Usability of Open Data for Smart City Applications – Evaluation of Data, Development of Application and Creation of Visual Dashboards

Stephan Lehmler, Syed Monjur Murshed, Louise Ansart, Yao Shen

(Stephan Lehmler, EIFER; Karlsruhe, DE, lehmler@eifer.org)
(Dr. Syed Monjur Murshed, EIFER; Karlsruhe, DE, monjur.murshed@eifer.org)
(Louise Ansart, BSc, CentraleSupélec; Gif-sur-Yvette, France, louise.ansart@supelec.fr)
(Yao She,n KIT; Karlsruhe, DE, yao.shen@student.kit.edu)

1 ABSTRACT

Today different sources of information on urban areas are becoming openly available at various spatial and temporal resolutions and extents. They are crucial for driving towards “Smart Cities”. Many smart city relevant applications, to understand and predict certain phenomena such as mobility, air quality, etc., depend on large amounts of readily available good quality data. Many datasets related to such topics are already publicly available. However, the appropriate use of these datasets must be ensured by checking the quality of data in a systematic way. Under quality of data, one not only evaluate the number of missing or false data points but also determine data characteristics such as resolution, frequency and ease of use, etc. Therefore, the objectives of this paper are to evaluate the open data available in different portals (80 in total) with special consideration to these factors and to evaluate their usability in some of the smart city applications. In this regard, an extensive literature review is carried out. We observed that especially official government data portal often lack these qualities. This could have occurred due to the lack of concrete examples of how cities and citizens can profit from the applications created with the appropriate kind of data. Some civil servants might have experienced some levels of mistrust regarding the abstract ideas of ‘Smart City’ and ‘Open Data’. We therefore illustrate three possible applications, e.g. (a) use of high-resolution low- cost sensor data around Europe (b) GPS trajectories of a large number of taxis monitored inside the city and (c) land-use and accesability data from voluntered geographic information. In this regard, other open source spatial data portals (such as Open Street Map APIs) and open source software such as python and relevant libraries are also utilized. For each application, we elaborate the data characteristics and the detailed methodological steps (e.g., analytics methods) as well as communicate the results with an easy to operate dashboard having strong visualisation and analytical aids (maps, graphs, statistical summary, etc.). The dashboards help to understand the significance of open data and to support decision makers in creating services for the citizens with the context of “Smart City”. Finally, we conclude with the limitation and further recommendations to the city officials regarding their role of shaping the future of (smart) cities with the right open data policy.

Keywords: Smart City, Open data, Data Analytics, Dashboard, Python

2 INTRODUCTION

One of the common characteristics of the smart cities are to ensure intelligent information and communication technologies (ICT), which allows constant flow of spatial and non-spatial data at very detailed spatial-temporal resolution and extent (often in four dimensions). The availability of various open source data and data portals, as well as propritary data (often gathered by smart meters, different types of sensors) offer huge potentials to inspect many problems in the city such as road congestion, reduced air quality, increase of extreme events, urban heat islands, etc. Some of the problems in the cities can be studied through data driven and model-based development of software and tools. In this regard, predictive modelling or machine learning algorithms can be exploited to the streams of data collected from different sources. This way, the topics of big data analytics and geo analytics will play a crucial role. On the other hand, cities are very complex (systems) and understanding them require appropriate approach that is able to deal with the dependencies, relationships or interactions among cities sub-systems (e.g., network, building, industry, etc.) [1]. Therefore, appropritae method and simulation can be utilized to the stream of data to understand any particular system (or sub systems) in the city in an efficient way. This way data can be transformed into information which again can be communicated to the decision makers and citizens through efficient visulaization and tools. In this regard, web based dashboards and mobile apps as well as immersive technologies such as Augmented Reality and Virtual Reality are developed.

We perform all these activities within the Smart City Lab, which is an innovation lab in the European Institute for Energy Research (EIFER). One of the main objectives of the lab is to serve as a platform for research based experiment and quick prototyping to illustrate solutions to the different problems in the cities. We try to address the problems in a holistic manner – by introducing new methods and algorithms in one hand and using/exploiting ICT infrastructure and technological advancement to analyse and visualize them in an efficient manner.

The goal of this paper is to evaluate the usuability of the open data in some of the smart city applications. Many such datasets related to smart cities are already publicly available. However, the appropriate use of these datasets must be ensured by checking the quality of data in a systematic way. We perform a rigorous quality check of 80 different datasets and portals considering the data characteristics such as resolution, frequency and ease of use, etc. Afterwards, we develop some prototypes considering several factors such as (a) use of open source tools and scripts. Every aspect of the prototype should be reproducible without paying for data access or tools. (b) creation of value of the data by applying analytical methods. In some cases, we incorporate other sets of open data to create additional value in prototype development. (c) creation of strong visual aids such as easy to use dashboard to let users to interact with the data and methods. The dashboards clearly demonstrate ideas of how open data could be utilized for improved decision-making. However, we believe that the proposed approaches and tools can later be used with the incoming stream of other data to demonstrate further smart city applications.

2.1 Definition of Open Data and Smart Cities

“Smart City” is a difficult term to define as it is used by a wide range of actors from different areas with differing focus and applications. So it is not surprising that some papers only focused on the elaboration of the terminology exist [2-5], without one fixed definition ever emerging. Perhaps there is indeed no need for one widely used and accepted definition as long as every actor clearly states the aspect of the fuzzy concept ‘Smart City’ they seem worth focusing on. Within the scope of this research, we therefore define a smart city as a city, where ICT is utilized to (a) offer clear thematic, up-to-date data and information on every aspect of the city to its stakeholders (administrators, citizens, businesses etc.) and (b) foster interaction and communication between these stakeholders for improved decision-making.

For the other main topic of this paper - “Open Data” – we can refer to the standard definition derived from the Open Definition [6]: ‘A piece of content or data is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike’.

Availability of data is obviously the deciding factor in creating smart cities [7]. Openly available data on city related issues has therefore long been described as one of the main driving factors fueling the transition towards smart cities (e.g. [8, 9]). Open Data would allow all kinds of innovators accessing and utilizing otherwise underused data for a wide range of applications, enabling before unthought-of business models [7]. Opening up datasets also allow holding competitions from which innovative new applications can emerge [10]. The common reasoning goes that Open Data creates a “bottom-up”, or market-driven approach to smart cities [11]. This is seen as favorable compared to the still common top-down approach to urban planning, as (smart) cities are organic systems and everyone participating in its ecosystem has to work together for the generation of creative and innovative solutions and applications [12].

By the term “Open Data in Smart Cities”, one can get the image of Open Data as a continuous stream of knowledge in which researcher and entrepreneurs can tap into to power their innovative ideas. Despite this assumed huge potential, especially open government data initiatives still struggle to overcome initial hurdles [13]. To understand this gap between assumed and realized potential, possible drawbacks from the viewpoint of data disseminators have to be understood. Zuiderwijk and Janssen [14] offered a comprehensive summary of the main problems related to Open Data. We agree with Martin et al. [15] when they pointed out that uncertainty regarding return and investment and the benefits for the public results in reluctance of civil servants to contribute to Open Data. It should also be noted that not all kinds of data is of equal value. Therefore our aim in this paper is to specify the possible values of openly available data. We will discuss the features these data ought to have and present three simple prototype applications.

2.2 Data Review

Before beginning prototype development, we reviewed openly available data sources, distributed and made accessible by different entities. We reviewed numerous data sources, but looked deeper into a subset of them that seemed the most promising. We rated their usability for creating visual dashboard and simple analytic models and we also evaluated their potential for later advanced usage. In total, we reviewed 80 different sets of data and data portals in the context of energy and mobility related smart city applications.

The parallel coordinates plot in Figure 1 shows how the attributes considered (such as coverage, resolution etc.) distribute in relation to specific topics of interest.

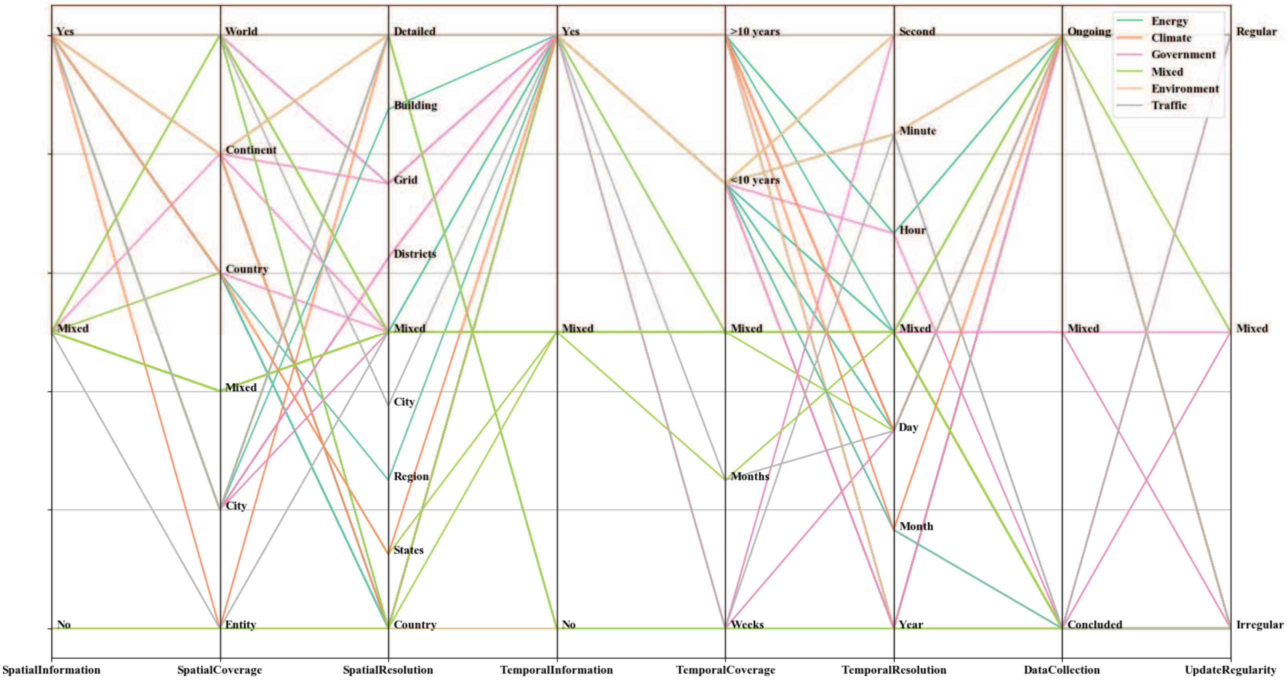


Fig. 1: Parallel Coordinates Plot showing the attributes reviewed for each dataset, grouped by topic.

For many smart city applications, spatial-temporal information is a core requisite. Data without spatial or temporal aspect may have some value as a general factoid, but for us it does not belong to the domain of a smart city. We therefore evaluated the data with regard to their spatial and temporal coverage and resolution.

While most of the reviewed data sources contained spatial and/or temporal information, a large part (especially data portals) offered this information for only a part of their data. We noted these sources as being mixed and therefore difficult to work with. Moreover, the spatial and temporal coverage of each dataset is noted. A wider geographic and temporal coverage makes datasets more useful for analysis as it allows, e.g. to compare the development of different cities or time points. The spatial coverage is satisfactory high for most datasets, meaning that most datasets cover areas wider than single cities and many cover the whole world. The temporal coverage is more mixed and therefore uncertain.

For most open data however, high coverage appears to be negatively correlated with resolution. For example, many of the datasets offering information on the whole world deliver nothing more than country averages. Some level of aggregation is often needed for privacy reasons (e.g. offering detailed lat/lon for energy consumption could infringe data privacy laws.). Nevertheless, a too high level of aggregation makes the set useless for most interesting applications.

Another point to note about existing open data sources is the actuality of the data collection activities. For nearly half the sources reviewed, the data collection is already concluded. Meaning, it is published as self-contained dataset, that won't be updated in the future. Real business opportunities will likely be relying on up-to-date city data. Published data that won't be updated can still be used for building demos and prototype applications, but its value for real life applications is limited. Another important aspect related to the data collection is the update regularity. Some data sets are/were updated in quite regular intervals (e.g. government data, sensor data) while others can have quite irregular updates (e.g. citizen science projects, voluntered geographic information).

Based on our analysis, we identified the 3 promising sets of data for further investigation and the creation of some prototypes dashboard (see section 3).

2.3 Description of Tools and Software

We used differet types of scripting languages (e.g. Python, HTML), libraries/APIs (e.g. Pandas, Bokeh, etc.) Software (QGIS, Pycharm, etc.) and data formats (such as Shape files, GeoJSON, etc.) to develop the prototypes with the open data (Table 1).

Category	Sub-Categroy	Version	Purpose of Use
Scripting Languages	Python	3.6	Main language used for accessing, processing and visualize data
	HTML	5	Markup Code for writing dashboard templates
	CSS	3	Markup Code for styling the dashboard
Libraries/APIs	Pandas	0.23.4	Python library for easy data handling and analysis
	Geopandas	0.4	Extension of Pandas supporting spatial data formats
	Shapely	1.6	Python library fpr analyzing geometrical shapes
	Pyproj	1.9.6	Python library for cartographic transformations
	Requests	2.21	Python library for writing HTTPS requests
	Bokeh	1.0.3	Python library for interactive data visualizations via web browser
	Overpy	0.4	Python library for accessing the Overpass API
	Overpass API	0.7.55	API for accessing Open Street Map data (OSM)
	Nominatim API	3.2	Geocoding API for OSM
Software	Spyder	3.3.2	Scientific Python Development Environment for writing code
	Pycharm	2018.3.3	Integrated Development Environment for Python
	QGIS	2.18.3	Geographic Information System
Data formats	CSV		Common table-based format in which many open data is delivered
	GeoJSON		Format in which OSM data is retrieved from Overpass API
	Shapefile		Format in which some OSM data is manually downloaded

Table 1: Description of tools, software and data formats for prototype development.

3 DEVELOPMENT OF PROTOTYPES WITH OPEN DATA

We have developed 3 prototypes on air quality, traffic and land-use to illustrate the usability of open data in smart city context. Each of them covers a specific topic of importance to citizens, urban planners and city officials. They were developed with the goal to give a visual example of value that could be generated with the proper utilization of the power of open data. A focus lies therefore on the visual presentation of the data and their possible usages. The visualizations of all applications have been generated with the python package Bokeh [16].

3.1 Air Quality

3.1.1 Data

For this prototype, we queried data from the citizen science initiative “luftdaten.info” [17]. The data is gathered from a huge amount of self-built low-cost air quality sensors set up by volunteers around the world. There already exists a professional online visualization of the current sensor-readings aggregated on a lower-resolution grid. It does however only provide a limited amount of temporal information. We therefore aim to offer an alternative visualization of the data focusing on the long-term temporal dynamics of the sensor measurements.

3.1.2 Method

We downloaded the achieved sensor-measured from an ftp-server provided by “luftdaten.info” using python and the http-library “requests”. After gathering all the csv files containing sensor readings in a 30-second interval, we prepared the data for visualization. Amongst others, we had to calculate aggregated daily averages for each sensor for a more lucid visualization. We then set up the browser-ready dashboard with the bokeh library. For geocoding user input, we use the OSM nominatim API. Figure 2 shows the schematic workflow of the application development.

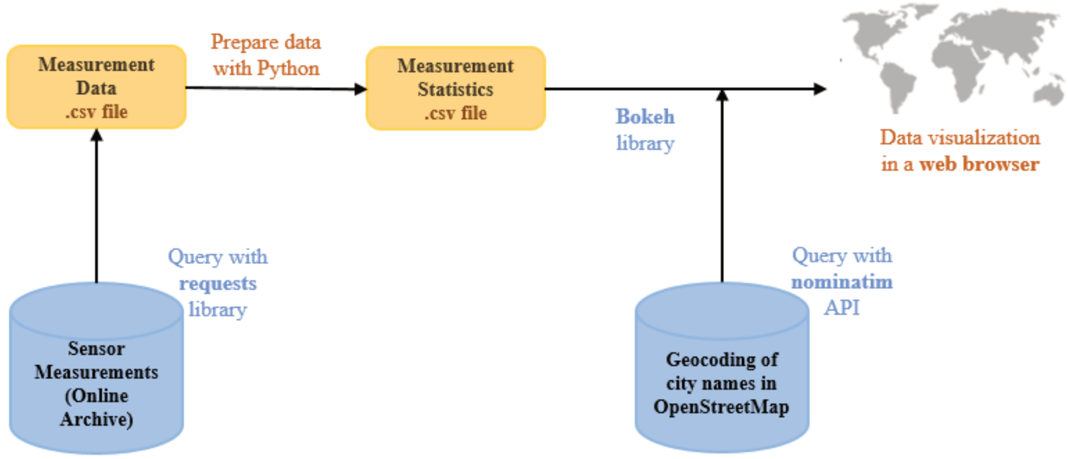


Fig. 2. Schematic workflow for air quality dashboard.

3.1.3 Visualization

The prototype consists of a web-based, interactive dashboard, showing each sensor and its daily mean measurement of a chosen pollutant (PM10 or PM2.5) on a map. Users can choose a specific date and select sensors for which a detailed plot of past measurements will be shown. They can also insert any city name into the search bar and the map will focus around that city. This feature is implemented using the nominatim geocoding API [17] for OpenStreetMap (OSM) [18]. Figure 3 shows the dashboard to inspect the low cost sensor data around the world. The color of the dots signifiesthe amount of air pollution measured.

3.1.4 Remarks

Using the slider to see the different sensor placements shows nicely how such a citizen science project can grow and evolve over time. We can also see the daily measurements and dynamics of air pollutant over time. This could be extended with predictive capabilities using time series forecasting. We faced however challenges concerning data access and data quality. The data has to be manually downloaded and cleaned as some of the sensors produce faulty measurements. The general quality of low-cost sensor readings has to be evaluated for each use case (as e.g. in [19]).

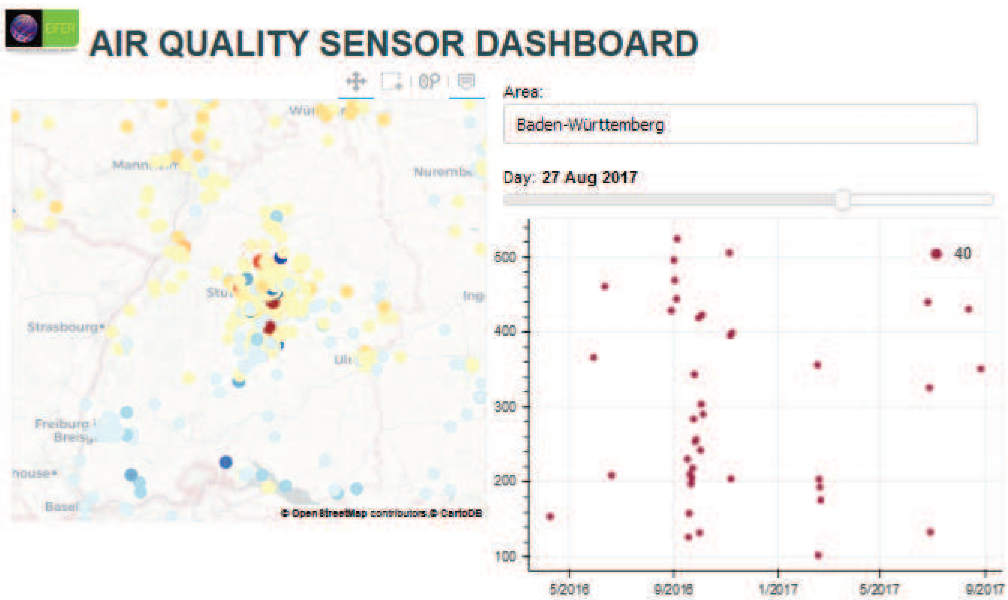


Fig. 3. Air quality dashboard to inspect the low cost sensor data around the world.

3.2 Taxi Data Beijing

3.2.1 Data

For this prototype, we used the T-Drive trajectory data sample [20], published by the Microsoft Research Lab Asia. It contains GPS trajectories of a large number of taxis (10357), monitored inside the city of Beijing for one week for every second. The full dataset is not freely available and covers 30000 taxis during 3 months and was used by Microsoft amongst others for the T-Drive Driving Directions Service [21, 22]. For our study, we used the free dataset.

3.2.2 Method

The developed Dashboard focuses on using the GPS trajectories for representing the traffic volume in the city. Figure 4 describes the schematic workflow of the application. For this simple model, we defined areas of interest and counted the number of taxis around these areas for each time step. As an example, we decided to look at street junctions, due to their important function in traffic flow. We used OSM to get the position of street junctions in the city. We noticed that this approach does not return every junction, as apparently only bigger street junctions where tagged in the OSM of Beijing. This is one common pitfall when working with volunteered geographic information (VGI).

After downloading and extracting the csv files containing the taxi trajectories, we processed it with Python to create a table containing the observed numbers of taxis on each motorway junction per hour. To get the coordinates of each junction, we used the OSM overpy API. The pre-calculated table is loaded by a separate script, preparing the visualization with bokeh.

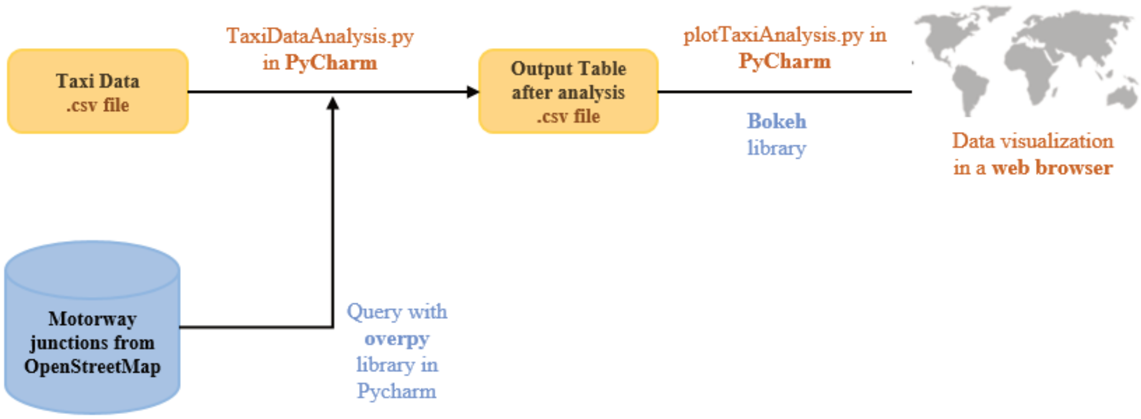


Fig. 4. Schematic workflow for traffic dashboard.

3.2.3 Visualization

As with the Air Quality prototype, we allow the user to use a slider to move through the time-steps and select a group of junctions for comparison. In addition to the timeline, we also can look into the relatedness of different junctions in terms of their correlation. Figure 5 shows the created dashboard. The red colored dots describe the junctions where most traffic exists at a certain time.

3.2.4 Remarks

It would be easy to implement a real-time traffic monitoring system in this way, if live data of such GPS trajectories is available. The analysis of correlation between areas demonstrates how even this kind of simple dashboard could aid urban planners to discover unknown spatial relationships among different junctions. As the disseminated data is only a sample of the full source, it lacks the temporal coverage for more complex analysis.

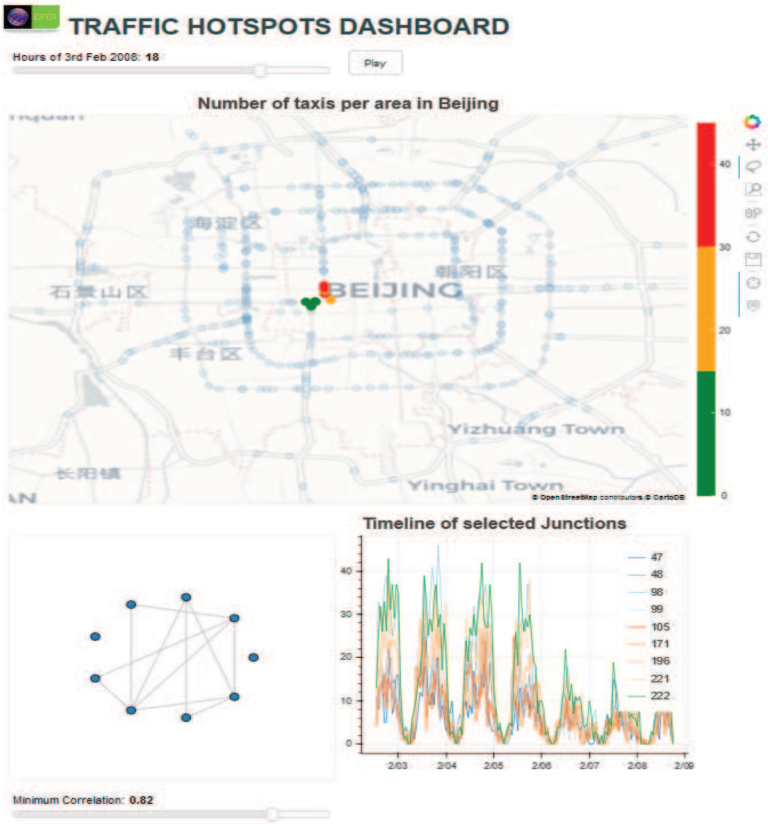


Fig. 5. The traffic dashboard displays the visual and analytical capabilitiy to understand the congested junctions.

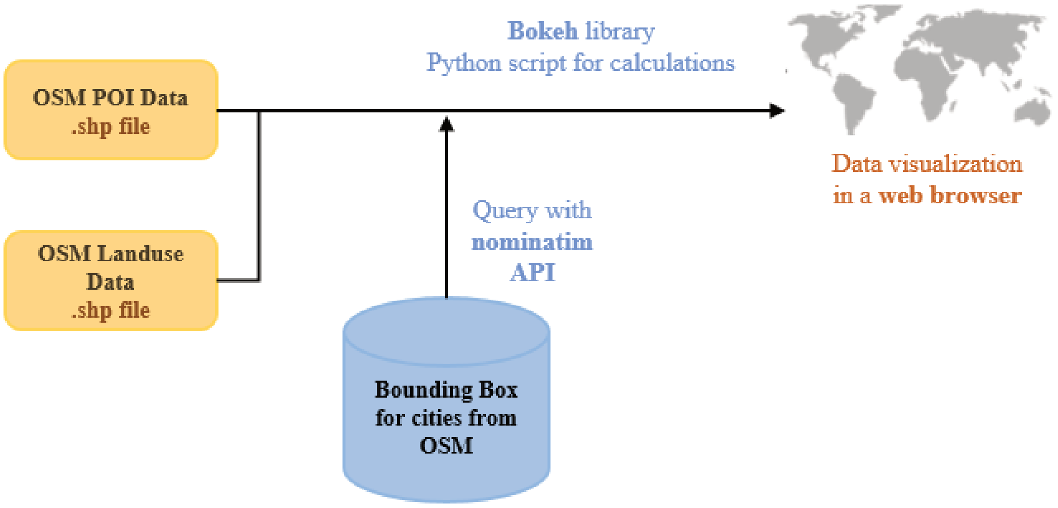
3.3 Land-use and Accessibility

3.3.1 Data

For the last protoype, we focused on VGI. We extracted and utilized data from OpenStreetMap (OSM) to make quick data analysis and to support the decision-making process in the smart city context using efficient visualization's techniques. More specifically, the information regarding the walkability and land-use situation of urban areas in Germany is explored.

3.3.2 Method

For this prototype, we downloaded the OSM dataset from Geofabrik's free download server. The downloaded data is in shapefile format when unzipped and is saved locally for this project. We chose cities of Germany as our samples to do the analysis as well as visualization. The analysis is done using python and the geopandas library, while the dashboard is visualized again with bokeh. Figure 6 describes the schematic workflow of the application.



ISBN 978-3-9504173-6-4 (CD), 978-3-9504173-7-1 (print)
Editors: M. SCHRENK, V. V. POPOVICH, P. ZEILE, P. ELISEI, C.BEYER, J. RYSER

Fig. 6. Schematic workflow for accessibility dashboard.

3.3.3 Visualization

At first, we built grid across the chosen city areas, and each grid cell becomes a study area. For points of interes (POIs), the area (in square metres) of each feature for every grid cell is calculated. For the distribution of land-uses, we considered the whole city area as a study area. Then the color code and the statistical graphic pie chart are used as visualization's tools to display the information for specific cities. For the dynamic visualization of areas, the nominatim API is used for selecting cities by name and getting the bounding-box of city for the illustration and calculation.

With this prototype, we allow users to choose a specific city from the select widget and the map will focus around that city, the land-use distribution of the chosen city will be shown using the pie chart. Moreover, users can also choose different POIs that they are interested in, and the color code will provide the information about the area of specific POI in each grid of the city and help people e.g. to pick a travel destination. Figure 5 shows the visualization of accesibility and land-use for one city.

3.3.4 Remarks

This prototype shows how VGI's can be utilized for analyzing some key indicators relevant for city planners and citizens. It also serves as an example of how thematic dashboards and data analytics can help to extract relevant information from already known rich data sources.

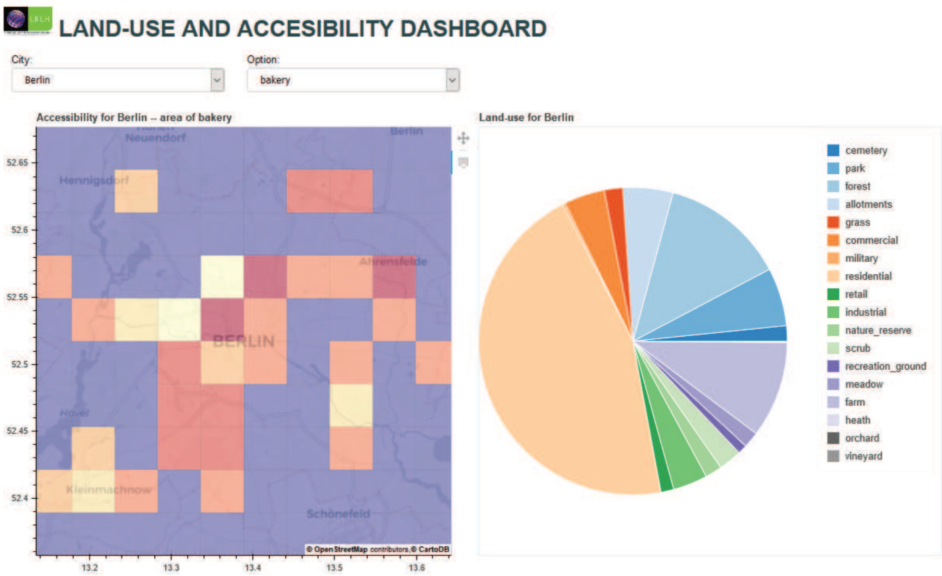


Fig. 7. Land use and accesibility dashboard to illustrate the land-use distribution in a given area.

4 CONCLUSION

In this paper, we have illustrated three potenital smart city applications of open data, e.g. (a) use of high-resolution low- cost sensor data around Europe (b) GPS trajectories of a large number of taxis monitored inside the city and (c) land-use and accesability data from VGI. In this regard, other open source spatial data portals (such as Open Street Map APIs) and open source software such as python and relevant libraries are also utilized. For each application, we elaborated the data characteristics and the detailed methodological steps (e.g., analytics methods), based on that, we communicated the results with an easy to operate dashboard having strong communication aids (maps, graphs, statistical summary, etc.). The dashboards help to understand the significance of open data and their application. This supports decision makers in creating services for the citizens with the context of "Smart City".

These applications might be relatively easy to create, given the right kind and quality of data. However, there is still a long way to go for open city data before we achieved any of the eager ideals proposed by the literature. Many of the available data lack most of the attributes deemed important by us. The datasets we used for our prototypes carried some of the necessary attributes to be applied for Smart City applications. If we would however have only focused on data provided by (local) governments, it would have been more difficult to find usable data. While we did not provide a detailed review of these government portals, most published open data appears to consist of highly spatial and temporal aggregated data with a slow update frequency. This may provide some level of transparency and insights, but does not appear to be capable to fuel innovative applications.

We therefore would conclude with a need for not simply “more” data as in its current form, but “better” data. Enforcing the publications of official statistics and protocol has its value, but it does not provide the citizens benefits or economic returns, often talked about in the context of smart cities. It therefore might lead to dissatisfaction of civil servants and the public with the continuous push to open data. This is why we advise to focus more on the creation of continuously updated, widely employed and standardized high-resolution data on specific topics.

It would be beneficial for future open data and smart city initiatives to be on a larger scale than only single cities. At the moment, it is difficult to uphold long lasting data collecting activities. It might be easier to conduct such activities in the long run, when multiple cities are involved. Additionaly, data might become more meaningful, if it is gathered on a wider scale, and more economic value could be created from it. As it is also difficult to create economic value from data if it is only available for one city. In the future, we also hope to see cities as active players in the deployment of sensor-networks and the creation of citizen science or VOI activities, as these seem to produce the most valuable data. Given the right kind of data, the work done so far could be greatly improved. Especialy opening up the access to real time sensor data and deploying sensors currently rarely used in city contexts (such as Lidar and other movable sensors). Instead of keeping these most valuable sources of information proprietary, it should be seen as a public good. They should open exactly this kind of interesting data, which is often kept closed by smart city initiatives created by public-private partnerships.[1] This way, cities and citizens can keep the control over their data, autonomously improve on data privacy issues while at the same time encourage participation and innovation. At least, this leads to a new public enthusiasm for the idea of the Smart City and helps to eliminate some common fears related to the concept.[2]

This paper reviewed some aspects of the current state of open data for smart cities and discussed the not yet realized potential. If cities are interested in our vision of an open smart city, there still remains a lot of work to be done. There currently is a lack of systematic review on open government data portals and on privacy issues regarding proprietary and open data sets. It would also be interesting to carry out qualitative studies on the attitude government officials hold towards questions of open data and smart cites. For us, it would be especially interesting to learn directly from citizens and city officials, what vision they have for the future of their smart city.

5 SUPPLEMENTARY MATERIALS

A video demonstration of the presented dashboards is available online: https://youtu.be/Fv-BYqmZzBA

6 REFERENCES

1. Becerra, G. and J.A. Amozurrutia, Rolando García’s “Complex Systems Theory” and its relevance to sociocybernetics. Journal of Sociocybernetics, 2015. 13(1): p. 18-30.
2. Cocchia, A., Smart and Digital City: A Systematic Literature Review, in Smart City: How to Create Public and Economic Value with High Technology in Urban Space, R.P. Dameri and C. Rosenthal-Sabroux, Editors. 2014, Springer International Publishing: Cham. p. 13-43.
3. Nam, T. and T.A. Pardo, Conceptualizing smart city with dimensions of technology, people, and institutions, in Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times. 2011, ACM: College Park, Maryland, USA. p. 282-291.
4. Dameri, R.P. and A. Cocchia. Smart city and digital city: twenty years of terminology evolution. in X Conference of the Italian Chapter of AIS, ITAIS. 2013.
5. Allam, Z. and P. Newman, Redefining the smart city: Culture, metabolism and governance. Smart Cities, 2018. 1(1): p. 4-25.
6. OpenKnowledgeInternational, Open Definition 2.1. Online. 2018.
7. Zotano, M.A.G. and H. Bersini, A Data-driven Approach to Assess the Potential of Smart Cities: The Case of Open Data for Brussels Capital Region. Energy Procedia, 2017. 111: p. 750-758.

[1] And can hinder the progress of smart city activities, e.g. https://www.engadget.com/2018/10/26/sidewalk-labs-ann-cavoukian-smart-city/?guccounter=1

[2] For example, just last year, the term ‘Smart City’ was awarded with the german mock prize ‘Big Brother Award’ (https://bigbrotherawards.de/en/2018/pr-marketing-smart-city)

8. Jaakola, A., et al., Open data, open cities: Experiences from the Helsinki Metropolitan Area. Case Helsinki Region Infoshare www. hri. fi. Statistical Journal of the IAOS, 2015. 31(1): p. 117-122.
9. Janssen, M., R. Matheus, and A. Zuiderwijk. Big and Open Linked Data (BOLD) to Create Smart Cities and Citizens: Insights from Smart Energy and Mobility Cases. 2015. Cham: Springer International Publishing.
10. Hielkema, H. and P. Hongisto, Developing the Helsinki Smart City: The Role of Competitions for Open Data Applications. Journal of the Knowledge Economy, 2013. 4(2): p. 190-204.
11. Walravens, N., J. Breuer, and P. Ballon, Open data as a catalyst for the smart city as a local innovation platform. COMMUNICATIONS & STRATEGIES, 2014. 96.
12. Cosgrave, E., K. Arbuthnot, and T. Tryfonas, Living Labs, Innovation Districts and Information Marketplaces: A Systems Approach for Smart Cities. Procedia Computer Science, 2013. 16: p. 668-677.
13. Mishra, A., et al. Assessment of Open Government Data Initiative - A Perception Driven Approach. 2017. Cham: Springer International Publishing.
14. Zuiderwijk, A. and M. Janssen. 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. 2014. ACM.
15. Martin, S., et al., Risk analysis to overcome barriers to open data. Electronic Journal of e-Government, 2013. 11(1): p. 348.
16. Bokeh Development Team, Bokeh: Python library for interactive visualization. 2018.