Exploratory Analysis of Public Transportation Data in Curitiba

Nádia P. Kozievitch[1], Tatiana M. C. Gadda[1], Keiko V. O. Fonseca[1],
Marcelo O. Rosa[1], Luiz C. Gomes-Jr[1], Monika Akbar[2]

[1] Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, PR – Brazil

[2]University of Texas at El Paso, El Paso, Texas, USA

`{nadiap,tatianagadda,keiko,mrosa}@utfpr.edu.br,`

`gomesjr@dainf.ct.utfpr.edu.br, makbar@utep.edu`

Abstract. *Smart transportation systems have been providing more data over time (such as bus routes, users, smartphones, etc.). Such data provides a number of opportunities to identify various facets of user behavior and traffic trends. In this paper we address some of the urban mobility challenges (already discussed by the Brazilian Computer Society), from a number of different perspectives, including (i) pattern discovery, (ii) statistical analysis, (iii) data integration, and (iv) open and connected data. In particular, we present an exploratory data analysis with GIS for public transportation toward a case study in Curitiba, Brazil.*

1. Introduction

The concept of smart cities is usually linked to efficiency in the use of natural resources [Souza et al. 2015]. Public transportation is a critical factor for the functioning of a city. It provides mobility to the masses and helps to mitigate traffic and pollution. With the advent of smart technologies, urban transportation systems are able to capture a lot of useful data. Such data can be used to shed light on a number of factors, including user trends and traffic patterns. These items are essential for urban planning, optimizing the transportation system (e.g., fuel, time), reducing environmental impacts of mass transport (e.g., noise pollution, air pollution).

Public transportation is one of the most critical areas of smart cities. In Brazil, the vehicle fleet in major cities grew more than the road structure [1]. Mobility challenges have already gained attention of computer science discipline in Brazil [2]. In particular, these challenges can be grouped in the following areas: (i) discovery of patterns, (ii) data statistics, (iii) data integration, (iv) location and tracking, (v) open and connected data, (vi) contextual information, (vii) security and privacy, (viii) energy and management, (ix) use of cloud resources, (x) trajectories with semantic information, among others.

In this paper we address the mobility challenge (by the public transportation view), from the top 5 perspectives previously mentioned, for a case study in Curitiba. Curitiba,

---

[1]   http://www1.folha.uol.com.br/cotidiano/2014/08/1503030-frota-de-veiculos-cresce-mais-rapido-que-a-estrutura-viaria-no-pais.shtml – Last visited on 05/03/2015.
[2]   http://www.sbc.org.br/documentos-da-sbc/send/141-grandes-desafios/802-grandesdesafiosdacomputaonobrasil – Last visited on 05/03/2015.

a city located in the south of Brazil, with 1.8 million people in a total area of 430,9 km^2, according to the Brazilian Institute of Geography and Statistics (IBGE) [3]. This area encompasses 75 districts, and is surrounded by other 29 cities, known as the metropolitan region of Curitiba (with the Portuguese acronym *RMC*).

In particular, Curitiba was already considered one of the world's smartest cities [4] and also belongs to a group of cities, the *C40 cities* [5], which set ambitious targets to improve urban life quality and protect their environment. Curitiba has developed and implemented mass transport corridors, densification of land-use along these corridors, and mobility solutions using Bus Rapid Transit (BRT) systems in the 1970s, where one main feature of the success of the system is its complex network of feeder lines [Duarte et al. 2016].

Routes run in exclusive lanes using high capacity bi-articulated buses, which transport up to 250 passengers per bus. Most of the buses in this system are currently powered by diesel engines. However, due to environmental concerns and sustainable principles, they should be gradually replaced by buses that use other sources of energy, like BEV buses [Sebastiani et al. 2016].

The city has also been participating in the open data initiative, through several government stakeholders, such as Instituto de Planejamento de Curitiba (IPPUC)[6] and the Municipality of Curitiba[7]. In the academy, although it has been studied under distinct perspectives [Zannin et al. 2002, Calixto et al. 2003, Zannin et al. 2003, Sebastiani et al. 2016], we believe that new clues about mobility can be derived from the analysis of new domains.

Briefly, the efforts of integrating different datasets towards exploring mobility can be addressed from different directions, including (i) from the GIS perspective, the challenge is how to integrate different sources, with data having single semantic along with efficiency (and to enable any future integration with other systems); (ii) from the integration perspective, the challenge is not to have a bottleneck; (iii) from the data perspective, the challenge is to detect limitations and standardize a unified view; and (v) from the pedagogical perspective, the challenge is to explore all this structure and have a interdisciplinary learning environment.

The objective here is not to understand the specific details (such as hardware [Suresh et al. 2009] , networks [Park and Kim 2011] or integrate environment-related parameters, but to understand the sources and obstacles which create impact from a general overview, along with a general understanding of the behavior of the system. The main goal is to explore the mobility challenge with an exploratory data analysis with GIS for public transportation toward a case study in Curitiba, Brazil.

The rest of this document is organized as follows: Section 2 presents the related work. Data is presented within Section 3. Behavioral trends are presented in Section 4.

---

[3] `http://www.ibge.gov.br` – Last visited on 14/05/2015.
[4] `http://www.forbes.com/2009/12/03/infrastructure-economy-urban-opinions-columnists-smart-cities-09-joel-kotkin.html` – Last visited on 15/04/2016.
[5] `http://www.c40.org` – Last visited on 30/05/2015.
[6] `http://ippuc.org.br/Lastvisitedon15/05/2015` – Last visited on 05/03/2015.
[7] `http://www.curitiba.pr.gov.br/DADOSABERTOS/` – Last visited on 15/05/2015.

Conclusions and future work are presented in Section 5.

2. Related Work

2.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a philosophical approach to data analysis [NIST/SEMATECH 2012]. The posed question is: what data can tell us about certain relationships, properties or structures. There are no imposed techniques to apply to the data set, but graphical visualization plays an important role in this approach [Hartwig and Dearing 1979]. The non-inferential approach in data analysis encourages the openness perspective required to integrate new domains. The availability of processing power and data storage provide new tools for handling massive amounts of data processing, allowing flexible search of evidences in the available data through designed experiments [Martinez et al. 2010].

2.2. GIS and GIS-T
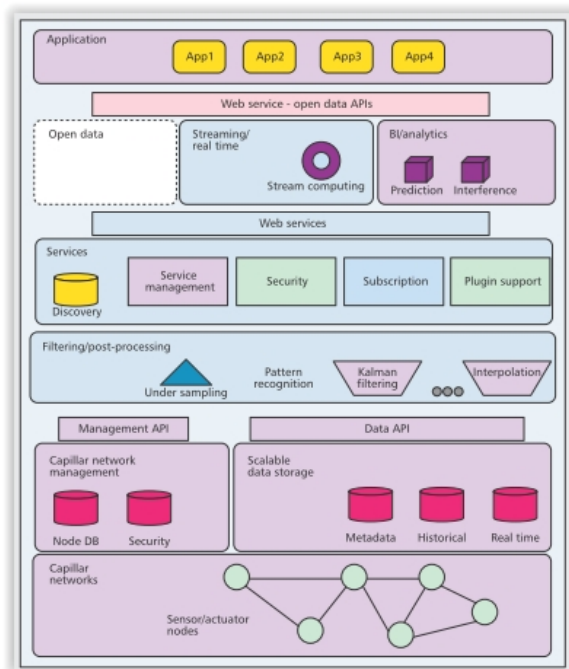
Figure 1. Generic platform showing the main components to support heterogeneous smart city applications [Vilajosana et al. 2013].

Geographic Information System (GIS) stores data from heterogeneous sources in several formats in the form of geodatabases representing spatial features, storing latitude and longitudinal positions. Geodatabases are increasing nowadays generating huge volume of data from satellite images, day traffic, among other sources. Many of these application systems are based on global positioning systems (GPS) [Stenneth et al. 2011], using specific data (such as the real time bus locations, spatial rail and spatial bus stop information) and specific techniques (such as spatial data mining [Mennis and Guo 2009]). An enterprise geodatabase can help solve common transportation challenges [Butler 2008]: the many origins, destinations, paths, and conveyances that may be present, along with their locations.

In particular, geographic information systems for transportation (GIS-T) are interconnected hardware, software, data, people, organizations and institutional arrangements for collecting, storing, analyzing, and communicating particular types of information about the earth [Butler 2008].

Within transportation systems, the GIS database for transportation is organized on the basis of the road network map. Each link can have as attributes topographic (e.g. nodes UTM coordinates, total length), toponomastic (street names), physical (traffic directions, number of lanes), transport (road typology by means of the speed-flow curves) and transit (description of public transport lines and corresponding frequencies) information [Arampatzis et al. 2004]. Here, generally origin destination matrices express the traffic demand.

Several models can be used in order to model the traffic: a deterministic model to simulate road traffic, the representation of the multi-modal transport problem (through simultaneous elastic demand-traffic assignment concepts model), the behavior of public transport users (modeled by using the concept of transit hyper-paths on a generalized network), among others.

Examples of applications include SIGGESC [de Oliveira et al. 2012], which associates the spatial and alphanumeric data of the bus lines with the occupation of the territory, mobility, census information, etc., available to the public through a web portal. General, data for these applications can explore open data, through different categories, such as public transit tracking (such as Chicago Transit Authority Bus Tracker [8]), stakeholders information (such as Chicago Transit Authority [9]), mobile phones [Stenneth et al. 2011], among others. Within the GIS-T challenges, we can mention standards, the representation within databases, unambiguous communication of locations, economic models, response to new technologies, and applications to other fields [Goodchild 2000].

2.3. Smart Cities

Smart Cities was already defined as the integration of technology into a strategic approach to sustainability, citizen well-being, and economic development [10]. Smart city platforms (such as Figure 1) are designed to provide and agglutinate heterogeneous services to support a variety of application domains. In particular, the data integration (GIS, historical, real time) includes sensor nodes, actuators, software components, and other devices that gather data from the city infrastructures and citizens [Vilajosana et al. 2013]. The delivery of the information can take advantage of open data models, without charging any fee to developers and contributors. In Brazil, several communities and projects have been created toward open data [11], smart cities [12], and smarter use of resources [13].

From IT perspective, there are still challenges regarding the architecture of this infrastructure [Chourabi et al. 2012], including (i) lack of integration across government

---

[8] http://www.ctabustracker.com/ – Last visited on 04/04/2016.

[9] http://www.transitchicago.com/about/facts.aspx – Last visited on 04/04/16.

[10] http://www.navigantresearch.com/media/smart-cities-5-security-areas-cios-should-watch – Last visited on 07/04/2015.

[11] http://br.okfn.org/category/dados-abertos/ – Last visited on 15/04/2016.

[12] smartcitiesforumbrasil.com.br/ – Last visited on 15/04/2016.

[13] http://redesinteligentesbrasil.org.br/o-projeto.html – Last visited on 15/04/2015.

systems; (ii) restrictions regarding the integration capabilities of existing internal systems; (iii) lack of knowledge regarding interoperability; and (iv) availability and compatibility of software, systems, and applications.

The objective of this paper is toward this direction, exploring the mobility challenge with an exploratory data analysis with GIS for public transportation toward a case study in Curitiba, Brazil.
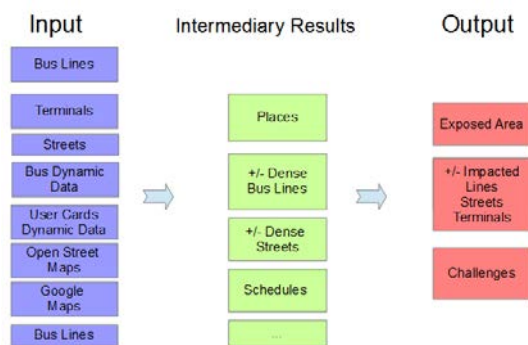

Figure 2. Overview of the dataset.

3. Data

The datasets used in this paper come from IPPUC, the Municipality of Curitiba, along with data from Open Street Map [14]. Figure 2 shows the input sources, divided between stationary and dynamic data. General statistics are presented in Figure 1 [15]. Details are listed below.

Description	Quantity
Official Number of Bus Vehicles	1500
Official Number of Daily Passengers	1.620.000
Official Number of Bus Routes	250
Number of Bus Terminals	23
Number of Tube Bus Stations	342
Average Bus Routes KMs by day	480.000
Average Age of the Fleet	7
Number of Bus Vehicles Detected with data sample	1368
Number of Bus Stops Detected with data sample	9940
Average number of Bus User Cards using GIS Online Tracking	349.729
Average number of Bus Routes using GIS Online Tracking	269
Average number of Bus Vehicles using GIS Online Tracking	1522
Number of Roads within Curitiba	9.135

Table 1. General Statistics of the Data and Bus Network.

Stationary Data. Stationary (or static) data includes georreferenced data from bus lines, bus stops, terminals, and streets that do not change often.

Bus Routes. The city has 482 bus routes distributed within 11 categories (Metropolitano, Linha Direta, Expresso Ligeirão, Interbairros, Expresso, Troncal, Convencional, Turismo, Circular Centro, Interhospitais, Alimentador), as shown by green lines in Figure 3. The categories Alimentador and Convencional have the majority of lines, with 265 and 65 units respectively.

---

[14] http://www.openstreetmap.org – Last visited on 14/05/2015.

[15] https://www.urbs.curitiba.pr.gov.br/ –Last visited on 04/04/2016.

Bus Stops. The city has 9940 bus stops detected within the stationary data. Curitiba has an additional bus stop category named Tube Bus Stations (in portuguese, *Estação Tubo*). The tube stations (officially 342) are bus stops which look like tubes, for specific bus routes, such as Expresso and Linha Direta. The districts named CIC and Centro have the majority of them, with a total of 1628 and 667 units each one.

Bus Terminals. The city has 23 terminals (buses) and one terminal which also use trains, as shown by red points in Figure 3. The oldest one is named Guadalupe, from January 1st, 1956. The districts CIC and Boqueirão concentrate their majority, with 3 and 2 units each one.

Streets. The city has 9.135 streets, divided among 17 categories (Anel Central, Central, Coletora 1, Coletora 2, Coletora 3, Externa, Linhão, Normal, Outras Vias, Pedestre, Prioritária 1, Prioritária 2, Rodovia Estadual Duplicada, Rodovia Estadual Simples, Rodovia Federal Duplicada, Rodovia Federal Simples, and Setorial). The districts CIC and Sítio Cercado have their majority, with 1217 e 524, each one.

Figure 5 presents the density of how bus lines are present in the city. Gray lines present streets with an average of 1-17 lines, the green lines represent an average of 17-37 lines, the pink lines represent an average of 37-55 lines, and the red lines represent streets with 55-73 lines. The zoom at the same figure presents the density of bus lines within the downtown district, having each edge the respective number of lines. For the same region with zoom within Figure 5, we calculated the *Eigenvector Centrality* [da F. Costa et al. 2007] of the the bus lines, as shown in Figure 6. The Eigenvector Centrality (also the basis for the PageRank algorithm) captures the influence of nodes in a network, taking into account that nodes with influential neighbors should have higher centrality values. In a bus line network, high centrality values indicate points of (potentially) intense traffic. Identifying the points with high centrality score in the Figure reaffirms our observation that traffic within the Guadalupe Bus Terminal (the green triangle at the Figure) concentrates intense traffic.

Dynamic Data Dynamic data includes data from bus lines, and user cards, transmitted at an average frequency of 5 minutes. In particular, the data selected was between October 19th, 2015 and October 26th, 2015.

User Cards. This set of data represents the commuting information comprising trip data per user card: the vehicle id, line code, the user card number, and the date of the trip. Within the date range selected, it was detected 349.729 different user cards, 269 different bus lines were present, divided among 1522 vehicles. Part of the vehicles and lines represented here have geolocated data, presented within the following set.

Vehicles and Respective Bus lines. This set of data represents the daily itinerary, having as data the vehicle id, line code, the datetime, and the latitude and longitude. Within the data range selected, it was detected 1623 distinct vehicles and 703 different bus lines.

The complete dataset were inserted in a PostGIS [16] database. Different sources were created as different tables in the database. Later, specific tablespaces and indexes were created in order to optimize the access. Since semantic errors were present (different sources presented different street names for the same location, for exam-

---

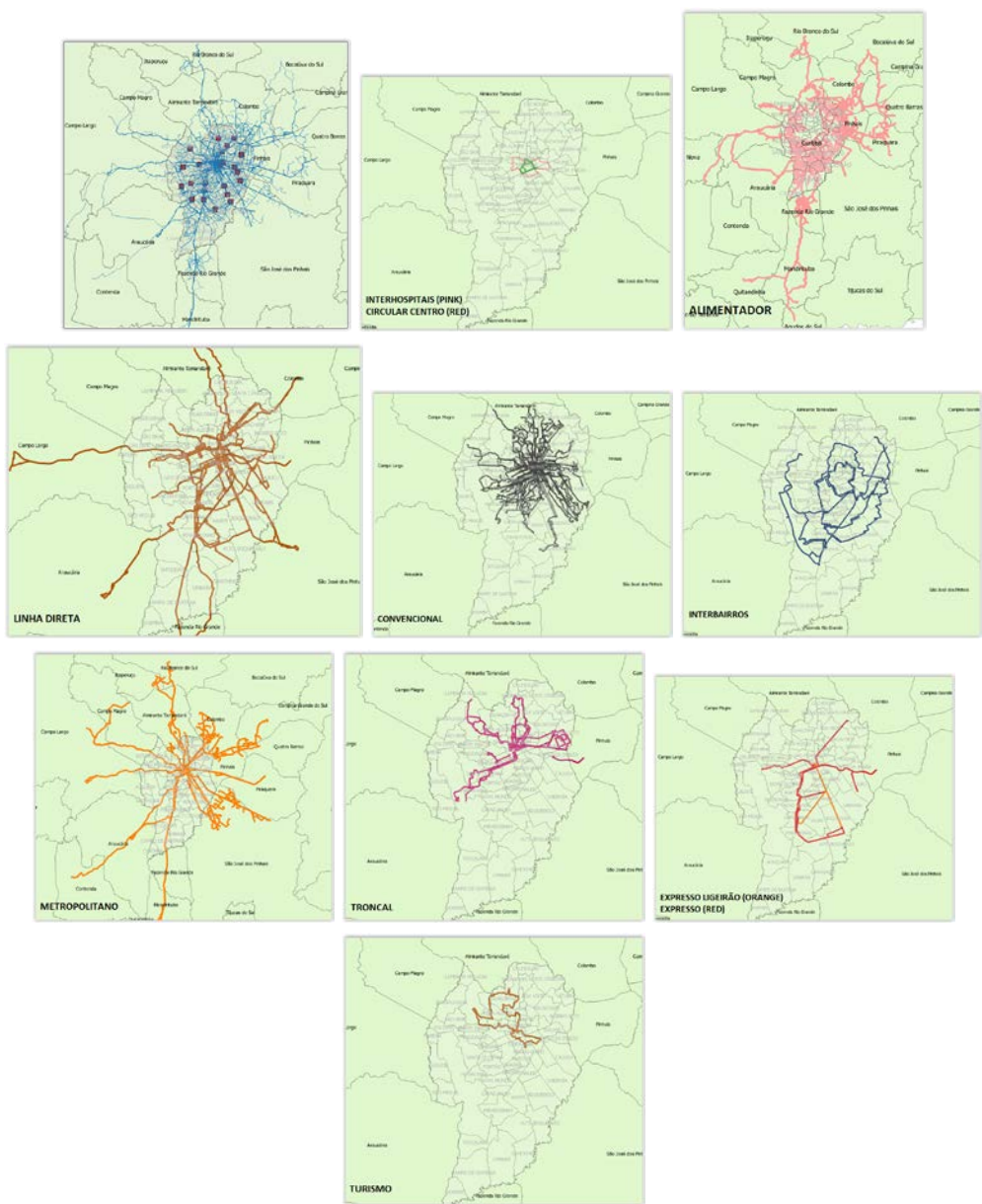[16] `http://www.postgis.net` – Last visited on 15/05/2014.

Figure 3. Representation of bus lines (blue) and bus terminals (red) on upper left, followed by the eleven bus categories.

ple [Barczyszyn 2015]), geolocation and specified time range were used to correlate the data.

4. Behavioral trends

The analysis of dynamic data from user cards indicated some trends: **(i)**: Over the week of data analyzed, there is an average of 350 bus lines, an average of 1500 vehicles, and an average of 250.000 user cards (Figure 7) per day. Sundays presented the lowest number of user cards. **(ii)**: The analysis of user cards by hour indicated that the peak hours for traffic are 7 AM, 5 PM, and 6 PM (Figure 8). Friday (October 23rd, 2015) presented the highest amount of different user cards (264.224), and Sunday (October 25th, 2015) presented the lowest amount of user cards (53.510). **(iii)**: The top 5 bus lines which had the majority of the user cards were during the time range were "Oper s/Linha", "Op. Contingencia",
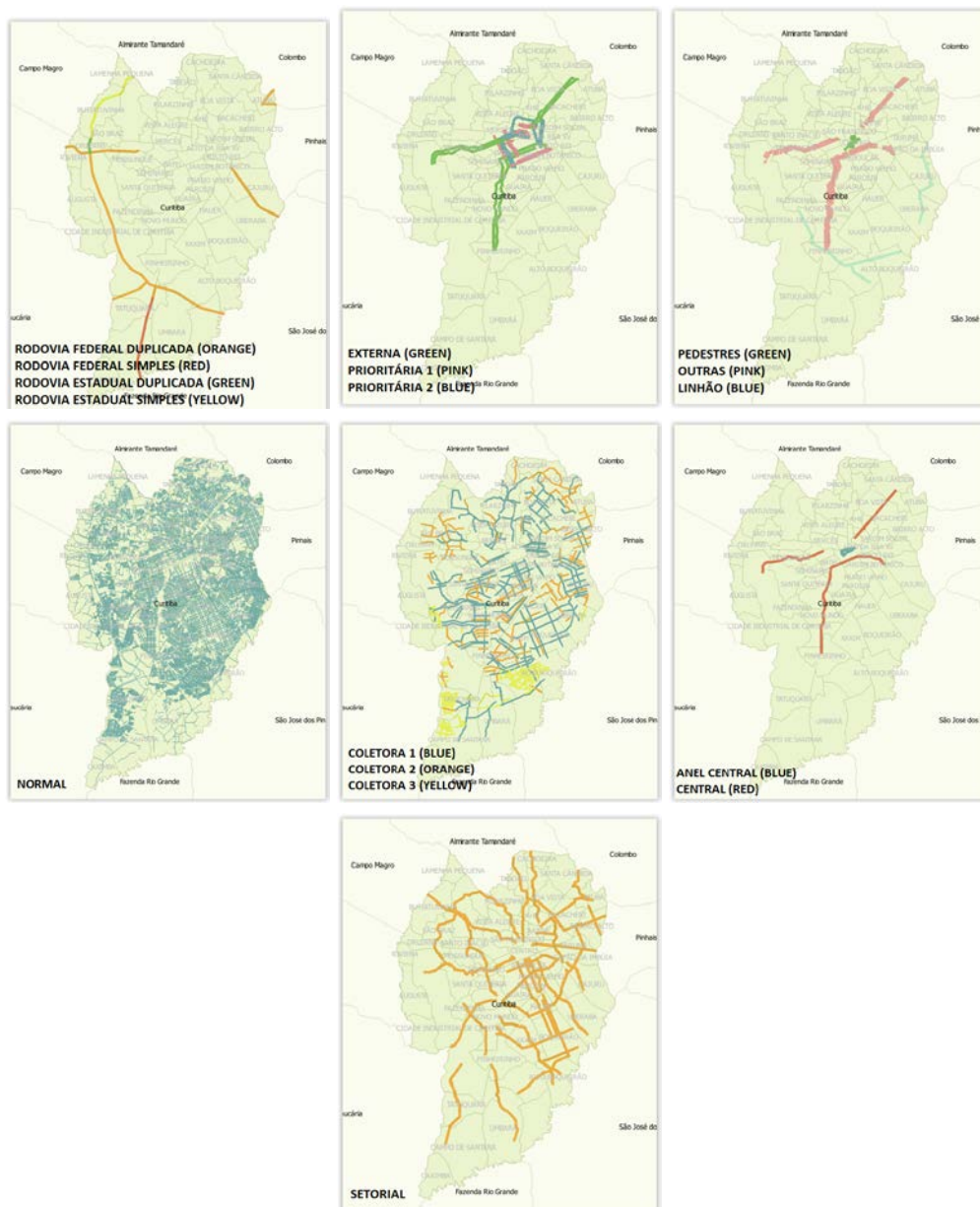
Figure 4. Representation of streets, by each category.

"Sistema Araucária", "Interbairros IV" and "Interb. II Anti H". The first one indicated bus operating without a specific line. The second one indicated extra buses to the already existing lines. The third one connects the metropolitan region to Curitiba. The last two lines operate within the Interbairros category. These lines connect several districts.

The analysis of the dynamic data from bus lines also presented particular characteristics:

- The analysis of Interbairros IV bus line data indicated that the vehicles were also used to other lines at the same category (as indicated in Figure 9: instead of just using the blue line indicated in the figure, vehicles were also using other lines).
- Routes which are not included within the respective line, were also present (Region A in Figure 9). The analysis of which streets are used by vehicles to connect different lines and how they are used during the peak hours could be used for the
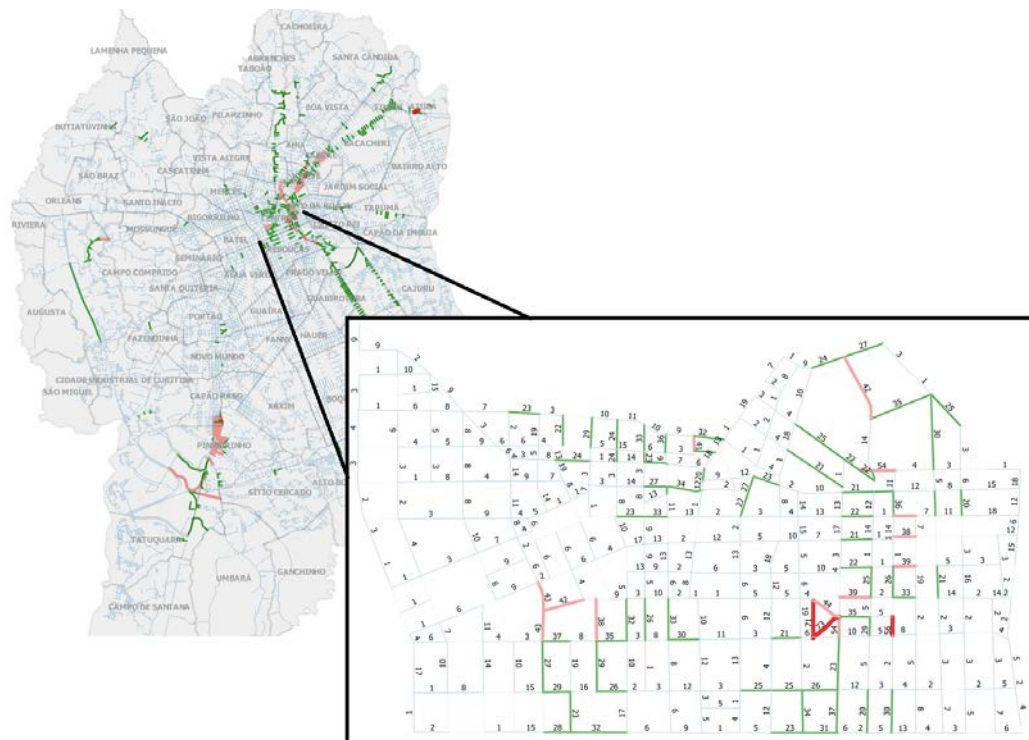
Figure 5. Density of Bus Lines within streets in Curitiba, with zoom in the central district. Pink and red streets present higher bus densities.
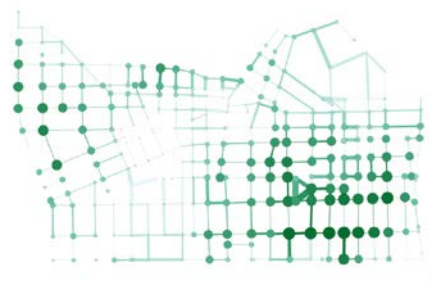
Figure 6. Eigenvector Centrality of the bus lines within the Central District. Larger and darker circles indicate higher centrality scores.

planning of the routes to optimize various parameters of the transportation system of the city.

- In parallel, routes which are regularly accessed, but are not present by main route were also detected (Region B in Figure 9). The behavior was present within both the peak and off-peak hours.

- The busiest route (Interbairros IV - blue line at Figure 9 - left) surrounds the central district, connecting several different neighborhoods and five bus terminals (Santa Felicidade, Campo Comprido, Fazendinha, CIC e Pinheirinho). The data indicates peaks for six A.M. (average of 902 users), seven A.M. (average of 938 users) and six P.M. (average of 604 users) within the bus line.

In summary, some of the detected patterns and associated challenges are as follows: (i) **Data Integration**: the different coordinate systems, formats, and information along the several government institutions keep the data integration as a challenge, even considering only Brazilian cities (such as São Paulo - through the Municipality of São
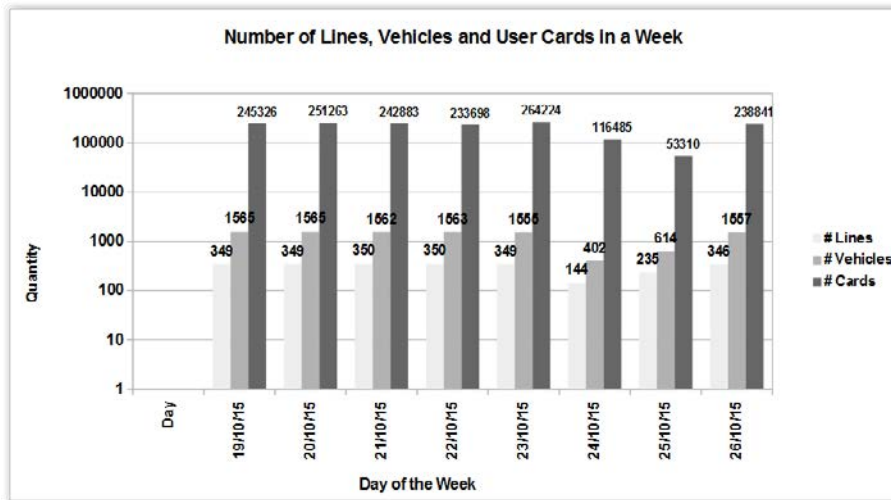
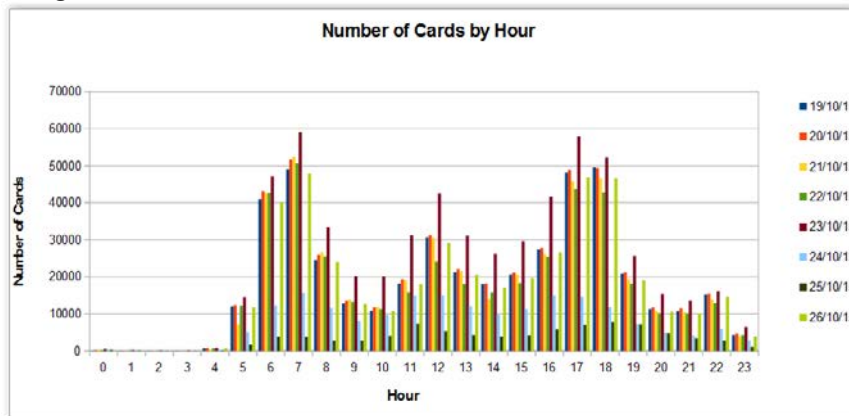**Figure 7. Number of Lines, Vehicles, and User Cards in a Week.**

Figure 8. Number of user cards by hour.

Figure 9. Representation of Interbairros IV line, during all day (left), seven A.M. (center) and five P.M. (left).

Paulo[17]), Recife (through the Municipality of Recife[18]) and Rio de Janeiro (through Sistema Municipal de Informações Urbanas - RJ [19])). In general, formats such as CSV, Excel,

---

[17] http://dados.prefeitura.sp.gov.br/ – Last visited on 02/06/2015.

[18] http://dados.recife.pe.gov.br/dataset – Last visited on 15/05/2015.

[19] http://portalgeo.pcrj.opendata.arcgis.com/ – Last visited on 15/05/2015.

Json and Shapefile are preferred (some use APIs and KML). Not all of them provide metadata or the data visualization. Some services still have issues (such as the download of spatial files from the Municipality of Rio de Janeiro). In particular, the data within the use case had text, excel and shapefile formats; (ii) **Data Statistics**: initial data statistics indicated impacts by peak hours, days of the week, contingency and metropolitan vehicles, among others. Visualization indicated vehicles use among several bus lines, with intermediary routes to connect them; (iii) **Open and Connected Data**: the use case presented used Open Data resulted from decree 1135/2012, from law 12.527 [20]. Besides this policy, the municipality is also using the data for competitions named Hackatons[21], as well as integration of partners for future projects (regional and international)[22][23].

5. Conclusion

While the number of vehicles operating in a city increases rapidly, the change in infrastructure is a rather slow process. In order to accommodate the users and their transportation needs, a city must carefully analyze data sources to determine the user needs and possible changes in transportation to support those needs. This paper presented an initial investigation in order to identify scenarios and implications from the public transportation data from the city of Curitiba, integrating several sources, including stationary and dynamic open data. Detected characteristics can be used for further analysis in order to optimize the transportation system. This work will be extended to provide recommendation and optimization for alternative and existing bus routes that minimizes various parameters including travel time and number of transfers. Lastly, we will investigate various approaches for integrating other data sources.

Acknowledgment

References

Arampatzis, G., Kiranoudis, C., Scaloubacas, P., and Assimacopoulos, D. (2004). A gis-based decision support system for planning urban transportation policies. *European Journal of Operational Research*, 152(2):465 – 475. New Technologies in Transportation Systems.

Barczyszyn, G. L. (2015). Integração de dados geográficos para planejamento urbano da cidade de Curitiba. Universidade Tecnológica Federal do Paraná.

Butler, J. A. (2008). *Designing Geodatabases for Transportation*. Esri Press.

Calixto, A., Diniz, F. B., and Zannin, P. (2003). The statistical modeling of road traffic noise in an urban setting. *Cities*, 20:1–74.

---

[20] `http://multimidia.curitiba.pr.gov.br/2014/00147194.pdf` – Last visited on 07/04/2015.

[21] `http://hackathon.curitiba.pr.gov.br/` – Last visited on 15/05/2015.

[22] `http://www.curitiba.pr.gov.br/noticias/curitiba-e-holanda-vao-trabalhar-juntas-em-projetos-de-ciclomobilidade-para-a-cidade/37601` – Last visited on 14/04/2016.

[23] `http://multimidia.curitiba.pr.gov.br/2015/00166636.pdf` – Last visited on 14/04/2016.

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., Pardo, T. A., and Scholl, H. J. (2012). Understanding smart cities: An integrative framework. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2289–2297.

da F. Costa, L., Rodrigues, F. A., Travieso, G., and Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.

de Oliveira, T. H. M., Painho, M., and Henriques, R. (2012). A spatial decision support system for the portuguese public transportation sector. In *IWGS '12*, pages 84–90, New York, NY, USA. ACM.

Duarte, F., Gadda, T., Luna, C. A. M., and Souza, F. T. (2016). What to expect from the future leaders of bogotá and curitiba in terms of public transport: Opinions and practices among university students. *Transportation Research Part F: Traffic Psychology and Behaviour*, 38:7 – 21.

Goodchild, M. F. (2000). Gis and transportation: Status and challenges. *Geoinformatica*, 4(2):127–139.

Hartwig, F. and Dearing, B. (1979). *Exploratory Data Analysis*. 07. SAGE Publications.

Martinez, W., Martinez, A., and Solka, J. (2010). *Exploratory Data Analysis with MATLAB, Second Edition*. Chapman & Hall/CRC Computer Science & Data Analysis. Taylor & Francis.

Mennis, J. and Guo, D. (2009). Spatial data mining and geographic knowledge discoveryan introduction. *Computers, Environment and Urban Systems*, 33(6):403 – 408. Spatial Data Mining-Methods and Applications.

NIST/SEMATECH (2012). *E-Handbook of Statistical Methods, available at http://www.itl.nist.gov/div898/handbook/*.

Park, H.-S. and Kim, J.-D. (2011). Modeling and Analysis of DTN in Metropolitan Bus Network. In *ICUIMC '11*, pages 20:1–20:10, New York, NY, USA. ACM.

Sebastiani, M. T., Luders, R., and Fonseca, K. V. O. (2016). Evaluating electric bus operation for a real-world brt public transportation using simulation optimization. *IEEE Transactions on Intelligent Transportation Systems*, PP(99):1–10.

Souza, R., Oliveira, I. P., Junior, F., Sales, L., and Ferraz, F. (2015). Beyond efficiency: How to use geolocation applications to improve citizens well-being. In *The Fourth International Conference on Smart Systems, Devices and Technologies*, pages 37 – 40.

Stenneth, L., Wolfson, O., Yu, P. S., and Xu, B. (2011). Transportation mode detection using mobile phones and gis information. In *GIS '11*, pages 54–63, New York, NY, USA. ACM.

Suresh, D. C., Agrawal, B., Yang, J., and Najjar, W. (2009). Energy-efficient encoding techniques for off-chip data buses. *ACM Trans. Embed. Comput. Syst.*, 8(2):9:1–9:23.

Vilajosana, I., Llosa, J., Martinez, B., Domingo-Prieto, M., Angles, A., and Vilajosana, X. (2013). Bootstrapping smart cities through a self-sustainable model based on big data flows. *IEEE Communications Magazine*, 51(6):128–134.

Zannin, P. H. T., Calixto, A., Diniz, F. B., and C., J. A. (2003). A survey of urban noise annoyance in a large brazilian city: the importance of a subjective analysis in conjunction with an objective *Environmental Impact Assessment Review*, 23:245–255.

Zannin, P. H. T., Diniz, F. B., and Barbosa, W. A. (2002). Environmental noise pollution in the city of Curitiba, Brazil . *Applied Acoustics*, 63(4):351 – 358.