation**: This activity can be considered as the final goal of the Data Value Network, though it does not signify the Data Value Network is finished. It involves consuming the data as a product.

Within an urban environment such as a city, the Data Value Network can have major impacts on the citizens, especially where a data product is used in a decision-making process. This aspect is considered to be part of a smart city. The *decision-making process* is a very broad term used to encompass the practice of familiarising oneself with the relevant information before taking a particular decision. This concept was discussed as early as the 1970s, where Montgomery [25] describes the use of information systems to aid in the planning and decision-making processes within a marketing environment.

While there are various definitions of a smart city throughout literature, we consider a city to be smart *where investments in human and social capital* *and traditional (transport) and modern (ICT) communication infrastructure fuel sustainable economic growth and a high quality of life, with a wise management of natural resources, through participatory governance* [7]. There is increasing effort worldwide to transform cities into smart cities, particularly through the release of government data to the public, as well as through the exploitation of this data. Examples include Rio de Janeiro in Brazil[6], Dublin in Ireland[7], and London in the United Kingdom[8]. Whilst earlier attempts at "smartifying" a city mostly concerned the automation of some routine functions, more recent attempts are focusing on improving the management and sustainability of a city through monitoring and analysing the relevant data, with the intention of improving the quality of life if its citizens [3]. Smart cities can hence impact various dimensions in a citizen's life, for example:

- Transportation: The analysis of traffic data can aid citizens to check the best time to use certain roads, public transport can be better managed through better prediction of arrival times, whilst the government can attempt to lessen traffic by providing alternative transportation options. For example, a live view of the car boarding areas for the ferry between the islands of Gozo and Malta is streamed[9] in order to enable citizens to check if there is currently a long queue and plan their travels accordingly. Moreover, traffic supervisors can be dispatched to control and manage the boarding process.
- Energy Consumption: The use of smart meters and other sensors can help in reducing energy consumption through monitoring use in real-time. For example, an initiative throughout the European Union is currently ongoing with the aim of controlling energy consumption and providing for a more sustainable environment[10].
- Weather Emergencies: Weather information can be used to predict if a weather-related emergency is incumbent, such as flooding, landslides, earthquakes, etc. This prediction can be used to issue warnings or evacuation orders in time. The city of Rio de Janeiro is a good example of this use case, as an operations centre[11] was established with the aim to prevent weather-related disasters (amongst other aims).
- Health: Patient data can be used to generally monitor a patient during an ongoing treatment or to issue reminders when check ups or vaccinations are due. The Immunize India initiative[12] is an example of the latter.

4 Value Creation Techniques

Table 1 shows the various Value Creation Techniques within the Data Value Network. While not comprehensive, we included the most popular and

[6] http://www.centrodeoperacoes.rio.gov.br/.

[7] http://www.dublinked.ie/.

[8] http://citydashboard.org/london/.

[9] http://www.visitgozo.com/en/content/live-ferry-queue-streaming-beta-1538/.

[10] http://my-smart-energy.eu/.

[11] http://centrodeoperacoes.rio/.

[12] http://www.immunizeindia.org/.

Table 1. Value Creation Techniques categorised according to the Data Value Network

Government data life cycle processes	Value creation techniques
Data Discovery	Generation Procuring Gathering Selection
Data Curation	Organisation Integration Merging Linking Quality aspects
Data Interpretation	Serialisation Analysis Information/Knowledge extraction
Data Distribution	Storing Publishing Sharing
Data Exploitation	Question answering Decision-making Visualisation Service creation

frequently-used techniques from various stakeholders participating in the Data Value Network. The aim of all these techniques is to create or improve upon a data product, resulting in data that is (more) ideal to be used in the required application and increasing its value and re-use potential.

Data is produced in the day-to-day administration of a governing entity. The simple **generation** of this data is the first step towards its (re) use as a data product. As opposed to data generation, data **procurement** involves obtaining data generated by a different entity through performing some sort of negotiation. Data **gathering**, on the other hand, refers to the aggregation of data from different entities or locations. Finally, data **selection** requires the stakeholder in question to choose a subset of available data and extract it, potentially for the use in a different use-case then what the data was originally generated for. In order for the best value potential, all generated, procured, gathered, or selected data, need to be complete. This means a record has all the information required for an accurate representation of the described data.

The value creation techniques falling under the Data Curation activity have the purpose of making the data more usable. Data **organisation** requires the structuring of data in such a way that the data is more understandable, or that the data follows some pattern; for example government budget data can be organised by year. Data **integration** has the purpose of enriching an existing dataset with new data, possibly with the intention to use the data in an unprecedented use. For example, the integration of weather data to accident information can be done by an insurance company to check the legitimacy of a claim. Another example is adding user feedback to product data in order to identify product faults. Data **merging** is somewhat similar, where different datasets are merged in order to obtain further information. For example, the merging of population data with geographical data can be used to obtain population density. On the other hand, the **linking** of different datasets is done in order to provide context, for example linking geographic data to textual descriptions about the locations in question. Finally, data **quality** involves the assessment and (if necessary) improvement and cleaning or repairing of data, such as removing duplicate data, ensuring the data is consistent, complete, timely, and trustworthy, and adding provenance data. This technique gives the data a higher level of quality and encourages its re-use. Similarly, metadata also enhances a datasets re-use potential. By enriching a dataset's metadata, a dataset is made more easily discoverable by potential users [31].

The Data Interpretation activity involves some sort of reasoning where the data in question is made more understandable. In the simplest way, data **serialisation** involves the conversion of data into semantically richer or lower formats, such as PDF to RDF, or CSV to RDB. This conversion enables stakeholders with different backgrounds to still be able to exploit the data in question to its highest potential. Moreover, the use of non-proprietary, machine-readable formats will increase the value creation potential of the data in question. The implementation of **analysis** techniques, such as data mining, pattern identification, and trend analysis, enables stakeholders to identify any existing patterns, which can eventually aid actors in the Data Value Network in actions such as decision-making. **Information/knowledge extraction** has a similar purpose, where raw data is interpreted manually (non-machine), and along with the available context information and the knowledge from the stakeholders in question can be used to arrive to particular conclusions.

Techniques such as storing, publishing, and sharing, all have the purpose of adding the potential of the data to be distributed to different entities and re-used. The **storing** of data enables actors to re-use the data in question without requiring a local copy. By **publishing** the data in an open manner, and making it **shareable**, it is also made available to many more external stakeholders. This publishing process creates value simply by making data available for re-use. The data distribution activity is a vital node within the Data Value Network, as data that is not made available publicly is very limited in its re-use potential. Therefore, data that is provided in a timely manner (data is provided in a reasonable amount of time after creation/generation), without discrimination on its consumers (not requiring any registration), and made accessible for all, has the best value creation potential. Moreover, the addition of metadata enables the data to be more discoverable, thus enhancing this potential.

Popular methods of publishing data include SPARQL[13] endpoints and Application Program Interfaces (APIs). Licensing is also vital here, as it has the purpose of declaring if and how data can be used. In the case of government data it is preferable that licences are of an open nature.

The Data Exploitation activity encompasses any value creation technique that involves consuming the data to solve a particular problem. **Visualisation** can be considered as an example of *passive* exploitation, where an actor consumes the data as information or knowledge. Visualisations involve a visual representation of data that, similar to data interlinking and data analysis, can provide us with a new insight. Visualisations can also be used to provide 'stories', since they are more easily interpreted than raw data. An example of a more *active* consumption of the data can be the use of data to influence **decision-making**, for example, a government might consider citizens' feedback before taking a decision. **Question answering** and **service creation** are other examples of active consumption of data. In the former data is collected and analysed in order to solve a specific question, whilst service creation is the provision of a service through the use of existing data, for example a mobile public transport timetable application.

4.1 Stakeholders: Beneficiaries, Contributors, and Their Roles

Government data, or public sector information, is a resource holding great potential for a large number of stakeholders. Governmental agencies, citizens, non-profit organisations, and businesses, are but a few of the potential stakeholders who, through the exploitation of open government data, can reap substantial benefits. Since the efforts of the latter stakeholders remain largely uncoordinated, their motivations, levels of expertise, and priorities differ. In this section we proceed to identify and explore the various stakeholders who, either through value creation or other means of consumption, use open government data.

The most obvious role of **governments** in open government data initiatives is the role of a data provider. Yet, public entities are also the direct beneficiaries of their own published data. Through transparency as a motivation, the publishing of data can increase accountability, and moreover inhibits corruption. In turn this increases citizens' trust in their government. The analysis of government data, such as budget data, has the potential of increasing efficiency and influencing decision-making. Innovations based upon such data can also be used to provide more personalised public services, thus increasing the quality of the interactions between governments and their citizens.

Through publishing government data, **citizens** are given the possibility of participating in governance processes. Apart from being able to make more informed decisions, citizens are sometimes given the opportunity to take part in participatory governance. For example, in a participatory budget effort citizens are given a say as to how, or for what, budget should be prioritised. Citizens can also participate in open government initiatives by being data *prosumers*.

[13] http://www.w3.org/TR/rdf-sparql-query/.

By this we mean citizens who both produce and consume data. For example, the Fix My Street[14] application provides a platform where anyone can submit an existing problem in a street, in order to indicate the problem areas to the government. In this crowdsourced co-production of value, we have geographical data consumption, and street issues data production. Open government data certainly has the potential of increasing citizens' quality of life.

Non-profit organisations, such as non-governmental organisations (NGOs) or Civil Society initiatives, can have a huge difference in their goals. Examples of such organisations include the Sunlight Foundation[15] and the Open Knowledge Foundation[16], present in various countries. Organisations such as the latter usually share the goals of demonstrating the benefits of opening governmental data both to the general public and to the governments themselves. They also play a vital role as intermediaries who can identify key datasets that have the potential of being very valuable if published as open data.

Table 2. The activities in which each actor participates within the Data Value Network

	Data Discovery	Data Curation	Data Interpretation	Data Distribution	Data Exploitation
Data Producer	✓				
Data Enhancer		✓	✓		
Data Publisher	✓			✓	
Service Creator		✓	✓		✓
Facilitator		✓	✓	✓	
Data Consumer			✓		✓

Private companies, small to medium enterprises (SMEs), entrepreneurs, and other **businesses**, have the potential of not only making an economic profit through using government data, but can also create more jobs, and (depending on the nature of the service) also provide innovative services that increase the beneficiaries' quality of life and indirectly impact job creation in this field. While the sole access to data does not provide competitive advantage, private entities can innovate upon the available data to provide value-added services.

Whatever the stakeholder's nature (citizen, governmental entity, NGO, etc.), we identify six roles in which they can participate to create value, and in Table 2 we show how each role participates within the Data Value Network.

- **Data Producer:** A data producer is the entity that creates, obtains, or generates the data. The role of a data producer can be considered as one of the

[14] https://www.fixmystreet.com/.
[15] http://sunlightfoundation.com/.
[16] https://okfn.org/.

most important roles within the Data Value Network, as any activity or action in the network depends on the available data. If the data producer does not obtain relevant data for the use case at hand, then the Data Value Network will not reach its target to obtain the intended value out of this data.
- **Data Enhancer:** This role involves creating value through the actual manipulation of the data in a way that it is more usable for the target aim. A data enhancer can influence the outcome of the Data Value Network by adapting the data so that its highest value potential can be exploited.
- **Data Publisher:** This role involves the discovery and distribution of the data product. This distribution process enables other stakeholders to discover potentially useful data products.
- **Service Creator:** A service creator entity has the task of using open government data to provide a service. This can take the shape of a website, a mobile application, information access points, etc.
- **Facilitator:** This role involves entities that, in some way or another, aid the other stakeholders in using, re-using, or exploiting, open government data. This can be done through the provision of software, services, or other technologies. For example, the creator of a government data portal is facilitating the use and re-use of government data from other stakeholders by organising heterogeneous government data in a single location.
- **Data Consumer:** The data consumer role can be considered the final role in the Data Value Network, however, this is not always the case. For example, when a consumer gives feedback, the feedback can in turn be used as a data product by the product manufacturer. In the case of crowdsourcing, the data consumer also has the role of a curator, blurring the lines between both roles. Actors in the role of a data consumer can exploit the data product in many ways, as defined in the *Data Exploitation* activity.

4.2 Barriers, Enablers, and Impacts of Value Creation

Within the Data Value Network, value creation is both dependent on a number of dimensions, and also results in impact on other dimensions. Based on efforts in the primary studies (See Sect. 2), and other literature such as [8,16,39,41], we identify the dimensions with the strongest impact. Figure 2 maps their relationship, where a number of dimensions act as *enablers* or *barriers* towards value creation. In turn, the value creation process impacts a number of other dimensions. The stakeholders, while they give input for value creation, are also impacted through the results of their efforts.

4.2.1 Value Creation Enablers/Barriers:

The latter dimensions have a great impact on value creation in that they control to what extent value is created.

The **Technical Dimension** mostly regards aspects concerning the data itself. The format of the data is an essential aspect. Two of the eight Open Government Data Principles[17], in fact, regard the format in which data is made

[17] http://opengovdata.org/.

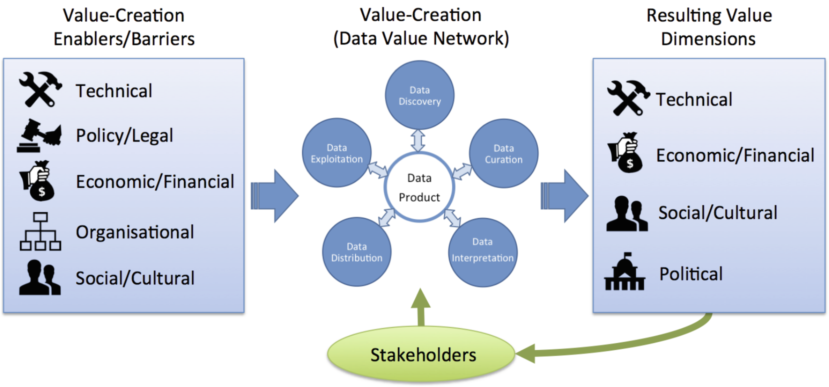


Fig. 2. Dimensions impacting, and impacted by, Value Creation

available to the public. They state that such data should be available in a *machine-processable* format which is *non-proprietary*. Such data would enable easier and un-restricted use of the data for value creation. Furthermore, if a format such as Resource Description Framework (RDF) is used, data ambiguity is reduced due to the format's expressivity, making the data more *understandable*. Additionally, the use of common schema aids to reduce interoperability issues caused by the large heterogeneity of the existing data. In order to encourage its use, data must also be easily *discoverable*. This is possible through the use of good quality metadata. The implementation of agreed-upon standards would aid reduce some, if not most, of the issues within this dimension.

The **Policy/Legal Dimension** regards issues with existing laws or policies that, through their ambiguity or due to being out-of-date, prevent data from being used to create value. On the other hand, well thought out policies encourage and enforce the creation of value, for example the publishing of data as Linked Data. Fortunately, there are growing efforts towards amending such laws and policies, but there is still a long way to go. Copyright and licensing of data can inhibit its unrestricted use. The incompatibility of licences, due to the data being created by various entities, further aggravates the issue. Privacy and data protection is another important aspect. Data providers need to strike a balance between making data freely available, whilst respecting the right to privacy.

The **Economic/Financial Dimension** is about aspects related to monetary issues and mainly concern the data provider and the data publisher roles. Being a relatively new concept, there might not be any budget allocation specifically for open government data efforts. In order to foster value creation, governmental entities cannot solely rely on existing data created in their day-to-day functionalities. Commitment is required, and hence also finances, for identifying and opening datasets with a high value creation potential.

The **Organisational Dimension** is concerned with the strategic aspects of the involved stakeholders. This dimension is especially relevant for governmental institutions. Considering there probably isn't an institution specifically in charge of open government data initiatives, data can get lost in the various hierarchical levels of a government. Adequate workflows need to be put in place for all the processes within a government data life cycle.

Finally, the **Social/Cultural Dimension** regards the feeling of the public towards open government data. While efforts are well under way to increasing awareness about the potential of open government data, not all stakeholders are ready to jump on the bandwagon. Workers within governmental entities might not understand the value of the data they are gathering/creating. This results in lack of motivation towards providing this data to the public. Stakeholders can also have misconceptions about the opening of public data. While open data can be considered as unfair competition for private entities (who invested to create their own data), public entities might consider the commercial appropriation of public open data unfair. The public also needs to be further informed on the advantages of public participation in creating value.

4.2.2 Impacts of Value Creation:

As already discussed in the previous sections, value creation has a number of different dimensions of impact, which in turn affect the stakeholders. The term *public value* is used to define "what adds value to the public sphere" [4], where the public sphere is used to broadly indicate all of the following dimensions:

Technical Value is simply generated through the implementation of standards and the creation of services. As more value is created upon government data, the available data will be of better quality, and value creating services will increase.

Economic Value is defined as the worth of a good or service as determined by the market [17]. Value creation upon data enables the data itself to be considered as a product. Therefore, opening government data encourages its re-use in value creation, in turn stimulating competitiveness in the participating stakeholders and also encourages economic growth. For example, Mastodon C (a big data company) used open data to identify unnecessary spending in prescription medicine[18]. This will result in potentially huge savings from the National Health Service in the UK.

Social/Cultural Value is created first and foremost through the engagement of the public in open government data initiatives. The opening of data allows stakeholders to scrutinise the data and provide feedback on it. If the governmental entities exploit this feedback, it can result in improvement of citizen services. This sort of participation also increases citizen social control. Social value is also generated through creating innovative services based on open government data. For example, the Walkonomics Application[19] uses open data to enable users to identify potential dangers in a street, such as fear of crime or road safety.

Political Value is created through the stimulation of democratic dialogue. Through participatory governance, citizens can gain a better insight as to

[18] http://theodi.org/news/prescription-savings-worth-millions-identified-odi-incubated-company.

[19] http://www.walkonomics.com/.

how the governing process works. Stakeholders can possibly also participate in improving the policy-making process. Besides, the efforts of governmental entities to be more transparent and accountable increases citizens' trust in their government.

Through value creation, stakeholders are hence affected through all of the above dimensions. In line with the most relevant motivations behind open government data initiatives[20], namely transparency, releasing social and commercial value, and participatory governance, we identify four main levels of impact that are affected by the above dimensions and can be tangibly felt by the involved stakeholders.

1. Access to Information - Once data is re-used, the most directly tangible impact is access to information. The innovation and creation of services upon government data provides all stakeholders with more and more data and information that they can create value upon. In turn, the increase in availability of data products no only creates more jobs, but also affects the stakeholders' quality of life. This level of impact is directly affected through the Technical and Economic dimensions.
2. Transparency - By enabling stakeholders to create value upon government data, there can be a considerable increase in transparency. This is directly impacted by the Social/Cultural and Political Dimensions. Citizens are not only able to scrutinise data, but also create value upon it by providing relevant feedback. This sharing of responsibilities will allow them to interact with the government more actively, providing them with an opportunity to further exercise their duty and right of participation.
3. Accountability - Similarly to transparency, the creation of value on government data allows stakeholders to assess the legitimacy and effectiveness of the government's conduct. This helps citizens to establish a trusting relationship with the government. Affected by the Social/Cultural and Political dimensions, accountability enables citizens to be aware of how they are being governed, and have the relevant justifications.
4. Democratic Governance - Value creation on open government data not only promotes transparency and accountability, but also democracy. By participating in an open government initiative, stakeholders can provide feedback. The latter not only informs the governmental entity of the public opinion, but can also be used to improve service delivery. Affected by the Economic and Political dimensions, democratic governance essentially provides citizens with more social control.

5 Linked Data

In recent open government data initiatives, Linked Data practices are being followed by an increasing number of data publishers/providers such as data.gov.uk and data.gov. Yet, the use of Linked Data in open government initiatives is still

[20] http://opengovernmentdata.org/.

quite low [35]. This might be due to a number of reasons, as the use of Linked Data is a process involving a high number of steps, design decisions and technologies [40]. We here investigate the advantages and benefits of using Linked Data practices in an open government data initiative.

The term *Linked Data* is used to refer to a set of best practices for publishing and connecting structured data on the Web [5]. Therefore, Linked Data is published on the Web in a machine-readable format, where its meaning is explicitly defined. It is also linked to and from external datasets. This has the potential of creating the *Web of Data* (also known as Semantic Web); a huge distributed dataset that aims to replace decentralized and isolated data sources [13]. The benefits of applying Linked Data principles to government data as covered in literature include [10,18]:

- Simpler data access through a unified data model;
- Rich representation of data enabling the documentation of data semantics;
- Re-use of existing vocabularies;
- Use of URIs allow fine-grained referencing of any information;
- Related information is linked, allowing its unified access.

While significant efforts in literature cover advantages of using Linked Data (for example [11,14,35,36]), there is no evident effort targeted towards the benefits of using Linked Data specifically in open government data value creation. We here therefore proceed to focus on the value creation techniques described in Sect. 4 and the benefits provided through the use of Linked Data. While still having similar barriers, enablers, and impacts, as described in Sect. 4.2, the use of Linked data can result in different levels of impact, since the use of Linked Data techniques directly reduces some barriers of the technical level.

5.1 Linked Data as a Basis for Value Creation

Linked Data and Semantic Web technologies have the potential of solving many challenges in open government data, as well as possibly lowering the cost and complexity of developing government data-based applications.

Starting from the most common starting point of creating value, in general, data **generation** is the least impacted from the use of Linked Data since essentially the data is still being created. Data **procurement** is similarly not impacted to a high level. Yet, the data **gathering** process can be enhanced through the use of Linked Data. Consider the example of providing feedback based on a linked open dataset consisting of budget data. The use of Linked Data enables feedback providers to have further context on the available data through the links. This would aid them in making a more informed decision. Furthermore, the high level of granularity of Linked Data has the potential of providing a deeper insight on the resource at hand. Also, since the data publisher is not necessarily the data provider, Linked Data will enable the access to primary data through the use of provenance information located within the metadata. In the case of data **selection**, the use of Linked Data is particularly useful in querying for subsets of an existing dataset. Query languages such as SPARQL enable actors to generate complex queries and get very specific subsets of data.

The value creation techniques within the Data Curation activity are some of the highest impacted techniques within the Data Value Network through the use of Linked Data. Linked Data is based on models (schema) or ontologies that are best suited to represent the data at hand. In this way, the **organisation** of data is very easily achieved through the manipulation of the model at hand. If an entity is working with Linked Data, we can safely assume the data is represented in a semantically rich, machine-processable format. Hence, links with or between other datasets are more easily identified through the implemented models, and thus, the data **linking** process is simplified. Thereafter, data **integration** and **merging** follow easily through joining the existing models. Through the use of the standards required to obtain Linked Data, the *fitness for use* of data, and hence its **quality**, is immediately increased. For example, data ambiguity is decreased through the use of a semantically rich format, and data consistency can be ensured through the implemented data model. Moreover, in some instances, the quality assessment of data (and the ensuing data repairing/cleaning) can be more easily executed. For example, having a model for a linked dataset enables a stakeholder to assess the schema completeness for the dataset. Linked Data also enables (semi) automated cleaning and repairing of datasets through the use of reasoners. In this way, the violation of logical constraints is easily identified through the dataset's underlying model. Through the use of metadata, a consumer can also check the provenance of the data, and ensure that it is a reliable source. Timeliness and versioning information can be obtained in the same manner.

Having Linked Data means that the available data already conforms to some standards with regards to formatting, however this does not necessary make it easier to **serialise** to other formats. Yet, the use of agreed-upon standards positively affects the accessibility, discoverability, and re-usability potential of the data in question. Since Linked Data standards demand the use of a semantic representation such as RDF, Linked Data is automatically more accessible than other standards such as CSV or PDF. Data **analysis**, is also enhanced through the use of Linked Data. As explained above, Linked Data enables easier integration and merging of datasets, which in turn affect the implementation of analysis techniques. Moreover, through the existence of links it is easier to get further context and information on the data at hand, enhancing pattern identification. Similarly, the use of Linked Data in **information/knowledge extraction** also provides further insight and context to actors through links between the datasets, and within datasets themselves. This increased information directly affects the data interpretation process, as the data consumer can interpret the data in a more informed manner, and generate knowledge from the existing information.

The aim of the value creation techniques within the Data Distribution activity is to make the data more accessible as a data product. As mentioned above, the use of Linked Data standards automatically makes the data more accessible and discoverable. Hence, **stored** or **published** Linked Data has the potential to be easily accessed and manipulated through a variety of manners, such as RESTful APIs and public endpoints (queryable through SPARQL). This means that while Linked Data alternatives might require a consumer to download a data dump, the use of Linked Data enables the same consumer to access the specific subset of data he/she needs, and manipulate it easily. Additionally, each data resource is dereferenceable, i.e. the resource URI can be resolved into a web document on the Web of Data. The **sharing** of data is also impacted through the use of Linked Data technologies, as the links in between different datasets make them more easily discovered through the crawling of web resources, which potentially could lead to the addition of the dataset to the more known LOD cloud[21].

Data Exploitation is possibly the activity that has the highest impact from the use of Linked Data. Similarly to the knowledge/information extraction process, **question answering** and **decision-making** are enhanced through the existence of links and the provision of further context. Hence a more informed stakeholder is more capable of making the best decision, or obtaining the best answer for the problem at hand. The creation of **visualisations** is also affected through the existence of links between multiple datasets. Visualising a dataset against a related dataset has the potential of providing the consumer with a new and different understanding of the data. Finally, **service creation** on top of Linked Data has the advantage of easier data consumption (through the use of standards), and more interoperability.

The above benefits of using data for value creation are only a few, yet they collectively encourage and enhance the exploitation of open (government) data. Of course, this does not mean the implementation of a Linked Data approach does not have its challenges. Various efforts in literature, such as [36], provide discussions on the topic.

5.2 Use Case of Linked Open Government Data

publicspending.net is a data portal created with the scope of demonstrating the power of economic Linked Open Data in analysing the situation with regards to market, competition conditions, and public policy, on a global scale. The creators of this portal consume and create value upon public spending data of seven governments around the world. Results of the analysis led on the data are then published on the portal as tables, graphs, and statistics. The stakeholders here participate through all six value-creating roles described in Sect. 4.1 and execute value creation processes accordingly. Firstly, the public spending data is produced by the various governments (Data Producers). The data is then subject to pre-processing and data-preparation. Through the role of a Data Enhancer, the stakeholders here homogenise and link the data through the Public Spending Ontology and other widely used vocabularies such as Dublin Core and FOAF. The resulting data in RDF is then published (Data Publisher) on the portal

[21] http://lod-cloud.net/.

and is available both as bulk datasets and through a SPARQL endpoint. The Data Facilitator Role and the Service Creator Role are then fulfilled through the application built on top of the data. These stakeholders use the internal data, along with other cross-referenced and external data, to provide a portal acting as an information point. Finally, the Data Consumer can view and exploit the provided data in a myriad of ways, including exploring and scrutinising spending data that giving them a good insight as to what is being spent, where, and by whom. Such an open government data initiative enhances accountability and prevents corruption since it aids citizens to be more informed about how their country is being led, and if it is being led in a suitable manner. This can also help them decide who to vote for in an upcoming election.

6 Risks of Open Government Data

Whilst there are certainly numerous benefits and advantages of opening government data and creating value upon it, there still are a number of challenges that deter such initiatives from being successful and reaching their full potential, such as this discussed in Sect. 4.2. Moreover, if an open government data initiative is not implemented properly, the opening of data might also pose risks to some of the involved stakeholders. Within itself, this deters stakeholders from participating within an open government data initiative. We here proceed to outline some of the major risks of opening government data and creating value on it.

Conflicting regulations: Open government data initiatives have only become popular in recent years. Whilst there is certainly an increasing effort towards establishing policies, many open government data initiatives still belong to existing legal frameworks concerning freedom of information, re-use of public sector information, and the exchange of data between public entities. The risk here lies in the uncertainty of how such initiatives can interact. This issue concerns both data consumers, who are unsure how the available data can be used, and the data producers, who end up being sceptical of fully opening up their institutions' data, even if it is covered by a clear legal framework [33].

Privacy and Data Protection: Data protection and the right to privacy have some essential conflicts with the aims behind an open government data initiative and its motivations of transparency and accountability [23,33,41,42]. Published data can certainly be anonymised, yet the merging or linking of different datasets can still possibly result in the discovery of data of a personal nature. For example, if garbage collecting routes are published, along with the personnel timetable, a data consumer would be able to identify the location of a particular employee. This issue requires more research in order to come up with guidelines that can provide a solution to this conflict, however a plausible approach would be to employ access control mechanisms which regulate data access. However, this restricts the openness level of such data.

Copyright and Licensing: The issue here lies with the incompatibility of used licences and copyright inconsistencies. Efforts in open government data initiatives strive towards publishing data in an open format, allowing the free and unrestricted use, re-use, and distribution of data. Since there are no agreed-upon standards, this results in a myriad of licenses that although all are of an open nature, they can be incompatible between them as they might contain restrictions that prevent data with different licences from being merged. Unclear dataset ownership resulting from data sharing, for example between different levels of public entities, results in copyright inconsistencies that hinders data from being published, as the rightful owner of the data is unclear [8,42].

Competition: There are two perspectives to this risk: (i) open data can be considered as unfair competition for private entities, and (ii) public entities might consider the commercial appropriation of public open data unfair [33]. In the first perspective consider business entities who invested in creating their own data stores. If the same data they created is made public through government open data initiatives, these companies will obviously deem it to be unfair competition as there is the possibility of new competitors who did not need to invest anything but could get the freely available open data. Thus, management mechanisms need to be applied in order to ensure that private companies do not suffer financial consequences due to opening up their data. On the other hand, public entities might be reluctant to publish their data openly due to not wanting data belonging to the public (and paid by taxes) to be used for commercial gain. A possible approach for the latter issue is to provide the data for a nominal fee. Yet, this limits the openness of the data in question.

Liability: Mainly, this risk is limited to data providers. The latter, in the context of this paper governmental entities, fear being held liable for damage caused by the use of the provided data due to it being stale, incorrect, or wrongly interpreted [12,33]. To cater for this fear, many public entities either do not publish their data or otherwise impose restrictions on its use, resulting in data which is not truly open. In the worst case, due to fears of data being used against the publishing entity, such data might not even be collected/generated any longer [42]. A possible solution for these issues is to enable social interaction with regards to the data in question. A community of stakeholders within the data platform where the data is published can aid data consumers to better interpret and exploit the published data.

Considering the above risks or negative impacts, it is vital to find a trade-off for open government initiatives. One must keep in mind the numerous benefits associated with open data, but also cater and prepare for any risks, challenges and issues.

7 Value Creation Assessment Framework

In order to assess the success of open government data initiatives, there exist a large number of assessment frameworks that aim to evaluate the effectiveness of an initiative in achieving its goals and objectives. Yet, rather than assessing the resulting impacts of such an initiative, real-life assessments, as documented in literature (See Sect. 2), mostly involve checking whether open government data initiatives are obeying existing policies and regulations [34]. Since the latter are not necessarily up to date with current technologies and approaches, this assessment is not really representative of the success of an initiative.

Consider the example of a government publishing the data in PDF. While the entity would be obeying existing laws requiring opening up such data, the use of PDF makes it pretty inconvenient for re-use and re-distribution. In this case, one could argue that the open government initiative is not really a success. For this reason, a number of assessment frameworks analyse open government data initiatives based on different criteria [6,21]. The latter include nature of the data, citizen participation, and data openness. In [2] we give a more in depth overview of existing assessment frameworks in literature. While there is still the problem that there is no agreed-upon assessment framework to evaluate open government initiatives, there is also limited literature (such as [38]) that focuses on the *impact of value creation*. Considering many resulting benefits of open government data depend on the creation of value (through the execution of one or more value creation techniques), we deem it essential to assess open government data initiatives on their potential for enabling value creation.

In Fig. 3 we provide an overview of commonly evaluated aspects (in blue) of an open government data initiative extracted from our primary studies. These mostly concern implementation aspects, such as the format of the data, and how the initiative respects the requirements set from existing laws and policies. The bottom part of the figure portrays the missing aspects (in red), i.e. those that are not considered when evaluating the success of an open government data initiative. We propose the latter aspects (together with a couple of aspects that are already being assessed) as part of a *Value Creation Assessment Framework*. The aim of this framework is to provide a guideline as to what aspects of an open government data initiative should be assessed to determine the potential of an open government data initiative to enable value creation, and thus exploit open government data to its highest potential. Here we briefly describe the aim of each aspect.

- *Data Format:* Formats such as CSV and RDF are much more usable then PDF. This is because they allow easier re-use of the represented data.
- *Data Licence:* Other than allowing for reasonable privacy, security, and privilege restrictions, data has the highest value creation potential if it is not subject to any limitations on its use due to copyright, patent, trademark or other regulations. Hence, data with an open licence has the best value creation potential.
- *Data Ambiguity:* Data ambiguity is reduced when a representationally rich format (e.g. RDF) is used.
- *Data Accuracy:* The extent to which data accurately represents the respective information.
- *Data Completeness:* Data is complete when all required information is available, for the representation of the data in question.

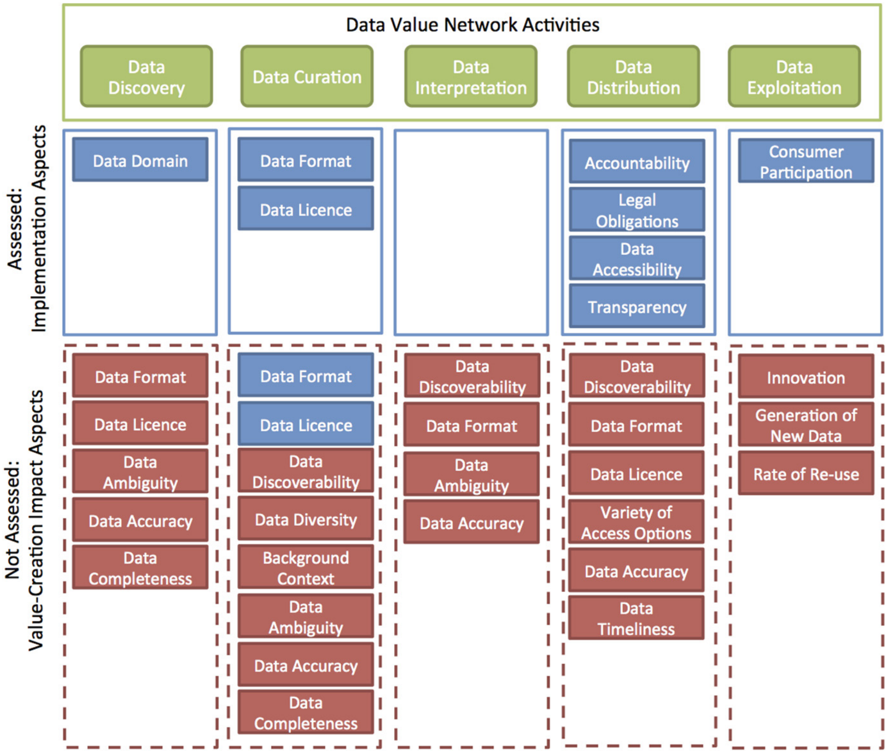


Fig. 3. Aspects assessed in existing frameworks (blue), aspects for Value Creation Assessment Framework (Red) (Color figure online)

- *Data Discoverability:* This aspect depends on the metadata annotating the data in question, and enables stakeholders to more easily find data that is relevant to their needs. Data Discoverability is also affected by the search functions provided by a government portal or catalogue.
- *Data Diversity:* In the Linking value-creation process, the use of diverse datasets has the potential of releasing new insights or unforeseen results.
- *Background Context:* The linking of datasets provides further context to the data in question, enabling stakeholders to have a deeper understanding.
- *Use of Standards:* Using agreed-upon standards throughout the life-cycle of government data encourages data re-use and integration.
- *Variety of Access Options:* Providing various access options to the available data, such as APIs and SPARQL endpoints, encourages stakeholders to create value upon the data as they are able to access the data in their preferred manner.
- *Data Timeliness:* Certain data might only be valuable if it is made openly available shortly after its creation.

- *Innovation:* Creating new products (data or otherwise) based on open government data is a direct impact of value-creation. Innovations include services and applications.
- *Generation of New Data:* The value-creation techniques in the Data Exploitation Process can result in the generation of new data, such as visualisations, that provide new interpretations or insight on the existing government data.
- *Rate of Re-use:* The participation of stakeholders in consuming the data is essential for value-creation. There is no use in having data made openly available if it is not exploited. The rate of re-use of open government data is directly indicative of the value-creation potential in the assessed initiative.

Since one of the major aims of open government initiatives is the release of social and commercial value, we deem that the proposed aspects are vital to determine the success of an initiative. Hence, these *value creation impact aspects* are used to assess the potential value that can be created through the use of the data product created as a result of each step within the Data Value Network.

7.1 Value Creation Assessment Framework in Action

In this section we implement the proposed assessment framework on two open government data initiatives, namely www.govdata.de and www.gov.mt, in order to portray its relevance and applicability in the context of value creation on open government data. Keeping in mind that this implementation is acting as a proof of concept, we restrain our metrics to assess the portal on a high level, as we consider a through and more accurate implementation to require significant more research. We therefore base the provided metrics on ground research. In Table 3 we provide a description of the metrics used, and the results of the portals[22]. We assign marks according to the assessed aspect, and where relevant we average the marks out based on the number of available datasets. For example, to assess the data format of eight datasets, if four datasets are in RDF and linked to other datasets (4 × 5 marks) and four datasets are in CSV 4 × 2 marks), then the result for the data format aspect is 3.5 marks.

Having a value-creation potential of 13.56 marks out of 20, www.govdata.de can do with some improvements, especially with regards to the use of RDF and the linking to other documents. The portal could also benefit from enabling users to both create new innovations or data through the portal itself, and also from providing some sort of documentation to both portray any innovations based on the data in question. In summary, www.govdata.de is on the right track towards the opening of governmental data, however it definitely requires more effort towards encouraging stakeholders to create value upon the published data.

On the other hand, www.gov.mt does not really excel in publishing government data. Apart from providing very few datasets, some require logging in with a government-issued e-id to download, and others are not even available

[22] As per 29th of December 2015.

Table 3. Value-creation assessment framework metrics and results

Value-creation impact aspects	Assessment metrics	Results govdata.de	Results gov.mt
Data format	5 star scheme for LOD: 1–5 marks according to format	2.71 out of 5	2.39 out of 5
Data licence	0 marks if no licence specified, 1 mark if licence has some restrictions, 2 marks if open and enabling re-use	1.85 out of 2	0 out of 2
Data ambiguity	1 mark if using semantically rich formats (e.g. RDF)	0 out of 1	0 out of 1
Data accuracy	Requires use of a gold standard[a]	-	-
Data completeness	Requires use of a gold standard[a]	-	-
Data discoverability	1 mark if metadata is available, 1 mark if portal offers search functions on the data (2 marks max)	2 out of 2	0 out of 2
Data diversity	1 mark if there is more than one dataset on a specific domain	1 out of 1	1 out of 1
Background context	1 mark if datasets are linked to other external datsets	0 out of 1	0 out of 1
Variety of access options	1 mark if more than one access option is available	1 out of 1	0 out of 1
Data timliness	1 mark if data has a timestamp, 1 mark if recently updated data is available (2 marks max)	2 out of 2	0 out of 2
Innovation	1 mark if portal provides innovations based on published data, 2 marks if different innovations are provided (e.g. services, applications) (3 marks max)	3 out of 3	0 out of 3
Generation of new data	1 mark if portal enables users to generate new data (e.g. visualisations)	0 out of 1	0 out of 1
Rate of re-use	1 mark if portal provides links and information on re-use of the published data	0 out of 1	0 out of 1
Total		13.56 out of 20	3.39 out of 20

[a] This aspect cannot be assessed on a high level as it requires the use of an algorithm that analyses each dataset in a portal and compares it to a gold standard.

(404 error given). Moreover, no search functions are provided to aid a user search within the provided datasets, such as a faceted browser. Whilst there is a statement encouraging stakeholders to innovate upon the data, no actual data licence is provided, leading room towards uncertainty.

8 Concluding Remarks

The main challenge in public value is that open data has no value in itself, yet it becomes valuable when it is used. In our information society, value creation processes have the potential of extracting the maximum value from data by building on its intelligent use. All stakeholders of value creation can participate through different roles, yet they have one common goal; that of creating a data product. Different dimensions impact the creation of such a product, namely technical, policy/legal, economic/financial, organisational, and cultural. Some of these dimensions are in turn also impacted by value creation. The use of Linked Data in creating value enhances the process, and also aids us to gradually proceed through various degrees of data products: starting with data, to information, and ultimately to knowledge. In order to truly assess the value creation process of an open government initiative, we propose an assessment framework that focuses on the potential impact achievable from a data product generated through a value creating process, and implement it on a high level on two government data portals. As future work we intend to further explore more accurate metrics that can be used to assess the suggested aspects within the framework. Step by step the vision of having open government data exploited to its full potential can be acquired.

References

1. Elements of a data value chain strategy | Digital Agenda for Europe | European Commission. https://ec.europa.eu/digital-agenda/en/news/elements-data-value-chain-strategy
2. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Gov. Inf. Q. **32**(4), 399–418 (2015). http://www.sciencedirect.com/science/article/pii/S0740624X1500091X
3. Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., Portugali, Y.: Smart cities of the future. Eur. Phys. J. Spec. Top. **214**(1), 481–518 (2012)
4. Benington, J.: From private choice to public value. In: Benington, J., Moore, M. (eds.) Public Value: Theory and Practice. Palgrave Macmillan, Basingstoke (2011)
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semant. Web Inf. Syst. **5**(3), 1–22 (2009)
6. Bogdanović-Dinić, S., Veljković, N., Stoimenov, L.: How open are public government data? An assessment of seven open data portals. In: Rodríguez-Bolívar, M.P. (ed.) Measuring E-government Efficiency. Public Administration and Information Technology, vol. 5, pp. 25–44. Springer, New York (2014)
7. Caragliu, A., Bo, C., Nijkamp, P.: Smart cities in Europe. J. Urban Technol. **18**(2), 65–82 (2011)
8. Conradie, P., Choenni, S.: Exploring process barriers to release public sector information in local government. In: Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2012, pp. 5–13. ACM, New York (2012)
9. Crié, D., Micheaux, A.: From customer data to value: what is lacking in the information chain? J. Database Mark. Customer Strategy Manag. **13**(4), 282–299 (2006)

10. Cyganiak, R., Maali, F., Peristeras, V.: Self-service linked government data with dcat and gridworks. In: Proceedings of the 6th International Conference on Semantic Systems - I-SEMANTICS 2010, p. 1. ACM, New York, September 2010
11. DiFranzo, D., Graves, A., Erickson, J.S., Ding, L., Michaelis, J., Lebo, T., Patton, E., Williams, G.T., Li, X., Zheng, J.G.: The web is my back-end: creating mashups with linked open government data. In: Wood, D. (ed.) Linking Government Data, pp. 205–219. Springer, Heidelberg (2011)
12. Eckartz, S.M., Hofman, W.J., Van Veenstra, A.F.: A decision model for data sharing. In: Janssen, M., Scholl, H.J., Wimmer, M.A., Bannister, F. (eds.) EGOV 2014. LNCS, vol. 8653, pp. 253–264. Springer, Heidelberg (2014)
13. Heath, T.: How will we interact with the web of data. IEEE Internet Comput. **12**(5), 88–91 (2008)
14. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, San Rafael (2011)
15. Janssen, K.: The influence of the PSI directive on open government data: an overview of recent developments. Gov. Inf. Q. **28**(4), 446–456 (2011)
16. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. **29**(4), 258–268 (2012)
17. Jetzek, T., Avital, M., Bjørn-Andersen, N.: Generating value from open government data. In: Baskerville, R., Chau, M. (eds.) Proceedings of the International Conference on Information Systems, ICIS 2013, Milano, Italy, December 15–18, 2013. Association for Information Systems (2013)
18. Kalampokis, E., Tambouris, E., Tarabanis, K.: A classification scheme for open government data: towards linking decentralised data. Int. J. Web Eng. Technol. **6**(3), 266–285 (2011)
19. Latif, A., Us Saeed, A., Hoefler, P., Stocker, A., Wagner, C.: The linked data value chain: a lightweight model for business engineers. In: Proceedings of International Conference on Semantic Systems, pp. 568–576 (2009)
20. Lee, C.C., Yang, J.: Knowledge value chain. J. Manag. Dev. **19**(9), 783–794 (2000)
21. Lourenço, R.P.: Open government portals assessment: a transparency for accountability perspective. In: Wimmer, M.A., Janssen, M., Scholl, H.J. (eds.) EGOV 2013. LNCS, vol. 8074, pp. 62–74. Springer, Heidelberg (2013)
22. Marchionini, G.: Exploratory search. Commun. ACM **49**(4), 41 (2006)
23. Meijer, R., Conradie, P., Choenni, S.: Reconciling contradictions of open data regarding transparency, privacy, security and trust. J. Theoret. Appl. Electron. Commer. Res. **9**, 32–44 (2014). http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-18762014000300004&nrm=iso
24. Miller, H.G., Mork, P.: From data to decisions: a value chain for big data. IT Prof. **15**(1), 57–59 (2013)
25. Montgomery, D.B., Urban, G.L.: Marketing decision-information systems: an emerging view. J. Mark. Res. **7**(2), 226–234 (1970)
26. Morgan, L., Feller, J., Finnegan, P.: Exploring value networks: theorising the creation and capture of value with open source software. EJIS **22**(5), 569–588 (2013)
27. Peppard, J., Rylander, A.: From value chain to value network. Eur. Manag. J. **24**(2–3), 128–141 (2006)
28. Wang, P., Hua, H.: A model of government information value-added exploitation based on cloud computing. In: 2011 International Conference on Business Management and Electronic Information, vol. 2, pp. 518–522. IEEE, May 2011
29. Porter, M.E.: Competitive Advantage: Creating and Sustaining Superior Performance, vol. 15. The Free Press, New York (1985)

30. Rayport, J.F., Sviokla, J.J.: Exploiting the virtual value chain. Harvard Bus. Rev. **73**, 75 (1995)
31. Reiche, K.J., Höfig, E.: Implementation of metadata quality metrics and application on public government data. In: COMPSAC Workshops, pp. 236–241 (2013)
32. Rojas, L.A.R., Bermúdez, G.M.T., Lovelle, J.M.C.: Open data and big data: a perspective from Colombia. In: Uden, L., Oshee, D.F., Ting, I.-H., Liberona, D. (eds.) KMO 2014. LNBIP, vol. 185, pp. 35–41. Springer, Heidelberg (2014)
33. Dulong de Rosnay, M., Janssen, K.: Legal and institutional challenges for opening data across public sectors: towards common policy solutions. J. Theoret. Appl. Electron. Commer. Res. **9**, 1–14 (2014)
34. Sandoval-Almazan, R., Gil-Garcia, J.R.: Towards an evaluation model for open government: a preliminary proposal. In: Janssen, M., Scholl, H.J., Wimmer, M.A., Bannister, F. (eds.) EGOV 2014. LNCS, vol. 8653, pp. 47–58. Springer, Heidelberg (2014)
35. Shadbolt, N., O'Hara, K.: Linked data in government. IEEE Internet Comput. **17**(4), 72–77 (2013)
36. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., Schraefel, M.: Linked open government data: lessons from Data.gov.uk, May 2012
37. Shah, S., Horne, A., Capella, J.: Good data won't guarantee good decisions - harvard business review. Harvard Bus. Rev. **90**(4), April 2012
38. Susha, I., Zuiderwijk, A., Janssen, M., Grönlund, Å.: Benchmarks for evaluating the progress of open data adoption: usage, limitations, and lessons learned. Soc. Sci. Comput. Rev. **33**(5), 613–630 (2015)
39. Ubaldi, B.: Open Government Data, May 2013
40. Villazón-Terrazas, B., Vilches, L., Corcho, O., Gómez-Pérez, A.: Methodological guidelines for publishing government linked data. In: Wood, D. (ed.) Linking Government Data, Chap. 2, pp. 27–49. Springer, New York (2011)
41. Zuiderwijk, A., Janssen, M.: Barriers and development directions for the publication and usage of open data: a socio-technical view. In: Gascó-Hernández, M. (ed.) Open Government. Public Administration and Information Technology, vol. 4. Springer, Heidelberg (2014)
42. Zuiderwijk, A., Janssen, M.: The negative effects of open government data - investigating the dark side of open data. In: Proceedings of the 15th Annual International Conference on Digital Government Research, dg.o 2014, pp. 147–152. ACM, New York (2014)

Collaborative Construction of an Open Official Gazette

Gisele S. Craveiro, Jose P. Alcazar, and Andres M.R. Martano(✉)

School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, Brazil
andres@inventati.org

Abstract. Given the potential use of open data and the obstacles for implementing Open Government Data (OGD) initiatives, this paper aims at describing the strategies adopted for preparing the implementation of an open Official Gazette at the municipal level. It is important to emphasize the potential value of the Official Gazette as a source of information, since it is perhaps the most detailed and comprehensive report the society can have on government daily activities. However, the data are mostly unstructured, and this fact, combined with the size of the database, makes any attempt to analyze it a non-trivial matter. Publishing the Official Gazette as OGD certainly does not address all the problems related to its use, but hopefully barriers can be overcome to allow more groups to make use of it. In this paper, three research methods are combined; a bibliographical review, documentary research, and direct observation. This paper describes the strategies and activities put into effect by a public body and an academic group in preparing the implementation of the open Official Gazette. It also analyses the outcomes of these strategies and activities by examining the tool implemented, the traffic and the reported uses of the Open Gazette. The paper concludes by reflecting on the main challenges that are raised in implementing open data initiatives at a local level in a developing country, and proposing an agenda for future research.

Keywords: Open government data · Official gazette · Coproduction

1 Introduction

The technological advances are changing the way the citizens have access to public information. Not only the commoditization of computers and other information and communication technologies, but also the access to public information is more and more disseminated in open format, allowing greater reuse driving innovation in the public sphere. This change is happening on the way we create and exchange knowledge and culture, as well as how we participate in civil life [1].

One of the main sources of public information are the official gazettes, since it is through them that the official acts are not only made public, but are also considered in force. That is, they are only applicable from the moment they are

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 111–128, 2016.
DOI: 10.1007/978-3-662-53416-8_7

published, or in cases specified on the acts themselves, in periods counted from these publication dates. For instance, in Brazil, this practice goes back to 1808 and, since then, the daily publications seek to ensure universal access of citizens to the public acts, as well as their historical record.

Naturally, this type of publication suffered the impact of the arrival of new information and communication technologies, and the official gazettes are now made available both in printed paper and in digital media. Usually, the digital version is offered in a PDF format, which has inhibited or even prevented the automatized reuse of the published information. Therefore, there is an excellent opportunity for publishing this documents in an open format and expand the current notion of universalization of access existing in public administration.

There also are great challenges to provide the availability and consumption of such information. The main ones are related to the unstructured or in some cases semi-structured nature of information, as well as the lack of standardization and controlled vocabulary on nomenclature used in public acts. Also, consultations and analyses on the daily volume of several years of publication require strategies for organizing and making the official gazette available in open format to all citizens.

This paper presents the process of developing an open 'official gazette' in Brazil. Although there is a growing interest in how to make governments more transparent, and in particular how to make government data open to a broader public, there is a lack in the scientific literature describing the effective development of such iniciatives. We aim to give contributions to the discussion about process and the system architecture from a real experience implemented in the local government. This study describe the process of opening the Official Gazette in the city of São Paulo, performed via partnership between the municipal public administration and a group of researchers. The methodology for constructing this initiative, named Diário Livre (Free[1] Gazette) is described, both in collecting the demands and expectations for the project and also in its technical implementation. Its initial impacts will be discussed, stating some benefits noticed from some of its consumers and the challenges faced during its implementation, as well as its maintenance. Finally, the paper will be concluded by presenting the future actions and final considerations.

2 Background

2.1 Official Gazette

Official gazettes are government newspapers for disseminating information, created to give publicity to measures taken by the government. They can have several names and have a wide coverage. In Brazil, the origins of the official gazettes go back to the time when the Portuguese Court was transferred to this country. It was in 1808, when the Royal Printing Press was established by

[1] "free" as in freedom.

decree in Rio de Janeiro and was granted exclusive rights to print all the official legislation and administrative measures from the government. Currently, the responsibility for publishing the Official Gazette is of the National Press, a body that has close ties with the Presidential Office of the Republic. There is a decree determining the scope of what subjects can be published in the Federal Official Gazette (DOU) [2].

In the case of the municipality of São Paulo, the Diário Oficial Municipal (DOM) (Official Gazette of the City) is under the responsibility of the Secretariat of Planning, a secretariat in the So Paulo City Hall, and published by the Official Printing House of the State of São Paulo.

In accordance with the requirements of the DOM site [3], the publication is divided into 7 sections and covers: decisions about processing, authorization and contracts, exemptions, nominations, appointments, competitions, expenses statements, responsibility and salary reports, balance sheets and more.

Thus, it should be stressed that the material supplied by DOM is very important to ensure transparency and government accountability, since it displays information about servers, public expenditure and decisions that affect the community as a whole.

2.2 Open Government Data

In a context where there are increasing talks on government innovation being made together with the society and not only for the society, the open government data is the raw material in co-creation processes between government and the civil society. The open government data (OGD) are described by eight principles [4]. These principles, together with the 5-star model [5], seek to allow new uses for the data produced. Therefore, their publication is not usually seen as an end per se, but as a means for possibly producing a positive impact on society [6].

It must also follow the principles found in Open Definition [7] "open data are data that can be freely used, reused and redistributed by any person". This involves the online publication and sharing of information in open, machine-readable formats, that can be freely and automatically reused by the society. The former will ease transparency and democratic control, citizen empowerment, innovation, improvement in government services and the discovery of new things from the analysis and mining of data [8].

There are many political benefits from open government data. These include the following: an increase in transparency, an upgrading of public services, an ability to combat corruption, an opportunity for innovation and an increase in government efficiency. This paves the way for greater involvement of the people, and the creation of new markets which can make use of the available data and generally improve decision-making, as discussed by [6,9,10]. However, although this trend has gathered momentum in the last few years and influenced governments throughout the world (by encouraging the adoption of a large number of similar initiatives), the real effects are still relatively unknown [10,11]. However, preliminary studies suggest that the effects may be negative as well as positive, contrary to the more optimistic expectations [12].

As discussed earlier, a wide range of factors including technical constraints, can mean that only special sectors in society can make use of OGD. This asymmetry of access can thus lead to greater inequality rather than reduce it [12–14]. Given this situation, it is clear that any decision to implement an information system that can provide open data, may also influence the policy making that can bring about its appropriation.

As a result, some of the measures recommended for the publication of DGA, which seek to encourage their use for positive ends, entail a greater involvement of society in ensuring that those groups that lack technical qualifications, are assisted and supported in their attempts to make use of the data. With regard to this, [10] stresses the importance of empowering those that are already carrying out this method of using DGA, as well as connecting users, developers and public managers, since it is generally found that managers are unable to visualize the best ways of employing the data. This interaction between groups is essential in deciding which data should be open and in what way it should be published. But, regrettably this kind of interaction is not the rule, and can, for example, enable many managers to publish data in a way that they believe to be open but which, in reality, does not allow a series of uses.

The three documents address the question of political feasibility and show its importance. The related recommendations involve working with the prevailing culture of the organization by employing prototype projects for the experiment carried out with open data and to demonstrate its benefits. With regard to the licences for the data, there is a general consensus that they should be open. This belief is linked to the fact that the three documents also support the idea of exploiting these data for commercial purposes – something which can be constrained by the more restrictive kinds of licences.

Two of the documents seem evidently to support a more participative opening process. This is made clear in several points. First of all, in arguing that the choice of data should be open, they state that this is the responsibility of a consultative group that is outside the control of the public administration. Following this, they suggest forming a catalogue of data which can enable the community to know what kind of data exist so that they can have a greater sense of proprietorship with regard to the prioritization of the opening.

Another key factor which is also an advantage with regard to participation, is the support provided by the use of open formats and free software, which allows a degree of collaboration in the development of the tools that are used for a greater interaction with, and appropriation of, the data made available. In addition, and also linked to participation, the three documents agree that some kind of partnership should be formed with the intermediaries to assist in the process as a whole, or in some particular phase.

Finally, with particular regard to the data consumption, the documents also discuss the idea of fostering this through events (such as hackathons or courses) or partnerships. It should thus be noted that, on the whole, there are no wide divergences between these three documents but only some differences of standpoint and that they can operate in a complementary way.

2.3 Related Work

The open government data initiatives are generally published as numerical data and are not strictly documents like those that constitute the Official Gazette. Although there are some works in the literature that tackle the question of machine readability applied to the Official Gazette [15–17], there are hardly any studies of this kind and they fail to address the question in its socio-technical totality. As a confirmation of the importance of this case study, not many studies were found about how to publish Public Gazettes as open data. One of the few studies that was found commented on an initiative concerning open data in the Philippines. This study stated that the first data to be published by the initiative consisted in digital versions of the Public Gazette for that country. However, the quality of its metadata has been criticized [18].

In another article, the authors had to examine the Brazilian Federal Gazette in order to make semantic annotations on the articles and link them to each other [17]. One of the results obtained was the ability to establish which acts had been annulled or superseded by others, although the only domain addressed was confined to the Treasury. The Brazilian Senate LexML was used as the basis for identifying the acts in the text of the articles and it seems this has greatly assisted the procedure. However, it was very hard to identify the signatures on the acts, since there is no defined vocabulary for handling the names of the people concerned. The authors did not make it clear what issues had been found while extracting the text from the PDF files or whether this process may have had an effect on the results. However, they state that the text in the PDF was organized in several columns which must have caused a great deal of difficulty in their extraction.

Finally, the last article found discussed the publication of data (through SPARQL) regarding the laws of Chile [16]. But, although the architecture chosen for supplying and documenting data has been outlined in detail, it was not clear which format was initially used for the database.

The award-winning initiative Federal Register 2.0 [15] is also very interesting, which seeks to convert printed material into machine-readable XML data. Its goal is "to make the *Federal Register* more searchable, more accessible, easier to digest, and easier to share with people and information systems". Despite being mentioned in some works as an example to be followed [19, 20], we were not able to find details on the project and implementation in scientific literature.

3 Methodology

This project was developed through a partnership between the public sector and the researchers, and used several methodologies to support its development. As well as the bibliographical review on official gazettes and their availability on the internet for seeking related experiences, a documental and experimental research was also performed.

The methodology employed in this work is inspired by the Action Research (AR) design. This choice was made because AR predicts and supports the intervention in the process, something which has occurred in this study because of the partnerships. According to [21], this strengthened approach includes a period for the establishment of the research environment and then the cyclical iteration of 5 phases: **Diagnostic**: formulation of theories about the causes of organizational problems; **Action planning**: the definition of the measures to be taken to tackle the problems based on the theoretical framework and register of projected aims; **Action**: implementation of planned actions, either personally or through third parties; **Assessment**: if they are successful, an assessment of whether the changes can really be of value and if there is a failure, if the framework and theories can be adapted; **Specific Learning**: structuring of the acquired knowledge, whether or not it has been a successful experiment.

In the domain of Information Systems, methodologies of this type have been employed by [21,22], both claiming to have obtained good results. The first study supports the view that, within the area of this research, the AR methodology is justified, especially when human organizations interact with information systems owing to the complexity of this interaction. However, neither of the studies devotes much space to the political context or the power relations involved in the implementation of these systems, which are factors that are fully discussed by [23]. In the opinion of this author, the AR should not just seek to obtain better results that are based on practice, but also act in a way that can reveal the power relations which can bring about subversion. In this way, the methodology can empower the participation of the process through a real inclusion.

Bearing in mind what has been outlined in the previous paragraph, this project has sought to employ an AR methodology. In the first meetings between the academic group and the policymakers, the initial scope of the project was set out. A team was appointed to carry it out by forming a partnership. In the course of several meetings, this took care of the planning cycles, action and reappraisals. Although there were some smaller cycles, in general terms the project took place as follows: (a) a stage to conduct a survey of the required conditions, (b) the construction of a prototype and its launching, and finally (c) a stage for gathering the impressions of the users and players involved.

It should be stressed that, owing to the inherent features of the methodology employed, these stages were not followed rigidly and there were some interactions between them. It can be said that these stages and the process as a whole, were mainly influenced by: discussions held during the periodic meetings of the team; the internal context of the municipal authority (coordination with secretaries, disputes between members of the inner group, availability of the data, etc.); capacities of the team (availability of time, technical knowledge); expectations of members of the team; contributions made during the first public event.

In addition, as a means of having a better knowledge of the context, the following results were achieved which could act as guidelines for the activities of the project and to ensure some degree of participation: an in-person questionnaire at the first event; an online questionnaire at the tools site which was during the whole process; a collection of statistics for access to the tool; interviews with the managers responsible for the database; a second public event for the launching of the tool and gathering of impressions.

For the survey of the current scenario and demands for the project, over 10 meetings were held with public managers from at least three secretariats related to the collection, organization and availability of public information gathered in the official gazette. As well as the continuous alignment of expectations with the execution of the implementation between public administration and researchers, the demand for opening and expansion of social participation in the construction of this software artifact has led to the existence of two events open to the public.

The first event was held on the dissemination of the project for collecting the demands of the citizenship, and the second public event was performed both for accountability purposes, as well as for the broad dissemination to reach the social actors who are interested in and have the means to consume and reuse the volume of data now available in open format. The information obtained in these meetings (restricted and open) were gathered through questionnaires, semi-structured interviews and observation records.

All information collected supported the design of the requirements and later project and implementation of the tool, subdivided in extraction, transformation and load modules. The technological development process first lead to standardization procedures of files made available by the public management related to eleven years of daily publications. After the initial treatment, information was organized on a base for the possibility of later creating indexes for the individual articles. On this base, some functionalities were implemented, such as search tools and different forms of publication. Due to the evolution of the negotiation process involved in obtaining data from public power, the implementation of the automatic extraction stage for daily updating of data can only be completely integrated to the system at the end of its implementation. Aiming to increase the replicability of the process, ease its adaptation and reduce costs, all software used in this project are open software.

The resulting tool, named Diário Livre (Free Gazette) was delivered as a proof of concept to the public management, and was made available from the research group infrastructure. Its official launching happened in October, 2014, and since then, the service is provided uninterruptedly, automatically collecting data from the public power and making them available through a web application.

After its launching in a public event and in the media, the number of accesses has been monitored, as well as the integration performed through e-mails or events where it is disseminated. It is important to mention that one of the main public consumers consists in the civil servants and thus, a message disseminating the tool and requesting its evaluation was sent to over 200 thousand government employees in the city of São Paulo.

An important qualification needs to be made about what has been outlined here. This study was carried out through a partnership with an administrative public body. This made it possible to conduct an in-depth analysis of the internal public mechanisms and automatically add a group to those responsible for making decisions about the project – the policymakers. On the other hand, this partnership imposed constraints on attempts to participate with other groups, either because this was the will of the policymakers involved or on account of the restrictions that were self-imposed. These factors added to the difficulty of employing the AR in the form supported by [23]. Another important point that needs to be underlined is the fact that the authors were only able to employ the methodology by intervening in the process – something anticipated and supported by the AR methodology. However, as stated earlier, there was no complete control of the procedure.

4 Development

In this section, the project development stages are presented in greater detail. First it describes the scenario found when the project was started, presenting which data was available to be worked on. It then lists which were the requirements surveyed for an availability that would improve the initial scenario. It also presents and justifies the architecture adopted for the availability of data. And finally, it describes how the architecture was implemented to meet the requirements of the project.

4.1 Scenario Found

The Official Gazette from the City of São Paulo has a curious flow. Its first stage is open, with information stored on a file in the ".txt" format. However, for its printed publication, through the Official Printing House, or even for online availability with legal value, a "PDF" file is generated. Due to its nature, this file format is closed, that is, it does not allow its information to be easily copied, handled or researched. The flow from the generation of the governmental information until its availability to the public is detailed below.

The Municipal Secretariat of Management (Secretaria Municipal de Gesto - SMG) is responsible for publishing the Official Gazette at the City Hall of São Paulo. This secretariat is responsible for hiring the Official Printing House (Imprensa Oficial - IO), the same company responsible for publishing the State Official Gazette, to publish the Municipal Official Gazette (Diário Oficial Municipal - DOM). Figure 1 represents the initial flow for publishing the DOM before the tool for publishing open data.

In order to a member of staff from any department in the city hall to be able to publish anything in the municipal Official Gazette following the guidelines from the IO itself, the text has to be written and saved as TXT. The name of the saved TXT file must contain a code, known as *retranca*, which represents two metadata: the content of the article and which public body has written it. This file is then sent to *Pubnet*, the IO system for collecting articles. Images, named *calhau*, follow a different process, being saved as PDF and not as TXT before being sent to publication [24].

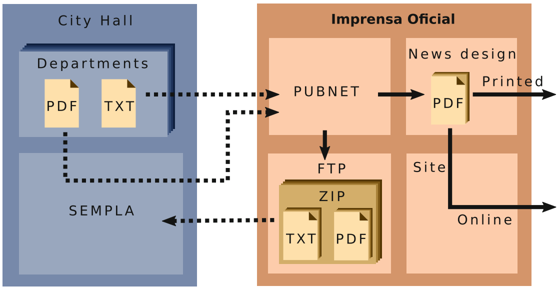


Fig. 1. The initial flow of data.

Once sent to IO, the material is saved and fully available, through FTP protocol, back at the city hall. This material, constituted by one ZIP file per day, is important, since it allows later checking if, for instance, the city hall decides there is an error made by the IO in publishing the Official Gazette. The material is also sent for editing, where it is formatted to be published. The latter is made on paper (printed newspapers) and online, making the PDFs available at the IO website.

The PDFs keep the same visual structure of the printed newspaper, very different from the structure of the initial TXTs. Such fact hinders the automatic extraction of text, but eases the comparison between the printed and digital version.

The form for textual search on the site, which indexes DOMs since 2005, allows filtering by publishing date, but does not allow filtering by publisher or published content, metadata available in the initial TXTs.

Even though the full database form the Municipal Official Gazettes is not currently available for download at the IO website, it is possible to write programs to automate the extraction of the PDFs, page by page. From them, the text from the articles can be extracted. However, such process is not precise, since the PDFs are not structured. Such procedure is performed, for instance, by companies that provide services of mining data on the extracted base.

4.2 Requirement Analysis

The requirements of the tool were raised through meetings with the managers involved in the partnership, in events with the community, from literature on open government data and from the answers to the questionnaires. The main requirements surveyed are described below, divided into three categories.

The first category consists in the requirements that are provided by the official site and were important to be kept in the new tool: **Daily updates:** the DOM is published every day and the site must be updated with new data with the same frequency; **A fast search:** even after being processed, the database included approximately 10 gigabytes of texts. An efficient tool for textual search was required in order to obtain the results and display them on the site in few seconds; **User notifications:** allow the users to be notified when there are new publications containing certain key words. Therefore, the user can be notified when, for instance, his name or any subject he is interested in are mentioned in a new article in the Official Gazette.

The second category consists in the requirements that are provided by the official site and should be improved in the new tool: **Search filters:** it was important to allow the data to be filtered through the available metadata, such as, for instance, publisher and date of publication.

And finally, the third category includes the requirements that are inexistent in the official website, but should be implemented in the new tool: **Text visualization:** many users complained about the difficulty of copying texts from the PDFs published by IO or about the loss of formatting when pasting copied texts. The site must return the articles as text in a common HTML page; **Access via API:** it was important for the site to offer some kind of API to allow easy access to the data through other applications; **Full database availability:** this allowed the use of the database as a whole, either for research or for applications that demand a greater control of data; **Unique URLs for articles:** it was important for each article to be accessible through a unique URL, so that it could be easily quoted.

4.3 Architecture

The *Diário Livre* (DL) tool was designed, implemented and launched on the basis of the requirements that were expressed. From the standpoint of a common user, it basically consists in a site where it is possible to search for words in the DOM from 2003 to the present day. Some filtering can be applied for the publisher, article content and date. The standard output format is a common page (HTML), although it is also possible to search for and visualize the articles in other formats.

The architecture adopted for the new system is represented at Fig. 2.

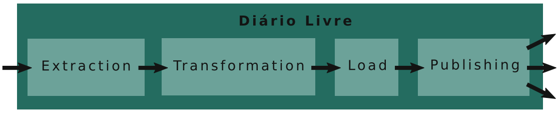


Fig. 2. Diário Livre architecture in layers and information flow.

On the extraction layer, the data are forwarded by the public management and inserted in the machine where the system is hosted in order to allow the processing of data. Due to the internal security policy at the city hall and agreement issues between it and IO, the system cannot collect data within the city hall network. Instead, the city hall staff need to forward the data to the publication system. And since, in a first moment, this transfer needed to be manually performed by the city hall staff themselves, the tool used in this stage needed to be usable by non-developers.

During transformation, the data are standardized and prepared for indexation and publication. The data provided by the city hall are compressed and include several file formats. They need to be treated and standardized, identifying the encoding of the text file contents and the metadata implicit in the name of these files, and only then, it is possible to index the data.

On the third stage, loading, the data are indexed to allow searches. Due to the size and characteristics of the database, it is necessary to have a tool that is able to index large amounts of text, perform textual searches in a timely manner and filter by metadata.

On the fourth stage, the data are finally published in several formats. These formats encompass the conventional web viewing, full download from the base and access through API, seeking to maximize the number of possible uses. The URLs used in the conventional web viewing or in the API allow individual access to the articles, and not by page, as it happens on the official IO system.

4.4 Implementation

Figure 3 represents the current flow of publication including DL. The two upper rectangles represent the flow between the city hall and IO prior to DL, and it continues to produce the official publication.

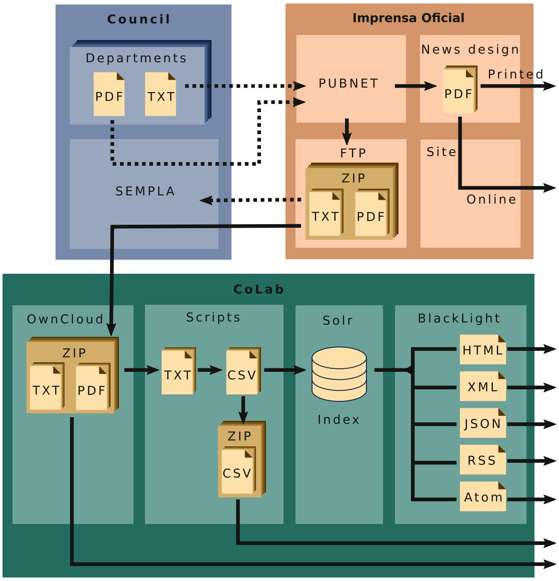


Fig. 3. The flow of data today.

The data, compressed in ZIPs, are sent by the city hall to the machine where the system is hosted. The TXT files are extracted, standardized and organized in CSVs, where each column has one metadata (date, content type and publishing body) and the whole content of the articles on the last one. These files are then indexed by the Solr [25] tool, allowing textual search and filtering by any of the metadata. Apache Solr, as presented in [26], is a NoSQL technology, that is, non-relational, which is optimized to resolve a specific class of problems for specific types of data. Solr is scalable and optimized for large volumes of data (data centralized in text) and returns results classified by relevance. The use of this tool allowed the execution of searches in a timely manner.

Once indexed, the data are published in several formats (HTML, JSON, XML, RSS and Atom) via the BlackLight [27] tool, which is used as a web interface to Solr and as an API for automated access to data.

As well as individual articles, two versions of the database are made available. One identical as the one forwarded to the city hall, containing all the files, but without treatment or standardization, and the other containing only the CSVs already treated. The first database, when decompressed, has approximately 50 GB, and the second, approximately 15 GB.

5 Results and Discussion

The difficulties in handling the official gazette in PDF has led to the offering of specialized services from several companies, which performed web scraping of data to sell solutions to clients. However, as well as being restricted to a privileged public, such solution is still prone to errors.

This way, both the citizens who wish to research basic information or even categorize and handle large amounts of data, as well as civil servants who need to monitor specific administrative acts, (such as, for instance, appointments, exonerations, waivers of public biddings, etc.) have difficulty in operating the traditional PDF version of the publication, since it is difficult to locate considerable part of the publications, and its handling is extremely impaired.

In face of the demand for democratization of the access and the need to automate searches for establishing internal control mechanisms, an experimental version of the Official Gazette in open format was developed, named Diário Livre. The prototype was developed and is working since October 2014, being used by citizens, civil society organizations and civil servants. At the time when this paper was written, the resulting site was available at: http://devcolab.each.usp.br/do

To a great extent, Diário Livre solves many of the difficulties that were previously presented. Since it receives information in open format, it makes such data available in the same manner. It is then possible to copy, extract and handle information from the Official Gazette, and now they can be read and processed by machines. Moreover, Diário Livre makes it possible to perform more qualified searches. Through it, information can be researched by Entity (secretariat of finance, hospital authorities, etc.), by Department (offices, directorates, etc.), by

Content type (dispatches, contests, bids, civil servants, etc.), and by Publishing Date. When accessing the results, it is also possible to classify them by relevance and data, automating the searches even further. It is also possible to download the entire database used in Diário Livre with information since 2003.

Considering the period of daily publications from 2003 to 2014, more than 1.4 million files in text format were obtained from the public power, containing a single article (public act) in each of them. Furthermore, approximately 61 thousand files in PDF and over 14 thousand files in .doc format were also received. In terms of volume of data, the files in text format totaled 13 GB and the files in the other non-text formats are a total of 10 GB. By analyzing the non-text files, it could be noticed that most of them contained images and layout proofs, and that the relevant content was in the text files, and for this reason, only these were considered to be part of the database. Nonetheless, the tool provides full disclosure of all files received from the public power, in all formats, through the bulk download. The tool also makes the version treated in the database used for indexation fully available through the bulk download.

The tool automatically collects the information generated on public acts from the municipal public power in São Paulo (executive, legislative and audit court), which guarantees the offer of updated information. This means that the average publication is of approximately 500 new articles per day and 127 thousand articles per year.

The published data by the Diário Livre were licensed as Creative Commons 4.0, (which is an open license). They are made available in machine-readable and non-proprietary formats. Each article is also made available in the formats stated above, and has its own URI. Thus, it can be said that according to [5], they deserve to be awarded 4 stars.

Functionalities aiming at the common user already existing in the official conventional platform, such as filter by data and textual search tool, are also offered in Diário Livre. Diário Livre also makes it simpler and more convenient to mark and copy texts, and allows filtering by categories and full access to raw files, making it easy to search the mass of data from a series of over ten years of publication, characteristics that are not observed in the conventional version. One of the functionalities offered, the automatic notification of terms, allows any person to receive, in a simple and free form, a service that is currently offered upon the payment of a subscription to the conventional official gazette.

But the characteristics that have no parallel in the official version are related to the potential reuse promoted by the automated consumption of these data. With the exporting of data through APIs, it is expected that Diário Livre eases or makes solutions for data mining viable, as well as new applications of social and economic interest in the city of São Paulo.

5.1 Dissemination And Repercussions

The system had two main dissemination moments. On the first moment, the launching, as well as a presence-based event, news was published on the websites belonging to the city hall and the university linked to the project. Each of these pieces of news was republished at least once by groups that were not directly involved in the project. During the same period, a member of the team was invited for an interview in a private radio broadcasted throughout the state. On the second moment, a dissemination e-mail was sent to all employees at the city hall. There was also an announcement about the system in the DOM itself.

A free software called Piwik was employed to monitor access to the site. Figure 4 shows the number of visitors per day based on the data collected. Two peaks can be seen that are related to the two moments aforementioned. The first moment occurred at the end of October and corresponded to the launching of the system. The second took place in the middle of December, corresponding to the sending of the dissemination e-mail and the announcement in DOM. The frequency fluctuation on the graphic is due to fewer visitors accessing the site on the weekend.

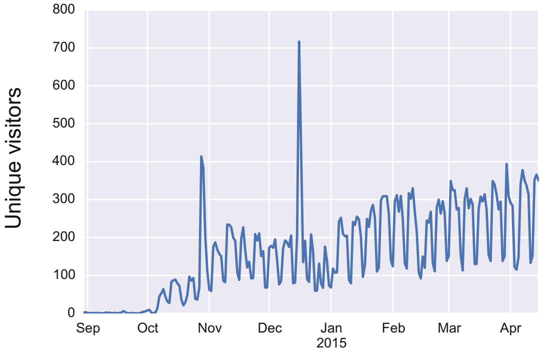


Fig. 4. Number of visitors per day.

Automatic visits, such as those made by search engines, are not included in the charts. It is interesting to note that this increase has started to take place even before the launching of the system. Considering all types of access (automated or not), on average, the system receives approximately 32 thousand accesses per day, which shows a certain robustness of the implemented solution.

The data collected show that most visitors accessed the site through searches in *Google*. On most occasions, the items sought by these visitors are names of people, companies or their official identifying numbers. Since these are items which are hardly ever found outside an official gazette, it meant that DL is among the few results that are returned from such searches.

Although PDFs are apparently subject to indexation by search engines, the official site of the publication of the DOM does not seem to be indexed by them. Bearing in mind that this site does not have a *robots.txt* (the file that can prohibit such indexation), it is believed that this did not occur, since there are no direct links to the PDFs. They can only be visualized by using the search interface on the site. In DL, the search filters are links that lead to the pages with the results, which in turn, lead to the published articles. This difference seems to have been essential to allow the indexation of the content.

5.2 Users Feedback

Regarding the online questionnaire on the DL website, 105 complete answers were obtained. From the 7 users who experienced difficulty with DL (about 7 % of the total), 4 stated that they had not been able to locate some of the features on the interface, 2 had problems with the security certificate and 1 did not specify the reason. From these 7, only 2 (2 %) stated that they had not had problems with the official IO website. In contrast, 39 users (37 %) had difficulties with the IO website and 33 of these (31 % of the total) did not have any difficulty with the DL. The problems found in the IO website included: seeking terms within the PDFs; copying the text from the PDFs; printing only part of the texts; the fact that the small font makes it difficult to read.

Finally, 65 respondents (62 %) were satisfied with the DL, as opposed to 13 (12 %) who were dissatisfied. These data support the hypothesis that DL is easier to use than the IO website.

Beside this, it is worth mentioning a few reports that provide some insight about the impact of Diário Livre on three distinct groups: civil servants, social movements and hacktivists. The first report illustrates the potential saving of resources (both financial and human) provided by using the Official Gazette in an open format. One civil servant from the legal area took approximately 30 min daily to read and select the subjects in the paper version of the Official Gazette. Using the DL API through a data-reading script, it was possible to automate the work and, in a few seconds, a report is produced and forwarded by e-mail to the interested parties.

Moving beyond the internal use of the City Hall, there are reports of usage by the organized civil society: a president of the neighborhood association in São Paulo stated he used Diário Livre to clarify his doubts and help him with the requests the neighbors in his county send to the association. He mentioned examples of complaints related to stores, since it is possible to easily seek information on operating licenses and their situations in the open version of the Official Gazette.

It is also important to report that there are examples of initiatives that are seeking to integrate applications with the data in Diário Livre. A group of local civil hackers integrated a tool on the platform for their project on budget transparency. This type of initiative shows there is a potential to stimulate the creative economy of applications from the development and improvement of Diário Livre.

6 Conclusion

This article described the joint initiative between the public power and academia to offer the Official Gazette of the city of São Paulo in a digital format that followed the principles of open data and thus foster the reuse of information by a broader range of possibilities.

The project was successful in reaching its primary objective, managing to make data available in real time and in different open formats, reaching the *4 stars of open data*. Although not yet an official portal, the initiative has become an important proof of concept that has not only demonstrated the technical viability of making an official gazette available in open format, but also offered a tangible example of the contribution of open format to a broader public.

The methodology employed for action research involved holding several meetings with the team and (together with the first event and related bibliography) made it possible to conduct a survey of the required conditions for the data publishing tool. This was then projected, implemented and launched with the same partnerships during the second event.

It was confirmed that the process fulfilled several of the recommendations made in the literature about open government data. This particularly applied to the tool produced and although it failed to satisfy all the possible uses for the data, it can be said that it met the essential requirements. Moreover, it was rated highly by the users and public administrations, as well as obtaining a growing number of accesses.

With regard to the factors in the initiatives that have motivated and attracted the key players, one can cite the search for greater transparency and compliance with related legislation. The challenge that arose largely originated from the fact that open government data initiatives at a local level are relatively recent and require a number of technological and cultural adjustments.

From the stand point of participation, an attempt was made to take this into account at every stage of the process. Despite the constraints imposed, which perhaps led to a lower level of participation than had originally been desired, it can be considered that in this respect, the final result was satisfactory. Two events were held, the first of which assisted in meeting the requirements of the final tool and the second helped to obtain a general equilibrium in the process. In addition, the online questionnaire also made a final means of contact possible and this was used by more than a hundred users.

The system had some repercussions in the media and received recognition from specialists in innovatory public management in the form of a reward at an official public event. The publication of the data also allowed some attempts to be made for their reuse both by the public administration and by people outside. These outcomes have already shown the potential of the database for obtaining non-systematized information at any place, although they have also revealed the challenges facing the question of the reuse of these data.

The implementation of the proof of concept and its consequent repercussion was a first step, and further works point towards directing the publication as linked data, with the crossing with other government database, such as, for instance, the one regarding public contracts and purchases.

Although the data contained in the Official Gazette deal with subjects of great diversity, which hinders the creation of a common vocabulary or an ontology to represent its content, it is possible to create ontologies or data connected to certain aspects dealt in the gazette. The creation of a system based on "Linked-Data" is being created to ease the recovery of information related to tenders, and the authors expect to obtain positive results both for public management and for citizenship in a broader context.

Finally, there is another question regarding the effects that these data can exert either on the public administration itself or on the diverse groups that comprise society – hackers, journalists, academics,OSCs, social movements, people with little familiarity with technology and others. To achieve this, we intend to conduct an analysis of the profile of the users with the aim of improving the presentation and attempting to design an adaptive version of our interface.

References

1. Benkler, Y.: The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press, London (2006)
2. BRASIL: Decreto n° 4.520. 16 de dezembro de 2002. http://www.planalto.gov.br/ccivil_03/decreto/2002/D4520.htm
3. Sempla: Diário Oficial da Cidade de São Paulo: manual de instruções (2010). http://www.prefeitura.sp.gov.br/cidade/secretarias/upload/chamadas/manual_de_instrucoes_do_diario_oficial_2010_1306171567.pdf
4. THE 8 PRINCIPLES OF OPEN GOVERNMENT DATA (2007). http://opengovdata.org
5. Berners-Lee, T.: Is Your Linked Open Data 5 Star? (2010). http://www.w3.org/DesignIssues/LinkedData.html
6. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. **29**(4), 258–268 (2012). http://www.tandfonline.com/doi/abs/10.1080/10580530.2012.716740. Accessed 11 Nov 2013
7. Open Knowledge. http://opendefinition.org/od
8. Lakomaa, E., Kallberg, J.: Open data as a foundation for innovation: the enabling effect of free public sector information for entrepreneurs. IEEE Access **1**, 558–563 (2013)
9. Ubaldi, B.: Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives (2013)
10. Halonen, A.: Being Open About Data: Analysis of the UK Open Data Policies and Applicability of Open Data. The Finnish Institute in London, London (2012)
11. Davies, T., Perini, F., Alonso, J.: Researching the Emerging Impacts of Open Data, vol. 20. World Wide Web Foundation, Washington, DC (2013)
12. Gurstein, M.: Open data: empowering the empowered or effective data use for everyone? First Monday, **16**(2) (2011). http://firstmonday.org/ojs/index.php/fm/article/view/3316/2764
13. Benjamin, S., Bhuvaneswari, R., Rajan, P.: Bhoomi: 'E–governance', or, an anti-politics machine necessary to globalize Bangalore? In: CASUM-m Working Paper (2007)
14. Grimmelikhuijsen, S.: A good man but a bad wizard. About the limits and future of transparency of democratic governments. Inf. Polit. Int. J. Gov. Democr. Inf. Age **17**(3/4), 293–302 (2012). http://search.ebscohost.com/login.aspx?direct=true/&db=afh/&AN=84341711/&lang=pt-br/&site=ehost-live

15. Richards, R.C. (2010). http://legalinformatics.wordpress.com/2010/07/26/federal-register-2-0-now-available
16. Cifuentes-Silva, F., Sifaqui, C., Labra-Gayo, J.E.: Towards an architecture and adoption process for linked data technologies in open government contexts: a case study for the library of congress of chile. In: Proceedings of 7th International Conference on Semantic Systems, pp. 79–86 (2011) http://doi.acm.org/10.1145/2063518.2063529
17. Brandao, S.N., Rodrigues, S.A., Silva, T., Araujo, L., Souza, J.: Open government knowledge base. In: ICDS 2013, 7th International Conference on Digital Society, pp. 13–19 (2013). http://www.thinkmind.org/index.php?view=article&articleid=icds_2013_1_30_10168
18. Davies, T.G.: Open data policies and practice: an international comparison (2014). SSRN 2492520, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2492520
19. Linders, D.: From e-government to we-government: defining a typology for citizen coproduction in the age of social media. Gov. Inf. Q. **29**(4), 446–454 (2012)
20. Dudley, L.R.: Federal Register 2.0: public participation in the Twenty-First Century. Legis. Pol'y Brief **3**, vii (2011)
21. Baskerville, R.L.: Investigating information systems with action research. Commun. AIS **2**(3es), 4 (1999)
22. Kim, P.H.: Action research approach on mobile learning design for the underserved. Educ. Technol. Res. Dev. **57**(3), 415–435 (2009)
23. Reid, C.: Advancing women's social justice agendas: a feminist action research framework. Int. J. Qual. Methods **3**(3), 1–15 (2004)
24. IMPRENSA OFICIAL: Manual de Conversão para PDF Envio de Arquivos ao Diário Oficial (2011). https://pubnet.imprensaoficial.com.br/pubnetii/manuais/ManualGeracaoPDF.pdf
25. Smiley, D., Pugh, E.: Solr 1.4 Enterprise Search Server. Packt Publishing Ltd., Birmingham (2009)
26. Grainger, T., Potter, T., Seeley, Y.: Solr in Action. Manning, Greenwich (2014)
27. DuPlain, R., Balser, D.S., Radziwill, N.M.: Build great web search applications quickly with Solr and Blacklight. In: Proceedings of SPIE, vol. 7740, pp. 774011–774011-12 (2010). http://dx.doi.org/10.1117/12.857899

A Solution to Visualize Open Urban Data for Illegally Parked Bicycles

Shusaku Egami(✉), Takahiro Kawamura, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga

Graduate School of Information Systems,
University of Electro-Communications, Tokyo, Japan
{egami.shusaku,kawamura,sei,tahara,ohsuga}@ohsuga.is.uec.ac.jp

Abstract. The illegal parking of bicycles is becoming an urban problem in Japan and other countries. We believe the data publication of such urban problems on the Web as Open Data will contribute to solving the problems. However, Open Data sets available for the illegally parked bicycles are coarse and in various formats, and then it is difficult to develop information services using the data. In this study, we thus build an ecosystem that generates Open Urban Data in Link Data format by socially collecting the data, complementing the missing data, and then visualizing the data to facilitate and raise social awareness of the problem. In our experiment, 747 pieces of information on the illegally parked bicycles in Tokyo were collected, and then we estimated the unknown number of the illegally parked bicycles with 64.3 % accuracy. Then, we published the data as the Open Data, and also a web application, which visualizes the distribution of the illegally parked bicycles on a map.

1 Introduction

The illegal parking of bicycles around railway stations is becoming an urban problem in Japan and other countries. An increase in awareness of health probmlems [1] and energy conservation [2] led to a 2.6 fold increase in bicycle ownership in Japan from 1970 to 2013. In addition to the insufficient availability of bicycle parking spaces, a lack of public knowledge of bicycle parking laws has meant that the problem of illegally parked bicycles is becoming more prevalent. The illegally parked bicycles block vehicle and foot traffic, cause road accidents, encourage theft, and disfigure streets. Furthermore, the broken windows theory [3] suggests that, by increasing urban disorders, they may lead to an increase in minor offenses.

Thus, in order to raise public awareness of this urban problem, we considered it necessary to publish data about the daily situation with respect to the illegally parked bicycles as Open Data. The Open Data is data that can be freely used, re-used and redistributed by anyone [4]. It is recommended that the Open Data should be structured according to the Resource Description Framework (RDF)[1], which is the W3C-recommended data model, and that relevant links

[1] http://www.w3.org/RDF/.

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 129–141, 2016.
DOI: 10.1007/978-3-662-53416-8_8

should be created between the data elements. This is called Linked Open Data (LOD) [5]. In recent years, the community that publishes LOD on the Web has become more active. The publication of urban problem data on the Web as LOD will allow users to develop information services that can contribute to solving urban problems. By using LOD about the illegally parked bicycles, for example, visualization of the illegally parked bicycles, suggestion of locations for optimal bicycle parking spaces, and removal of the illegally parked bicycles will be possible. However, Open Data sets available for the illegally parked bicycles are currently coarse, and it is difficult for services to utilize the data. In addition, other data concerning issues such as bicycle parking and government statistics, have been published in a variety of formats. Hence, a unification of data formats and definition of schema for data storage are important issues that need to be addressed.

In this study, we collect data about the illegally parked bicycles from Twitter and the attribute data describing attributes, which affect the number of illegally parked bicycles. In order to facilitate the reuse of these data sets which have different formats, we define schemata, unify the data formats, and publish the data on the Web as LOD. Moreover, we estimate the missing data (the number of illegally parked bicycles) using Bayesian networks. Our predictions take into consideration attributes such as time, weather, nearby bicycle parking information, and nearby Points of interest (POI). However, because there are cases that lack these attribute values, the missing attribute values are also complemented based on the semantics of the LOD. We thus use Bayesian networks to estimate the number of illegally parked bicycles for datasets, whose attributes have been complemented. These results are also incorporated to build LOD with a particular property. In addition, we develop a service that visualizes the illegally parked bicycles using the constructed LOD. This visualization service raises the awareness of the issue in local residents, and prompts users to provide more information about the illegally parked bicycles. Therefore, this study is divided into the following six phases. Phases (2) to (6) are executed repeatedly as more input data become available.

1. Designing LOD schema.
2. Collecting observation data and attribute data.
3. Building of LOD based on schema.
4. Complementing missing attribute values using LOD.
5. Using Bayesian networks to estimate the missing number of illegally parked bicycles at each location.
6. Visualization of illegally parked bicycles using LOD.

Thus, we build LOD while collecting data, and complementing the missing data. The service that visualizes the illegally parked bicycles will give local residents incentive to report infractions, as Open Data. In this manner, we aim to solve the problem of the illegally parked bicycles by building the ecosystem for Open Urban Data.

The remainder of this paper is organized as follows. In Sect. 2, an overview of sensor LOD and crowdsourcing is given. In Sect. 3, our techniques for data collection and building LOD are described. In Sect. 4, two approaches, which complement the missing attribute values, and estimate the illegally parked bicycle using Bayesian networks, are described. Also, we evaluate our results and summarize our findings. In Sect. 5, the visualization of the LOD is described. Finally, in Sect. 6, we discuss some possible directions for the future research that have arisen from our work.

2 Related Work

In most cases, LOD sets have been built based on the existing databases. However, there is little LOD available, which provides sensor data for urban problems so far. Thus, it is required to have methods for collecting new data to build Linked Open Urban Data. Data collection methods for building Open Data include crowdsourcing and gamification. A number of projects have employed these techniques. OpenStreetMap[2] is a project that creates an open map using crowdsourced data. Anyone can edit the map, and the data are published as Open Data. FixMyStreet[3] is a platform for reporting regional problems such as road conditions and illegal dumping. Crowdsourcing to collect information in FixMyStreet has meant that regional problems are able to be solved more quickly than ever before. Zook et al. [6] reported the case, where the crowdsourcing was used to link published satellite images with OpenStreetMap after the Haitian Earthquake. A map for the relief effort was created, and the data were published as Open Data. Celino et al. [7] have proposed an approach for editing and adding Linked Data using a Game with a Purpose (GWAP) and Human Computation. However, since the data concerning illegally parked bicycles are time-series data, it is difficult to collect data using these approaches. Therefore, new techniques are required and we propose a method to build Open Urban Data while complementing the missing data.

Also, there are studies about building of Linked Data for cities. Lopez et al. [8] proposed a platform, which publishes sensor data as Linked Data. The platform collects stream data from sensors, and publishes RDF in real-time using IBM InfoSphere Stream and C-SPARQL [9]. The system is used in Dublinked2[4], which is a data portal of Dublin, Ireland, and publishes information of bus routes, delay, and congestion update every 20 s. However, since embedding sensors is costly, this approach is not suitable for our study.

Furthermore, Bischof et al. [10] proposed a method for collection complementation, and republishing of data as Linked Data, as with our study. This method collects data from DBpedia [11], Urban Audit[5], United Nations Statistics Division (UNSD)[6], and U.S. Census[7], and then utilizes the similarity among such

[2] https://www.openstreetmap.org/.

[3] https://www.fixmystreet.com/.

[4] http://www.dublinked.ie/.

[5] http://ec.europa.eu/eurostat/web/cities.

[6] http://unstats.un.org/unsd/default.htm.

[7] http://www.census.gov/.

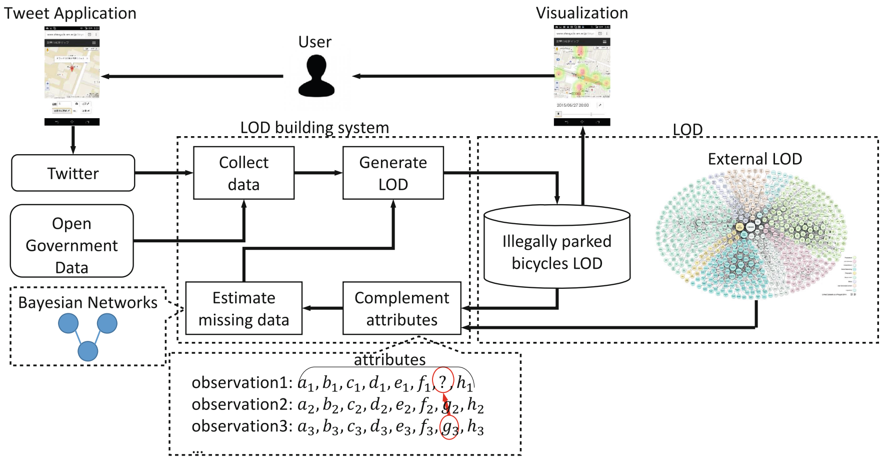


Fig. 1. Overview of this study

large Open Data sets on the Web. However, we could not find the corresponding data sets and thus apply the same approach to our study.

3 Collection of Observation Data and Building of LOD

Figure 1 provides an overview of this study. The LOD building system collects data about the illegally parked bicycles, builds LOD with a fixed schema, complements the missing attribute values, and estimates the missing data (the number of the illegally parked bicycles). This system builds sequential Open Urban Data generation, while integrating the government data and the existing LOD. The web application posts information about the illegally parked bicycles to Twitter and visualizes the distribution of them on a map.

3.1 Collection of Observation Data

We began by collecting tweets containing location information, pictures, hashtags, and the number of the illegally parked bicycles. However, obtaining the correct locations from Twitter is difficult, since mobile phones often attach incorrect location information. Mobile phones are equipped with inexpensive GPS chips, and so it is known that the accuracy will be inaccurate due to weather conditions and GPS interference area [14]. To address this problem, we developed a web application that enables users to post to Twitter after correcting their location information, and made an announcement asking public users to post tweets of illegally parked bicycles using this application. Figure 2 shows a screen shot of this application. After OAuth authentication, a form and buttons are shown. When the location button is pressed, a marker is displayed at the user's current location on a map. The marker is draggable, allowing users to correct

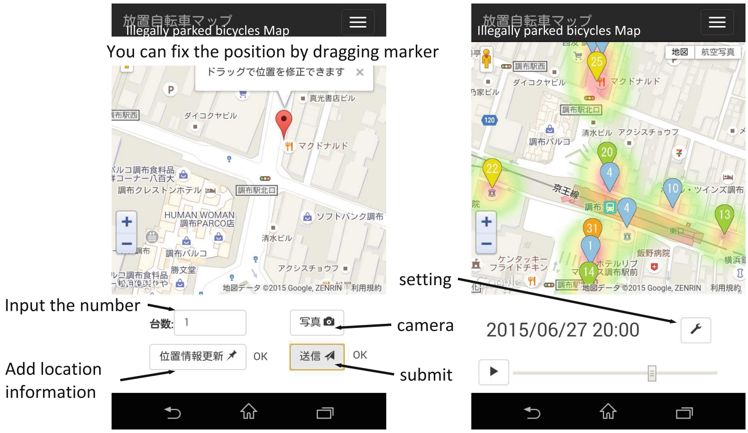


Fig. 2. Screenshots of the web application

their location information. When the users add their location information, enter the number of illegally parked bicycles, take pictures, and submit them, tweets including this information with a hashtag are posted.

The data were collected from January, 2015 until September, 2015. The LOD was built by the observation data. In order to estimate the number of the illegally parked bicycles using Bayesian networks, data for attributes considered as the causes of the illegal bicycle parking were also required. For this purpose, meteorological data were acquired from the website of the Japanese Meteorological Agency (JMA), and bicycle parking data also were acquired from the websites of municipalities.

3.2 Schema Design and Building of LOD

Illegally Parked Bicycles LOD Schema Design. When building LOD based on a well-known ontology, it becomes possible to make deductions based on that ontology. In addition, reduction in labor is possible when trying to understand the different data structures for each LOD. The observation data for the illegally parked bicycles resemble sensor data, since it is time-series data, which include location, date and time information. As a result, our schema for the illegally parked bicycles LOD was designed with reference to the Semantic Sensor Network Ontology[8]. Figure 3 shows part of the illegally parked bicycles LOD. Sensors and monitoring cameras are not used in this study, and then people observing illegally parked bicycles are considered to be virtual sensors and included as instances of the Sensor class. Since this LOD links to DBpedia Japanese[9] and GeoNames.jp[10], it is possible for people and programs to acquire

[8] http://www.w3.org/2005/Incubator/ssn/ssnx/ssn.
[9] http://ja.dbpedia.org/.
[10] http://geonames.jp/.

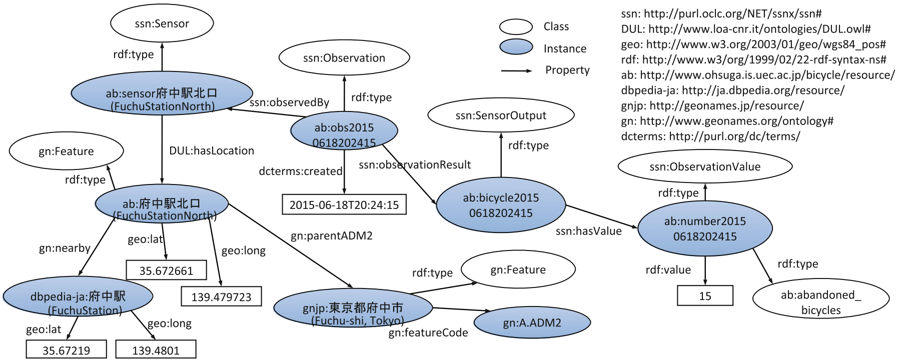

ssn:Sensor
rdf:type
ab:sensor府中駅北口 (FuchuStationNorth)
ssn:observedBy
ssn:Observation
rdf:type
ab:obs2015 0618202415
Class
Instance
Property
ssn: http://purl.oclc.org/NET/ssnx/ssn#
DUL: http://www.loa-cnr.it/ontologies/DUL.owl#
geo: http://www.w3.org/2003/01/geo/wgs84_pos#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
ab: http://www.ohsuga.is.uec.ac.jp/bicycle/resource/
dbpedia-ja: http://ja.dbpedia.org/resource/
gnjp: http://geonames.jp/resource/
gn: http://www.geonames.org/ontology#
dcterms: http://purl.org/dc/terms/
ssn:SensorOutput
rdf:type
gn:Feature
rdf:type
DUL:hasLocation
dcterms:created
ssn:observationResult
ssn:ObservationValue
rdf:type
ab:府中駅北口 (FuchuStationNorth)
2015-06-18T20:24:15
ab:bicycle2015 0618202415
ssn:hasValue
ab:number2015 0618202415
gn:nearby
geo:lat
geo:long
gn:parentADM2
35.672661
139.479723
rdf:type
gn:Feature
rdf:value
rdf:type
dbpedia-ja:府中駅 (FuchuStation)
gnjp:東京都府中市 (Fuchu-shi, Tokyo)
gn:A.ADM2
gn:featureCode
15
ab:abandoned_ bicycles
geo:lat
geo:long
35.67219
139.4801

Fig. 3. Part of the illegally parked bicycles LOD

additional information by conducting traces. DBpedia Japanese is the LOD of Japanese Wikipedia, and a hub of the LOD cloud. GeoNames.jp is the URI base of Japanese place names. Using that schema definition, it is possible to acquire longitude and latitude data as numerical values that are easy for programs to use. Moreover, it is possible to search specified time and area ranges using SPARQL Protocol and Query Language (SPARQL)[11]. In Fig. 3, the data that 15 illegally parked bicycles have been observed in front of Fuchu Station at 20:24:15 on June 18, 2015 are represented by an RDF graph.

Building of Illegally Parked Bicycles LOD. Collected data about illegally parked bicycles are converted to LOD based on the designed schema. First, the server program collects tweets containing particular hash-tags, location information, and the number of illegally parked bicycles in real-time. The number of illegally parked bicycles is extracted from the text of tweet using regular expressions.

Next, the program checks whether there is an existing observation point to a radius of less than 30 m using the latitude and the longitude of the tweet. If there is no observation point in illegally parked bicycles LOD, the point is added as a new observation point. In order to add new observation points, the nearest POI information is obtained using Google Places API and Foursquare API. A new observation point is generated based on the name of the nearest POI.

Then the address, prefecture's name, and city name are obtained using Yahoo! reverse geocoder API and then Links to GeoNames.jp are generated based on the obtained information. GeoNames.jp is a Japanese geographical database. This process is necessary for integration with other data.

After collecting tweets, the information about observation points is obtained using Web API, and then an RDF graph is added to the illegally parked bicycles LOD in real-time.

[11] http://www.w3.org/TR/sparql11-query/.

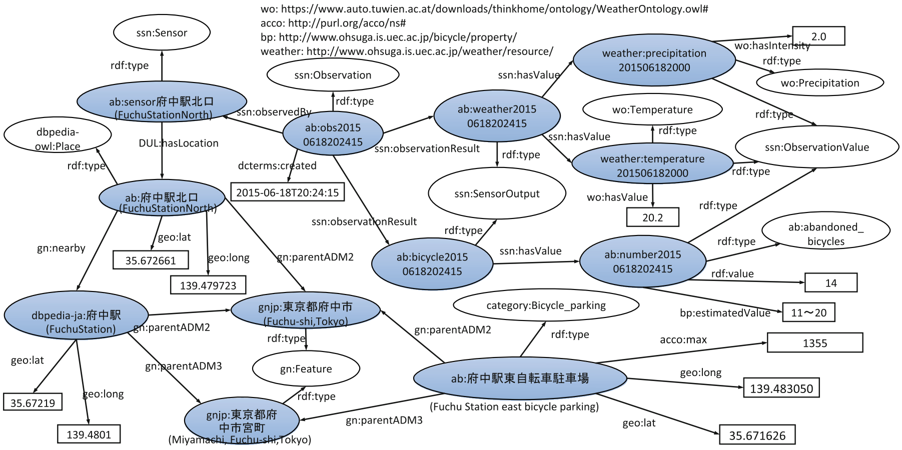


Fig. 4. Part of the integrated LOD

Building of LOD Based on for Attributes. Since the data sets acquired in Sect. 3.1 are in a number of different formats, they have poor reusability. Therefore, it was necessary to design the schema for these data and build the LOD. We designed the weather LOD schema with reference to the Weather Ontology[12]. Based on this schema, we converted the data acquired from the JMA website into LOD. We also designed a bicycle parking LOD schema and converted the data acquired from websites of municipalities into LOD. A bicycle parking resource is an instance of the "Bicycle_Parking" class, and has properties of location information, shape, and the maximum number of bicycles that can be accommodated. Furthermore, the weather LOD and the bicycle parking LOD were linked to the illegally parked bicycles LOD. Figure 4 shows part of the integrated LOD. Also, the LOD have been published via the SPARQL endpoint[13]. Thus, it is possible to link the number of illegally parked bicycles with the weather data and the data about nearby bicycle parking areas.

4 Complementing and Estimating Missing Data

Since we rely on public people to observe illegally parked bicycles, we do not have round the clock data for any place, and so there are the missing data in the illegally parked bicycles LOD. There are also the missing geographical data, since we do not have exhaustive knowledge of all the places, where bicycles might be illegally parked. Because the number of the illegally parked bicycles is influenced by several attributes, we estimate this missing data using Bayesian networks. We considered geographical features and weather to be major attributes affecting to

[12] https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/WeatherOntology.owl.

[13] http://www.ohsuga.is.uec.ac.jp/sparql.

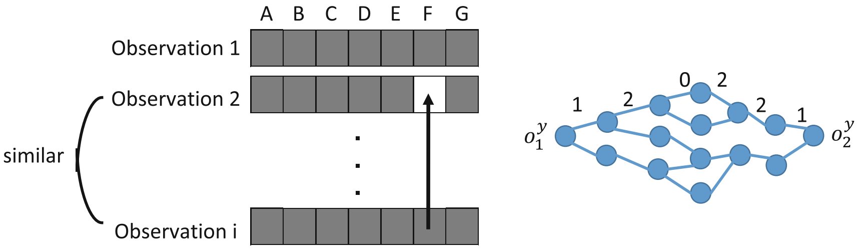


Fig. 5. Complementation of missing attribute values

the illegal parking of bicycles, and so we used data about these attributes in our estimation. However, there are also the missing attribute values. Thus, so we first complement these attribute values from similar observation data, which are found using SPARQL searches on the illegally parked bicycles LOD, the DBpedia Japanese, and the Japanese WordNet RDF [12]. Figure 5 illustrates the complementation process of the missing attribute values. After the complementation, the number of the illegally parked bicycles is estimated using Bayesian networks.

4.1 Complementing of Missing Attribute Values

In this paper, we consider seven attributes: day of week, time, precipitation (true = 1 or false = 0), the nearest POI, distance to the nearest station, distance to the nearest bicycle parking, and the maximum number of bicycles that can be accommodated in the nearest bicycle parking area. As an example, we explain our approach in the case, where the value of the maximum number of bicycles that can be accommodated is missing. Suppose the aggregates of each attribute are given by day of week $A = \{sun, mon, ..., sat\}$, time $B = \{0, 1, ..., 23\}$, precipitation $C = \{0, 1\}$, distance to the nearest station $D = \{0, 1, ...\}$, distance to the nearest bicycle parking $E = \{0, 1, ...\}$, category of POI $F = \{0, 1, ...\}$, the maximum accommodation number $G = \{0, 1, ...\}$, and the number of illegally parked bicycles $H = \{1, 2, ..., 6\}$, then the observation data are stored as an aggregate O of vectors $o \in A \times B \times ... \times H$. The number of parked bicycles is classified into six classes by the number of bicycles: 0–10, 11–20, 21–30, 31–40, 41–50, and 51–60. The missing attribute values are complemented using the corresponding attributes of the most similar data found in a search on the observation data. When the observation data including the missing attribute values is o_1, and the observation data that is a candidate from the complementary source is o_2, the similarity of o_1 and o_2 is calculated using the distance formula provided in Eq. 1.

$$Dist(o_1, o_2) = \sum_{x \in X} \frac{|sub(o_1^x, o_2^x)|}{max(x)} + \sum_{y \in Y} \frac{propCost(o_1^y, o_2^y)}{max(propCost(o_1^y, o_2^y))}, \qquad (1)$$

where X is the set of attributes with numerical values. If the value of the maximum accommodation number is missing, $X = \{B, C, D, E, H\}$. In addition, Y

is the set of attributes whose values are not numeric, so $Y = \{A, F\}$. Moreover, o_1^x denotes the value of the attribute x in o_1, $sub(o_1^x, o_2^x)$ is the value of the difference between o_1^x and o_2^x, and $max(x)$ is the maximum value of the difference in attribute x. Note that the maximum value of the difference in time is 12. The differences between the distances to the nearest station, and the differences between the distances to the nearest bicycle parking are given as numerical values in the range 0–11, where each unit corresponds to 20 m of distance. The variable $propCost(o_1^y, o_2^y)$ denotes the distance between o_1^y and o_2^y on DBpedia Japanese and Japanese WordNet RDF. Therefore, the value of $propCost(o_1^y, o_2^y)$ may be interpreted as the total cost required to travel from o_1^y to o_2^y on these LOD. The right side of Fig. 5 shows an example of this process.

Table 1. Semantics and costs of properties (owl: http://www.w3.org/2002/07/owl#, skos: http://www.w3.org/2004/02/skos/core#, dbpedia-owl: http://dbpedia.org/ontology/, wn20schema: http://www.w3.org/2006/03/wn/wn20/schema/, rdfs: http://www.w3.org/2000/01/rdf-schema#)

Semantics	Property	Cost
Synonymy	owl:sameAs	0
	owl:equivalentClass	
	skos:closeMatch	
	dbpedia-oql:wikiPageRedirects	
	wn20schema:inSynset	
Classification, Class-instance	dcterms:subject	1
	rdf:type	
is-a, part-of	skos:broader	1
	rdfs:subClassOf	
	wn20schema:hyponymOf	
	wn20schema:partMeronyOf	
Other	other	2

Furthermore, properties are classified into four semantics, and each of these semantics is allocated a cost. Table 1 shows these semantics and the corresponding costs of the properties in the classification. The variable $propCost(o_1^y, o_2^y)$ denotes the sum of the total costs of the properties, which are passed through from o_1^y to o_2^y. The maximum value of $propCost(o_1^y, o_2^y)$ is 12, which is the value obtained by multiplying the cost of the other properties by 6 based on the hypothesis of Six Degrees of Separation [13]. After searching a group (o_1, o_2) of the observation data, where $Dist(o_1, o_2)$ is minimized, o_2 is substituted for o_1.

Table 2. Statistics for observation data

Area	#observation points	Amount of observation data
Chofu-shi, Tokyo	18	601
Nerima-ku, Tokyo	5	50
Naka-ku, Yokohama-shi, Kanagawa	1	37
Fuchu-shi, Tokyo	5	19
Musashino-shi, Tokyo	4	16
Chuo-ku, Sapporo-shi, Hokkaido	9	14
Isogo-ku, Yokohama-shi, Kanagawa	2	3
Kokubunji-shi, Tokyo	2	3
Kita-ku, Sapporo-shi, Hokkaido	3	3
Inagi-shi, Tokyo	1	1

4.2 Estimating the Number of Illegally Parked Bicycles Using Bayesian Networks

We estimate the number of illegally parked bicycles when the number data is missing. The input dataset is the dataset complemented using the method described in Sect. 4.1. Bayesian networks are graphical models incorporating probabilities that represent a causal relationship between the variables of interest. Since we consider the number of illegally parked bicycles to be causally related to the day of the week, weather, and surroundings, we use Bayesian networks for our estimations. We use the Bayesian network tool, Weka[14] to estimate the unknown numbers of illegally parked bicycles. The input data is a set O, which consists of vectors with eight elements. There are 747 observation data. We used HillClimb as search algorithm, and also used Markov blanket classifier. The estimated data are added to the illegally parked bicycles LOD with a particular property. More details are described in experiments.

4.3 Evaluation and Discussion

747 pieces of observational data were collected in total from January 1 to September 20, 2015. The number of triples (records in DB) included in the illegally parked bicycles LOD was 98315. Table 2 shows statistics about the observation data. There are 237 pieces of the observation data that have the missing attribute values, and these missing attribute values have been complemented using the method discussed in Sect. 4.1. Furthermore, the number of the illegally parked bicycles for those datasets, whose attributes have been complemented from the input data, is estimated using the Bayesian networks. The attributes are day of week, time, precipitation, distance to the nearest station, distance to

[14] http://www.cs.waikato.ac.nz/ml/weka/.

Table 3. Detailed accuracy

The number of illegally parked bicycles	The amount of data	True Positive rate
0–10	612	0.930
11–20	101	0.245
21–30	26	0.060
31–40	6	0.030
41–50	0	0.000
51–	2	0.000
Weighted average		0.643

the nearest bicycle parking, category of POI, and the maximum accommodation number of bicycles that can be accommodated in the nearest bicycle parking area. As a result of a 10-fold cross-validation, the accuracy of the estimation for the unknown number of illegally parked bicycles was 64.3 %. Table 3 shows the detailed accuracy. The most of observation data were in the range of 0–10, and the accuracy of estimation was high. The accuracy of the other ranges is, however, relatively lower, since there were not sufficient amount of the observation data. Thus, it affected the overall accuracy.

Also, we selected ten observation points randomly, and then estimated the number of the illegally parked bicycles at unobserved times. As a result, the data for six observation points were correctly estimated. Since Open Data is in the early stages of the diffusion, we believe data collection and expansion are of great importance, as well as the accuracy of data.

Moreover, the accuracy of the estimated data in this study was lowered for the following reasons. The amount of the observation data was less than the amount required, and it was imbalanced. The observations used in this experiment were not equally distributed over all observation points, and the quantity of data obtained from each observation point was different. As a result, the quantity of data obtained was not sufficient to accurately estimate each conditional probability for the number of the illegally parked bicycles.

Furthermore, the accuracy may have been lowered by restricting the number of nearby POIs to a single location. In many cases, several stores and establishments are close to an observation point, and they affect to an increase in the number of illegally parked bicycles. Therefore, we can improve the accuracy of our results by allowing multiple POIs and incorporating weights for each type of POI.

5 Visualization of LOD

Data visualization enables people intuitively understand data contents. Specifically, it is possible to raise the awareness of an issue among local residents by providing a visualization of data pertaining to the urban problem. Furthermore, it is expected that we shall collect more urban data. In this section, our visualization method of the illegally parked bicycles LOD is described.

The illegally parked bicycles LOD are published on the web, and SPARQL endpoints[15] are set. Consequently, anyone can download it and use it as APIs via the SPARQL endpoint. As an example of the use of this data, we developed a web application that visualizes illegally parked bicycles. The application can display time-series changes of the distribution of the illegally parked bicycles on a map. Also, the application has a responsive design, and so it is possible to use it on various devices such as PCs, smartphones, and tablets. When the start and end times are selected, and the play button is pressed, time series changes of the distribution of the illegally parked bicycles is displayed. The right side of Fig. 2 shows a screenshot of an Android smartphone, on which the web application is displaying such an animation near Chofu Station in Tokyo using a heatmap and a marker UI. In this study, we designated the point of illegally parked bicycles according to user's tweets. However, the ranges or scales of the areas vary and thus it is difficult to display the exact ranges of the illegally parking areas. Therefore, in the current visualization, the point of the marker and the center of the heatmap are located at the center of the observation points, and the range of the heatmap is fixed in 30 m radius. Also the concentration of the heatmap is proportional to the logarithm based on the number of illegally parked bicycles. This visualization application and the tweet application in the left side of Fig. 2 are hosted on the above website, and so it is possible to see the visualized information just after tweeting. Thus, users are given the instant feedback of posting new data.

6 Conclusion

In this paper, building and visualization of Open Urban Data was described for a solution of illegally parked bicycles problem. The techniques proposed were data collection from Twitter, an illegally parked bicycles LOD based on a schema design, complementing and estimating the missing data, and then visualization of the LOD. Thus, we expect that it increases public awareness of local residents to the problem, and also encourages them to post more data.

In the future, we will increase the amount of observation data and attributes in order to improve the accuracy of the estimation. Moreover, we will visualize statistics of the illegally parked bicycles LOD, and clarify the problems caused by illegally parked bicycles in cooperation with local residents. Also, we will evaluate the growth rate of illegally parked bicycles LOD.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Numbers 24300005, 26330081, 26870201.

[15] http://www.ohsuga.is.uec.ac.jp/sparql.

References

1. Nishi, N.: The 2nd Health Japan 21: goals and challenges. J. Fed. Am. Soc. Exp. Biol. **28**(1), 632.19 (2014)
2. Ministry of Internal Affairs and Communications. Current bicycle usage and bicycle-related accident (2015). (in Japanese). http://www.soumu.go.jp/main_content/000354710.pdf. Accessed 10 Sept 2015
3. Wilson, J.Q., Kelling, G.L.: Broken windows. Atl. Monthly **249**(3), 29–38 (1982)
4. Open Knowledge Foundation. What is Open Data? http://opendatahandbook.org/guide/en/what-is-open-data/. Accessed 10 Sept 2015
5. Tim Berners-Lee (2006). Linked Data http://www.w3.org/DesignIssues/LinkedData.html. Accessed 10 Sept 2015
6. Zook, M., Graham, M., Shelton, T., Gorman, S.: Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian Earthquake. World Med. Health Policy **2**(2), 7–33 (2010)
7. Celino, I., Cerizza, D., Contessa, S., Corubolo, M., DellAglio, D., Valle, E.D., Fumeo, S., Piccinini, F.: Urbanopoly: collection and quality assesment of geospatial linked data via a human computation game. In: Proceedings of the 10th Semantic Web Challenge, November 2012
8. Lopez, V., Kotoulas, S., Sbodio, M.L., Stephenson, M., Gkoulalas-Divanis, A., Aonghusa, P.M.: QuerioCity: a linked data platform for urban information management. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part II. LNCS, vol. 7650, pp. 148–163. Springer, Heidelberg (2012)
9. Barbieri, D.F., Ceri, S.: C-SPARQL: SPARQL for continuous querying. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1061–1062 (2012)
10. Bischof, S., Martin, C., Polleres, A., Schneider, P.: Collecting, integrating, enriching and republishing open city data as linked data. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9367, pp. 57–75. Springer, Heidelberg (2015). doi:10.1007/978-3-319-25010-6_4
11. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
12. Koide, S., Takeda, H.: RDFization of Japanese electronic dictionaries and LOD. In: Proceedings of the 2nd Workshop on Linked Data in Linguistics, pp. 64–69 (2013)
13. Milgram, S.: The small world problem. Psychol. Today **2**(1), 60–67 (1967)
14. Hwang, S., Yu, D.: GPS location improvement of smartphones using built in sensors. Int. J. Smart Home **6**(3), 1–8 (2012)

An Intelligent Hot-Desking Model Based on Occupancy Sensor Data and Its Potential for Social Impact

Konstantinos Maraslis[1(✉)], Peter Cooper[1,2], Theo Tryfonas[1], and George Oikonomou[1]

[1] University of Bristol, Bristol, UK
{k.maraslis,theo.tryfonas,g.oikonomou}@bristol.ac.uk
[2] Arup, London, UK
peter.cooper@arup.com

Abstract. In this paper we develop a model that utilises occupancy sensor data in a commercial Hot-Desking environment. Hot-Desking (or 'office-hoteling') is a method of office resource management that emerged in the nineties hoping to reduce the real estate costs of workplaces, by allowing offices to be used interchangeably among employees. We show that sensor data can be used to facilitate office resources management, in our case desk allocation in a Hot-Desking environment, with results that outweigh the costs of occupancy detection. We are able to optimise desk utilisation based on quality occupancy data and also demonstrate the effectiveness of the model by comparing it to a theoretically ideal, but impractical in real life, model. We then explain how a generalisation of the model that includes input from human sensors (e.g. social media) besides the presence sensing and pre-declared personal preferences, can be used, with potential impact on wider community scale.

Keywords: Hot-desking · Optimisation

1 Introduction

Due to the increasingly digital world we live in, we tend to derive value and knowledge from as many sources of data as possible. Apart from any sociological parameters [1], there are two key factors that enabled that trend.

Firstly, it is the Internet of Things (IoT) or in other words the idea of providing internet connectivity, not only to established IT devices such as phones and computers but also to more 'traditional', seemingly non-IT devices such as air conditioners, fridges, chairs, locks etc. [2].

Secondly, it is the rise of the so-called Big Data (BD). The constantly increasing amount of connected devices is generating an exponentially growing amount of data. This, in conjunction with the more and more sophisticated methods of analysing data and extracting knowledge, is bound to change the way we live [3].

Nowadays, numerous industries collect and analyse data for multiple purposes. From organisation with environmental mindfulness that try to measure and mitigate the impact of modern life-style on environment [4] to businesses that are after the most effective methods to reduce costs and increase profits.

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 142–158, 2016.
DOI: 10.1007/978-3-662-53416-8_9

These are only some technological trends, among the many that use data-harnessing concepts, often labelled as 'Smart'. Due to their ubiquity, we can only expect similar examples to become more and more popular.

1.1 Smart Buildings

Today, the notion of Smart Cities is popular, profitable and academically thriving. The underlying notion that a proliferation of connectable infrastructure, distributed, personal sensors and big data could create efficient, enjoyable and sustainable cities has become one of the defining schemes of the current age [2, 5, 6].

The application of the same notions and fundamentals within the bounds of a building instead of the whole city (i.e. Smart Buildings) has a relatively smaller growth although it is actually an essential part of the applications in a city level [7].

The existing work in the field of Smart Buildings, research tends to be more aligned with more traditional concepts such as 'smart energy', 'smart structures', 'smart lighting' etc.

1.2 Hot-Desking

After the rise of the service sector in developed western economies, new large office workplaces were built by a new and increasingly diverse wave of consultancies and financial services. This, in conjunction with the rising rental costs in the large cities where these offices needed to be located [8] generated the issue of excessively high real estate costs for the companies.

As such, minimising the cost of large office areas became increasingly important. A popular idea emerged in the late 90s to replace territorial working systems - whereby each individual is directly associated with a specific desk - with an allocation system whereby those who attend the office on a specific day are given a free desk from a pool. The key value driver of this was that office sizes could be reduced up to 30 % [9] depending on the tendency of the business to visit clients and collaborators outside the premises. A rise in part time working [10] further improved the benefit of non-territorial desk systems.

Today, the form of hot-desking that is usually met is simply employee-led: on attendance to the workspace, an employee chooses a free desk and claims it for the day. However, such schemes have had mixed success [11]. Literature's criticisms on that can be categorised into three key aspects: (a) Ineffective management applying slow and inconsistent methods of distributing desks that can often even lead to misunderstandings about whether or not a desk if free [12], (b) Loss of working synergies which actually consists of the loss of collaboration and exchange of ideas due to not placing staff working on similar projects in close proximity, and (c) cultural and behavioural barriers which could include but not limited to the personalisation of an office (which is mostly lost in Hot-Desking environments) that could make the individual more comfortable and therefore more productive [13]. None of these parameters should look insignificant since even small variations (for example 1 % decrease) in productivity have significant impact on even the smallest scales [14].

1.3 Intelligent Hot-Desking

The rise of 'Smart' enablers provides a unique opportunity to fundamentally alter the nature of Hot-Desking by utilising increased data about the workplace, its occupants and their intentions and preferences. There is a considerable literature base that highlights that an employee's position, both in an absolute sense and in relation to other employees, has a strong impact on their behaviour and happiness in the workplace [15].

In principle, rather than a 'pegs into a slot' approach (i.e. simple linear desks assignment in a first-come-first-served basis), intelligent Hot-Desking would evaluate the best position for an employee to work based on an algorithm combining a number of weighted inputs. These inputs could include, but are not limited to:

Noise level [16] of workplaces, derived from acoustic sensors distributed across the office. There are workgroups that due to their work subject can only tolerate minimum noise (and usually produce minimum noise too) while other groups can work effectively in a noisy environment as well. The inability to effectively manage noise-sensitive and noise-making workgroups in an office can be one of the top 3 factors preventing their company from being more profitable [17].

Duration of stay derived from calendar data, or asked for at an on-arrival desk requests. Smaller 'touch down desks' can be useful for individuals staying for exceptionally short periods of time. This may further improve the floor area savings of traditional Hot-Desking.

Nature of work [18], which in the case of a very large staff group, could be derived from a system, where keywords for the type and project of work could be requested from individuals for a given day or calendar period. This element will enable workgroups of individuals with similar subjects and possibly similar goals to be formed which is proven to lead in greater productivity. Similar benefits would be realised for smaller projects too.

Environmental preferences [15] derived from various datasets, that could be generated, among others, from temperature and light sensors across the office. Many small but psychologically significant issues could be tackled this way. For example, individuals with a preference to warmer office environments could be placed further away from colder areas, whereas those with a mood that is more influenced from daylight on could be placed closer to the window.

Desk configuration, derived from asset location and management information and could include office equipment such as multiple monitors etc.

There could also be other kinds of personal preferences that could be, derived from occupant feedback (like for example level of satisfaction about previous desks given). Of course, the most appropriate combination of all the aforementioned parameters will always be heavily context-dependent.

1.4 Purpose

While it is apparent from the outset that distributing desks intelligently is indeed possible, little research exists on how optimization might look in practice, or the value it could bring to the workplace.

Within this study we will explore the potential for Intelligent Hot-Desking to result in superior working conditions (in the form of increased productivity) in comparison to a Traditional Hot-Desking Systems.

To demonstrate this we will use the distribution logic of 'work theme' within a demonstrator context of an engineering consultancy's commercial office, facilitated by primary data.

As such our objectives are as follows:

1. Establish a modelling framework, context and distribution algorithm for our scenario.
2. Observe the practical workings of an Intelligent Hot-Desking System throughout a simulated day.
3. Deduce an estimate for the improvement in productivity that Intelligent Hot-Desking Systems could bring over Traditional Hot-Desking Systems.
4. Discuss the potential barriers and enablers to implementation of Intelligent Hot-Desking Systems.
5. Explore the potential for expanding the model to inter-organisational scenarios and professional social networks.

2 Related Work

The bibliography that is related to Hot-Desking can be mostly categorised into three main research topics. Firstly, it is the topic about the impact of Hot-Desking on the health status of the employees. The second category is related to the examination of the evolution of the workspaces throughout the years. Finally, the third one is about the importance of the workplace for the employees and its impact on their productivity or even on the mind-set and their sense of team spirit. Existing studies were not found to have similarities to this one. Related work that is presented here is about different use cases that the concept of Hot-Desking is used for and although they can be seen as somewhat similar to our work (by various criteria that are explained below) they are still remote enough.

It is worth mentioning that the definition of Hot-Desking is somewhat vague and therefore some conflicts can often occur among different authors [19, 20]. However, the term 'hot desks' is most commonly used in order to express 'desks that can be used each time by a different user' and this is the definition that we will use in this work.

It is often due to this controversy on the definition, that the topic of Hot-Desking is related to Sit-and-Stand desks and therefore to employees' health. Authors of [21] for example relate hot desks with standing desks and they look into the impact that this kind of desks has on the sedentary work time in an open plan office. According to the findings, these desks did not have a great impact on the sitting working time of the employees.

In a similar fashion, the effectiveness of sit-stand workstations in terms of their ability to reduce employees' sitting time is studied in [22]. However, the findings from this 'Stand@Work randomised controlled trial pilot' differ significantly from the previous one since that study shows that these kind of desks can indeed reduce sedentary work times in the short term. It should be mentioned though that authors note the necessity of larger scale studies on more representative samples in order for the exact impact of sit-stand workstations on the health of individuals to be more accurately determined.

In [23], an attempt for results of six related pieces of research to be compared is made. All six of them are about the effect that some interventions at the workplace can have on the sitting habits of the employees during their working hours. The interventions vary from one another and in all of them, sitting time had not a significant decrease due to the aforementioned interventions.

Authors of [24] relate hot desks with sit-stand desks. These are desks that are considered 'hot' according to the definition that we adopt, with the specificity of being used in a standing position. The objective here was to examine whether the use of these desks along with awareness regarding the importance of postural variation and breaks would manage to cause better sedentary habits for the employees. The results showed that the adoption of these desks led to a better sedentary behaviour.

In a fashion similar to the previous works that were presented, authors of [25] experiment on the effect that the installation of sit-stand workstations could have on the reduction of worker's sitting times. In this study the results were very encouraging since the adoption of the sit-stand workstations was astonishing with huge impact on the sitting times ('Sitting was almost exclusively replaced by standing'). However, although the strong acceptability of these workstations, there were some design limitations that should be considered in future attempts.

All the aforementioned pieces of research belong to the first of the three categories that the bibliography can be summed up to (i.e. the impact of Hot-Desking on the health status of the employees). Below, we present characteristic representatives of the remaining two categories. Representatives of the second category (i.e. examination of the evolution of the workspaces throughout the years) followed by the ones related to the importance of the workplace and its impact on the productivity, mind-set and team spirit of the employees, which is the third category.

The evolution of the workplaces is examined at [19]. In particular, its authors investigate the rate of adoption of modern-type workplaces, including but not limited to Hot-Desking. It is interesting though that the authors define 'hot desks' as 'desks which workers have to book in advance to use' while the definition we adopted resembles more the definition that authors use for 'collective office' which according to them is 'facilities that are shared and used on an as needed basis'. Combining many sources of evidence, authors conclude that although workplaces tend to differ more and more from the typical conventional ones that were used in the past almost exclusively, this is happening with a slower rate than some claim. The findings of this study are mostly confirmed by the findings of [13]. According to the evidence of the latter, office work is increasingly differentiated from the traditional workplaces although for the majority of employees, work still corresponds to a designated place.

In [26] we meet once more the concept of Stand@Work, but this time it is not its impact to the sedentary patterns that is investigated. Instead, the objective was to qualitatively evaluate the willingness of the employees to adopt new types of workplaces, the feasibility of such a venture and the general perception of employees about the use of sit-stand workstations. The whole scheme was generally perceived as both acceptable and feasible although studies with different populations and settings need to be made.

Another study [26], considers Hot-Desking within the grand scheme regarding the societal changes in the ownership of space. The aim of this study is to sociologically analyse the emergent sociospatial structures in a Hot-Desking environment where space is used by more than one users exchangeably. The study results in two interesting findings. Firstly, the find that the perception of mobility may not be spread evenly among the employees, resulting in two different groups of them: the settlers (i.e. the most resistive to change) and the 'hot-deskers'. Secondly, according to the findings, the routine of mobility itself can generate additional work and a motion of marginalisation to the adopters.

For the third and final category of related studies, we can include [11] as well, although it belongs to the previous category too. That is because its findings are related not only to the evolution of workplaces but also to the impact that this has on the adopters, from multiple perspectives.

Apart from that study, there is also [14] which examines the impact of Hot-Desking on organisational and team identification. The study tested the level up to which the organisational and the team identity are affected by the way desks are assigned and secondly the impact that physical arrangements have on the level of engagement with the organisation. According to the results, team identity is more salient than organisational identity when a traditional desks assignment is applied whereas organisational identity is more salient when Hot-Desking is applied. The findings also denote that physical arrangements not only have significant impact on the level of engagement of the employees, but also on the on the type and focus of organisational participation.

2.1 Elements of Originality

It is obvious from the related work that is presented, that research in the field is relatively undeveloped, especially when we consider when these studies were made. But most importantly, there is a big gap in the bibliography when it comes to the research of the connection between the Hot-Desking and the productivity of the adopters. As shown already, studies on that connection are very scarce and even then it is only an indirect connection that researchers usually study. Now researchers almost always examine the implications of Hot-Desking on health, or more specifically on the sedentary habits of the adopters. Even the study on profitability, which is one of the reasons that Hot-Desking was initially developed, has been ignored due to the aforementioned approaches.

Furthermore, the nature of the existing approaches is such that no modelling is performed in order to utilise Hot-Desking in the best possible way, both in terms of organisation's profitability and employees' productivity.

What we offer is a different approach. It is a model that based on occupancy data of the employees, calculates in real time and suggests which desk has to be assigned to every employee at the time they arrive at the organisation. The model decides which desk will make the employee as productive as possible based not only on the project that they are working on but also on the projects that all the remaining employees are working on at that period of time. That way, not only employees find themselves working in the most productive environment possible without having to decide the sitting arrangements themselves (with any disadvantages that this would entail in terms of inter-employee relationships) but also the organisation will have a double benefit as it will make profit not only due to the number of desks that will not need to use anymore (desks will be less than the employees while still covering their needs), but also due to the fact that all employees will work under optimal productivity conditions.

3 Modelling

We will principally address the situation as a discrete events simulation handling the grid of desks as a grid of slots, each of which can be occupied by only one employee at a time and is either free or occupied.

On an employee's arrival, the model will decide the best desk for the individual to be assigned to. Every individual's productivity affects and get affected by individuals close to them (more on the notion of productivity are explained below). The positioning will be such that the total productivity of the grid of desks, which is the sum of the productivities of the employees of all desks, is maximised.

The simulation will run for 1 day during which many properties of the employees are logged (their desk, their individual productivity, the time they spent in their offices and the total productivity of the grid). The time is so accurately measured that is practically impossible that two incidents (each of which can be either an arrival or a departure) can happen at exactly the same time.

3.1 Individuals

For the behaviour of individuals we will be using primary observational data collected from an anonymised office of an engineering consultancy. The observed scenario has the following characteristics:

- Office grid: 144 desks (12 x 12)
- Total number of employees: 180

In practice, the time spent in the office will vary distinctly among individuals. Support staff, such as HR and Accounting are unlikely to ever leave for off-site work. Low and middle-ranking general employees are likely to attend client sites on occasion, and high-ranking staff, whose role include client relation management and thought-leadership, are likely to regularly leave, and be, out of office. These are of course generalisations and the exact spread and nature of office attendance will depend on organisational size, office size, industry and organisational culture.

By observation, supported by reasonable assumption, we can see that flow to the office in our scenario is a combination of (a) traditional morning and evening peaks for entrance and exiting to the office and between these (b) a lesser, broader flow of assorted leaving and re-entering of the office for various business engagements.

The first is relatively simple to model; the latter will require considerable simplification. Fitting normal distributions, we will estimate the probability of an individual entering the office over the course of the day and the probability of an individual who is in the office, leaving an office, as the sum of the following weighted distributions:

*Arriving: w1*A + w2*B; w1 + w2 = 1*
A: Norm (8.5, 1), w1 = 0.7
B: Norm (13, 5), w2 = 0.3
*Leaving: y1*A + y2*B; y1 + y2 = 1*
A: Norm (18, 1), y1 = 0.7
B: Norm (13, 5), y2 = 0.3

Figure 1 displays this graphically. These estimates will serve as a reasonable assumption for a generic context. However, variation will exist between different companies and different industries.

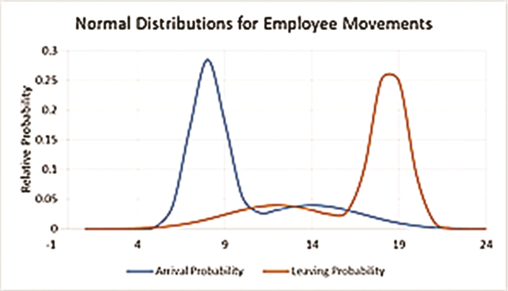


Fig. 1. Graphs of the distributions that describe arrival and departure times of employees

We will also simplify as to there being no inter-relation between arrivals and departures of individuals. In other words, if an individual arrives late to the office, they are just as likely to leave for a meeting as someone who has been there since early. We will deem this an acceptable simplification. Furthermore, employees will only be able to enter and leave the premises once. The probability distributions will in effect simulate real return visits as new individuals.

Lunch and other temporary breaks have been ignored as observation demonstrates that desks remain allocated during these periods.

In the wider group of staff from which our sample is taken, there were five work types. The distributions of these work types (i.e. the probability of random employee to belong in any of these types) in our primary data are thus: Type A: 40 %, Type B: 30 %, Type C: 15 %, Type D: 10 %, Type E: 5 %. If some day, other than the one we model, employees change workgroups, this probabilities will change too. This specific distribution may not be the reality in all samples. However, our research suggests this is not unusual for the industry from which the examined organisation is from [27].

3.2 Productivity

There is no documented method for assessing the level or quality of interaction between two individuals in the workplace and the distance between their desks. As discussed, research has simply shown that the quality, with respect to pragmatic business ends, appears to be higher when ‘the right’ individuals are in a ‘close proximity’ since the ability to speak to one another is regularly cited as a beneficial consequence of sitting near another individual [7]. Thus, we will use the behaviour of noise to model these relationships since noise levels can determine the quality of the aforementioned communication. In particular the square law will be used to describe noise impact with respect to distance from the noise generator.

As such, we depict individuals as being able to have a positive productivity influence, following square-law decay, to other relevant (i.e. of the same work type) individuals in their proximity. Of course, when individual *i* influences individual *j*, *j* also affects *i* in the same way, since this is only distance and workgroup dependent. Influences will sum linearly when several sources of influence are combined. We model irrelevant staff as having neither positive nor negative effect.

We assume desk units have a size of 2.5 m boundary from observation in our scenario, and that noise values are measured 0.5 m from the centre of the unit – again, a realistic point of seat from observation of scenario. We will then use basic square law as an estimate:

$$I_2 = \left(\frac{d_1}{d_2}\right)^2 \times I_1, I_3 = \left(\frac{d_1}{d_3}\right)^2 \times I_1 \tag{1}$$

For simplicity we ignore diagonal inaccuracies. Value of n will start at 25, to produce the values that are demonstrated below:

1st Row Proximity: 1
2nd Row Proximity: 0.25
3rd Row Onwards: (neglected for simplicity)

This is also depicted in Fig. 2 which uses the square law formulas (1). What this practically means is that every employee has a zero productivity when arriving the premises and after the algorithm has assigned a desk to that individual, every employee's productivity becomes:

$$Prod(emp) = 1 \times n_1 + 0.25 \times n_2 \tag{2}$$

Where, n_1 is the number of employees of the same workgroup that occupy desks (out of the 8 in total) neighbouring to the employee whose productivity we measure (i.e. first row neighbours) and n_2 is the number of employees of the same workgroup that occupy desks (out of the 16 in total) that are next to the neighbours of the employee whose productivity we measure (i.e. second row neighbours).

It can be easily observed that if an employee is surrounded in the first row on all sides by other employees of the same workgroup, a value of 8 (8 x 1) is achieved. If the same applies for the second row then a value of 12 is achieved (8 x 1 + 16 x 0.25), which is the maximum achievable productivity for any individual. Therefore, there is obviously a synergy: when there are two individuals, they improve each other's working environment, so the total 'quality of environment' increases from 0 (with one person) to 2 (with both).

According to formula (2), when an employee is not surrounded by an employee of the same workgroup, then the productivity of this particular employee is equal to zero.

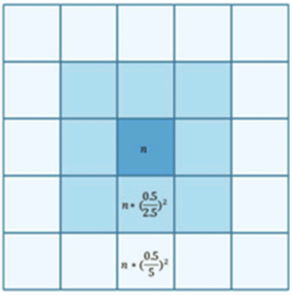


Fig. 2. Representation of the effect of one employee to the productivity of their neighbours

That does not mean, that this employee is not contributing at all, but what this model computes is the best way to allocate employees to desks and for that reason, even assuming that every employee's initial productivity (i.e. when there is no employee of the same workgroup around) is equal to $prod_{init}$, if it is the same for everyone (and it cannot be assumed that the individuals are not equally productive when working alone without any measurements to support that), then that would offer nothing to the model and the optimal solutions would be exactly the same. That would happen because when comparing between two possible allocations A_1 and A_2 of h employees where

$$TotalProductivity(A_1) = P_1,\ TotalProductivity(A_2) = P_2 \tag{3}$$

Then, with the addition of $prod_{init}$, we would have

$$TotalProductivityNew(A_1) = P_1 + h \times prod_{init} \tag{4}$$

$$TotalProductivityNew(A_2) = P_2 + h \times prod_{init} \tag{5}$$

Therefore, it is obvious that the result of the comparison between *TotalProductivity*(A_1) and *TotalProductivity*(A_2) would be always the same as the result of the comparison between *TotalProductivityNew*(A_1) and *TotalProductivityNew*(A_2).

3.3 Intelligent Hot-Desking Distribution Process

Possible methods by which we could evaluate the distribution of the desks in this system include:

- *On-arrival, Current-State Individual Optimisation*: In a system where no pre-advice is given as to who will be in and who shall not, desks are allocated aiming to maximise the productivity of the arriving individual based on information for the exact moment they enter, hoping conditions stay favourable.
- *On-Arrival, Current-State Group Optimisation*: In a system where no pre-advice is given as to who will be in and who shall not, desks are allocated aiming to maximise the total productivity of all currently in the office, based on information for the exact moment they enter, hoping conditions stay favourable.

- *Full-Term, Group Optimisation*: In a system where pre-advice is given as to who will and will not be in (including duration of stay), desks are allocated aiming to maximise the total productivity of all individuals intending to arrive that day.

It is clear that the more advanced the system, the more ideal the seating locations and the higher the productivity overall. For purposes of computational simplicity, and to avoid reviewing a distribution process with significant cultural barriers to implementation, we will use the second method in this instance.

By observation it can be considered that systems 1 and 2 will struggle with early arrivals as many permutations are identical – yet their decision will strongly influence the rest of the day. As such a tie-breaker logic is required. After experimentation of several tie-breaker systems, the most effective was chosen. The first (out of the ones that are present at the premises; not including the ones that have left) representative of every workgroup that arrives will be sent as close to a predefined extremity of the office that has been preassigned to that workgroup as possible. These will be the four corners (workgroup A at top left, B at top right, C at bottom left and D at bottom right) and the centre of the grid for workgroup E. In effect, the distribution has a disposition to form colonies with enough space to expand before starting interfering with each other, plan that will lead to high total productivity.

3.4 Variations of the Model

The model under testing has actually four versions which can be perceived as four different models. All the aforementioned characteristics are common across all models. Their differences are the following:

- *Model 1:* When an employee arrives, the algorithm assigns an empty desk to them. If there is no free desk, the employee leaves the premises and does not return the same day. When the employees leave the premises, either because it is time for them to leave or because there is no free desk, they do not return the same day.
- *Model 2*: When an employee arrives, the algorithm assigns an empty desk to them. If there is no free desk, the employee goes at the end of a First-In-First-Out queue. The employee leaves the queue if it is time to leave the premises or if there is a free desk for them (whichever comes first). When the employees leave the premises, either because it is time for them to leave or because there is no free desk (or both), they do not return the same day.
- *Model 3*: When an employee arrives or when an employee departs, all the employees (apart from the one that is leaving, in the case of departure) are reassigned (possibly different) desks of the grid, so that the maximum possible productivity can be achieved with the given employees at that time. When an employee arrives and there are no free desks, the employee leaves the premises. When the employees leave the premises, either because it is time for them to leave or because there is no free desk, they do not return the same day.
- *Model 4*: When an employee arrives or when an employee departs, all the employees (apart from the one that is leaving, in the case of departure) are reassigned (possibly different) desks of the grid, so that the maximum possible productivity can be

achieved with the given employees at that time. When an employee arrives and there are no free desks, the employee goes at the end of a First-In-First-Out queue. The employee leaves the queue if it is time to leave the premises or if there is a free desk for them (whichever comes first). When the employees leave the premises, either because it is time for them to leave or because there is no free desk (or both), they do not return the same day.

It is worth clarifying that an employee can leave the premises while waiting in the queue, for the same reasons that they could leave while being in a desk (i.e. external business commitments etc.)

Model 1 has been actually tested at [28] when it was compared to the following three variations:

- Individuals come in and are allocated a desk randomly among the free desks, with no logic applied. If there is no free desk, they leave the premises and do not return the same day.
- Individuals come in and are given a desk in a 'closest desk free' (to the top left of the office) system. Essentially, this is the linear, 'pegs into a slot' distribution that has already been discussed. If there is no free desk, they leave the premises and do not return the same day.
- For means of understanding its influence, we will simulate a distribution that simply has the 'extremities' tie-breaker logic only, and aims to throw individuals as close to the predefined extremities, and does none of the evaluation in the intelligent system. If there is no free desk, they leave the premises and do not return the same day.

As a result of that comparison, Model 1 was found to be the best (i.e. leads to a distribution of employees with higher total productivity than the total productivity of the distribution that the remaining three variations lead to).

The aim of this work is to take that previous study one step further and compare Model 1 with variations like Model 2, Model 3 and Model 4. Although it is obvious that Model 3 and Model 4 are not applicable in real life, they are still useful for comparison because they represent the ideal models. That is because these two models solve an inevitable problem that Model 1 and Model 2 have. Although Model 1 encourages the creation of colonies by employees from workgroups A, B, C, D and E (which is the best way to result in a high total productivity since individuals increase their productivity when they are close to other individuals of the same workgroup), inevitably there will be times where a colony will have a free desk in it, due to a departed employee of that colony, which will be occupied by an employee of another workgroup who cannot be placed closer to their own workgroup because there are not any free desks close to that group. That will create desk grids with individuals that are not placed in the most optimised way. However, this is inevitable unless all employees are rearranged frequently during the day, which is impractical and inapplicable in real life. However, it is useful to check how much better the results of Model 3 and Model 4 are when compared to Model 1 and Model 2 respectively, because if the difference is small that would mean that Model 1 and Model 2 are actually very close to the absolute optimal and therefore work great.

4 Results

In this section the results of Model 1, as described before, will be demonstrated, analysed and compared to Model 2, Model 3 and Model 4. Figure 3 depicts the impact of all models on the total productivity of the organisation throughout the whole day.

The equivalence of the aforementioned models to the ones on Fig. 3 is: Model 1 = Hotdesk, Model 2 = Queue, Model 3 = New and Model 4 = NewQueue. Judging by this figure, we can tell that the addition of queues not only has very small impact on the productivity, but also that slight impact is not always positive (it is not easily visible in this size of the figure but it is positive sometimes) but it can also be negative. That may not always be the case with queues, but even in this case it should not be seen as an unorthodox fact. The reasoning behind that phenomenon can be explained with the following example. Since the employees are less than the desks, there can be times where all desks are occupied and employees keep arriving. In the scenario that includes queues, if employee e_1 arrives and there are no free desks, e_1 will go last in the queue. If employee e_2 arrives later and there are still no free desks, e_2 will go last in the queue, behind e_1 (providing that e_1 has not left the queue because it was time to leave). By the time there is a free desk for e_2, it can be the case that e_2 has already left while some other employees, like e_1 for example, may have found a desk by then. Therefore, due to the queues, employee e_1 was advantaged compared to e_2. However, if there were no queues, there would be higher chances for e_2 to find a desk on arrival because if some other employee, like e_1, had arrived before e_2 and had not found a free desk, they would have left, instead of waiting of waiting in a queue in front of e_2. Thus, in the case of queues, e_2 would be disadvantaged compared to e_1 even if e_2 had more to offer than e_1 to the total productivity. This example demonstrates situations that can occur and lead to Model 2 resulting in less productivity than Model 1 (and Model 4 less than Model 3, respectively) for some periods of time. To sum up, queues maintain the first-come-first-served logic of the desks assignment whereas absence of queues can break that rule (like in the example where e_2 could have found a desk before e_1, if e_1 had departed just after their arrival) which can sometimes be beneficial for the total productivity.

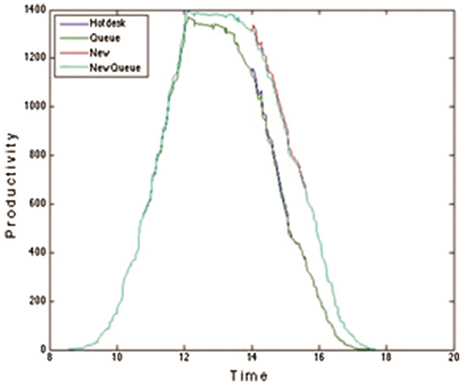


Fig. 3. Comparison of all 4 models with respect to the productivity they result in

However, the most important finding that comes out of this figure is the fact that Models 3 and 4 do not produce significantly better total productivity than Models 1 and 3, respectively, throughout the biggest part of the day. In other words, the, not applicable in real life, Models 3 and 4 that produce the best possible total productivity, seem to perform only slightly better than Models 1 and 3, respectively. The only periods of time, that Models 3 and 4 outperform Models 1 and 2 significantly is towards the end of the day when not many employees are still at their desks and if they have been arranged according to Models 1 or 2 then they will most probably be disorderly spread. And still, this difference is significant more in percentage terms and less in absolute numbers That is a huge success for Models 1 and 2 and a very good indicator that there is not much room for improvement of the algorithm, providing that the fundamental assumptions of the model remain the same. A possible and simple way to make Model 1 (resp. Model 2) almost equivalent to Model 3 (resp. Model 4) is to rearrange all employees only once (which is viable) in the afternoon, when the impact of the many departures is already apparent. After that time, although Models 3 and 4 will continue to perform better than 1 and 2, the difference will be even smaller. Figure 4 actually demonstrates that idea in practice for Model 1 (‘Hotdesk’) compared to Model 3 (‘New’). The reassignment occurs at 3 pm and its result is demonstrated on Fig. 5.

In order for the difference between Model 1 and Model 3 to be seen in practice, snapshots from the distribution of employees among the desks is provided at 3 pm, when a significant amount of employees has already departed and since there are not many that are still to come, most of the workgroups are not optimally spread across the desks, in case of Model 1, but are still optimally spread in case of Model 3. This is not a contradiction to the previous explanation of Fig. 5 because it is expected that the snapshot at 3 pm of the modified version of Model 1 (with one rearrangement at 3 pm) will be the same as the snapshot of Model 3, at the same time (3 pm).

Converting the gain in productivity into gain in profitability is not always straightforward. One of the reasons is that the gain in productivity will lead to gain in working time which is not always sure if it will invested on productivity again and in what percentage. Making very austere assumptions about the percentage of the saved working

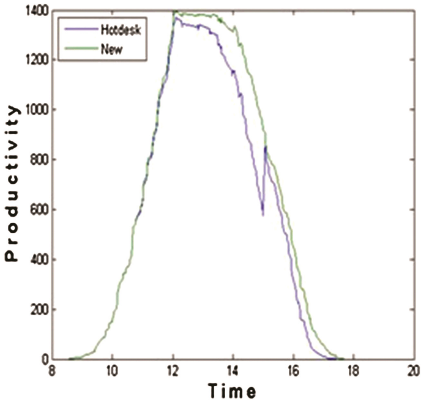
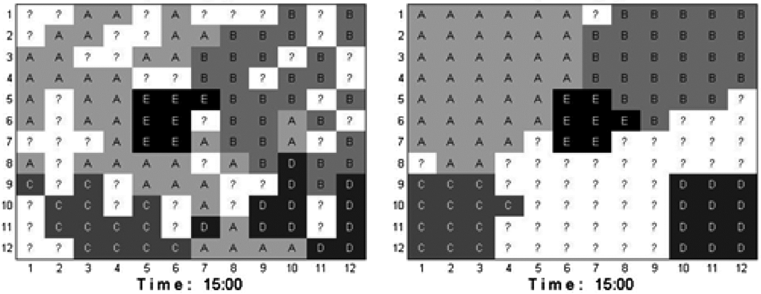


Fig. 4. Comparison of Model 1 with a rearrangement at 3 pm (‘Hotdesk’) to Model 3 (‘New’)



	1	2	3	4	5	6	7	8	9	10	11	12
1	?	?	A	A	?	A	?	?	?	B	?	B
2	?	A	A	?	?	A	A	B	B	B	?	B
3	A	A	?	?	A	A	B	B	B	?	B	?
4	A	A	A	A	?	?	B	B	?	B	B	?
5	A	?	A	A	E	E	E	B	B	B	?	B
6	A	?	A	A	E	E	?	B	B	A	B	?
7	?	?	?	A	E	E	A	B	B	A	?	B
8	A	?	A	A	A	A	?	A	B	D	B	B
9	C	?	C	?	A	A	A	?	?	D	B	D
10	?	C	C	?	C	?	A	?	D	D	?	D
11	?	C	C	C	C	?	D	A	D	D	?	D
12	?	?	C	C	C	C	A	A	A	A	D	D

Time: 15:00

	1	2	3	4	5	6	7	8	9	10	11	12
1	A	A	A	A	A	A	?	B	B	B	B	B
2	A	A	A	A	A	A	B	B	B	B	B	B
3	A	A	A	A	A	A	B	B	B	B	B	B
4	A	A	A	A	A	A	B	B	B	B	B	B
5	A	A	A	A	A	E	E	B	B	B	B	?
6	A	A	A	A	A	E	E	E	B	?	?	?
7	A	A	A	A	?	E	E	?	?	?	?	?
8	?	A	A	?	?	?	?	?	?	?	?	?
9	C	C	C	?	?	?	?	?	?	D	D	D
10	C	C	C	C	?	?	?	?	?	D	D	D
11	C	C	C	?	?	?	?	?	?	D	D	D
12	C	C	C	?	?	?	?	?	?	D	D	D

Time: 15:00

Fig. 5. Snapshots of workgroups allocation for Model 1 (left) and Model 3 (right) at 3 pm (? = Free)

Table 1. Correspondence of productivity increase (%) to actual annual profit

Percentage productivity increase	Annual value	Investment repayment time (Years)
0.1 %	£ 15 502	0.97
0.2 %	£ 31 004	0.48
0.3 %	£ 46 505	0.32
0.4 %	£ 62 007	0.24
0.5 %	£ 77 509	0.19
0.6 %	£ 93 011	0.16
0.7 %	£ 108 512	0.14
0.8 %	£ 124 014	0.12
0.9 %	£ 139 516	0.11
1.0 %	£ 155 018	0.10
2.0 %	£ 310 036	0.05
3.0 %	£ 465 053	0.03
4.0 %	£ 620 071	0.02
5.0 %	£ 775 089	0.02

time that will be reinvested in productivity (0.1 %–5 %) and based on Table 1 [29], we calculate the years that will need in order for the investment of installing the system to run the aforementioned models to be fully repaid. It is worth mentioning that the cost of such an investment is considered to be in the neighbourhood of £15 000 [30].

5 Conclusions and Future Work

Out of the three methodologies that were described earlier (i.e. (a) On-arrival, current-state individual optimisation, (b) On-arrival, current-state group optimisation and (c) Full-term, group optimisation) we modelled the second one. That is because it is more sophisticated than the first methodology and there are only specific applications where this could potentially be preferred. The third methodology, would require even more data and forecasting on the arrival and departure times which means that there would be the danger of resulting in big inaccuracies. Furthermore, an adjustment period is required before such a model can be trusted. Using the second methodology we managed to provide a realistic and productivity-oriented way of assigning desks to individuals at a workplace. Additionally, not only did we confirm that this method can outperform other common ways of desk assignment, but we demonstrated that its effectiveness is comparable with a model that was designed to result in the optimal outcome. Finally, the profit implications for the corresponding organisation were analysed and the adoption of the model was found to be an easily repayable investment.

However, we aspire to use this modelling for greater social impact that transcends organisational boundaries. At the heart of our model is the assumption that sensing data and personal preferences can feed into an intelligent platform that will bring together the most suitable co-workers under their preferred working conditions. But there is no constraint to assume that these persons must be working within the same organisation. In fact, if we apply this model in facilitating the desk allocation in the scenario of a business incubator, it could bring together complementary skills and expertise as well as personality types. To that effect we intend to develop the model further to include inputs from human sensors (e.g. social media updates), besides the ‘hard’ sensing data which may include e.g. presence and location, as well as predefined personal preferences and maybe calendar entries. We have planned an amendment that will be able to foster meaningful clustering in an incubator setting and hopefully facilitate co-working between entrepreneurs with compatible ideas and complementary skills.

References

1. Armour, S.: Generation Y: they’ve arrived at work with a new attitude. USA Today **6**, 2005 (2005)
2. Townsend, A.M.: Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia. WW Norton & Company, New York (2013)
3. Ricci, D.: Big data management. In: Barcelona SMart City World Congress 2014, Barcelona (2014)
4. IBM-A Smarter Planet-Smarter Cities-Ideas-New Zealand, 11 April 2015. http://www.ibm.com/smarterplanet/nz/en/smarter_cities/ideas
5. Webb, M., Finighan, R., Buscher, V., Doody, L., Cosgrave, E., Giles, S., et al.: Information marketplaces, the new economics of cities. The Climate Group, Arup, Accenture, Horizon (2011)
6. Glaeser, E.: Triumph of the City: How Our Greatest Invention Makes US Richer, Smarter. Healthier and Happier. Pan macmillan, london, Greener (2011)
7. Cole, R.J., Bild, A., Oliver, A.: The changing context of knowledge-based work: consequences for comfort, satisfaction and productivity. Intell. Buildings Int. **4**, 182–196 (2012)
8. Jones, C., Orr, A.: Spatial economic change and long-term urban office rental trends. Reg. Stud. **38**, 281–292 (2004)
9. Harris, D.: Turning office desks into hot property. In: The Sunday Times, (ed.) (1992)
10. Stuart, C.: Change in the workplace what motivates people at work? (2014)
11. Höpfl, H., Hirst, A.: Settlers, vagrants and mutual indifference: unintended consequences of hot-desking. J. Organ. Change Manage. **24**, 767–788 (2011)
12. Halford, S.: Towards a sociology of organizational space (2004)

13. Felstead, A., Jewson, N., Walters, S.: The changing place of work. In: ESRC Future of Work Programme, Working Paper No, vol. 28 (2003)
14. Millward, L.J., Haslam, S.A., Postmes, T.: Putting employees in their place: the impact of hot desking on organizational and team identification. Organ. Sci. **18**, 547–559 (2007)
15. Westerman, J.W., Yamamura, J.H.: Generational preferences for work environment fit: effects on employee outcomes. Career Dev. Int. **12**, 150–161 (2007)
16. Leather, P., Beale, D., Sullivan, L.: Noise, psychosocial stress and their interaction in the workplace. J. Environ. Psychol. **23**, 213–222 (2003)
17. Anonymized: Issues affecting the office workplace in banking contexts by a director at major UK Banking Firm, ed (2012)
18. Sydow, J., Lindkvist, L., DeFillippi, R.: Project-based organizations, embeddedness and repositories of knowledge: Editorial. Organ. Stud. Berlin Eur. Group Organ. Stud. **25**, 1475 (2004)
19. Felstead, A.: Rapid change or slow evolution? changing places of work and their consequences in the UK. J. Transp. Geogr. **21**, 31–38 (2012)
20. Mirchandani, K.: "The best of both worlds" and "cutting my own throat: contradictory images of home-based work. Qual. Sociol. **23**, 159–182 (2000)
21. Gilson, N.D., Suppini, A., Ryde, G.C., Brown, H.E., Brown, W.J.: Does the use of standing 'hot' desks change sedentary work time in an open plan office? Prev. Med. **54**, 65–67 (2012)
22. Chau, J.Y., Daley, M., Dunn, S., Srinivasan, A., Do, A., Bauman, A.E., et al.: The effectiveness of sit-stand workstations for changing office workers' sitting time: results from the Stand@ Work randomized controlled trial pilot. Int. J. Behav. Nutr. Phys. Act **11**, 127 (2014)
23. Chau, J.Y., van der Ploeg, H.P., Van Uffelen, J.G., Wong, J., Riphagen, I., Healy, G.N., et al.: Are workplace interventions to reduce sitting effective? a systematic review. Prev. Med. **51**, 352–356 (2010)
24. Straker, L., Abbott, R.A., Heiden, M., Mathiassen, S.E., Toomingas, A.: Sit–stand desks in call centres: associations of use and ergonomics awareness with sedentary behavior. Appl. Ergonomics **44**, 517–522 (2013)
25. Alkhajah, T.A., Reeves, M.M., Eakin, E.G., Winkler, E.A., Owen, N., Healy, G.N.: Sit–Stand workstations: a pilot intervention to reduce office sitting time. Am. J. Prev. Med. **43**, 298–303 (2012)
26. Chau, J.Y., Daley, M., Srinivasan, A., Dunn, S., Bauman, A.E., van der Ploeg, H.P.: Desk-based workers' perspectives on using sit-stand workstations: a qualitative analysis of the Stand@ Work study. BMC Pub. Health **14**, 752 (2014)
27. Wiltch, D.: Arup Associate Director, ed. (2015)
28. Cooper, P., Maraslis, K., Tryfonas, T., Oikonomou, G.: An intelligent hot-desking model harnessing the power of occupancy sensing. In: Big Data for Facilities Management in the AEC sector [Under Review] (2015)
29. Suzuki, L., Cooper, P., Tryfonas, T., Oikonomou, G.: Hidden presence: sensing occupancy and extracting value from occupancy data. In: Marcus, A. (ed.) DUXU 2015. LNCS, vol. 9188, pp. 412–424. Springer, Heidelberg (2015)
30. Iraki, Y.E.: MySeat Occupancy Solutions (2015)

Characterization of Behavioral Patterns Exploiting Description of Geographical Areas

Zolzaya Dashdorj[1,2,3,4(✉)] and Stanislav Sobolevsky[5,6]

[1] University of Trento, Via Sommarive, 9, Povo, TN, Italy
[2] SKIL LAB - Telecom Italia, Trento, Italy
[3] DKM - Fondazione Bruno Kessler, Trento, Italy
[4] SICT - Mongolian University of Science and Technology,
Bayanzurkh District 22th, Khoroo, UB, Mongolia
dashdorj@disi.unitn.it
[5] New York University, 1 MetroTech Center, Brooklyn, NY, USA
[6] Massachusetts Institute of Technology, MIT,
77 Massachusetts Avenue, Cambridge, MA, USA
sobolevsky@nyu.edu

Abstract. The enormous amount of recently available mobile phone data is providing unprecedented direct measurements of human behavior. Early recognition and prediction of behavioral patterns are of great importance in many societal applications like urban planning, transportation optimization, and health-care. Understanding the relationships between human behaviors and location's context is an emerging interest for understanding human-environmental dynamics. Growing availability of Web 2.0, i.e. the increasing amount of websites with mainly user created content and social platforms opens up an opportunity to study such location's contexts. This paper investigates relationships existing between human behavior and location context, by analyzing log mobile phone data records. First an advanced approach to categorize areas in a city based on the presence and distribution of categories of human activity (e.g., eating, working, and shopping) found across the areas, is proposed. The proposed classification is then evaluated through its comparison with the patterns of temporal variation of mobile phone activity and applying machine learning techniques to predict a timeline type of communication activity in a given location based on the knowledge of the obtained category vs. land-use type of the locations areas. The proposed classification turns out to be more consistent with the temporal variation of human communication activity, being a better predictor for those compared to the official land use classification.

Keywords: Land-use · Cell phone data records · Big data · Human activity recognition · Human behavior · Knowledge management · Geospatial data · Clustering algorithms · Supervised learning algorithms

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 159–176, 2016.
DOI: 10.1007/978-3-662-53416-8_10

1 Introduction

Recent extensive penetration of digital technologies into everyday life have enabled creation and collection of vast amounts of data related to different types of human activity. When available for research purposes this creates an unprecedented opportunity for understanding human society directly from it's digital traces. There is an impressive amount of papers leveraging such data for studying human behavior, including mobile phone records [5,16,29–31], vehicle GPS traces [22,37], social media posts [20,21,25] and bank card transactions [38,39]. With the growing mobile phone data records, environment modeling can be designed and simulated for understanding human dynamics and correlations between human behaviors and environments. Environment modeling is important for a number of applications such as navigation systems, emergency responses, and urban planning.

Researchers noticed that type of the area defined through official land use is strongly related with the timeline of human activity [13,24,28,33,48]. But those sources of literature do not provide extensive analyses on categorical profile of the geographical areas. This limits the understanding of the dependency of human behaviors from geographical areas. Our analysis confirms this relation, however we show that land use by itself might be not enough, while categorical profile of the area defined based on OSM provides a better prediction for the activity timeline. For example, even within the same land use category, timelines of activity still vary depending on the categorical profile. In this paper, different from these works, we start from clustering the entire city based on area profiles, that are a set of human activities associated with a geographical location, showing that those activities have different area types in terms of the timelines of mobile phone communication activity. Further we show that even the areas of the same land use, which is formally defined by land-use management organizations, might have different clusters based on points of interest (POIs). But those clustered areas are still different in terms of the timelines. This will contribute to other works showing that not only the land use matters for human activity.

This paper uses mobile phone data records to determine the relationship between human behaviors and geographic area context [9]. We present a series of experimental results by comparing the clustering algorithms aiming at answering the following questions: (1) To what extent can geographical types explain human behaviors in a city, (2) What is the relationship between human behaviors and geographical area profiles? We demonstrate our approach to predict area profiles based on the timelines of mobile phone communication activities or vice versa: to predict the timelines from area profiles. We validate our approach using a real dataset of mobile phone and geographic data of Milan, Italy. Our area clustering techniques improve the overall accuracy of the baseline to 64.89 %. Our result shows that land-uses in city planning are not necessarily well defined that an area type is defined with one type of human activity. But growing and development of city structures enable various types of activities that are present in one geographical area. So this type of analysis and its application is important for determining robust land-uses for city planning. Also the hidden patterns and unknown correlations can be observed comparing the mobile phone timelines in relevant areas. The result of this work is potentially useful to improve the classifications of human behaviors for better understanding of human dynamics in real-life social phenomena and to provide a decision support for stakeholders in areas, such as urban city, transport planning, tourism and events analysis, emergency response, health improvement, community understanding, and economic indicators.

The paper is structured as follows Sect. 3 introduces the data sources we use in this research and the data-processing performed. The methodology is described in Sect. 4. We present and discuss the experimental results in Sect. 5. Finally, we summarize the discussions in Sect. 6.

2 Related Works

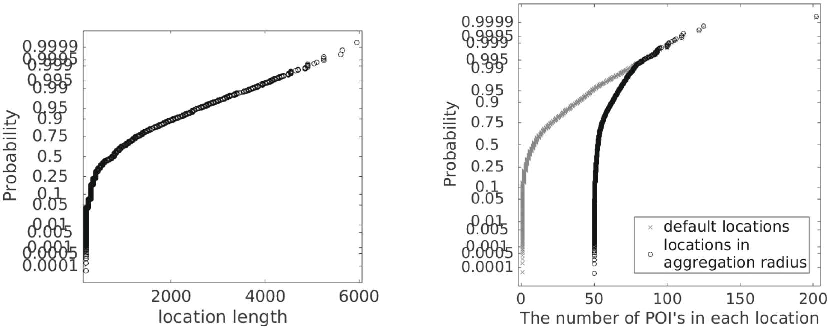
Human behavior is influenced by many contextual factors and their change, for instance, snow fall, hurricane, and festival concerts. There are number of research activities that shed new light on the influence of such contextual factors on social relationships and how mobile phone data can be used to investigate the influence of context factors on social dynamics. Researchers [2,4,15,28] use an additional information about context factors like social events, geographical location, weather condition, etc. in order to study the relationship between human behaviors and such context factors. This is always as successful as the quality of the context factors. The combination of some meteorological variables, such as air temperature, solar radiation, relative humidity, can effect people's comfort conditions in outdoor urban spaces [43], poor or extreme weather conditions influence peoples physical activity [45]. Wang and Taylor [49] exhibited high resilience, human mobility data obtained in steady states can possibly predict the perturbation state. The results demonstrate that human movement trajectories experienced significant perturbations during hurricanes during/after the Hurricane Sandy in 2012. Sagl et al. [35] introduced an approach to provide additional insights in some interactions between people and weather. Weather can be seen as a higher-level phenomenon, a conglomerate that comprises several meteorological variables including air temperature, rainfall, air pressure, relative humidity, solar radiation, wind direction and speed, etc. The approach has been significantly extended to a more advanced context-aware analysis in [36]. Phithakkitnukoon et al. [28] used POIs to enrich geographical areas. The areas are connected to a main activity (one of the four types of activities investigated) considering the category of POIs located within it. To determine groups, that have similar activity patterns, each mobile user's trajectory is labeled with human activities using Bayes Theorem in each time-slot of a day for extracting daily activity patterns of the users. The study shows that daily activity patterns are strongly correlated to a certain type of geographic area that shares a common characteristic context. Similar to this research idea, social networks [48] have been taken into account to discover activity patterns of individuals. Noulas et al. [24] proposed an approach for modelling and characterization of geographic areas based on a number of user check-ins and a set of eights type of general (human) activity categories in Foursquare. A Cosine similarity metric is used to measure the similarity of geographical areas. A Spectral Clustering algorithm together with K-Means clustering is applied to identify an area type. The area profiles enables us to understand groups of individuals who have similar activity patterns. Soto and Frias-Martinez [42] studied mobile phone data records to characterize geographical areas with well defined human activities, by using the Fuzzy C-Means clustering algorithm. The result indicated that five different land uses can be identified and their representation was validated with their geographical localization by the domain experts. Frias-Martinez et al. [13] also studied geolocated tweets to characterize urban landscapes using a complimentary source of land-use and landmark information. The authors focused on determining the land-uses in a specific urban area based on tweeting patterns, and identification of POIs in areas with high tweeting activity. Differently, Yuan and Raubal [50] proposed to classify urban areas based on their mobility patterns by measuring the similarity between time-series using the Dynamic Time Warping (DTW) algorithm. Some areas focus on understanding urban dynamics including dense area detection and their evolution over time [23,46]. Moreover, [14,32,41] analyzed mobile phone data to characterize urban systems. More spatial clustering approaches (Han et al. [19]) could group similar spatial objects into classes, such as k-means, k-medoids, and Self Organizing Map. They have been also used for performing effective and efficient clustering. In this research, we use spectral clustering with eigengap heuristic followed by k-means clustering. Reades et al. [33] and also [18,27] used eigengap heuristic for clustering urban land-uses. In many works [3,26,32,34,40,44] the authors analyzed mobile phone data activity timelines to interpret land-use type. Pei et al. [26] analyzed the correlation between urban land-use information and mobile phone data. The author constructed a vector of aggregated mobile phone data to characterize land-use types composed of two aspects: the normalized hourly call volume and the total call volume. A semi-supervised fuzzy c-means clustering approach is then applied to infer the land-use types. The method is validated using mobile phone data collected in Singapore. Land use is determined with a detection rate of 58.03 %. An analysis of the land-use classification results shows that the detection rate decreases as the heterogeneity of land use increases, and increases as the density of cell phone towers increases. Girardin et al. [17] analyzed aggregate mobile phone data records in New York City to explore the capacity to quantify the evolution of the attractiveness of urban space and the impact of a public event on the distribution of visitors and on the evolution of the attractiveness of the points of interest in proximity.

3 Collecting and Pre-processing the Data

We use two types of datasources for this experiment; (1) POIs from available geographical maps, Openstreetmap (2) Mobile phone network data (sms, internet, call, etc.) generated by the largest operator company in Italy. The mobile phone traffic data is provided in a spatial grid, the rectangular grid of dimensions 100×100, where each unit size of the grid is $235\,\mathrm{m} \times 235\,\mathrm{m}$. We use the grid as our default configuration for collecting human activity distribution and mobile network traffic activity distribution.

3.1 Openstreetmap

In [6–8,11], one of the key elements in the contextual description of geographical regions is the point of interest (POI) (e.g. restaurants, ATMs, and bus stops) that populates an area. A POI is a good proxy for predicting the content of human activities in each area that was well evaluated in [10]. Employing a model proposed in [10], a set of human activities likely to be performed in a given geographical area, can be identified in terms of POI distribution. This allows us to create area profiles of geographical locations in order to provide semantic (high level) descriptions to mobile phone data records in Milan. For example, a person looking for food if the phone call is located close to a restaurant. We exploit the given spatial grid to enrich the locations with POIs from open and free geographic information, Openstreetmap (OSM)[1]. We collected in total 552,133 POIs that refined into 158,797 activity relevant POIs across the locations. To have a sufficient number and diversity of POIs in each location, we consider the nearby areas for estimating the likelihood of human activities. The nearby areas are the intersected locations within the aggregation radius of the centroid point at each location. The aggregation radius is configured differently in each location, which satisfies the need for the total number of POIs in such intersected locations to be above the threshold h, see Fig. 1a and 1b where each location at least $h = 50$ number of POIs in the intersected locations. Across locations, the min, median, and max number of POIs are 50, 53, and 202.



(a) The location size (aggregation radius * 2) distribution

(b) Human activity relevant POI distribution considering aggregation radius

Fig. 1. The distributions of POIs and human activities across locations.

[1] http://www.openstreetmap.org.

In order to build area profiles of each location, a $n \times m$ dimensional matrix $A_{n,m}$ is defined for each location $n \in \{1, .., 10000\}$. Each element $A_{n,m}$ contains the weight of activity categories m in location n where the $m \in \{$eating, educational, entertainment, health, outdoor, residential, shopping, sporting, traveling, working$\}$, with the total number of 10 measurements of human activities per each location. The weight of each category of activities are estimated by the HRBModel which allows us to generate a certain weight for human activities that is proportional to the weight of relevant POIs located in each location. The weight of POIs in a given location, is estimated by the following equation of $tf - idf(f, l) = \frac{N(f,l)}{\underset{w}{\operatorname{argmax}}\{N(w,l):w \in l\}} * \log \frac{|L|}{|\{l \in L: f \in l\}|}$, where f is a given POI; $f \in F$, F=$\{$building, hospital, supermarket,...$\}$ and l is a given location; $l \in L$, L=$\{$location1, location2, location3,...$\}$, $N(f, l)$ is the occurrence of POI f and its appearance in location l and $\underset{w}{\operatorname{argmax}}\{N(w, l) : w \in l\}$ is the maximum occurrence of all the POIs in location l, $|L|$ is the number of all locations, $|\{l \in L : f \in l\}|$ is the number of locations where POI f appears.

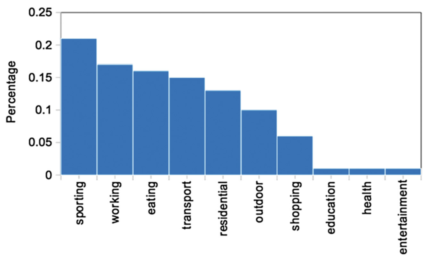


Fig. 2. The activity distribution in Milan

The activity distribution in Milan area is shown in Fig. 2. The sporting, working, eating and transportation types of activities are mainly performed in the city.

3.2 Mobile Phone Network Traffic

In this work, we used a dataset from "BigDataChallenge"[2] organized by Telecom Italia. The dataset is the result of a computation over the Call Detail Records (CDRs) generated by the Telecom Italia cellular network within Milan. The dataset covers 1 month with 180 million mobile network events in November, 2014 as November is a normal month without any particular events organized in Milan. The CDRs log the user activity for billing purposes and network management. There are many types of CDRs, for the generation of this dataset we considered those related to the following activities: square id (the id of the square

[2] http://www.telecomitalia.com/tit/it/bigdatachallenge.html.

that is part of the Milan GRID which contains spatially aggregated urban areas), time interval (an aggregate time), received SMS (a CDR is generated each time a user receives an SMS), sent SMS (a CDR is generated each time a user sends an SMS), incoming Calls (a CDR is generated each time a user receives a call), outgoing Calls (a CDR is generated each time a user issues a call), internet (a CDR is generate each time, a user starts an internet connection, or a user ends an internet connection).

By aggregating the aforementioned records, this dataset was created that provides mobile phone communication activities across locations. The call, sms and internet connection activity logs are collected in each square of the spatial grid for Milan urban area. The activity measurements are obtained by temporally aggregating CDRs in time-slots of ten minutes. But the temporal variations make the comparison of human behaviors more difficult. The standard approach to account for temporal variations in human behavior is to divide time into coarse grained time-slots. In Farrahi and Gatica-Perez [12], the following eight coarse-grained time-slots are introduced: [00–7:00 am., 7:00–9:00 am., 9:00–11:00 am., 11:00 am.–2:00 pm., 2:00–5:00 pm., 5:00–7:00 pm., 7:00–9:00 pm., and 9:00 pm.–00 am.]. Here, we aggregate the mobile phone network data in such coarse-grained time-slots to extract the pattern of 1 month network traffic volume in each location. For each location, we then aggregated the total number of call (outgoing and incoming call without considering a country code), and sms (incoming and outgoing), internet activity for each of those eight time-slots. Such time-slot based timelines can give us actual patterns of mobile network traffic activity. Then the dataset reduced to 2.4 million CDR each of which consists of the followings: square id, day of month, time-slot, and total number of mobile network traffic activity. We build a $n \times p \times d$ dimensional matrix $T_{n,p,d}$ to collect a mobile phone traffic activity timeline, where n is the number of locations in $[1, 10000]$, p is the time-slot divisions of the day $[1, 8]$ and d is the day in $[1, 31]$. To identify timeline patterns among those locations, we performed a normalization for the timelines based on z-score which transforms the timeline into the output vector with mean $\mu = 0$ while standard deviation σ is negative if it is below the mean or positive if it is above the mean. The normalized timelines by day is visualized in Fig. 3 which show a stable communication activity within the month. For this transformation, we used $T'_{i,j,k} = \dfrac{T_{i,j,k} - \mu_i}{\sigma_i}, i \in n, j \in p, k \in d$, where μ_i is the average value of the mobile phone activity traffic in location i, σ_i is the standard deviation of the mobile phone activity traffic in location i.

4 The Approach

We present our methodology for identifying the relation between geographical locations and human behaviors. Our methodology is divided into two phases: (1) clustering approaches for inferring categorical area types in terms of geographical area profiles (2) classification approaches for validating the observed area types by mobile phone data records. Clustering techniques are mostly unsupervised

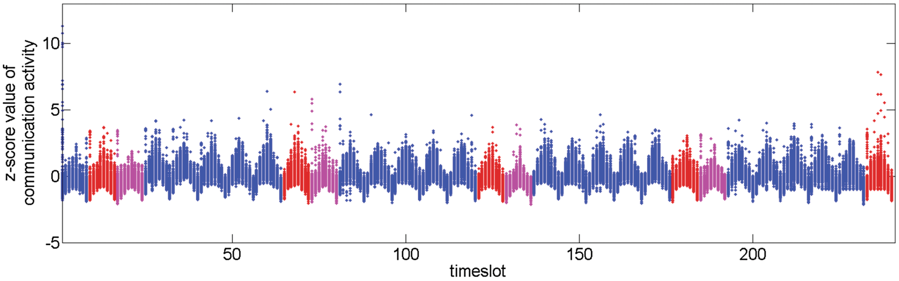


Fig. 3. The timelines for each time-slot of day, z-score normalization by day

methods that can be used to organize data into groups based on similarities among the individual data items. We use the spectral clustering algorithm which makes use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset.

We define a vector space model that contains a set of vectors corresponding to areas. Relevance between areas is a similarity comparison of the deviation of angles between each area vector. The similarity between the areas is calculated by the cosine similarity metric by estimating the deviation of angles among area vectors. For example, the similarity between area l_1 and l_2 would be $\cos\theta_{l_1,l_2} = \frac{\mathbf{l_1} \cdot \mathbf{l_2}}{\|\mathbf{l_2}\|\|\mathbf{l_1}\|}$ where l_i denotes the area or the features associated to the areas. We denote each area l_i with a set of corresponding features associated with a weight measure j. Having the estimation of similarity between the areas, we can now create a similarity graph described as the weight matrix W generated by the cosine similarity metrics and the diagonal degree matrix D is utilized by the spectral clustering algorithm which is the one of the most popular modern clustering methods and performs better than traditional clustering algorithms. We create the adjacency matrix A of the similarity graph and graph Laplacian LA, $LA = D - A$ (given by normalized graph Laplacian $LA_n = D^{-1/2}LAD^{-1/2}$). Based on eigengap heuristic [47], we identify the number of clusters by k-nearest neighbor to observe in our dataset as $k = argmax_i(\lambda_{i+1} - \lambda_i)$ where $\lambda_i \in \{l_1, l_2, l_3, .., l_n\}$ denotes the eigenvalues of l_n in the ascending order. Finally, we easily detect the effective clusters (area profiles) $S_1, S_2, S_3, ..., S_k$ from the first k eigenvectors identified by the k-means algorithms. We investigate the relation between geographical locations and human behaviors based on categorical area types. To do that, we use supervised learning algorithms to predict area profile of a given area if we train a classification model with training data, which are the timelines labeled with area types. In supervised learning, each observation has a corresponding response or label. Classification models learn to predict a discrete class given new predictor data. We use several of classifiers for learning and prediction. We prepare a test set for testing classification models by k-fold cross validation method.

5 Experiments and Results

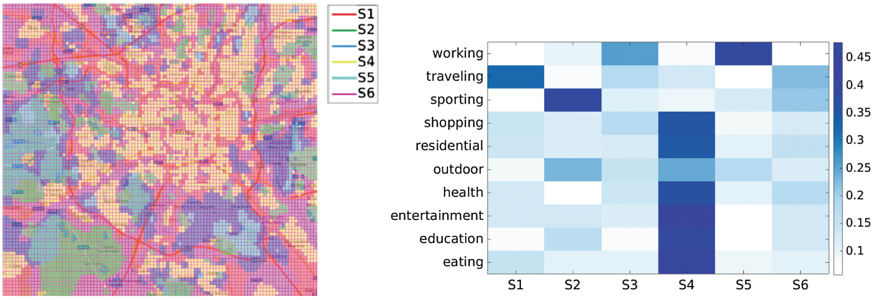
In this section, we demonstrate the identification of the relationships between locations and human behaviors in terms of two types of features in each location: (1) location contexts: categories of human activity estimated through types of available POI (2) mobile communication activity timeline: mobile communication activity in time-series of coarse grained time-slots. In other words, we estimate the extent to which human behaviors depend on geographical area types. To identify and quantify these dependencies, we perform two types of validations: (1) observed area type we defined vs human behavior (2) land-use type defined formally vs human behavior by estimating the correlations and prediction algorithms.

5.1 Observed Area Type Vs Human Behavior

We first check the two datasets can be clustered or randomly distributed using Hopkins statistic, $H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$. The distance between element p_i and its nearest neighbor in dataset D is $x_i = \min_{v \in D} \{dist(p_i, v)\}$ and the distance between element q_i and its nearest neighbor in $D - q_i$ is $y_i = \min_{v \in D, v \neq q_i} \{dist(q_i, v)\}$. The Hopkins statistic for the location context dataset is 0.02 and the mobile communication timeline is 0.04 that indicates that the datasets are highly clustered and regularly distributed. So we then analyze the correlations of location context and mobile phone communication timeline in order to understand if humans are attracted to location contexts through the area types (i.e., shopping, woking, and studying). To validate such relationship, we start with the geographical area clustering based on the location context by semi-supervised learning algorithms. We perform spectral clustering on the locations based on their similarity of human activity distribution $A_{n,m}$. Each location of the grid has a distribution of activity categories with relative frequency of their appearance. The spectral clustering with k-nearest neighbor ($k = 10$ based on cosine similarity metrics) approach allows us to classify geographical areas L based on such multi-dimensional features, $A_{n,m}$. We then observed significantly different six types of areas, that are geo-located in Fig. 4(a). The average values of the activity categories for those area types are presented in Fig. 4(b).

The table shows that categorical area type $S4$ contains high percentage values for residential, and eating activities. The center of the city including a residential zone were clustered into one area type. The area type $S3$ contains high percentage value on working activity. This classification can be refined if we increase the number of area types observations. For each area type, we are now able to extract and observe timelines $T_{n,p,d}$ from mobile phone data records in order to determine the correlation between the timelines and the area profiles for those area types.

The density of the clusters are almost uniform distributed except cluster S4 and S5, see Fig. 5(a). This unbalanced datasets for clusters could contribute to an acceptable global accuracy, but also to a (hidden) poor prediction for instances in



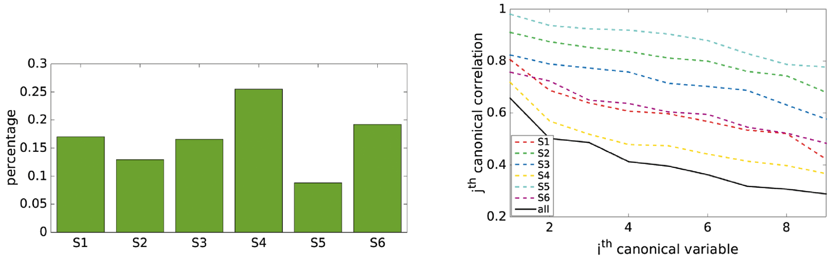
(a)

(b)

Fig. 4. (a) Observed area types of the geographical area of Milan based on the area profiles, $k = 6$, where $S1$ is red, $S2$ is lime, $S3$ is blue, $S4$ is yellow, $S5$ is cyan/aqua, and $S6$ is magenta/fuchsia. (b) The average values of the activity categories in categorical area types observed. (Color figure online)

minority classes. In this context, alternative metrics, such as per class accuracy will be considered. We estimate the accuracy per class using the two techniques (canonical correlation coefficients vs learning techniques). Figure 6 shows the actual volume of the mobile network traffic activities by the area types.

We illustrated the correlation between the area profiles $A_{n,m}$ and timelines $T_{n,p,d}$ based on the canonical correlation [1] (see Fig. 5(b)). The canonical correlation investigates the relationships between two sets of vectors by maximizing the correlation in linear combination. In order words, canonical correlation finds the optimal coordinate system for correlation analysis and the eigenvectors defines the coordinate system. While the overall maximum correlation coefficient ($j = 1$) is 65 % between the two vectors, the correlation coefficient by area types is high between 72 % and 98 %. For example, the correlation in area type $S5$ is stronger than other area types, in which working type of activities are more distributed. The maximum correlation in $S2$ containing high percentage of sporting activity is 82.38 %.



(a)

(b)

Fig. 5. (a) The density distribution of area types observed in Milan. (b) Canonical correlation between the two feature matrices for locations.

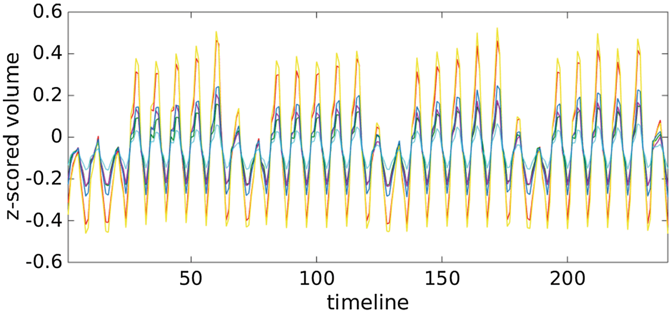


Fig. 6. The average timeline of mobile phone data by area types ($k=6$), where $S1$ is red, $S2$ is lime, $S3$ is blue, $S4$ is yellow, $S5$ is cyan/aqua, and $S6$ is magenta/fuchsia (Color figure online)

We also compared the distance between the two vectors (mean) of area types to investigate the similarity of the relevant area profiles can have the similar human behaviors. We observed linear correlation with a coefficient of $r = 0.61$ This result shows that as the distance between the area profiles is increased, the timeline difference increases, and human behaviors are strongly correlated to geographical area profiles. In second, we profiles the communication timelines with the cluster labels observed in each location that will be used to estimate the correlation by supervised learning algorithms. The prediction accuracy of timeline types in a given location could be an evaluation of the dataset. To that end, we train several predictive models (i.e., Bayesian algorithms, Decision Trees, Probabilistic Discriminative models and Kernel machines.) to measure the prediction accuracy by k-fold cross validation method ($k=10$), which is used to estimate how accurately a predictive model will perform. We need to prepare training and test data. The training data are the timelines labeled by area types through the location. This allows us to determine if timelines are clustered as geographical area profiles. The experimental results on our data are shown in Table 1. This classification of the predictive models is aimed at choosing a statistical predictive algorithm to fit in our analysis.

Among the considered techniques, the Random Forest and the Nearest Neighbor algorithms are resulted in the lowest error with high accuracy, in other words, if we take the area profile of the nearest-neighbor (the most common area profile of k-nearest-neighbors), that would give the right timeline type. The confusion matrix of the Random Forest classifier, and the precision, recall are estimated in the following Table 2. The receiver operating characteristic curve for visualizing the performance of the classifiers is described in Fig. 7. This result shows that the area type S5 is the well classified and compact by showing a strong correlation between the area activity categories and area timeline. The area types S1, S2, S3 and S4, S6 can be still refined in terms of the area activity categories.

Table 1. Results for the predictive models with the use of area types observed by spectral clustering algorithm

Algorithm	Cross validation	Overall ACC
Random classifier	0.83	16.7 %
Linear discriminant	0.5404	45.01 %
Quadratic discriminant	0.4649	52.90 %
Naive bayes (kernel density)	0.6748	20.38 %
K-NN (k = 5, euclidean dist)	0.3822	61.73 %
K-NN (k = 10, euclidean dist)	0.4068	59.26 %
Decision tree	0.4806	52.58 %
Random forest	0.3513	64.89 %
Multi-class SVM	0.4997	49.47 %

Table 2. Confusion matrix and precision, recall and f-measure in each area type defined for predicting timeline based on location context about categorical human activity by Random Forest classifier

Area type defined	S1	S2	S3	S4	S5	S6
S1	8.91%	0.20%	1.80%	4.47%	0.07%	1.57%
S2	0.10%	8.58%	0.70%	1.77%	0.47%	1.30%
S3	1.77%	0.43%	10.15%	1.70%	1.27%	1.23%
S4	2.34%	0.53%	1.13%	19.63%	0.33%	1.54%
S5	0.03%	0.23%	1.37%	0.53%	6.54%	0.07%
S6	2.40%	1.27%	1.67%	2.74%	0.07%	11.08%

Prec.	Recall.	F-measure.
52.35%	57.30%	54.71%
66.41%	76.26%	70.99%
61.29%	60.32%	60.80%
76.96%	63.64%	69.67%
74.52%	74.81%	74.67%
57.64%	66.00%	61.54%

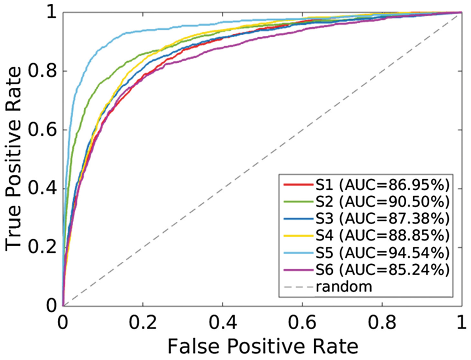
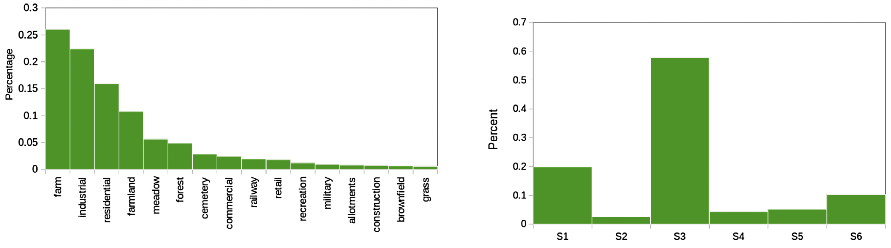


Fig. 7. Receiver operating characteristic to multi-class by random forest classifier



(a) (b)

Fig. 8. (a) The distribution of land-use classification in Milan. (b) The distribution of categorical activity clusters within commercial land-use area

5.2 Land-Use Type Vs Human Behavior

After we observed strong prediction accuracy of timelines based on categorical area types, we analyze the relation between the timelines and land use types which are formally defined by land-use management organizations. While many works try to predict activity based on land use, we perform a comparative study of the two approaches. We identify that even the area of the same land use might have different area types in terms of area profiles and those are still different in terms of human activity timelines quantified through mobile phone records, which validates significance of activity-based classification vs official land use. We predict the timeline type of a given area based on the land-use type using the Random Forest and the Nearest Neighbor classifiers. We used the land-use types from the OSM[3] for this prediction task (see the distribution of land-use types of Milan in Fig. 8(a)). The prediction accuracy of the Random Forest classifier is 53.47 %. This shows that predicting power of categorical types is higher compared to land use types.

We also match the area types we observed with the land-use types given officially. The result shows that even within the same land-use type, the timelines corresponding to different clusters are still different. For example, 58 % of the commercial land-uses matched with the area type S3 which followed by S1, S6 and S2, S3, S4, S5, shown in Fig. 8(b). The corresponding timelines to the different clusters within the commercial land-use type are illustrated in Fig. 9.

The timelines in the same area type observed, also in the same land-use officially defined, can be still refined, but the timeline pattern refinement will require more emphasis on the appropriate features, for example, timelines for weekday or weekend. The area profiles are semantically different concepts in terms of human activities performed in geographical areas. Further, it will allow us to identify a standard or exceptional type of mobile network activities in relevant areas, as well as to enable the identification of unknown correlations, or hidden patterns about anomalous behaviors.

[3] http://wiki.openstreetmap.org/wiki/Key:landuse.

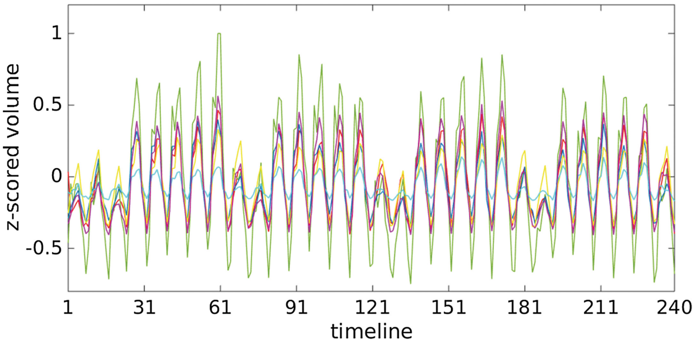


Fig. 9. The timelines belong to different clusters within the commercial land-use: $S1$ is red, $S2$ is lime, $S3$ is blue, $S4$ is yellow, $S5$ is cyan/aqua, and $S6$ is magenta/fuchsia (Color figure online)

6 Conclusion and Future Works

In this paper, we proposed an approach that characterizes and classifies geographical areas based on their anticipated (through POI distribution) human activity categorical types, such as working or shopping oriented areas. We concentrated on the analysis of the relationship between such spatial context of the area and observed human activity. Our approach compares the similarity between area activity categorical profiles and human activity timeline categories estimated through cell phone data records. We found an overall correlation of 61 % and canonical correlation of 65 % between contextual and timeline-based classifications. We observed six types of areas according to the area activity categories where we compared their human activity timelines with their area activity categories and the correlation (canonical) coefficient is between 72 % and 98 %. For example, the area type $S5$ related to working activity has a strong correlation of 98 % which followed by the area types, $S2$ related to sporting activity and $S3$ related to the human activities in the center of the city. The supervised learning approach validates possibility of using an area categorical profile in order to predict to some extent the network activity timeline (i.e., call, sms, and internet). For example, the Random Forest approach performs well with the accuracy of 64.89 %. So human behaviors' temporal variation is characterized similarly in relevant areas, which are identified based on the categories of human activity performed in those locations. Furthermore we found that the prediction accuracy based on the official land use types is only 53.47 %. So the official land-use types by themselves are not enough to explain the observed impact of area context on human activity timelines, also because even within the same land use type, different activity categorical types still demonstrate different activity timelines. Further, the semantic description of area profiles associated to mobile phone data enables the investigation of interesting behavioral patterns, unknown correlations, and hidden behaviors in relevant areas. We expect the approach to be further applicable to other ubiquitous data sources, like geo-localized tweets, foursquare data, bank card transactions or the geo-temporal logs of any other service.

Acknowledgments. The authors would like to thank the Semantic Innovation Knowledge Lab - Telecom Italia for publicly sharing the mobile phone data records which were provided for Big Data Challenge organized in 2013, Italy. We also would like to thank MIT SENSEable City Lab Consortium partially for supporting the research.

References

1. Härdle, W., Simar, L.: Canonical correlation analysis. In: Härdle, W., Simar, L. (eds.) Applied Multivariate Statistical Analysis, pp. 321–330. Springer, Heidelberg (2007)
2. Bagrow, J.P., Wang, D., Barabási, A.-L.: Collective response of human populations to large-scale emergencies. CoRR, abs/1106.0560 (2011)
3. Becker, R.A., Cceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A tale of one city: using cellular network data for urban planning
4. Calabrese, F., Pereira, F.C., Di Lorenzo, G., Liu, L., Ratti, C.: The geography of taste: analyzing cell-phone mobility and social events. In: Floréen, P., Krüger, A., Spasojevic, M. (eds.) Pervasive 2010. LNCS, vol. 6030, pp. 22–37. Springer, Heidelberg (2010)
5. Calabrese, F., Ratti, C.: Real time Rome. Netw. Commun. Stud. **20**(3–4), 247–258 (2006)
6. Dashdorj, Z., Serafini, L.: Semantic enrichment of mobile phone data records using linked open data. In: Proceedings of the 12th International Conference on Semantic Web Conference Poster and Demonstrations Track (2013)
7. Dashdorj, Z., Serafini, L.: Semantic interpretation of mobile phone records exploiting background knowledge. In: The Proceedings of the 12th International Conference on Semantic Web Conference Doctoral Consortium (2013)
8. Dashdorj, Z., Serafini, L., Antonelli, F., Larcher, R.: Semantic enrichment of mobile phone data records. In: MUM, p. 35 (2013)
9. Dashdorj, Z., Sobolevsky, S.: Impact of the spatial context on human communication activity. CoRR, abs/1506.03668 (2015)
10. Dashdorj, Z., Sobolevsky, S., Serafini, L., Antonelli, F., Ratti, C.: Semantic enrichment of mobile phone data records using background knowledge. arXiv preprint arXiv:1504.05895 (2015)
11. Dashdorj, Z., Sobolevsky, S., Serafini, L., Ratti, C.: Human activity recognition from spatial data sources. In: Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS 2014, pp. 18–25. ACM, New York (2014)
12. Farrahi, K., Gatica-Perez, D.: What did you do today? Discovering daily routines from large-scale mobile data. In: ACM International Conference on Multimedia (ACMMM). IDIAP-RR 08-49 (2008)
13. Frías-Martínez, V., Soto, V., Hohwald, H., Frías-Martínez, E.: Characterizing urban landscapes using geolocated tweets. In: SocialCom/PASSAT, pp. 239–248 (2012)

14. Fujisaka, T., Lee, R., Sumiya, K.: Exploring urban characteristics using movement history of mass mobile microbloggers. In: Proceedings of the Eleventh Workshop on Mobile Computing Systems & #38; Applications, HotMobile 2010, pp. 13–18. ACM, New York (2010)
15. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Identifying users profiles from mobile calls habits. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing, UrbComp 2012, pp. 17–24. ACM, New York (2012)
16. Girardin, F., Calabrese, F., Fiore, F.D., Ratti, C., Blat, J.: Digital footprinting: uncovering tourists with user-generated content. IEEE Pervasive Comput. **7**(4), 36–43 (2008)
17. Girardin, F., Vaccari, A., Gerber, R., Biderman, A.: Quantifying urban attractiveness from the distribution and density of digital footprints. J. Spat. Data Infrastruct. Res
18. Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., Ratti, C.: Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong. CoRR, abs/1406.4400 (2014)
19. Han, J., Kamber, M., Tung, A.K.H.: Spatial Clustering Methods in Data Mining: A Survey. Taylor and Francis, Milton Park (2001)
20. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C.: Geolocated twitter as proxy for global mobility pattern. Cartogr. Geogr. Inf. Sci. 1–12 (2014)
21. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM (2007)
22. Kang, C., Sobolevsky, S., Liu, Y., Ratti, C.: Exploring human movements in singapore: a comparative analysis based on mobile phone and taxicab usages. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, p. 1. ACM (2013)
23. Ni, J., Ravishankar, C.V.: Pointwise-dense region queries in spatio-temporal databases. In: Chirkova, R., Dogac, A., Ã-zsu, M.T., Sellis, T.K. (eds.) ICDE, pp. 1066–1075. IEEE, Piscataway (2007)
24. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In: The Social Mobile Web (2011)
25. Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C., González, M.C.: Urban magnetism through the lens of geo-tagged photography. arXiv preprint arXiv: 1503.05502 (2015)
26. Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., Zhou, C.: A new insight into land use classification based on aggregated mobile phone data. Int. J. Geogr. Inf. Sci. **28**(9), 1–20 (2014)
27. Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Zhou, C.: A new insight into land use classification based on aggregated mobile phone data. CoRR, abs/1310.6129 (2013)
28. Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C.: Activity-aware map: identifying human daily activity pattern using mobile phone data. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.) HBU 2010. LNCS, vol. 6219, pp. 14–25. Springer, Heidelberg (2010)
29. Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., Crowcroft, J.: Recommending social events from mobile phone location data. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), pp. 971–976 (2010)

30. Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., Strogatz, S.H.: Redrawing the map of great britain from a network of human interactions. PLoS One **5**(12), e14248 (2010)
31. Ratti, C., Williams, S., Frenchman, D., Pulselli, R.: Mobile landscapes: using location data from cell phones for urban analysis. Environ. Plan. B **33**(5), 727 (2006)
32. Ratti, C., Williams, S., Frenchman, D., Pulselli, R.M.: Mobile landscapes: using location data from cell phones for urban analysis. Environ. Plann. B Plann. Des. **33**(5), 727 (2006)
33. Reades, J., Calabrese, F., Ratti, C.: Eigenplaces: analysing cities using the space–time structure of the mobile phone network. Environ. Plann. B: Plann. Des. **36**(5), 824–836 (2009)
34. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: explorations in urban data collection. IEEE Pervasive Comput. **6**(3), 30–38 (2007)
35. Sagl, G., Beinat, E., Resch, B., Blaschke, T.: Integrated geo-sensing: a case study on the relationships between weather and mobile phone usage in Northern Italy. In: IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, ICSDM, Fuzhou, China, 29 June–1 July 2011, pp. 208–213 (2011)
36. Sagl, G., Blaschke, T., Beinat, E., Resch, B.: Ubiquitous geo-sensing for context-aware analysis: exploring relationships between environmental and human dynamics. Sensors **12**(7), 9800–9822 (2012)
37. Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S., Ratti, C.: Taxi pooling in New York City: a network-based approach to social sharing problems. arXiv preprint arXiv:1310.2963 (2013)
38. Sobolevsky, S., Sitko, I., Combes, R.T.D., Hawelka, B., Arias, J.M., Ratti, C.: Money on the move: big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in Spain. In: 2014 IEEE International Congress on Big Data (BigData Congress), pp. 136–143. IEEE (2014)
39. Sobolevsky, S., Sitko, I., Grauwin, S., Combes, R.T.D., Hawelka, B., Arias, J.M., Ratti, C.: Mining urban performance: scale-independent classification of cities based on individual economic transactions. arXiv preprint arXiv:1405.4301 (2014)
40. Sobolevsky, S., Szell, M., Campari, R., Couronn, T., Smoreda, Z., Ratti, C.: Delineating geographical regions with networks of human interactions in an extensive set of countries. PLoS ONE **8**(12), e81707 (2013)
41. Soto, V., Frías-Martínez, E.: Robust land use characterization of urban landscapes using cell phone data. In: The First Workshop on Pervasive Urban Applications (PURBA), June 2011
42. Soto, V., Frías-martínez, E.: Robust land use characterization of urban landscapes using cell phone data (2011)
43. Stathopoulos, T., Wu, H., Zacharias, J.: Outdoor human comfort in an urban climate. Build. Environ. **39**(3), 297–305 (2004)
44. Sun, J., Yuan, J., Wang, Y., Si, H., Shan, X.: Exploring space–time structure of human mobility in urban space. Phys. A: Stat. Mech. Appl. **390**(5), 929–942 (2011)
45. Tucker, P., Gilliland, J.: The effect of season and weather on physical activity: a systematic review. Pub. Health **121**(12), 909–922 (2007)
46. Vieira, M.R., Frías-Martínez, V., Oliver, N., Frías-Martínez, E.: Characterizing dense urban areas from mobile phone-call data: discovery and social dynamics. In: Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM 2010, pp. 241–248. IEEE Computer Society, Washington, DC (2010)

47. von Luxburg, U.: A tutorial on spectral clustering. CoRR, abs/0711.0189 (2007)
48. Wakamiya, S., Lee, R., Sumiya, K.: Urban area characterization based on semantics of crowd activities in twitter. In: Claramunt, C., Levashkin, S., Bertolotto, M. (eds.) GeoS 2011. LNCS, vol. 6631, pp. 108–123. Springer, Heidelberg (2011)
49. Wang, Q., Taylor, J.E.: Quantifying human mobility perturbation, resilience in hurricane sandy. PLoS ONE **9**(11), e112608 (2014)
50. Yuan, Y., Raubal, M.: Extracting dynamic urban mobility patterns from mobile phone data. In: Xiao, N., Kwan, M.-P., Goodchild, M.F., Shekhar, S. (eds.) GIScience 2012. LNCS, vol. 7478, pp. 354–367. Springer, Heidelberg (2012)

Analysis of Customers' Spatial Distribution Through Transaction Datasets

Yuji Yoshimura[1,2(✉)], Alexander Amini[1], Stanislav Sobolevsky[1,3],
Josep Blat[2], and Carlo Ratti[1]

[1] SENSEable City Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue,
Cambridge, MA 02139, USA
{yyoshi,amini,ratti}@mit.edu

[2] Department of Information and Communication Technologies, Universitat Pompeu Fabra,
Roc Boronat, 138, Tanger Building, 08018 Barcelona, Spain
josep.blat@upf.edu

[3] Center for Urban Science and Progress, New York University, 1 MetroTech Center, 19th Floor,
Brooklyn, NY 11201, USA
sobolevsky@nyu.edu

Abstract. Understanding people's consumption behavior while traveling between retail shops is essential for successful urban planning as well as determining an optimized location for an individual shop. Analyzing customer mobility and deducing their spatial distribution help not only to improve retail marketing strategies, but also to increase the attractiveness of the district through the appropriate commercial planning. For this purpose, we employ a large-scale and anonymized datasets of bank card transactions provided by one of the largest Spanish banks: BBVA. This unique dataset enables us to analyze the combination of visits to stores where customers make consecutive transactions in the city. We identify various patterns in the spatial distribution of customers. By comparing the number of transactions, the distributions and their respective properties such as the distance from the shop we reveal significant differences and similarities between the stores.

Keywords: Consumer behaviors · Transaction data · Human mobility · Urban studies · Barcelona

1 Introduction

The diversity of a retail shop and its density make an urban district attractive and unique, thereby enhancing the competition between shops and enticing external visitors from other districts both nearby and abroad [1]. Pedestrian exploration and their presence encourage other pedestrians to interact with one another, generating liveliness throughout the neighborhood [2]. Conversely, retailers believe a key driver of store performance is location [3], which collectively determines the way a customer transitions from shop to shop. This is greatly influenced by geographical accessibility to said shops: a central location is easier to be approached from anywhere, making it more

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 177–189, 2016.
DOI: 10.1007/978-3-662-53416-8_11

visible and popular to attract both people and goods [4]. “Constraints on mobility determine where we can go and what we can buy” [5].

The objective of this paper is to analyze customers’ spatial distribution considering their consecutive transaction activities through three large-scale department stores in the city of Barcelona, Spain. We study similarities in customers’ origin and destination locations between the same chains of these three stores, which are located in varying urban settings. Essential understanding of this area is largely related to how the power of attraction and distribution for each store affects both the customers as well as the holistic urban environment.

For this purpose, we employ a large-scale transaction dataset provided by one of Spain’s largest banks: Banco Bilbao Vizcaya Argentaria (BBVA). This dataset contains the geographic zip code of a shop where a customer made a transaction, timestamps, and monetary amount of said transaction (see Sect. 4 for more details). We extracted the combination of retail shops, where customers make consecutive transactions before or after any transactions in one of three large-scale department stores. This approach differs from that in previous studies, which use credit card transactions in the analysis of human behavior [6, 7]. Similarly, it is different from analyzing the predictability of human spending activities [5], because the latter utilizes detailed topological analysis whereas we use the physical spatial analysis.

The advantages of our dataset can be summarized as follows: contrary to the point of sales (POS) or the customer loyalty cards [8], BBVA’s credit cards are designed to be used with specific readers installed in over 300,000 BBVA card terminals in Spain [6]. This enables us to analyze spatial distributions of a customer’s sequential purchasing behavior between retail shops over the territory. In addition, the detection scale for the purchase location is smaller than the one for passive mobile phone tracking [9–13] RFID-based studies [14, 15] or Bluetooth sensing techniques [16–20]. This indicates that the attractivity analysis for each shop can be studied at a much finer grain of resolution than in previously recorded studies [21, 22].

Conversely, our research does present several limitations. The dataset consists solely of customers who hold BBVA’s credit or debit card and used it for the purchases we analyze. This suggests that our analysis contains a possible bias in terms of the type of customers we study (i.e., highly educated upper and middle class). In addition, our analysis is based on customers’ successive order of purchase behaviors between different retail shops, meaning that we cannot deduce their transition path or their purchase decisions when they don’t use BBVA’s card. Moreover, our dataset cannot reveal customers’ decision-making processes or value consciousness because it doesn’t contain their inner thought process typically derived from interviews, questionnaires or participatory observation. Furthermore, there is an inherent temporal sparseness present in the data with just a small fraction of all activities being recorded, although this provides enough of sample at the aggregated scale.

Within these limitations, we try to uncover the features of a customer’s transaction activities and the similarities of their spatial distribution through the city and the urban structure.

2 Context of the Study: Barcelona

The city of Barcelona is divided into 10 administrative districts, and 73 neighborhoods within those districts, each of which with its own unique identity.

Figure 1 shows the districts, major avenues, and plazas which determine the urban structure of the city of Barcelona. There are approximately 50,000 business entities throughout the city, including department stores, commercial centers, supermarkets, shopping streets with exclusive designer boutiques and international/local brands.

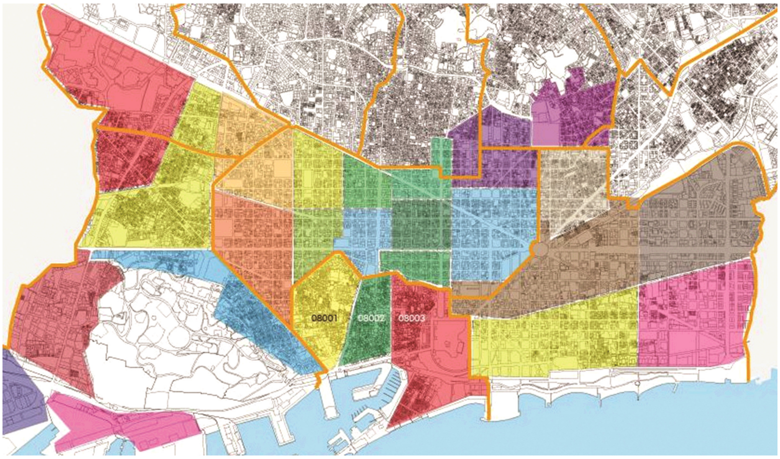


Fig. 1. The map of the city of Barcelona. The zip code, 10 districts and 73 neighborhoods.

This paper analyzes customer spatial distributions through analysis of their mobility, based on their consecutive activities made before and after visiting the same chain of a large-scale department store. They are located in one of three different neighborhoods in the city. We selected the same chain of large-scale department stores rather than small- and medium-scale shops because (1) we can expect a larger number of customer transactions because of the stores' higher attractivity, (2) customers can be derived from far locations as well as nearby, which enables us to analyze urban structure throughout a larger landscape and (3) the obtained dataset of customers can be more homogeneous rather than distorted and biased.

Each one of these stores attracts a large volume of customers and is therefore able to create expanded distributions of customers to other retail shops in surrounding neighborhoods. They can be considered one of the strongest hubs in the district, triggering a customer's sequential shopping movements. Thus, their presence has great spatial impact in the district in terms of the volume of attracted customers as well as the associated sequential movements.

The first shop (PC) is located in the city center, Ciutat Vella (old town). Ciutat Vella district is composed of four neighborhoods: El Raval, El Gòtic, La Barceloneta, Sant Pere, Santa Caterina i la Ribera. These neighborhoods are full of retail shops with the most famous brands in the wide commercial area between Pelayo and Portaferrissa streets, and the Portal de l'Àngel. Because of its scenic monuments, architectures and environment, this district attracts tourists as well as locals from all districts of the city.

The second one (AD) is located in Eixample district. This district is divided into six neighbhorhoods (El Fort Pienc, Sagrada Família, Dreta de l´Eixample, Antiga Esquerra de l´Eixample, Nova Esquerra de l´Eixample, Sant Antoni). This area is a business district surrounded by a variety of private companies. Therefore, customers are likely to be workers for these companies as well as people from the wealthy neighborhoods of Pedralbes, Sant Gervasi, and Sarrià.

The last one (PA) is located in Nou Barris district. The shop faces the comer of Sant Andreu and Avenida Meridiana, one of the biggest avenues in Barcelona. This area has a high concentration of immigrants and working-class citizens, as well as a high level of registered unemployment. The specific geographical location is at an entrance to the city of Barcelona and therefore attracts customers traveling from adjacent districts/villages.

By comparing consumer patterns for the same store located in different regions of the city, our analysis reveals dependencies on neighborhood features more clearly than if different shops has been analyzed.

3 Methodology

Our goal is to isolate transactions before and after visiting one of three shops in the city of Barcelona within a 24-hour window. We will refer to these three shops (PC, AD, PA) as the focal shops of our study. Specifically, we extracted consecutive sequential credit and debit transactions as customers moved between stores either before or after visiting the focal shops.

We define an incoming customer as one who makes a transaction in any shop before making a transaction in a focal shop. Similarly, we define a leaving customer as one who makes a transaction in any other shops after doing so in a focal shop.

Figure 2(a), (b), and (c) show the location of each shop. We aggregate the number of customers within a radius of 1 km from each store. This methodology permits us to aggregate customer spending behavior in terms of spatial dimension, where they come from, and where they move to before or after visiting one of those stores.

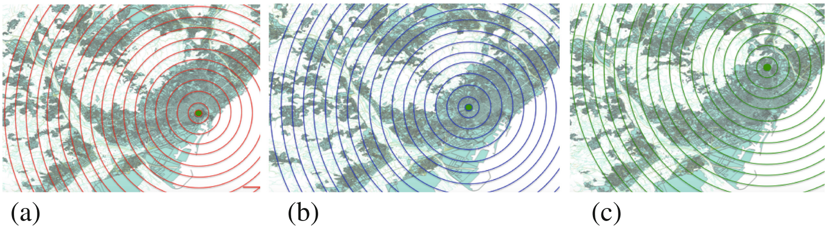


Fig. 2. (a) The location of the shop PC with radius of 1 km. (b) AD. (c) PA.

Within this framework, this paper assesses the spatial distribution based on customers' sequential movement around the large-scale department store located in Barcelona.

4 Data Settings

Data for this paper was provided by one of the largest Spanish banks–Banco Bilbao Vizcaya Argentaria (BBVA). The data consists of bank card transactions performed by two groups of card users: direct customers who hold a debit or credit card issued by BBVA and others who made transactions through one of the approximately 300,000 BBVA card terminals. Once customers make transactions with their debit or credit card, the system registers those activities. The information contains the randomly generated IDs of customers, and indication of a customer's residence and a shop where a customer made a transaction at the level of zip code, a time stamp, and each transaction denoted with its value. The datasets do not contain information about items purchased, and the shops are categorized into 76 business categories such as restaurants, supermarkets, or hotels. In addition, the location where a customer makes transactions is denoted as a zip code rather than the actual street address. The data is aggregated and hashed for anoymization in accordance to all local privacy protection laws and regulations. The total number of customers are around 4.5 million, making more than 178 million transactions totaling over 10 billion euro during 2011 (see [6] for more details).

5 Spatial Analysis

5.1 Customers Distribution in the Micro Scale

In this section, we analyze the spatial distribution based on customer mobility in the microscopic scale, considering their purchase behaviors. We focus on transactions at shops before or after visiting the three focal shops (AD, PA, PC) around the city of Barcelona. This reveals, on the one hand, each shop's customer mobility in the city of Barcelona, and, on the other hand, the degree of each shop's attracting power and distribution power and their customers' sequential movements around each one.

The volume of transactions against distance for the shop PA can be seen in Fig. 3(b). PA starts to attract customers from 1 km to 2 km (8.41 %), meaning their customers don't make transactions nearby (0–1 km, 0.00 %) before/after visiting it. In addition, almost no customers make transactions from proximate locations such as within 2–3 km (0.26 %), 3–4 km (0.00 %), 4–5 km (0.00 %), 5-6 km (0.00 %). The hot spot of customers' locations of origin can be found within 6–7 km (9.24 %), 10–12 km (14.67 %), 12–14 km (14.41 %) and 16–18 km (15.12 %).

Conversely, the shop AD attracts customers who make transactions nearby (0–1 km, 3.10 %). This distribution pattern is unique to AD. The number of customers increases with the distance until 6–7 km (i.e., 3–4 km, 3.71 %, 4–5 km, 4.53 %, 5–6 km, 8.12 %) and is maximized at 7–8 km. In addition, the locations far from the shop tend to show lower percentages of transactions (i.e., 8–9 km, 2.37 %, 9–10 km, 4.95 %, 10–12 km,

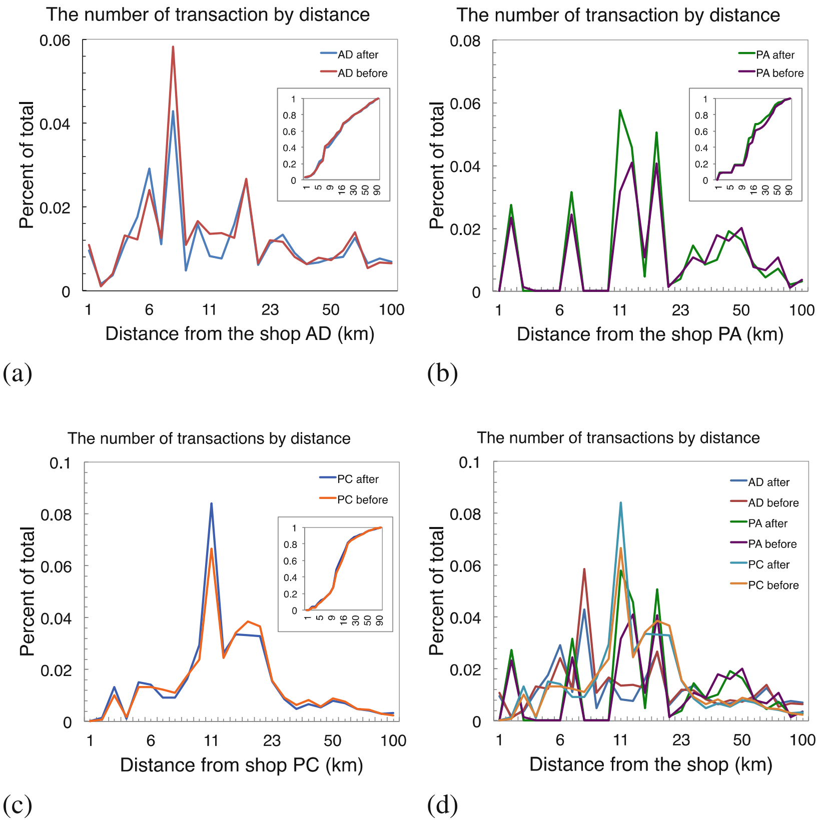


Fig. 3. (a) The distance against the frequency of transactions by the shop AD. (b) PA. (c) PC. (b) All shop.

4.58 %, 12–14 km, 4.44 %, 14–16 km, 4.34 %, 16–18 km, 8.05 %), indicating that the concentration of transaction volume for AD is intensified in proximal locations.

With respect to the shop PC, customer transactions appear within 2–3 km (3.03 %); meanwhile, there is almost no customer within 2 km (0–1 km, 0.00 %; 1–2 km, 0.29 %). The highest concentration of customer transactions occurs within the 10–12 km radius (19.50 %) with smaller aggregate transactions intervening (4–5 km, 3.69 %; 5–6 km, 3.57 %; 6–7 km, 2.80 %; 7–8 km, 2.66 %; 8–9 km, 4.39; 9–10 km, 6.90 %). The customers also increase positively toward 20 km (12–14 km, 6.60 %; 14–16 km, 8.77 %; 16–18 km, 9.33 %; 18–20 km, 8.98 %).

The following is an analysis of the overlap of those geographical locations between the three shops. Figure 4(a), (b), and (c) visualize the concentration of customer transaction to geographical locations. Figure 5(a), (b), and (c) show the overlap of those concentrations between PC and AD, and PC and AP, and PA and AD, respectively.

As we can see, PA's trading area is sometimes overlapped with that of shops AD and PC. For the former case, it is southwest of Barcelona, and for the latter case, it is northwest of Barcelona. This indicates that those two shops (i.e., PA and AD, and PA

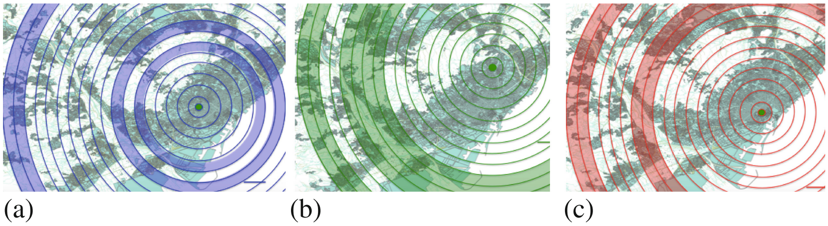


Fig. 4. (a) The visualization of the peaks of the number of transactions for the shop AD. (b) PA. (c) PC.

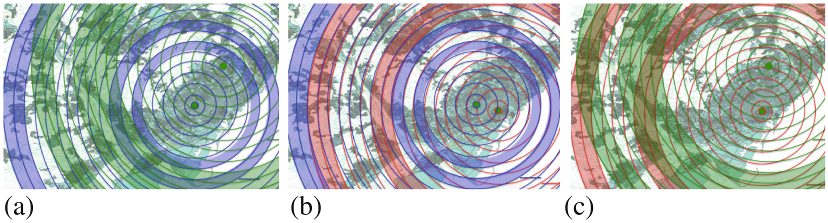


Fig. 5. (a) The visualization of the peaks of the number of transactions for the shop PC and AD. (b) PC and AP. (c) PA and AD.

and PC) compete for their trading area rather than complement each other in the city. Conversely, the trading areas between shops AD and PC are nonoverlapping. They are clearly separated, meaning that harmonious operations are achieved by each shop despite the proximity between them.

All these facts uncover the hidden structures of shops' trading areas and their similarities at the micro scale. Each shop has unique concentrations of customer transactions.

5.2 Customers' Spatial Distributions in the Macro Scale

This section analyzes the customers' origins and destinations for each store over the wider territory. The goal is to detect the macroscopic trading area through spatial analysis. The difference from the previous section is the scale. While the previous section examined it within the city of Barcelona, this section focuses on the wider territory over the city.

We compute the cumulative number of transactions made by the leaving and incoming customers against the distance from the focal shops. December, January and July show significantly larger number than other months for all three cases. This result coincides with previous studies where those three months mean a high season through a year in Spain. In addition, this result shows that an individual shop's attractivity seems dynamic rather than static depending on the season.

Conversely, we also compute the cumulative distribution of transactions against the distance from the shop (see Fig. 6(a)). They show that incoming and leaving customers of each shop have a particular pattern in terms of distributions of locations where customers make the consecutive transactions. For instance, shop PC and shop PA present the sudden increase in transactions around 14 km, while shop AD's happened at 7 km. With respect to shop PC, the slope starts to decrease at around 15 km, and 14 km in the case of shop PA. In addition, Fig. 6(b) presents that log-log plot of the number of transactions against the distance from the shop (Table 1).

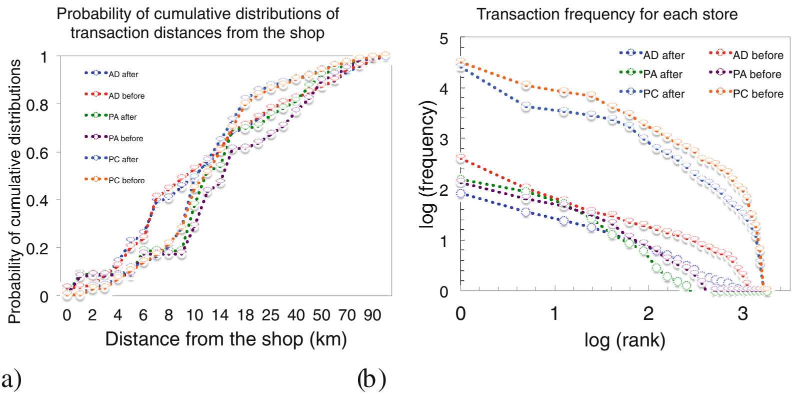


Fig. 6. (a) The distance from the shop where transactions are made against the cumulative frequency of the normalized number of transactions of leaving/incoming customers. (b) The transaction frequencies for each rank of distance from the shop.

Table 1. The slope of the line of best fit for each log ranked customers' frequency vs distance during the high seasons.

	January	July	December
AD after	−0.8682	−0.848	−0.7961
AD before	−0.864	−0.864	−0.7998
PA after	−1.0646	−1.0183	−1.1348
PA before	−1.0362	−1.0362	−1.0113
PC after	−1.3679	−1.2859	−1.2432
PC before	−1.2729	−1.231	−1.3701

Let's examine the log-log plot of the spatial distribution of the number of transactions in each month. Figure 6(b) presents transaction frequencies for each ranked distance bin for the entire period, and Fig. 7 presents the change of the slope of its rank plot by each month. We can observe that both pre- and posttransaction in shops AD and PA is nearly-1.0 in January, July, and December, which corresponds with the high seasons. This indicates that few locations have a much higher number of transactions, while most locations have very few transactions. And this tendency is even stronger in shop PA than in shop AD and PC. Most of PA's customers tend to derive from a minimal number of places and subsequently move to few locations. Conversely, the origin and destination shops for PC's customers become largely dispersed in January, July, and December compared to other months.

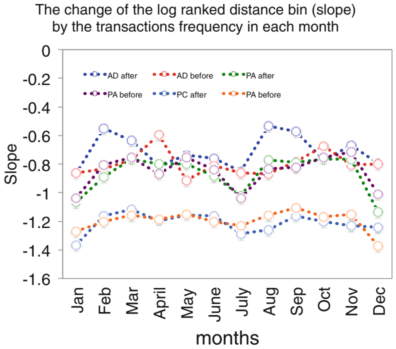


Fig. 7. The change of the log ranked distance bin (slope) by the transactions frequency in each month.

We can see from these results that customer transaction activities have unique patterns in terms of their spatial distribution, which are unique to each individual shop. We speculate that PA might attract local customers rather than tourists from far away. This explains that the origin as well as the destination of their customers is quite similar, and those few places are the main sources for their customers. Conversely, PC appears to attract tourists rather than local citizens, and this tendency is magnified during the high seasons of the year. The customer origin and destination become more dispersed throughout the discount season.

We tend to consider that high seasons increase the number of transactions since many drastic discounts cause customers to rush to shops even from abroad. Our result partially reveals this phenomenon in the case of shop PC, but this is not a consistent pattern among all stores. On the other hand, we showed that the number of transactions during the high season has the same proportion as the low season, meaning that the former portrays an increase in transaction volume compared to the latter. That is, the spatial distribution of transaction activities is exactly the same between the high and low seasons. However, the cause of this increase varies largely depending on the specific store and its location. In case of PA, this effect is not due to the increase of customers who come from other places but simply an increase of the quantitative volume from the same places. Contrary to this fact, in the case of PC, this effect is largely due to the ones deriving from other places, indicating that the simple increase of the same customers from the same locations does not apply in this case.

6 Conclusions

This paper uncovers customers' spatial distributions by analyzing their mobility patterns. We extract locations of consecutive transactions made by customers before and after going to one of the selected three focal shops.

These shops, PC, AD, and PA, are each located in a different urban context across the city of Barcelona, thereby uncovering unique characteristics of their customers as well as the area they are located in. The large-scale and anonymized credit card transaction dataset makes it possible to analyze the successive chains of a customer's purchase history between shops dispersed over the territory rather than an analysis inside a single unique shop.

Our findings reveal that the trading area of each store is largely distributed in a specific way. Customers of shops AD and PC derive from similar places, resulting in competition to attract said customers from each other. Conversely, customers of shops AD and PC share no overlap within the city, allowing them to coexist rather than compete.

In addition, we discover that some distributions of the number of transactions against the distance from the shop follows a power law. This reveals that few locations have higher frequencies of transactions, while most of them have very few transactions. This tendency is amplified even further in shop PA compared to AD or PC. Moreover, our analysis discloses how transaction volumes increase during high and low season. Specifically, customers during high seasons come from similar places rather than from different locations in the case of shop PA. The number of transactions in the former just increases from a similar place in proportion with the ones for the latter, meaning that the customer's spatial distribution is exactly the same for both. However, in the case of shop PC, the customer's mobility pattern is different. The origin and destination of shop PC's customers become dispersed during the high season rather than converged as in the low season.

The outcome is almost reversed between shops PC and PA, although they are the same chain of the large-scale department store. We speculate that this feature might be due to the geographical and sociocultural context of each store. While shop PA is situated in the suburban area with a higher rate of immigration, shop PC is located at the center of the city, which is one of the most popular touristic places.

We have an intuition that urban contexts and their differences cause the feature of stores and their customers to differ. For instance, the store located at a tourist setting may attract many more tourists compared to one in a business or suburban district, and vice versa. In spite of these beliefs, this paper reveals this difference quantitatively through the spatial analysis based on large-scale dataset.

All of these analyses were not possible prior to our research. The previous researchers have frequently used the Huff model [21, 22] to estimate the trading area of a shop in a macroscopic point of view. This merely reveals the homogeneous distribution of customer home locations and the strength of the shop's attractivity, since the model simply depends on the distance from and the size of the shop. Thus, the result of the analysis doesn't represent heterogeneous customers and their geographical features, or the temporal factors. Also, this information is not possible with active mobile phone tracking with or without GPS [23, 24], or with passive mobile phone tracking [12] and Bluetooth detection techniques [20]. The dataset collected by those methods just provide the users' locations without considering evidence of their purchases. Thus, we are only able to predict when purchases are made with a series of significant assumptions. The combination of RFID [14] and the POS system is proposed to reveal a relationship between sales volumes made by customers and their mobility patterns. However, it is possible only inside a single store or mall.

Our proposed methodologies should address these drawbacks. Our dataset permits us to analyze the customer's consumer behaviors across different retail shops, which are dispersed in the urban area; thus, we reveal subsequent purchase behaviors while considering their mobility aspects when they complete microscopic transaction activities. This means that our current research shows the locations of customer transactions rather than just customers passing through these shops. In addition, our methodology and analysis can reveal the individual shop's attractivity and its influences in the territory as trading areas in the micro scale. Furthermore, our methodology and extracted knowledge are extremely helpful in improving Christaller's urban centrality model [25] and reveal the urban structure as well as its hierarchy. Although spatial structure and hierarchy of cities by size and distance have been well studied [25, 26], "the regularity of the urban size distribution poses a real puzzle, one that neither our approach nor the most plausible alternative approach to city sizes seems to answer" (page 219 in [27]).

These extracted patterns help improve spatial arrangements and services offered to customers. Thus, retail shops and their districts can improve sales as well as their environment, thereby revitalizing the center of the urban districts. In addition, these findings are useful to urban planners and city authorities in revitalizing deteriorated districts or rehabilitating neighborhoods. Understanding customers' sequential movement with transaction activities enables us to identify potential customer groups and their geographical demographics spatially. Finally, city planners can consider optimizing the infrastructures and the locations of the retail shops to make the district more attractive and active by increasing the number of pedestrians. For instance, the customers' sequential movement between different retail shops facilitates collaboration between all shops in a district as a whole rather than individually, to organize planned sale periods. Based on our findings, neighborhood associations can organize discount coupons or advertisements in relevant and adequate places. This can serve as an efficient indicator as to when they are most likely to complete transactions as well as their successive locations.

Acknowledgments. We would like to thank the Banco Bilbao Vizcaya Argentaria (BBVA) for providing the dataset for this study. Special thanks to Juan Murillo Arias, Marco Bressan, Elena Alfaro Martinez, Maria Hernandez Rubio and Assaf Biderman for organizational support of the project and stimulating discussions. We further thank MIT SMART Program, Accenture, Air Liquide, The Coca Cola Company, Emirates Integrated Telecommunications Company, The ENEL foundation, Ericsson, Expo 2015, Ferrovial, Liberty Mutual, The Regional Municipality of Wood Buffalo, Volkswagen Electronics Research Lab and all the members of the MIT Senseable City Lab Consortium for supporting the research.

References

1. Jacobs, J.: The Death and Life of Great American Cities. Random House, New York (1961)
2. Gehl, J.: Life Between Buildings: Using Public Space. Island Press, Washington-Covelo-London (2011)
3. Taneja, S.: Technology moves in. Chain Store Age **75**, 136–138 (1999)
4. Porta, S., Latora, V., Wang, F., Rueda, S., Strano, E., Scellato, S., Cardillo, A., Belli, E., Càrdenas, F., Cormenzana, B., Latora, L.: Street centrality and the location of economic activities in Barcelona. Urban Stud. **49**(7), 1471–1488 (2012)

5. Krumme, C., Llorente, A., Cebrian, M., Pentland, A., Moro, E.: The predictability of consumer visitation patterns. Sci. Rep. **3**, 1645 (2013). doi:10.1038/srep01645
6. Sobolevsky, S., Sitko, I., Grauwin, S., des Combes, R.T., Hawelka, B., Arias, J.M., Ratti, C.: Mining urban performance: scale-independent classification of cities based on individual economic transactions (2014). arXiv:1405.4301
7. Sobolevsky, S., Sitko, I., des Combes, R.T., Hawelka, B., Arias, J.M., Ratti, C.: Money on the move: big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in spain. Big Data (BigData Congress) In: 2014 IEEE International Congress, pp. 136–143 (2014)
8. Leenheer, J., Bijmolt, Tammo, H.A.: Which retailers adopt a loyalty program? an empirical study. J. Retail. Consum. Serv. **15**, 429–442 (2008)
9. González, M.C., Hidalgo, C.A., Barabási, A.L.: Understanding individual human mobility patterns. Nature **453**, 779–782 (2008)
10. Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., Pujolle, G.: Estimating human trajectories and hotspots through mobile phone data. Comput. Netw. **64**, 296–307 (2014)
11. Kung, K.S., Greco, K., Sobolevsky, S., Ratti, C.: Exploring universal patterns in human home/work commuting from mobile phone data. PLoS ONE **9**(6), e96180 (2014)
12. Ratti, C., Pulselli, R., Williams, S., Frenchman, D.: Mobile landscapes: using location data from cell phones for urban analysis. Environ. Plan. B Plan. Des. **33**(5), 727–748 (2006)
13. Sobolevsky, S., Szell, M., Campari, R., Couronné, T., Smoreda, Z., Ratti, R.: Delineating geographical regions with networks of human interactions in an extensive set of countires. PLoS ONE **8**(12), e81707 (2013)
14. Kanda, T., Shiomi, M., Perrin, L., Nomura, T., Ishiguro, H., Hagita, N.: Analysis of people trajectories with ubiquitous sensors in a science museum. In: Proceedings 2007 IEEE International Conference on Robotics and Automation (ICRA 2007), pp. 4846–4853 (2007)
15. Larson, J., Bradlow, E., Fader, P.: An exploratory look at supermarket shopping paths. Int. J. Res. Mark. **22**(4), 395–414 (2005)
16. Delafontaine, M., Versichele, M., Neutens, T., Van de Weghe, N.: Analysing spatiotemporal sequences in Bluetooth tracking data. Appl. Geogr. **34**, 659–668 (2012)
17. Kostakos, V., O'Neill, E., Penn, A., Roussos, G., Papadongonas, D.: Brief encounters: sensing, modelling and visualizing urban mobility and copresence networks. ACM Trans. Comput. Hum. Interact. **17**(1), 1–38 (2010)
18. Versichele, M., Neutens, T., Delafontaine, M., Van de Weghe, N.: The use of bluetooth for analysing spatiotemporal dynamics of human movement at mass events: a case study of the ghent festivities. Appl. Geogr. **32**, 208–220 (2011)
19. Yoshimura, Y., Girardin, F., Carrascal, J.P., Ratti, C., Blat, J.: New tools for studing visitor behaviours in museums: a case study at the louvre. In: Fucks, M., Ricci, F., Cantoni, L. (eds.) Information and Communication Technologies in Tourism 2012, pp. 391–402. Springer, New York (2012)
20. Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardin, F., Carrascal, J.P., Blat, J., Sinatra, R.: An analysis of visitors' behaviour in The Louvre Museum: a study using Bluetooth data. Environ. Plan. B Plan. Des. **41**(6), 1113–1131 (2014)
21. Huff, D.L.: Defining and estimating a trade area. J. Mark. **28**, 34–38 (1964)
22. Huff, D.L.: A programmed solution for approximating an optimum retail location. Land Econ. **42**, 293–303 (1966)
23. Asakura, Y., Iryo, T.: Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. Transp. Res. Part A Policy Pract. **41**(7), 684–690 (2007)

24. Shoval, N., McKercher, B., Birenboim, A., Ng, E.: The application of a sequence alignment method to the creation of typologies of tourist activity in time and space. Environ. Plan. B Plan. Des. **42**(1), 76–94 (2013)
25. Christaller, W.: Central Places in Southern Germany. English edition, 1966, translated by Carlisle W., Baskin. Englewood Cliffs, Printice-Hall, New Jersey (1935)
26. Losch, A.: The Economics of Location (translated by Woglam W H). Yale University Press, New Haven (1954)
27. Fujita, M., Krugman, P., Venables, A.: The Spatial Economy-Cities, Regions and International Trade. MIT Press, Cambridge (1999)

Case Studies for Data-Oriented Emergency Management/Planning in Complex Urban Systems

Kun Xie[1], Kaan Ozbay[1(✉)], Yuan Zhu[1], and Hong Yang[2]

[1] Department of Civil and Urban Engineering, Urban Mobility and Intelligent
Transportation Systems (UrbanMITS) Laboratory, Center for Urban Science and
Progress (CUSP), New York University, Brooklyn, NY 11201, USA
{kun.xie, kaan.ozbay, yuan.zhu}@nyu.edu
[2] Department of Modeling, Simulation and Visualization Engineering, Old
Dominion University, Norfolk, VA 23529, USA
hyang@odu.edu

Abstract. To reduce the losses caused by natural disasters such as hurricanes, it is necessary to build effective and efficient emergency management/planning systems for cities. With increases in volume, variety and acquisition rate of urban data, major opportunities exist to implement data-oriented emergency management/planning. New York/New Jersey metropolitan area is selected as the study area. Large datasets related to emergency management/planning including, traffic operations, incidents, geographical and socio-economic characteristics, and evacuee behavior are collected from various sources. Five related case studies conducted using these unique datasets are summarized to present a comprehensive overview on how to use big urban data to obtain innovative solutions for emergency management and planning, in the context of complex urban systems. Useful insights are obtained from data for essential tasks of emergency management and planning such as evacuation demand estimation, determination of evacuation zones, evacuation planning and resilience assessment.

Keywords: Emergency management/planning · Complex urban systems · Big data · Evacuation modeling · Hurricane

1 Introduction

Hurricanes can have devastating effects on coastal areas due to flooding, high wind, and rainfall, resulting in serious loss of life and property. To reduce the losses caused by hurricanes, it is necessary to build effective and efficient emergency management/planning systems. Essential tasks of emergency management/planning include determination of evacuation zones (identify evacuation zones in a way to indicate its inhabitants whether or not they are prone to hurricane-related risk in advance of disaster impacts), evacuation demand estimation (estimate origins, destinations and numbers of evacuees based on evacuation zones, demographic features and evacuation behavior), evacuation planning (determine the evacuation time, destinations and routes based on

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 190–207, 2016.
DOI: 10.1007/978-3-662-53416-8_12

the evacuation demand), and resilience assessment (evaluate the recovery ability of transportation systems in the post-hurricane periods).

The complexity of urban systems creates challenges for emergency management/planning. The urban transportation systems are multimodal, generally composed of the highway system, the pedestrian system and the public transit system. The urban transportation systems are further complicated by the random occurrences of incidents such as accidents, disabled vehicles, debris, downed trees and flooding. Therefore, it is challenging to evaluate the carrying capacities of urban transportation systems, especially during the hurricane-impacted periods when hurricane-related incidents such as downed trees and flooding are more likely to happen. On the other hand, it is difficult to precisely estimate the evacuation demands which are closely related to the evacuation zone divisions and evacuation behavior. The determination of evacuation zones is associated with a variety of factors such as ground elevation, evacuation mobility and demographic features. Moreover, different evacuation behavior (e.g. whether to evacuate or not, how to evacuate and where to evacuate) is present among inhabitants who are prone to hurricane-related risks.

In the era of "Big Data", with increases in volume, variety and acquisition rate of urban data, there are a number of very exciting opportunities to implement data-driven emergency management/planning. Massive amounts of digitalized data such as evacuation zone maps, past incidents, geographical features, historical highway traffic volumes, public transit ridership can be available from multiple sources. Useful insights can be obtained from this big urban data for performing essential tasks of emergency management/planning. Therefore, this paper aims to present a comprehensive overview on how to use the big urban data to provide solutions and innovations for emergency management/planning in the context of complex urban systems.

New York City (NYC) is vulnerable to hurricanes. According to NYC Office of Emergency Management (OME), NYC has about 600 miles of coastline and almost 3 million people living in the areas at the risk of hurricanes [1]. In the morning of August 28th, 2011 hurricane Irene made landfall at Coney Island, NYC and in the evening of October 29th, 2012 hurricane Sandy landed in New Jersey. Hurricanes Irene and Sandy caused significant devastation to the east coast (especially to NYC), but also provide valuable data for the research on emergency management/planning. Moreover, New York City's open data policy makes a variety of datasets from government agencies available to the public. NYC and its surrounding regions are selected as the study areas.

2 Big Urban Data

Massive amounts of data from multiple sources are collected to support data-oriented emergency management/planning. The major datasets are classified into eight groups including evacuation management data, traffic incident data, taxi and subway trip data, traffic volume and demand data, evacuation survey data, geographical data, building damage data and socio-economic data. The sources and practical usage of those datasets are summarized in Table 1, and more detailed descriptions are introduced in the following subsections.

Table 1. Summary of sources and usages for datasets collected

Dataset	Source	Usage
Evacuation management	NYC Office of Emergency Management (OEM)	Estimate the demand for evacuation and destination choices of evacuees
Traffic incident	Transportation Operations Coordination Committee (TRANSCOM)	Estimate incident-induced capacity losses
Taxi and subway trip	NYC Taxi & Limousine Commission (TLC) and Metropolitan Transportation Authority (MTA)	Calibrate and validate the evacuation models as well as assess the resilience of transportation systems
Evacuation survey	Northern New Jersey evacuation survey	Analyze the behavior of evacuees
Geographical	National Elevation Dataset (NED)	Determine the division of evacuation zones
Building damage	Environment Systems Research Institute (ESRI)	Additional indicator for risk evaluation
Socio-economic	U.S. Census Bureau	Estimate the evacuation demand and the division of evacuation zones

2.1 Evacuation Management Data

NYC Office of Emergency Management (OEM) provides Hurricane Evacuation Zones Map[1] (downloadable as GIS shapefiles) to help residents make decisions on evacuation. Evacuation zone division was updated in 2013 after Hurricane Sandy, adding 600,000 New Yorkers not included within the boundaries of the former 2001 evacuation zones. The zone division is updated according to the empirical data during Hurricane Sandy and storm surge simulations which are based on the current climate situation. The 2013 evacuations zones are listed from zone 1 to zone 6, from the highest risk to the lowest risk. Evacuation centers which offer shelters to evacuees during hurricanes are also presented in the Hurricane Evacuation Zones Map. Evacuation zones can be used to estimate the demand for evacuation and the locations of evacuation centers are related with the destination choices of evacuees.

2.2 Traffic Incident Data

Incident data of the interstate, US and New York State highways in New York City and its surrounding areas from Oct. 1st 2012 to Jan. 31st 2013 were obtained from Transportation Operations Coordination Committee (TRANSCOM). More detailed description of this dataset is given in [2]. A total of 354 incidents occurred during the evacuation period (12 AM, Oct. 26th, 2012–12 PM, Oct. 29th, 2012) before Sandy's

[1] Source: http://maps.nyc.gov/hurricane/.

landfall. Those incidents can be classified as six different types including accident, debris, disabled vehicle, downed tree, flooding and others. Accidents and downed trees are the major incident types during evacuation the evacuation period, and account for over 50 % of all the incidents. The incident durations were computed using the fields of *create time* and *close time* in the incident records. Each incident was located in the GIS map according to its coordinates and then was matched to the highway where it was detected. Incident data can provide information on the highway capacity losses which are attributed to the occurrence of incidents right before and during hurricanes.

2.3 Taxi and Subway Trip Data

Taxi trip data of NYC is made available to public by NYC Taxi & Limousine Commission (TLC) [3, 4]. The dataset includes taxi trips from years 2010 to 2013 and it contains pick-up and drop-off time and location information. The taxi trips generated is approximately 175 million per year. Subway ridership data were obtained from Metropolitan Transportation Authority (MTA) turnstile dataset, which includes subway turnstile information since May, 2010 and is updated every week. The data is stored in txt format and available through an official data feed [5]. The data is organized by weeks, remote units (stations) and control areas (turnstiles). Each station can have multiple control areas, and for each turnstile, there are two increment counters used to record numbers of entries and exits. Typically, counter readings of each turnstile is recorded every four hours. Taxi and subway trip data are used to calibrate and validate the evacuation models as well as to assess the resilience of transportation systems.

2.4 Traffic Volume and Demand Data

NY Best Practice Model (NYBPM) [6], which covers 28 counties in the Tristate area and involves more than 22 million population, provide well-calibrated background traffic demand trip tables. In addition, the traffic volumes on the main interstate highways, US highways, and NY highways in the NYC and surrounding regions were obtained from TRANSCOM. The traffic volumes obtained from traffic sensors were used to build evacuation response curves [7] for critical corridors during evacuation period of Hurricane Sandy.

2.5 Evacuation Survey Data

A random digit dial telephone survey was conducted between August and October of 2008 in northern New Jersey [7]. It covers a large urban region consisting of Passaic, Bergen, Hudson, Morris, Essex, Middlesex and Union Counties. The total population of the region is approximately 4.5 million. In total, 2,218 households were interviewed with a set of questions related to their evacuation experience, disaster preparedness (including hurricane, industrial accident and catastrophic nuclear explosion), evacuation decision choices, evacuation destinations, and evacuation mode choices. In addition, a series of questions regarding the characteristics of the household and household members, such as income, vehicle ownership, family size etc. were asked. The evacuation survey data can be used to analyze the behavior of evacuees and thus more accurate evacuation demand can be obtained.

2.6 Geographical Data

Digital Elevation Model (DEM) data of NYC provides a representation of the terrain with elevations above the ground in a regular raster form. The DEM data of Manhattan was extracted from National Elevation Dataset (NED) developed by U.S. Geological Survey (USGS)[2]. The resolution of the DEM data is 1 arc second (about 90 feet) and the pixel values are elevations in feet based on North American Vertical Datum of 1988 (NAD83). The average elevation which is associated with the flooding risk was aggregated for each grid cell. Another geographic feature collected for each cell is the distance to the coast, since areas closer to the coast are more likely to be affected by the storm surges. Geographical data can be used to infer the division of evacuation zones.

2.7 Building Damage Data

The building damage record during Hurricane Sandy was achieved from the Environment Systems Research Institute (ESRI) datasets[3]. Federal Emergency Management Agency (FEMA) inspectors conducted field inspections of damaged properties and recorded relevant information such as location and damage level, when households applied for individual assistance. The number of damaged building was obtained by summarizing households in the same location, assuming they are from a single multi-family building. Buildings damaged in historical hurricanes can be used as an additional indicator for risk evaluation.

2.8 Socio-economic Data

The socio-economic data based on 2011 census survey was retrieved from U.S. Census Bureau[4]. The socio-economic data is composed of demographic features (e.g. total population, population under 14 and population over 65), economic features (e.g. employment and median income), and housing features (e.g. median value and household average size). The demographic features can be used to estimate the evacuation demand. In addition, socio-economic data can affect the division of evacuation zones. For example, the zones with large number of elderlies and children tend to be more vulnerable and should be given higher priority of evacuation.

[2] Source: http://ned.usgs.gov/.

[3] Source: http://www.arcgis.com/home/item.html?id=307dd522499d4a44a33d7296a5da5ea0.

[4] Source: http://factfinder.census.gov.

3 Data-Oriented Emergency Management/Planning

This section presents five case studies on how to use big urban data to gain useful insights for decision-making in emergency management/planning. The main purposes and key datasets used for each case study are listed in Table 2. Those five cases studies are all data-oriented and related with each other. The evacuation behavior analysis and evacuation zone prediction can be used to estimate the evacuation demand; while the incident analysis provide information on the uncertainties of capacity supply of transportation systems. Evacuation simulation is used to evaluate whether the capacity supply could accommodate the evacuation demand under different evacuation scenarios. Resilience assessment is post-evaluation on the recovery ability of transportation systems.

Table 2. Summary of case studies in data-oriented emergency management/planning

Case study	Main purpose	Key datasets used
Evacuation behavior analysis	Estimate evacuation demand	Evacuation survey data
Evacuation zone prediction	Identify evacuation zones	Evacuation management data, geographical data, building damage data, and demographic data
Traffic Incident analysis	Predict capacity related uncertainties	Traffic incident data
Evacuation simulation	Evaluate whether the capacity supply can accommodate the evacuation demand	Evacuation management data, taxi and subway data, traffic volume and demand data and demographic data
Resilience assessment	Evaluate the recovery ability of transportation systems	Evacuation management data and taxi and subway data

3.1 Evacuation Behavior Analysis

A key issue in evacuation studies is to understand the evacuation behavior of residents. Questions related to whether to evacuate, when to evacuate, how to evacuate, where to evacuate, etc. are critical in developing reasonable evacuation plans. Thus it is necessary to examine the factors that affect the evacuees' decisions regarding these questions. Questionnaires have been designed to interview the residents and aim to identify the underlying factors affecting their decision makings (please see the subsection "Evacuation Survey Data" for more details). Based on the surveyed results, statistical models such as logistic regressions, multinomial logit models, etc. have been developed to examine the key factors affecting the decisions. Factors such as the socio-economic and demographic characteristics of the evacuees, locations, and type of the extreme events (i.e. hurricanes/explosions) are often considered in the modeling process. The advanced models usually help improve our predictions for evacuation planning. However, in practices, many models were developed independently.

They did not account for the potential interactions among different evacuation behavior. In the decision-making process, many evacuees are likely to make their choices on a question conditional on the decisions for other questions. Thus there is necessity to examine the issue considering possible interactions among different evacuation behavioral responses.

As a pilot study, we have applied the dataset from the telephone survey [8] to investigate the relationship between evacuation decision (the preference to evacuate) and evacuation destination choices under the hurricane scenario. For the responses of evacuation decision, the ordered probit regression model has been proposed as the responses are ordered in terms of multilevel preference:

$$y_i^* = X_i^{'}\beta + \varepsilon_i$$
$$y_i = \begin{cases} 1 \quad if\, \tau_0 < y_i^* \leq \tau_1 \quad \text{(Response = very unlikely)} \\ 2 \quad if\, \tau_1 < y_i^* \leq \tau_2 \quad \text{(Response = not very likely)} \\ 3 \quad if\, \tau_2 < y_i^* \leq \tau_3 \quad \text{(Response = somewhat likely)} \\ 4 \quad if\, \tau_3 < y_i^* \leq \tau_4 \quad \text{(Response = very likely)} \end{cases} \tag{1}$$

where y_i^* denotes the latent variable measuring the evacuation decision of the i^{th} interviewed person; X_i is a vector of observed non-random explanatory variables; β is a vector of unknown parameters; and ε_i is the random error term. The latent variable y_i^* is mapped to the observed variable y_i, according to threshold parameters τ_j‘s, with $\tau_{j-1} < \tau_j$, $\tau_0 = -\infty$, and $\tau_J = +\infty$.

In addition, the choices on the potential evacuation destinations were modeled by the multinomial logit model. Given one choice as a reference (i.e., public shelter), the probability of each choice π_{ij} is compared to the probability of the reference choice π_{iJ}. For choices $j = 1, 2, \ldots J - 1$, the log-odds of each choice is assumed to follows linear model:

$$\eta_{ij} = \log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = Z_i^{'}\alpha_j \tag{2}$$

where $Z_i^{'}$ is a vector of explanatory variables and α_j is a vector of regression coefficients for each choice $j = 1, 2, \ldots, J - 1$. To identify the potential relationship between the evacuation decision and the choice of the evacuation destinations, we have proposed the use of the structural equation modeling, where the evacuation decision y_i is used as one of the explanatory variable in evacuation destination model (Eq. (2)). More detailed description of the proposed approach is reported in our recent work (Yang et al. [9]). An example of the structure equation modeling process is shown in Fig. 1. Though only two behavioral responses have been examined in the pilot study, the proposed method can be extend to examine more complicated interactions among multiple types of behavioral responses.

The key factors that affect the evacuation decision as well as the evacuation destination choices have been determined through a Bayesian estimation approach, which

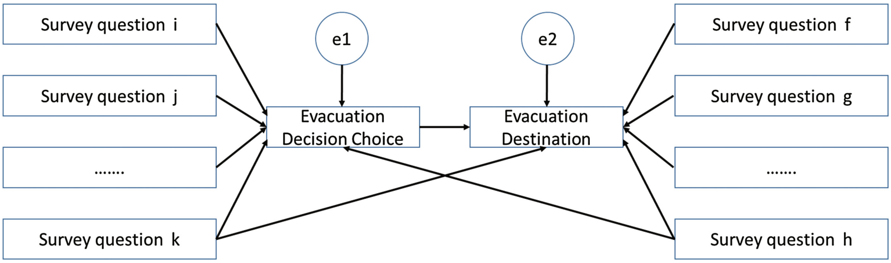


Fig. 1. Sample structural equation modeling process to explore multiple behavioral responses.

is not detailed here (See Yang et al. [9]). Other than the conventional factors such as age and distance to the shore, the modeling results suggest that there is only weak relationship between the evacuation decision choices and the evacuation destination choices. In other words, whether or not the individuals consider to evacuate, the decisions on choosing public shelters as well as other places as their evacuation destinations will not change notably based on the surveyed data.

3.2 Evacuation Zone Prediction

It is important for emergency planners to define evacuation zones which can indicate inhabitants whether or not they are prone to hurricane-related risk in advance of disaster impacts. The delineation of evacuation zones can be used to estimate the demand of evacuees, and thus it is helpful in developing effective evacuation management strategies. The evacuation zones defined currently cannot remain the same in the future, since the long-term climate change such as the rise of sea level would have major impacts on hurricane-related risks. One notable factor of climate change is global warming and the resulting rise of sea level. To manage emergency resources more efficiently, it is important to update the delineation of current evacuation zones to make it adaptable to the future hurricanes.

To predict future evacuation zones, traditional methods rely on the estimation of surge flooding using models such as the SLOSH (sea, lake, and overland surges from hurricanes) model and the ADCIRC (a parallel advanced circulation model for oceanic, coastal, and estuarine waters) model [10]. However, the implementation of the SLOSH and ADCIRC models can be really time-consuming and costly. We aim to develop a novel data-driven method which can promptly predict future evacuation zones in the context of climate change. Machine learning algorithms are used to learn the relationship between current pre-determined evacuation zones and hurricane-related factors, and then to predict how those zones should be updated as those hurricane-related factors change in the future.

The map of Manhattan, which is the central area of NYC, was uniformly split into 150 × 150 feet^2 grid cells (N = 25,440) as the basic geographical units of analysis.

Evacuation zone category (E1, E2, E3 and S)[5], geographical features (including average elevation above sea level and distance to coast), historical hurricane information (including building damage intensity), evacuation mobility (including distance to the nearest evacuation center, distance to the nearest subway station, distance to the nearest bus stop and distance to the nearest expressway), and demographic features (including total population, population over 65 and population under 14) in the current year were captured for each cell. A decision tree and random forest were trained to relate cell-specific features with current zone categories which could reflect the risk levels during storms. Ten-fold cross-validation was used to evaluate model performance and performance measures of the classification tree and the random forest are reported in.

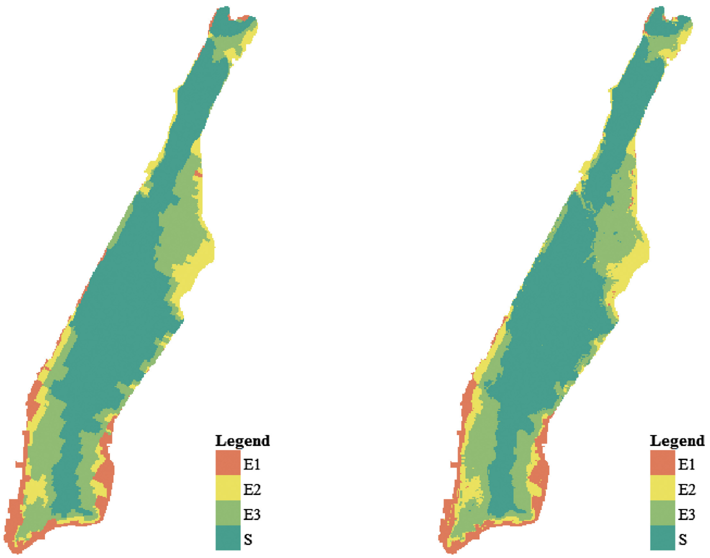
Table 3. It was found that the random forest outperformed the decision tree in term of the accuracy and Kappa statistic [11]. Regarding the better performance, the prediction outcomes of the random forest are visualized in the GIS map and compared with actual evacuation zones as presented in Fig. 2. It is found that the estimated evacuation zone division is quite similar to the actual one (accuracy = 94.13 %). It implies that the random forest succeeds in learning the potential pattern of delineating zones with different risk levels. More details on description and specification of the proposed models are presented in our recent work (Xie et al. [12]).

Table 3. Performance measures of the classification tree and the random forest

	Classification tree	Random forest
Correctly classified instances	22965	23947
Incorrectly classified instances	2475	1493
Total number of instances	25440	25440
Accuracy	90.27 %	94.13 %
Kappa statistic	0.8420	0.9049

The sea level rises in the future were also estimated based on emission scenario Representative Concentration Pathway (RCP) 8.5 [13]. The RCP 8.5 scenario assumes that little coordinated actions are made among countries, so that the climate radiative forcing to the atmosphere from anthropogenic emissions is as high as 8.5 watts per square meter over the globe. The upper 95 % bounds of sea levels are estimated to be 36.3 inches for the 2050s and 45.1 inches for the 2090s. As a result of climate change, the terrain elevation above the sea level is expected to decrease. This will lead to a higher flooding risk and thus the evacuation zone categories need to be updated accordingly. The proposed random forest is used to predict the evacuation zones for the 2050s and 2090s, based on the expected decrease in average elevation above the sea level and assumption that other hurricane-related characteristics are kept the same the future.

[5] "E1" corresponding to NYC 2013 evacuation zone 1, "E2" corresponding to NYC 2013 evacuation zone 2 and zone 3, and "E3" corresponding to NYC 2013 evacuation zone 4, zone 5 and zone 6, and "S" corresponding to the safe zone beyond the evacuation region.



(a) (b)

Fig. 2. Current evacuation zones (a) and predicted evacuation zones using the random forest (b).

The predicted future evacuation zones are presented in Fig. 3. Compared with the current zoning, the areas with need of evacuation are expected to expand in the future.

3.3 Traffic Incident Analysis

Incidents are defined here as any occurrence that temporarily reduce highway capacity such as accidents, disabled vehicles and downed trees. Capacity losses caused by incidents are closely related to the incident types, frequencies and durations. The section aims to investigate the characteristics of incidents in the context of hurricane Sandy, and to propose an approach to accommodate the uncertainty of roadway capacities due to incidents.

The incident data used is introduced in subsection “Incident Data” above. As shown in Fig. 4, the proportions of incident types vary greatly between the Sandy week (Oct. 26th, 2012 ~ Nov. 1st, 2012) and the regular time (time intervals before and after the Sandy week). In the Sandy week, the proportions of debris, downed trees, flooding and weather related incidents increased significantly. Meanwhile, there were fewer accidents and disabled vehicles compared with the regular time.

The relationship between incident frequency during the evacuation period of Hurricane Sandy (12 AM, Oct. 26th, 2012–12 PM, Oct. 29th, 2012) and highway characteristics such as road length and traffic volume was investigated. The incident frequency during evacuation for each highway section was obtained. Negative binomial (NB) models can accommodate the nonnegative, random and discrete features of

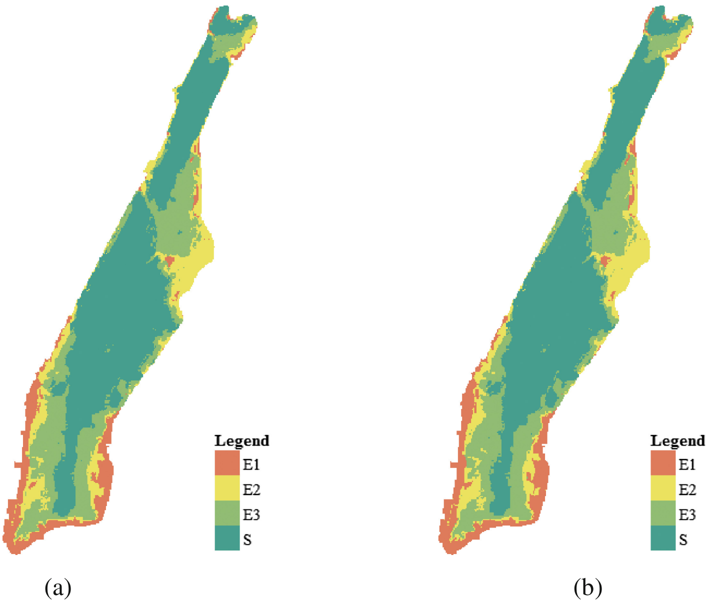


Fig. 3. Predicted evacuation zones for the 2050s (a) and the 2090s (b).

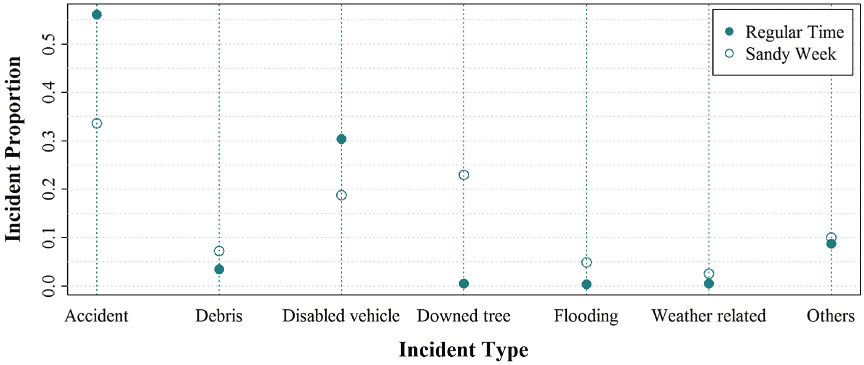


Fig. 4. Proportions of incident types in the regular time and the Sandy week. (Source: Xie et al. (2015) [2])

event frequencies and have been proved better to deal with the over-dispersed data by introducing an error term [14]. A NB model was used to replicate incident frequencies of highway sections, and it can be expressed as follows:

$$\begin{gathered} f_i \sim Negbin(\theta_i, r) \\ \ln(\theta_i) = \alpha X_i \end{gathered} \tag{3}$$

where f_i is the observed incident frequency for freeway section i, θ_i is the expectation of y_i, X_i is the explanatory variables, α is the vector of regression coefficients to be estimated, and r is the dispersion parameter. Results show that the logarithm of traffic volume and the logarithm of highway length are positively associated with the incident frequencies. In addition, more incidents are expected to happen in interstate highways compared with other highways. The developed incident frequency model can be used to predict the probability of incident occurrence for each highway section in the capacity-loss simulation.

Duration distributions vary for different incident types. The relationship between the incident type and duration can be explored using a lognormal model [2, 15]. A lognormal model assumes a linear relationship between the logarithm of incident durations and explanatory variables. It can be expressed as:

$$\begin{aligned} &\ln(d_j) \sim Normal(\mu_j, \sigma^2) \\ &\mu_j = \beta Z_j \end{aligned} \tag{4}$$

where d_j is the observed duration for incident j, μ_j and σ^2 are the mean and variance of the normal distribution, Z_j is the explanatory variables (dummy variables indicating the incident types), β is the vector of regression coefficients to be estimated. Accidents, debris and disable vehicles are expected to have shorter duration than other incidents; while duration of incidents such as downed tree and flooding tend to be shorter. These modeling results can be used to generate the duration for each incident in the capacity-loss simulation.

The incident type proportions, incident frequency and incident duration models developed are used as inputs for simulating incident-induced capacity losses for the whole study network (40442 links) during the evacuation period. Monte Carlo simulation method is used to generate observations randomly from specified distributions [16]. A detailed simulation procedure to generate capacity losses is introduced in our recent paper [17]. The main steps of this novel approach are summarized as:

Step 1: Use the incident frequency model estimate the expectation of incident frequency for each link

Step 2: For each incident, generate incident type according to the type proportions during evacuation period

Step 3: Use the incident duration model to estimate the duration for each incident

The results of the incident simulation can tell us the likely locations of incidents as well as their types and durations. Based on the incident simulation results, the capacity loss of each link can be estimated and used as inputs in the network-wide evacuation simulation.

3.4 Evacuation Simulation

Simulation of hurricane evacuation is an important task in emergency management/planning. However, this process has to face two challenges: (1) how to estimate evacuation demand based on socio-economic characteristics and evacuation zone division; and (2) how to deal with the uncertainty due to the roadway capacity losses because of highway incidents. The evacuation simulation model built in this study incorporates most recent hurricane experiences in the New York metropolitan area.

We propose an hour-by-hour evacuation simulation based on a large-scale macroscopic network model of the New York metropolitan area developed in the TransCAD Software [18]. This model reflects the latest traffic analysis zones (TAZs), road network configuration, and socio-economic data. The procedure for the network-wide evacuation simulation is shown on Fig. 5. Prior to traffic assignment, it is crucial to estimate evacuation demand and generate capacity losses for road network. For demand estimation, the first step is to identify the evacuation zones, then estimate the number of people that need to be evacuated based on the socio-economic data. Generated evacuation demand is distributed to each hour according to the empirical evacuation curve obtained from the traffic volumes observed. Unlike most of the previous studies that assume static highway capacities, we attempt to treat the highway capacities to be stochastic, based on the outcomes from incident-induced capacity loss simulation (as described in the previous subsection). The hour-by-hour capacity losses are simulated for the whole network. Three scenarios are developed, including one base and two evacuation scenarios (one considers incident-induced capacity losses and the other doesn't). Under the base scenario, the trip tables are constructed from the background traffic in the regular time, while under the two evacuation scenarios, the trip tables consist of both assumed background traffic and additional evacuation demand.

We run network assignment model using the quasi-dynamic traffic assignment method described in Ozbay *et al.* [19] for each hour based on different scenarios and obtain results including the performance of network links and evacuation times between each O-D pairs of the study network. At last assignment results are analyzed to determine evacuation times from evacuation zones to safe zones and the performance of the network with and without consideration of capacity losses. Figure 6 shows the zonal travel times for two evacuation scenarios and observed taxi trips. It can be seen that travel times for Harlem and downtown areas are lower than Midtown, and travel times for east side of Manhattan is shorter than the east side for all scenarios. Compared

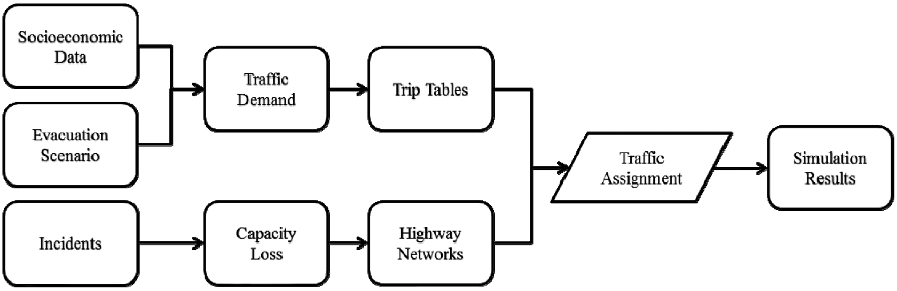
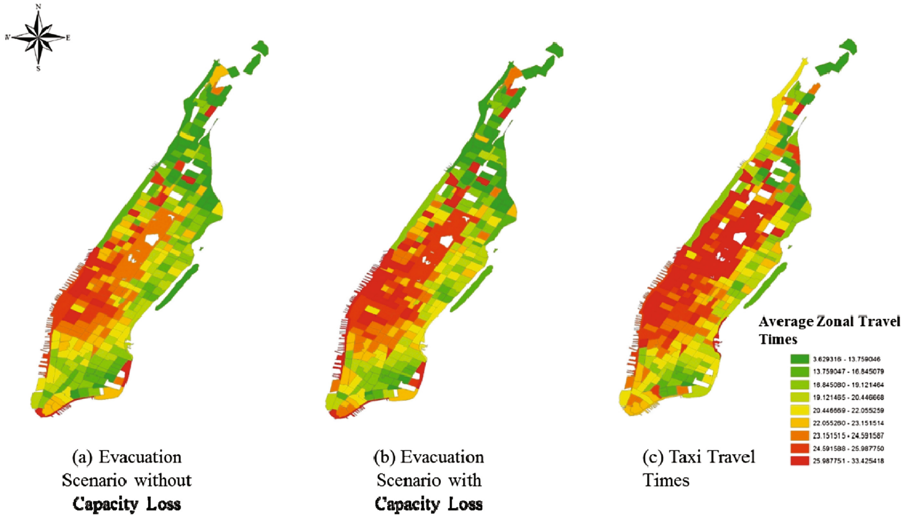


Fig. 5. Network-wide modeling methodology for hurricane evacuation combined with capacity losses due to incidents.

with the scenario with full capacity, the evacuation travel times for capacity loss scenario are significantly higher, and closer to the ones observed from the empirical taxi trip data.

3.5 Resilience Assessment

This subsection evaluates the resilience of roadway and transit systems in the aftermath of hurricanes using large-scale taxi and transit ridership datasets during Hurricanes Irene and Sandy. Recovery curves of subway and taxi trips are estimated for each zone category (evacuation zones 1 ~ 6 and safe zone).



(a) Evacuation Scenario without **Capacity Loss**

(b) Evacuation Scenario with **Capacity Loss**

(c) Taxi Travel Times

Fig. 6. Zonal travel times of Manhattan for (a) evacuation scenario without capacity losses, (b) evacuation scenario with capacity loss and (c) observed taxi trips.

The logistic function is used in modeling process, since characteristics of logistic model resembles evacuation and recovery activities, which are shown to follow an S-shape. Basic logistic function is shown in Eq. (5):

$$P_t = \frac{1}{1 + e^{-\alpha(t-H)}} \tag{5}$$

where P_t represents zonal recovery rate by time t, α is the factor affecting slope of the recovery rate, and H is half recovery time (the time when half of the lost service capacity is restored). According to Yazici and Ozbay [20], α can be regarded as the parameter that controls behavior of evacuees whereas H controls total clearance time ($2H$). So α and H together can be used to determine two factors of resilience, namely, severity of outcome and time for recovery.

Empirical and model estimated recovery curves are visualized in Fig. 7. For more detailed parameter estimates, please refer to a recently study by Yuan et al. [21]. X axis of each subplot range from 0 to 11, which stands for the days elapsed from hurricane impact to the end of the study period. For Hurricanes Irene and Sandy, starting days are August 28, 2011 and October 30, 2012, respectively. As shown in Fig. 7, during Hurricane Irene, the curves for roadway recovery reached one in two days for nearly all the zones. Full recovery of the subway system took longer than the roadway system for most zones. Compared with Hurricane Irene, Hurricane Sandy recovery for both modes required much longer recovery time. Subway system recovery in the case of Sandy is also slower than roadway system. Spatial patterns are also presented in Fig. 7, roadway curves were not fully recovered at the end of study period for zones 1 to 4. For zone 5, roadway system recovered on day 10, zone 6 and Safe zone recovered on Days 6 and 5, respectively. Subway recovery curves remain flat for high-risk zones. With decreasing rates of zonal vulnerability, subway curves become steeper. For zone 1 (refer to subsection "Evacuation Management Data" for zone division details), only 25 % of subway recovery was completed on day 11. Patterns for all other zones are similar, and subway ridership recovered on day 10 or 11.

The above results show that the process of multi-mode post-hurricane recovery can be captured by using logistic functions. The initial recovery rate of zones which are prone to hurricane-related risk such as zone 1 is lower than those of others, and it takes longer time for such zones for full recovery. Road network is found to have better resilience than subway network, since subway recovery has later initial starting point, lower initial

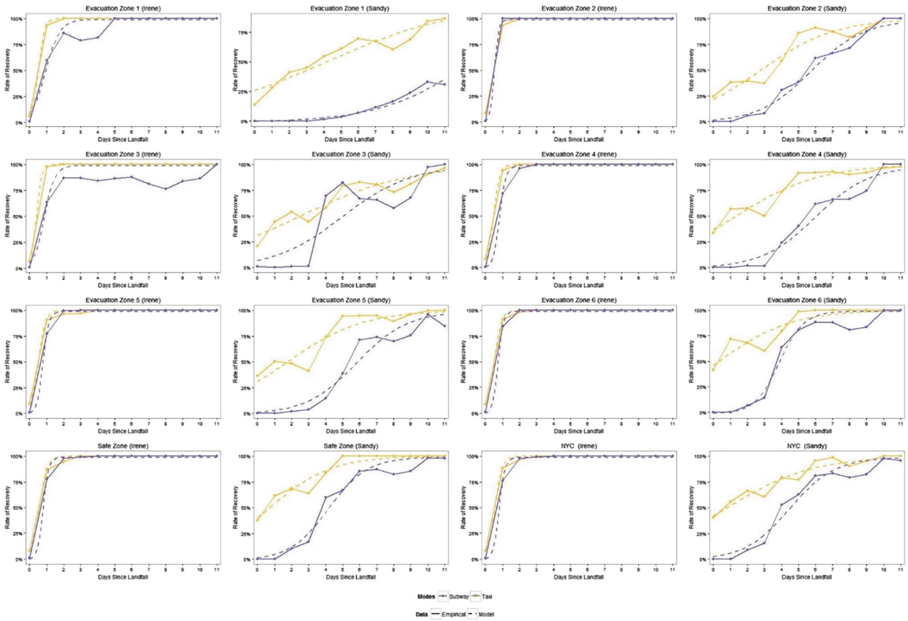


Fig. 7. Empirical and modeled response curves.

percentage and longer recovery period. One of the possible reasons is that failure of one single subway station/line always influences the entire system, whereas this is not the case for the roadway system due to the availability of more alternative routes.

4 Conclusion

This paper provides a comprehensive overview of data-oriented emergency management/planning in the complex urban systems by summarizing five case studies conducted using the big urban data of New York/New Jersey metropolitan area. There are great opportunities for the development of data-driven methods to obtain innovative solutions to the problems of emergency management and planning. The main findings from these case studies conducted by the research team are as follows:

(1) Evacuation behavior analysis
The use of the structural equation modeling is proposed to identify the potential relationship between the evacuation decision and the evacuation destination choices. A weak relationship is found between the evacuation decision and the evacuation destination choices based on the survey data.

(2) Evacuation zone prediction
The random forest has better performance in learning the relationship between current pre-determined evacuation zones and hurricane-related factors. The evacuation zones in the 2050 s and 2090 s are predicted using the random forest and are expected to expand along with the sea level rises.

(3) Traffic incident analysis
It is found that the proportion of debris, downed trees, flooding and weather related incidents increases significantly during the hurricane-impacted period. Based on developed incident frequency and incident duration models, a Monte Carlo simulation method is used to simulate the incident-induced capacity losses for the whole road network during the evacuation period.

(4) Evacuation simulation
An hour-by-hour evacuation simulation model is proposed based on a large-scale macroscopic network model, with consideration of incident-induced capacity losses. Compared with the scenario with full capacity, the evacuation travel times for capacity loss scenario are significantly higher, and are closer to the ones calculated from the historical taxi trip data in the same period.

(5) Resilience assessment
The process of multi-modal post-hurricane recovery can be captured by using logistic functions. The initial recovery rate of evacuation zones which are prone to hurricane-related risk is found to be lower than those of others. It is also found that road network has better resilience than subway network due to its operational, physical and topographical characteristics.

Acknowledgments. The work is partially funded by New York State Resiliency Institute for Storms & Emergencies, Urban Mobility & Intelligent Transportation Systems (UrbanMITS) laboratory, Center for Urban Science and Progress (CUSP), Civil and Urban Engineering at New

York University (NYU) as well as University Transportation Research Center (UTRC) at City College of New York (CUNY). The contents of this paper reflect views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the paper do not necessarily reflect the official views or policies of the agencies.

References

1. Gregory, K.: City Adds 600,000 People to Storm Evacuation Zones. http://www.nytimes.com/2013/06/19/nyregion/new-storm-evacuation-zones-add-600000-city-residents.html. Access 21 July 2015
2. Xie, K., Ozbay, K., Yang, H.: Spatial analysis of highway incident durations in the context of Hurricane Sandy. Accid. Anal. Prev. **74**, 77–86 (2015)
3. Donovan, B., Work, D.: Using coarse GPS data to quantify city-scale transportation system resilience to extreme events. In: Transportation Research Board 94th Annual Meeting, Washington DC (2015)
4. Work, D., Donovan, B.: 2010–2013 New York City taxi data. http://publish.illinois.edu/dbwork/open-data/
5. Metropolitan Transportation Authority, MTA turnstile data. http://web.mta.info/developers/turnstile.html
6. New York Metropolitan Transportation Council, Best Practice Model. http://www.nymtc.org/project/bpm/bpmindex.html
7. Li, J., Ozbay, K.: Empirical evacuation response curve during Hurricane Irene in Cape May County, New Jersey. Transp. Res. Rec. J. Transp. Res. Board **2376**(1), 1–10 (2013)
8. Carnegie, J., Deka, D.: Using hypothetical disaster scenarios to predict evacuation behavioral response. In: Proceedings of the Transportation Research Board 89th Annual Meeting (2010)
9. Yang, H., Morgul, E.F., Ozbay, K., Xie, K.: Modeling evacuation behavior under hurricane conditions. In: Transportation Research Board, Washington, DC (2016)
10. Wilmot, C., Meduri, N.: Methodology to establish hurricane evacuation zones. Transp. Res. Rec. J. Transp. Res. Board **1922**, 129–137 (2005)
11. Viera, A.J., Garrett, J.M.: Understanding interobserver agreement: the kappa statistic. Fam. Med. **37**(5), 360–363 (2005)
12. Xie, K., Ozbay, K., Zhu, Y., Yang, H.: A data-driven method for predicting future evacuation zones in the context of climate change. In: Transportation Research Board, Washington, D.C (2016)
13. Van Vuuren, D.P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G.C., Kram, T., Krey, V., Lamarque, J.-F.: The representative concentration pathways: an overview. Clim. Change **109**, 5–31 (2011)
14. Xie, K., Wang, X., Huang, H., Chen, X.: Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accid. Anal. Prev. **50**, 25–33 (2013)
15. Garib, A., Radwan, A., Al-Deek, H.: Estimating magnitude and duration of incident delays. J. Transp. Eng. **123**(6), 459–466 (1997)
16. Mooney, C.Z.: Monte carlo simulation, Sage (1997)
17. Zhu, Y., Ozbay, K., Xie, K., Yang, H., Morgul, E.F.: Network modeling of hurricane evacuation using data driven demand and incident induced capacity loss models. In: Proceedings of the Transportation Research Board, Washington, D.C (2016)
18. Caliper, TransCAD - Transportation planning software. http://www.caliper.com/tcovu.htm

19. Ozbay, K., Yazici, M., Iyer, S., Li, J., Ozguven, E.: Use of regional transportation planning tool for modeling emergency evacuation: case study of northern New Jersey. Transp. Res. Rec. J. Transp. Res. Board **2312**, 89–97 (2012)
20. Yazici, M.A., Ozbay, K.: Evacuation modeling in the United States: does the demand model choice matter? Transport Rev. **28**(6), 757–779 (2008)
21. Zhu, Y., Ozbay, K., Xie, K., Yang, H.: Using big data to study resilience of taxi and suway trips for Hurricanes Sandy and Irene. In: Transportation Research Record, Washington, D.C (2016)

Author Index