ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Processing, mining and visualizing massive urban data

Pierre Borgnat [1], Etienne Côme[2] and Latifa Oukhellou [2]

1- Laboratoire de Physique de l'École normale supérieure de Lyon,
CNRS, Université de Lyon,
46 allée d'Italie, F-69364 Lyon Cedex 7, France

2- Université Paris-Est, COSYS, GRETTIA, IFSTTAR,
F-77447 Marne-la-Vallée, France

Abstract. The development of smart technologies and the advent of new observation capabilities have increased the availability of massive urban datasets that can greatly benefit urban studies. For example, a large amount of urban data is collected by various sensors, such as smart meters, or provided by GSM, Wi-Fi or Bluetooth records, ticketing data, geotagged posts on social networks, etc. Analysis of such digital records can help to build decision-making tools (for analytical, forecasting and display purposes) with a view to better understanding the operating of urban systems, to enable urban stakeholders to plan better when extending infrastructures and to provide better services to citizens in order to assist the development of the city and improve quality of life. This paper will focus on three main domains of application: transportation and mobility, water and energy.

1 Introduction

In recent decades, demographic growth, urban sprawl, increasing road congestion, and the desire to reduce environmental pollution have been responsible for the emergence of new sustainable urban policies in a variety of areas such as mobility, energy, water and air quality. The public authorities play a significant role in this, giving impetus to sustainable urban practices which aim to develop more efficient resource demand and management. In the domain of mobility, urban policies aim to reduce the use of private cars and increase the use of sustainable modes of transport such as public transport, walking, cycling, bike-sharing, car-sharing and car-pooling. These modal choices on the part of citizens are not only the outcome of economic determinants (the pricing of public transport, the cost of fuel, the price of parking), but also partly due to town planning decisions which impact the spatial configurations of population and job density at metropolitan level, road design and local transport supply. As far as resources such as energy and water are concerned, public authorities are also persuing policies that aim to rationalize water use in a context of scarcity and to ensure efficient planning and optimization of energy resources, in particular through the use of renewable energy. Within this context, based on the development of smart technologies, massive urban datasets are collected by a variety of urban sensors. This data can be used to monitor urban systems in order to address urban issues in a number of areas such as planning, transportation, the environment, energy, water and health.

In the domains of energy and water, cities and electricity/water companies are implementing many programs to equip buildings with smart meters which can provide hourly consumption readings. In the domain of transport, digital records generated during citizens' trips can provide new insights for the analysis of urban mobility. There are many such sources of data, for example ticketing logs, GSM, Wi-Fi or Bluetooth records and the geo-tagging of posts that are generated on social networks during trips. Although the technologies in question were not initially designed for the analysis of mobility, their usefulness is obvious. Moreover, the different kinds of data collected in a city can be cross-mined to analyze urban phenomena. With regard to air quality, a considerable amount of research has investigated the impact of road traffic on air pollution. In this case, traffic data and air quality data are simultaneously analyzed to reveal the underlying phenomena responsible for urban air pollution. More generally, the availability of heterogeneous data can be helpful in order to monitor urban systems and create high value-added services for citizens.

With regard to operational aspects, mining and visualizing such massive amounts of urban data can help when building decision-making tools for urban stakeholders to allow them to better monitor urban systems, optimize the working of the city, better meet future needs through improved forecasting and better match the supply of urban services to citizen demands. One of the main challenges of the "city of tomorrow" is the ability to achieve real-time control of urban systems in order to optimize the working of the city. One possible example of this real time control is cross-combining weather with transport services in order to provide citizens with transport services that match weather conditions and particular events. Another example would be to match transport supply to real time demand.

The rest of this paper is organized as follows. In Section 2, we shall describe the issues to be addressed and the datasets that are often used for some applications. In Section 3, we shall describe some typical data mining and visualization objectives. Section 4 concludes the paper by outlining some possibilities for the future.

2 Urban Data: some applicative examples

With reference to three urban systems, namely transport and mobility, energy and water, this section sets out to describe the general application-oriented issues to be tackled by the mining of collected urban data.

2.1 Transport and Urban Mobility

Until now, most urban mobility analysis has been based on stated preference surveys. Such surveys have a number of advantages, for example, they cover all transport modes and all trip purposes and they also contain meta-data about the respondents (gender, socio-occupational group ...). However, they are expensive and consequently performed fairly infrequently (typically, public surveys in France are carried out every five or ten years), which means they do not allow close monitoring of ongoing developments and the public policies that aim to influence them. The analysis of daily travel can make use of the increasing number of digital records that are generated during citizens' trips because the availability of portable technologies means that most people carry passive mobile sensors in urban areas and leave digital traces during their trips. Such data sources are numerous, for example ticketing logs, GSM, Wi-Fi or Bluetooth records and geo-tagged posts that are generated on social networks. Although the devices used were not initially designed for the analysis of urban mobility, their usefulness is obvious. These datasets can be used to identify mobility footprints and the development of such new sensors therefore greatly benefits urban mobility studies. This leads to the emergence of a new field of research, known as urban computing [1].

Digital records remove some of the difficulties that exist with surveys by covering a large proportion, if not the quasi-totality, of citizens. The large volume of data and the diversity of the situations covered will mean that by comparing these digital records with the actual supply of transport services it will be possible to reveal the impact of the different characteristics of supply (price, frequency, reliability ...) on the main travel decisions made by citizens (route, departure time, transport mode...). Monitoring users over a period of time will even make it possible to conduct disaggregated analysis of these decisions, by studying how a given individual reacts to variations in transport supply (the variability of congestion, incidents or works on the route, ride-sharing systems ...). Moreover, it will be possible to provide citizens with personalized recommendations based on historical knowledge of their customary travel practices and the urban context, resulting in the provision of smart services.

Depending on the type of dataset and the goals of the application, a large number of varied case studies have been performed. Mobile phone data have been used either to extract mobility patterns as in [2], [3] or to provide on-demand public transportation as proposed in [4]. A taxi-sharing system that reacts in real-time to ride requests received from smartphones has been proposed in [5]. Transport ticketing data from, for example, metro, bus and shared mobility systems has also been used to identify trip purposes, reveal mobility patterns or better understand and predict passenger flows. A number of research studies have been carried out on the bike-sharing systems which have been deployed in many large cities worldwide. Some of them set out to optimize bike reallocation techniques [6], [7], while others attempt to identify patterns in system usage [8], [9] or forecast station usage [10]. In the field of public transport, a considerably amount of research has been undertaken on the mining of ticketing logs. Some recent studies in this area include those by [11, 12] and [13] which focused on the partitioning of network passengers into groups based on their transportation network activity, not forgetting the work carried out by [14] that aims to analyze multi-week activity patterns and propose a representation of activity sequences.

2.2 The domains of energy and water

Smart energy and water management in cities is one of the challenges we need to face in order to achieve efficient and responsible resource control. In the domain of energy, public stakeholders, which include electricity suppliers, distribution network operators and city managers, have to cope with a rapid increase in electricity demand, diversify the sources of energy production and be able to balance production, supply and demand. At the same time, there is a growing trend towards vehicle electrification, which raises various issues such as the location of charging stations to better respond to demand, the prediction of the charging needs of electric vehicles and the management of a fleet of electrical vehicles. Water resource management is also a major concern for countries which are exploring measures to rationalize water management in order to cope with scarcity and improve quality. Accurate monitoring of drinking water systems and citizens' consumption is required to permit preventive anomaly detection and to offer personalized water management to citizens.

Resource management can benefit from the development of smart technologies that allow the collection of a large amount of data. For example, smart meters permit hourly consumption readings for both residential and industrial activities. Such datasets can be used to identify typical demand patterns corresponding to consumer behaviors, to predict consumption in both the short and long terms and also at different spatial scales (building, city, country). Dataset analysis of this type may also be useful for developing personalized management strategies. For example, one common policy is to assign greater value to interrupting consumption during peak periods. Profiling and targeting consumers who are likely to respond positively to such a pricing policy can be achieved more effectively by using knowledge extracted from the datasets.

In the domain of energy, several studies have been undertaken to analyze an open dataset collected by the European Commission for Energy Regulation (CER) during a smart meter installation trial. The dataset contains the electricity consumption information from smart meters installed in households and small or medium- sized enterprises (SME) located in Ireland during 2010 [15, 16, 17, 18]. Clustering methods were applied to classify the customers into clusters on the basis of their daily electricity consumption patterns. The authors also examined cross-tabulating the clustering results with household information collected about customers with the aim of identifying potential targets within the clusters for a demand response program. Supervised methods have also been applied to infer household characteristics from energy consumptions. In the field of water, similar issues have been addressed specifically to better characterize demand [19, 20, 21].

3 Key challenges and main urban data mining and visualization approaches

We can distinguish between two types of urban data, depending on the field of application. These are data designed to monitor a specific urban system as is the case for the domains of energy, water and air quality, and data gathered from passive sensors in urban areas including mobile phones, ticketing data and GPS devices which were not initially designed for urban monitoring. The rapid growth in data storage capabilities means that historical data are now available in many domains. The availability of data sources of this type highlights how important it is to develop novel approaches based on engineering and computer sciences. A preprocessing step is often required to handle these datasets which can be enormous. Classical approaches such as filtering and normalization can be applied to clean such datasets, but, depending on the application, the preprocessing can be much more radical. The preprocessing of ticketing logs which do not store destination in order to recover this missing information and to reconstruct trip chains are typical examples of such complex preprocessing steps. When the dataset has been cleaned and is ready for analysis all the main types of machine learning techniques may be valuable in order to deal with different types of problems. We shall review some of them with some of their applications in the paragraphs below.

One of the main focuses of urban data mining is *unsupervised learning* to obtain useful insights into the sizeable datasets. This involves using classical algorithms or developing new methodological approaches in order to identify a reduced subset of typical patterns that can easily be interpreted while retaining the detailed nature of the data. A trade-off has to be made between easy interpretation of the results and the descriptive power of the subspace. This projection can either be performed on the users (passengers or consumers) to highlight a reduced set of typical patterns in their behaviors [12, 14, 21, 16] or on spatial entities like transport stations or neighborhoods [9, 10]. Crossing the results with the socio-demographic characteristics of citizens and geographical variables that relate to a city is often performed to highlight the important features of the city that explain variations in demand. Subsequently, these may help in the design of new predictive demand models that can be used by urban planners to optimize or extend the existing urban systems. In the domain of transport, clustering is usually performed on the stations of a transport network or on the flows between stations with respect to their usage profiles or the way flows are exchanged between the stations. In this way, it is easier to assign a latent activity corresponding to the station usage to each of the clusters.

Another type of method commonly used in urban data analysis is *novelty / anomaly detection.* This type of approach has already been used to monitor road networks [22] and subway networks [23]. Different types of settings are possible for this task. A supervised dataset of normal/abnormal behaviors may be built and machine learning approaches can be applied to discriminate between the different operating classes. Learning can be performed in a completely unsupervised way. The tools in question can be used to monitor urban systems and detect abnormal behaviors in order to take corrective actions as soon as possible. They are also of great interest for urban data analysis as witnessed by the two papers on the subject at this special session [23] [24].

Prediction is also a major area of investigation with regard to the mining of urban data. The aim is to develop forecasting models to predict the operation of the urban system [19, 6]. Forecasting demand is in fact a central problem for the organization of urban systems since it provides a way of anticipating demand and planning appropriate supply beforehand. One can distinguish between two different goals, the first relates to long-term prediction, which can serve, for example, for the planning of urban systems, whereas the second relates to short-term forecasting with a view to matching supply with demand. If we take the case of public transport, long-term prediction might be valuable for planning the transport network, while short-term demand prediction could be used for transit operation purposes to match supply with demand. More formally, this problem can be tackled by a variety of methods and tools whose formalization requires the relevant temporal and spatial horizons to be specified. Time series models, graph models for dynamic complex networks, generative mixture models or machine learning models can be used to carry out these tasks. Due to the availability of sizeable historical datasets, deep networks are obvious candidates for performing this type of task.

When the problem to be solved more closely resembles one of real-time control, *reinforcement learning* is a good candidate for optimizing control. Approaches of this type are already being investigated for applications such as traffic signal control or real-time speed limit control [25, 26]. In such cases, reinforcement learning is classically used in conjunction with simulation and/or prediction. The application of such approaches in real settings is rare at the moment, but as real-time data monitoring capacity increases it will certainly become important in the near future.

As we have already seen, machine learning can play an important role in the study of urban data. However, the real problems to be solved that may benefit from new urban data are often not well defined. An appropriate way of tackling such ill-defined problems may be to combine automatic processing performed by machine learning tools with human exploration and analysis using *visualization* tools. Let us consider, for example, the choice of a new layout for a bus network. There are a number of aspects to this problem that are difficult to formalize completely and a combination of automatic analysis and data exploration may assist the experts engaged in design. Such approaches have already been used, for example, to design the new bus network in Moscow [27]. Visualization is thus particularly useful for tackling fuzzy problems that require an exploratory approach. To quote [28], it "facilitates discovery, contemplation and presentation amongst other roles". To be effective, however, visualizations must deal with the complexity of urban datasets which commonly contain geographical and temporal components in addition to multiple variables. Dedicated tools must be developed to handle these aspects as described in one the papers presented at this

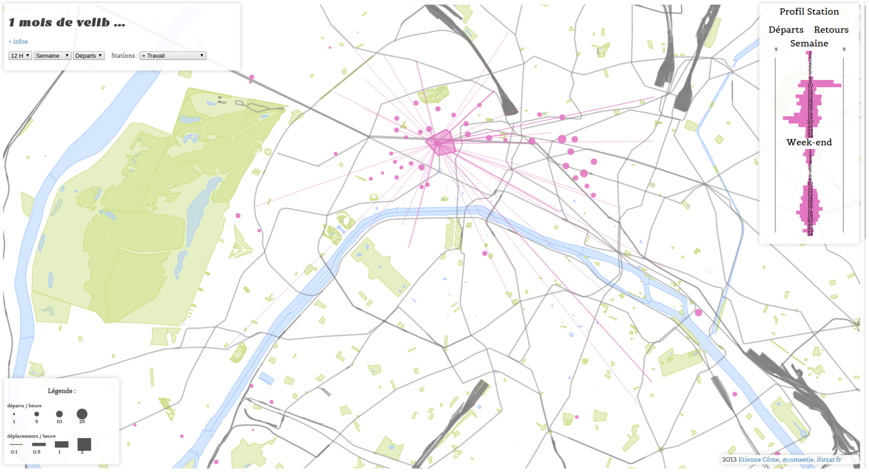


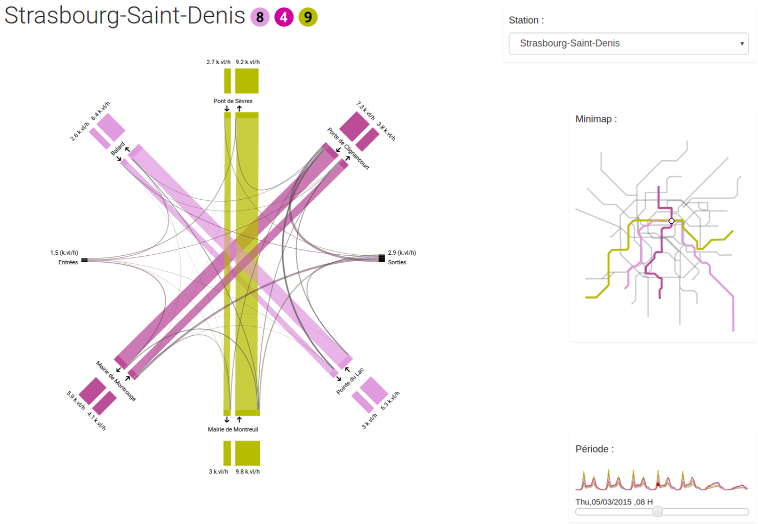
Fig. 1: Visualization tools for exploring the spatio-temporal dynamics of the origins and destinations for the Paris bike-sharing system (the tool combines clustering and visualization to facilitate dataset exploration).

special session [29] and in the Figures 1 and 2 which present two visualizations developed to facilitate exploration of two massive urban datasets (see caption for details).

Another important issue is related to *privacy* issue. The growing amount of both public and non-public heterogeneous data now appears as a meaningful way to monitor urban systems and to create high value-added services for citizens. One of the main challenge is to be able through the computation of different sources of data to protect personal privacy while creating services for individuals and more globally for the society. The availability of large amount of urban datasets also raises an *ethical* issue. Currently, most data are collected by private companies which make some datasets available on-line. Given that the society could benefit from collected urban data, recommendations are made to consider data as a common good that has to be shared as much as possible. Including citizens in the process of collecting and sharing the data and increasing their awareness of potential individual and societal benefits might help to gain acceptance for the massive urban data. The question of equal access and potential discrimination of access that can be raised by such analysis and tools is another important concern for urban data analysis.

4 Conclusion

The development of smart technologies is leading to the collection of massive urban datasets by a variety of urban sensors. This data can be used to monitor urban systems in order to address several issues that affect cities in such areas



Station : Strasbourg-Saint-Denis

Thu,05/03/2015 ,08 H

Fig. 2: Visualization tools for exploring the temporal dynamics of subway network flows at station level (the tool combines data-enrichment and visualization to facilitate dataset exploration).

as urban planning, transportation, the environment, energy, water and health. This paper has examined the main issues to be tackled in three domains of application: mobility and transport, energy and water. The final goal is to build decision-making tools for urban stakeholders to allow them to better monitor urban systems, optimize the working of the city, better meet future needs through improved forecasting and better match the supply of urban services to citizen demands. The main approaches of urban data mining and visualization have been surveyed as have crucial privacy and ethical issues. One of the main challenges is to be able to protect personal privacy and respect ethical standards when the different sources of data are processed to create high value-added services for individuals and communities.

References

[1] Z. Yu, L. Capra, W. Ouri, and Y. Hai. Urban computing concepts methodologies and applications. *ACM Transaction on Intelligent Systems and Technology*, 2014.

[2] F. Calabrese, M. Dia, G. Di Lorenzo, J. Ferreira, and C. Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C*, 26:301–313, 2013.

[3] Z. Wangsheng, L. Shijian, and P. Gang. Mining the semantics of origin-destination flows using taxi traces. In *proceedings of ACM Conference on Ubiquitous Computing (UbiComp)*, pages 943–949, 2012.

[4] T. Altshuler, Y. Shiftan, R. Katoshevski, N. Oliver, A-S. Pentland, and Y. Altshuler. Mobile phones for on-demand public transportation. In *proceedings of NetSci*, 2014.

[5] S. Ma, O. Wolfson, and Y. Zheng. Real-time city-scale taxi ridesharing. *IEEE Transactions on Knowledge Discovery and Data Engineering*, 27:1782–1795, 2015.

[6] L. Yexin, Z. Yu, Z. Huichu, and C. Lei. Traffic prediction in a bike-sharing system. In *proceedings of ACM SIGSPATIAL*, 2015.

[7] R. Nair, E. Miller-Hooks, R.C. Hampshire, and A. Bušić. Large-scale vehicle sharing systems: Analysis of vélib'. *International Journal of Sustainable Transportation*, 7:85–106, 2012.

[8] M. Ahillen, D. Mateo-Babiano, and J. Corcoran. The dynamics of bike-sharing in washington, d.c. and brisbane, australia: Implications for policy and planning. *International Journal of Sustainable Transportation*, 2015.

[9] E. Côme and L. Oukhellou. Model-based count series clustering for bike sharing system usage mining: A case study with the vélib system of paris. *ACM Trans. Intell. Syst. Technol.*, 5, 2014.

[10] P. Borgnat, Robardet C., RouquierJ.B., P. Abry, E. Fleury, and P. Flandrin. Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14:1–24, 2011.

[11] M. Xiao-lei, W. Yao-Jan, W. Yin-hai, C. Feng, and L. Jian-fengtitle. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.

[12] A-S. Briand, E. Côme, M.K. El Mahrsi, and L. Oukhellou. A mixture model clustering approach for temporal passenger pattern characterization in public transport. *International Journal of Data Science and Analytics*, pages 1–14, 2016.

[13] M. Poussevin, N. Baskiotis, V. Guigue, and P. Gallinari. Mining ticketing logs for usage characterization with nonnegative matrix factorization. In *proceedings of SenseML 2014 – ECML Workshop*, 2014.

[14] G. Goulet Langlois, Koutsopoulos H.N., and J. Zhao. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, 64:1–16, 2016.

[15] Y. Wang, Q. Chen, C. Kang, and Q. Xia. Clustering of electricity consumption behavior dynamics toward big data applications. *IEEE Transactions on Smart Grid*, 7:2437–2447, 2016.

[16] S. Haben, C. Singleton, and P. Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid*, 7:136–144, 2016.

[17] C. Beckel, L. Sadamori, T. Staake, and S. Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, 2014.

[18] F. McLoughlin, A. Duffy, and M. Conlon. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An irish case study. *Energy and Buildings*, 48:240–248, 2012.

[19] K. Aksela and M. Aksela. Demand estimation with automated meter reading in a distribution network. *Journal of Water Resources Planning and Management*, 137, 2011.

[20] S.A. McKenna, F. Fusco, and B.J. Eck. Water demand pattern classification from smart meter data. *Procedia Engineering*, 70:1121–1130, 2014.

[21] N. Cheifetz, A. Same, Z. Noumir, a.C. Sandraz, C. Feliers, and V. Heim. Extracting urban water usage habits from smart meter data: a functional clustering approach. In *proceedings of the* 25^{th} *European Symposium on Artificial Neural Networks*, 2017.

[22] Z. Zhou, P. Meerkamp, and C. Volinsky. Quantifying urban traffic anomalies. In *proceedings of Data for Good Exchange*, 2016.

[23] E. Tonnelier, N. Baskiotis, V. Guigue, and P. Gallinari. Anomaly detection and characterization in smart card logs using nmf and tweets. In *proceedings of the* 25th *European Symposium on Artificial Neural Networks*, 2017.

[24] P. Laharotte, R. Billot, and N. El Faouzi. Detection of non-recurrent road traffic events based on clustering indicators. In *proceedings of the* 25th *European Symposium on Artificial Neural Networks*, 2017.

[25] E. Van der Pol and F. A. Oliehoek. Coordinated deep reinforcement learners for traffic light control. In *In proceedings of NIPS'16 Workshop on Learning, Inference and Control of Multi-Agent Systems*, 2016.

[26] E. Walraven, M.T.J. Spaan, and B. Bakker. Traffic flow optimization: A reinforcement learning approach. *Engineering Applications of Artificial Intelligence*, 52:203–212, 2016.

[27] Urbica Design. Data-driven transit design. `https://medium.com/@Urbica.co/data-driven-transit-design-9454bc9ed059`, 2016.

[28] A. Chua. *The Role of Data Visualisation in Urban Studies.* PhD thesis, KU Leuven, 2016.

[29] E. Côme and A. Remy. Multiscale spatio-temporal data aggregation and mapping for urban data exploration. In *proceedings of the* 25th *European Symposium on Artificial Neural Networks*, 2017.