Revista Brasileira de Computação Aplicada, July, 2021

DOI: 10.5335/rbca.v13i2.12351
Vol. 13, Nº 2, pp. 16–27
Homepage: seer.upf.br/index.php/rbca/index

ORIGINAL PAPER

Origin-Destination Data: a prototype and related scenarios

Yussef Parcianello [1], Nádia P. Kozievitch [1], Keiko V. O. Fonseca [1], Marcelo Rosa [1]

[1]Federal University of Technology - Paraná

*yussef.parcianello@ifsc.edu.br; nadiap@utfpr.edu.br; keiko@utfpr.edu.br; mrosa@utfpr.edu.br

Received: 2021-03-14. **Revised:** 2021-03-29. **Accepted:** 2021-06-22.

Abstract

The Public Transportation System and its operation management require the processing of large amount of data (like bus routes, user data and bus schedules). In particular, origin-destination data serves to indicate citizens' travel patterns, providing insights related to the dynamic of the urban space occupation. Given this scenario, this paper presents a prototype of origin-destination data visualization, based on queries associated with a set of trips (and related attributes), analysis of trips and services for Curitiba, clustering of georreferenced data (for visualization) and a local study of Origin-Destination from the Public Transportation of Curitiba. The novelty relies on visualization through clustering of georreferenced data, allowing the analysis of different regions of interests (neighborhood, regionals or mathematical regions using K-means algorithm). We demonstrate the prototype through several scenarios, and interviews done to local citizens. Challenges related to meaningful presentation of results are discussed under the perspective of visualization and analytics.

Keywords: Data Visualization; GIS; Origin-Destination; Public Transportation System.

Resumo

O Sistema de Transporte Público e sua gestão operacional requerem o processamento de grande quantidade de dados (como rotas de ônibus, dados de usuários e horários de ônibus). Em particular, os dados de origem-destino servem para indicar os padrões de viagem dos cidadãos, fornecendo insights relacionados à dinâmica da ocupação do espaço urbano. Diante desse cenário, este artigo apresenta um protótipo de visualização de dados de origem-destino, baseado em consultas associadas com um conjunto de viagens (e seus atributos relacionados), análise de viagens e serviços para Curitiba, clusterização de dados georreferenciados (para visualização) e um estudo local de Origem-Destino do Transporte Público de Curitiba. A novidade está na visualização por agrupamento de dados georreferenciados, permitindo a análise de diferentes regiões de interesse (vizinhança, regionais ou regiões matemáticas por meio do algoritmo K-means). Demonstramos o protótipo por meio de vários cenários e entrevistas feitas a cidadãos locais. Os desafios relacionados à apresentação significativa dos resultados são discutidos sob a perspectiva da visualização e análise.

Palavras-Chave: GIS; Origem-Destino; Sistema Público de Transporte; Visualização de Dados.

1 Introduction

The growth of urban centers brings challenges related to urban mobility that impact the well-being of the population. In this scenario, the public administration has sought to follow the Open Data trend, making city data available in its Open Data Portals (such as Paris[1], Nova Iorque[2] and Moscow[3]). In Brazil, the city of Curitiba followed this trend, by making its data available through

[1] opendata.paris.fr
[2] opendata.cityofnewyork.us
[3] data.gov.ru

several sources (City Hall Portal[4], Instituto de Pesquisa e Planejamento Urbano (IPPUC)[5], Urbanização de Curitiba (URBS)[6], TransitFeeds[7] and GTFS [8]).

In Curitiba, each day 1229 vehicles carry over 1.365.615 passengers (where 60% of them use smartcards) through 251 different routes and 24 bus terminals and resulting, on average, 14.166 bus trips on 300.373 kms by day over 9940 bus stops. If we consider one month of data from the user smartcards pick-ups (Sep 30th 2017 - Oct 30th 2017) on Fig. 1, weekends (such as Sept 30th, Oct 1st, Oct 7th, Oct 8th) have a lower average compared to week days, with a total of 8.710.082 use of smartcards in the period. The number of routes and vehicles are also impacted by weekends.

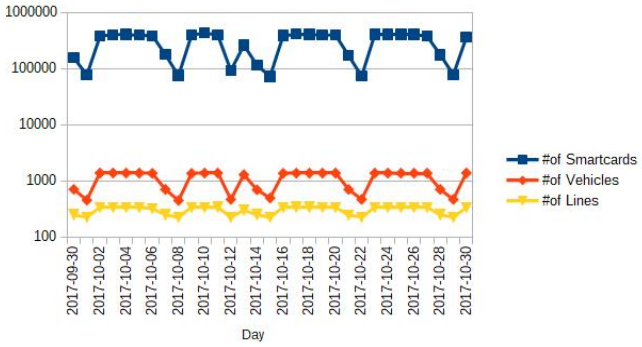

Figure 1: Number of Smartcards/Vehicles/Lines by day in one month (Sep 30th 2017 - Oct 30th 2017).

From a formal perspective, a transport network has been defined by Añez et al. (1996) as a set of directed links of the form $N = (V, L)$, where $V = \{v_1, v_2, ..., v_n\}$ is a set of vertices and $L = \{v_1, v_2, ..., v_n\}$ is a set of links, such that $L = (v, w, Q^{vw}), v, w \in V$, where v and w are origin and destination nodes, and Q^{vw} is a set of attributes of each link such as distance, capacity, number of passengers, vehicle speed. Since the origin and destination are node numbers, once all such calculations have been performed, the model can be transformed back in the original graph notation, without any additional computational effort. The same model can be extended to multidimentional networks, to represent transit routes that share a physical link. Alternative methods include arc-vertex and forward-star (Añez et al., 1996). If we consider the definition above along with the data from Curitiba, L could be presented as shown in Table 1, where v and w could be a bus stop or a bus terminal, and Q^{vw} is a set of parameters available from the local bus administration.

From GIS and data perspective, the public transport system has geographic and temporal components, as well as data on its dynamics, for example, about travel. A bus trip (where a user goes from an origin to a destination and comes back to an origin) can be characterized by dates, departures and arrivals (namely origin-destination data) associated with a bus line (predetermined route) at passenger pick-up and drop-off points. A trip can also have information such as name and line code, vehicle code, user smartcard id, gender, date of birth, latitude and longitude of the bus (for example, every five minutes). In GIS perspective, bus stops, terminals, bus trips, passenger pick-up an drop-offs are stored in low-level features such as points, lines and polygons, with known latitude and longitude. The analysis can be focused on places, tasks, phenomena, cause-effect, OD flows and simulations, among others. In order to understand the OD dynamics, in general, users and Public Transportation System (PTS) managers are interviewed and reports emmited about the information collected (Pichiliani, 2017). For manipulation and processing of the referred data it is possible to use tools such as ArcGIS, QGIS and Excel, where generally only part of the data are analyzed. More complex analysis require knowledge of specific languages (such as SQL queries within QGIS), in time consuming tasks.

In this direction, this paper presents a visual tool to support spatial-temporal queries over OD data, having the following as requirements: 1) to be able to visualize queries associated with a set of bus trips (and related attributes) for non experts (Ferreira et al., 2013), 2) to visualize details of analysis of bus trips and services focusing on particularities of Curitiba (Diniz Jr., 2017), 3) to be able to scale the visualization using clustering of georreferenced data (Vila et al., 2016), and 4) to enhance the OD analysis from (Pichiliani, 2017), along with its particular characteristics (such as having an interface similar to Tableau) and variables (similar division for age, time, regions and week aggregation). The novelty relies on visualization through clustering of georreferenced data (Vila et al., 2016), allowing the analysis of different regions of interests (neighborhood, regionals or mathematic regions using K-means algorithm). From the data perspective, the visualization can provide both overview and details, maintaining the spatial and temporal contexts. We demonstrate our tool through several scenarios motivated by a local study of Origin-Destination (Pichiliani, 2017), and interviews done to local citizens.

This paper is organized as follows. Section 2 gives an overview of related work. We present the prototype at Section 3. Section 4 presents the case studies. Finally, we present our conclusion at Section 5.

2 Related Work

Public transportation is one of the most critical areas under the city perspective. Mobility challenges have already gained attention of computer science community in Brazil[9]. In particular, these challenges could be grouped in the following areas (Vila et al., 2016): (i) discovery of patterns, (ii) data statistics, (iii) data integration, (iv) location and tracking, (v) open and connected data, (vi)

[4] www.curitiba.pr.gov.br/dadosabertos/
[5] ippuc.org.br
[6] urbs.curitiba.pr.gov.br/
[7] https://transitfeeds.com/l/388-brazil/
[8] https://developers.google.com/transit/gtfs/reference

[9] http://www.sbc.org.br/documentos-da-sbc/send/141-grandes-desafios/802-grandesdesafiosdacomputao/no-brasil

Table 1: Set L, considering origin and destination data for Curitiba.

v	w	Q^{vw}
(-25.460766;-49.345255)	(-25.436778;-49.274341)	("Linenumber=INTERBAIRROS IV", "Vehicle=MC304";"Smartcard=0000558750"; "DateTime=2017-10-04 06:20:27", "Birthdate=1987-09-17";"Gender=F")
(-25.38542;-49.28037)	(-25.417515;-49.246675)	("Linenumber=INTERBAIRR II","Vehicle=DR113","Smartcard=0000558806", "DateTime=2017-10-03 07:26:24", "Birthdate=1987-09-16";"Gender=F")
...		
(-25.422613;-49.300685)	(-25.430181;-49.2724)	("Linenumber=BIGORRILHO","Vehicle=BC852";"Smartcard=0000559715"; "Datetime=2017-10-04 14:38:16", "Birthdate=1982-11-04";"Gender=F")

contextual information, (vii) security and privacy, (viii) energy and management, (ix) use of cloud resources, and (x) trajectories with semantic information, among others.

If we consider legislation, NBR ISO37120:2017 ("Desenvolvimento sustentável de comunidades - Indicadores para serviços urbanos e qualidade de vida"[10]) could be cited as the Brazilian technical standard for sustainable cities. This standard is a translation and adaptation of the standard ISO37120:2014 - "Sustainable development of communities - Indicators for city services and quality of life"[11], elaborated by TC-268 (*Technical Committee*)[12]. NBR ISO37120:2017 proposed a standardization of indicators related to the city (urban services offered and quality of life, among others). The objective was to provide opportunities for comparative analyzes of different communities, favoring the exchange of experiences and good practices (Couto, 2018).

From the information perspective, several efforts could be listed. Curitiba and New York, for example, were analyzed (Parcianello et al., 2018), in a comparative study based on open data. This study identified similarities and contrasts between the existing systems and highlighted the importance of seeking to implement an inter and multimodal public transport system.

Different transportation telematic services were proposed in Diniz Jr. (2017) (such as location without bus routes, average bus crowdness, alert for different routes and speeding alerts), along with efficiency (Braz et al., 2018), using the same data proposed in this paper.

The comparison of open data on road network, demographics, territorial extension and other urban indicators from 645 cities of the state of São Paulo was proposed in Spadon et al. (2018). The study also applied complex network concepts and clustering algorithms to classify such cities by similarity from different perspectives.

The public transportation system from Rio de Janeiro was used in the study of Cruz et al. (2018), where the identification and classification of anomalies (using open text and data mining) using the Apriori algorithm.

The comparative analysis using complex network models in Chicago and Melbourne was presented in Saberi et al. (2017), as a first step towards a better understanding of the structure, interactions, and evolution of travel demand networks in cities. They suggested that the underlying processes in travel demand, viewed as a network, are also driven by the interaction strength between places (or nodes).

Images in GIS is an important task, but in transportation, the movement is crucial (examples in Andrienko and Andrienko (2013)) . As already mentioned in Ferreira et al. (2013): much of the work is based in trajectory data, where the location of moving entities is recorded. In contrast, multi-variate OD data has only the start and end positions, together with attributes associated with the movement. At the same paper, taxi data visualization is studied, and a solution is proposed for non experts, with different visualization approaches.

Regarding data apresentation, several projects might be mentioned, such as DataViz[13], VisualComplexity[14] or CityGeographics[15]. On the other way, open data have already coming with interactive solutions, such as Manhattan Population Explorer[16] or Transit Accident Dashboard from IPPUC[17].

In Vila (2016), a web-mobile solution was proposed, allowing spatiotemporal use of georeferenced data. For the presentation of results, graphic resources were used as markers, thermal map and clustering of markers.

Studies involving OD can benefit from the adoption of different data visualization techniques Andrienko et al. (2020), Itoh et al. (2016), Palomo et al. (2016), Wood et al. (2010), Guerra (2011).

According to Guerra (2011), OD studies aim to identify the amount of displacements made and the profile of the citizen who travels over a period of time from a home zone to a destination zone. These zones can be defined from geographic divisions (based on neighborhoods and macroregions, for example), via mathematical divisions (via data clustering techniques), among others. A general framework for using visual analytics techniques and workflows in place connectedness studies is presented by Andrienko et al. (2020), using place-centered tasks, link-centered tasks, interemediate-level tasks along with their costs. In Itoh et al. (2016), unusual phenomena (ex. marathons) and their propagation (cause/effect) on a spatio-temporal space is presented through visualization, using visualizations such as heatmaps, AnnimatedRibbon and TweetBubble. In Palomo et al. (2016) is presented a visual exploration tool (composed by trip and stop explorer), developed to identify, inspect and compare spatio-temporal patterns for planned and real transportation service. In Wood et al. (2010), OD vectors are mapped as cells rather than lines, using a hash grid spatial data structure for enhance scalability to large collection of vectors.

[10] abntcatalogo.com.br/norma.aspx?ID=366389
[11] iso.org/standard/62436.html
[12] iso.org/committee/656906.html
[13] datavizproject.com
[14] www.visualcomplexity.com
[15] citygeographics.org
[16] manpopex.us
[17] ippuc.org.br/mapasinterativos/AcidentesDeTransito/dashboard.html

The main challenges toward the use of several approaches for OD data include: 1) the performance of processing and visualizing a huge amount of data; 2) the integration of different technologies (not all of them are compatible); 3) neither the available software is free nor the source code is not easy to understand and integrate with other software, among others (free datasets, metadata).

Several theoretical approaches for Public Transport Systems can be mentioned, such as graphs (Silva et al., 2016, Chapleau and Morency, 2005, Zhang et al., 2015), Marey's graphs (Palomo et al., 2016), matrix (Diniz Jr., 2017) and hash grids (Wood et al., 2010).

Clustering algorithms can also be used to analyze PTS. According to Cassiano (2014), data clustering (or cluster analysis) is a multivariate data mining technique that aims to group the n database cases into k groups called clusters. Data clustering can also be defined as the process as a grouping of information, considering: (i) the existence of a strong similarity between the elements belonging to the same group; (ii) existence of a weak similarity of elements belonging to different groups (Zaiane et al., 2002). In literature, the clustering can also be called cluster analysis, Clustering, Q-analysis,Typology, Classification Analysis or Numerical Taxonomy.

In particular, K-means clustering algorithm Silva et al. (2016), Osama et al. (2015) was used to partionate the regions into K groups within this paper, using their geographic position (given by latitude and longitude coordinates). After setting centroid coordinates randomly for each group, the algorithm basically consists of alternating between two steps: the assignment step where each region is assigned to its nearest group (considering the group centroid coordinates), and the update step where the group centroid coordinates are updated according to its assigned region. The K value as 500, for example, is used to present smaller regions of the city.

3 Prototype

3.1 Requirements.

For the prototype, a questionnaire was submitted to users of the PTS in academic community. This questionnaire was applied in April 17 to 24, 2019 [18] and it was composed by questions such as: a) if the citizen had already any contact with studies related to the use of the Curitiba PTS, b) if it would be interesting to develop a solution that would allow visualizing, quantifying and exploring data from the Curitiba PTS, c) which search filters the solution should offer and d) what would be the possibilities of using this solution. From the analysis of the answers obtained, we notice that: (a) all interviewed citizens had already had some contact with studies involving the use of the PTS Curitiba; (b) a solution that would allow them to view, quantify and explore data related to the use of public transport would be interesting; (c) filters should be an important feature for the solution. As for the search filters that the solution should have (listed in Fig. 2), the most highly rated were: 1) trips by drop-in and drop-off regions, 2) trips by age group of users and 3) trips based on passenger gender. Users also indicated that such a solution would be applicable 1) to the study of public transport demand, 2) to user profile analysis and 3) to study the dynamics of urban space occupation. These answers subsidized the development of a prototype designed to allow users to easily perform OD analysis.

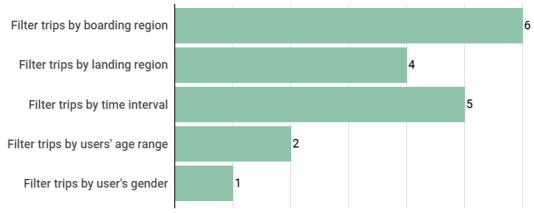

Figure 2: The main data filters requested by users.

The report for OD data in Curitiba, from IPPUC (Pichiliani, 2017) was also used in order to understand which analysis were necessary and which variables and classifications were used, such as gender ("F" or "M"), day shift classifications (00:00 to 05:59, 06:00 to 11:59, 12:00 to 15:59, 16:00 to 23:59), age classification (0 to 5 years, 5 to 12 years, 12 to 18 years, 18 to 65, greater than 65 years), along with the classification by neighborhood an regionals. Note that similar characteristics will be used for visualization in this prototype, such as having an interface similar to Tableau, and variables aggregation (similar division for age, time, regions and week aggregation). In particular, the age range "0 to 5 years" and "greater than 65 years" do not pay any fee for using the buses in Curitiba. The age range "5 to 12 years" is used for children, and "12 to 18" used for adolescents, according to the Statute of Children and Adolescents in Brazil[19]. The age range "18 to 65" is used for other users from the PTS. Note that the traditional way of OD analysis is by interviews.

In particular, this proposal led to the following requirements: 1) to be able to visualize queries associated with a set of bus trips (and related attributes) for non experts (Ferreira et al., 2013), 2) to visualize details of analysis of bus trips and services for Curitiba mentioned in Diniz Jr. (2017), 3) to be able to scale the visualization using clustering of georreferenced data(Vila et al., 2016), with further options other than districts, and 4) to enhance the OD analysis from Pichiliani (2017), along with its particular characteristics (such as Tableau[20], providing a visualization of both overview and details) and variables.

[18] Available at: https://drive.google.com/file/d/1unZsrXWMTmb3LZdJ1z8Cnf5ZfUnCyppc/view?usp=sharing

[19] http://www.planalto.gov.br/ccivil_03/leis/l8069.htm

[20] http://www.tableau.com

3.2 Technologies.

The database used PostgreSQL 9.5 x64[21], PostGIS 2.1.1[22] in a two cores of an AMD EPYC 7401 with 1.5G of memory, running Debian 9 runs with 5.1 kernel. For the application, the server was Intel Core i7 3632QM 4-core 8-theads x64 2.2GHz, 8GB RAM, with Windows 7 64 running the following software: Apache Server v. 2.4.37[23], PHP v. 7.2.12[24], Open Street Map, Leaflet v. 0.7.7[25], Leaflet Markerclusterer v. 0.5.0[26], Leaflet Heatmap v. 0.7.7[27], Leaflet Draw v. 0.4.2[28], JQuery v. 3.3.1[29], Bootstrap v. 3.3.7[30], Bootstrap Datepicker v. 1.6.4[31], Bootstrap Datetimepicker v. 4.17.47[32] and Chart.js v. 2.7.3[33]. Fig. 3 presents how the tecnologies were combined within the prototype.

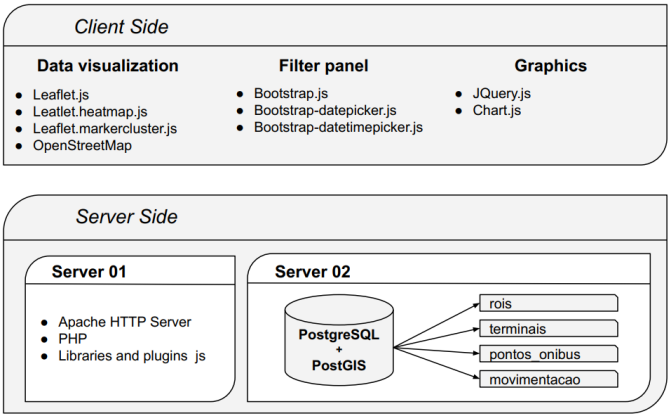

Figure 3: Technologies used in the prototype.

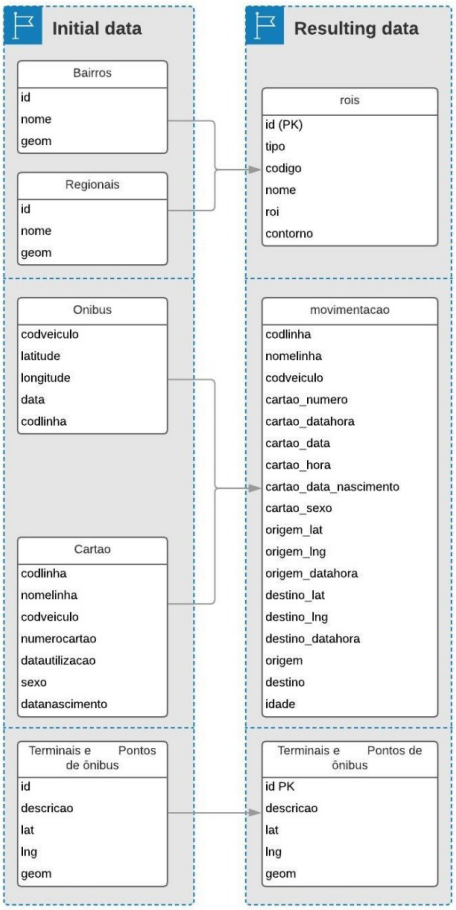

Figure 4: Data structure used in the prototype: the initial data (left) along with the final tables (right).

3.3 Data

Initially, files (CSV and SHP formats) with a total size of 26.8 GB of data were used. The datasets used in this paper come from IPPUC[34] and the Municipality of Curitiba[35]. Fig. 4 presents the data used, as listed below (additional details are present in Vila et al. (2016), Parcianello et al. (2018)):

Bus Stops. The city has 9940 bus stops, provided in a shapefile format.

Bus Terminals. The city has 23 terminals (buses) and one terminal which also use trains (data provided in a shapefile format).

NeighborhoodCity Regionals. The city has 75 distinct neighborhoods and 9 regionals, provided in shapefile format.

User Cards. The anonymized data was provided as zipped csv files, with 8,710,082 tuples from October 2017.

Bus Locations. The data was provided as zipped csv files, with 114,602,481 tuples from October 2007.

Movement. The data from User Cards was combined with Bus Location providing the movement data or OD data. The final table presented 1,378,733 tuples using the same criteria for selecting OD data present in Diniz Jr. (2017). The resulting data from this table can also be presented as shown in Table 1.

ROIS. The possible regions of interests (ROIS) were defined as geometric definitions of neighborhood, city regionals or K-Means. The mathematic regions are using k-means (K equals 5, 10, 20, 30, 40, 50, 100, 250, 500 or 1000) were inspired by Silva et al. (2016), Osama et al. (2015).

[21] www.postgresql.org/download/linux/ubuntu/
[22] postgis.net/2013/11/08/postgis-2.1.1/
[23] archive.apache.org/dist/httpd/
[24] php.net/releases/7_2_12.php
[25] leafletjs.com/download.html
[26] github.com/Leaflet/Leaflet.markercluster
[27] github.com/Leaflet/Leaflet.heat
[28] github.com/Leaflet/Leaflet.draw
[29] jquery.com/
[30] getbootstrap.com
[31] github.com/eternicode/bootstrap-datepicker
[32] github.com/Eonasdan/bootstrap-datetimepicker
[33] www.chartjs.org/
[34] http://ippuc.org.br
[35] http://www.curitiba.pr.gov.br/DADOSABERTOS/

The overall objective of the prototype here was not only having traditional areas from the city (regionals, neighborhoods) but also regions based in mathematic division of the city (which might be bigger or smaller than the traditional ones). Fig. 6 presents all the possible regional divisions for the city of Curitiba. The best K parameter for the clustering was based on Elbow Method (Tibshirani et al., 2001, Stolfi et al., 2017), using seven days of data. On the other hand, the clusterization enhanced the visualization of the aggregation groups.

In order to improve performance, btree indexes were used in non GIS columns (gender, dates, smartcard number, among others), and gist indexes were used in all GIS columns. The biggest table (Bus Locations) used table partitioning by day and indexes in order to increase performance. The table movement (the biggest used by the interface with query in Fig. 5) used three non GIS indexes and 2 GIS indexes (origin and destination). Tests selecting a date range from 7, 15 and 30 days showed a 57% better average answer after creating such specific indexes (the biggest test query with 32,773 tuples returned in 480 seconds, compared to 1020 seconds without indexes). Note that the majority of related work cited in this paper do not use GIS databases in order to store the OD data.

3.4 The prototype.

With the requirements checked, the first step was to build the filter panel (listed in modules in Fig. 3): start and end date, the ROIS (regionals, neighboorhoods or mathematical, via K-Means algorithm division), the origin and destination areas, terminals, bus stops, gender and age (as listed in Fig. 7 - left). Internally, the selected filters are processed through specific indexes and queries (such as Fig. 5) in the database. As defined in Ferreira et al. (2013), each query has a set of temporal, attributes and spatial constraints. Such constraints can also be mapped to Peuquet's Triad Framework (Peuquet, 1994): spatio-temporal data—space (where), time (when) and objects (what).

The results are presented in two panels: the OD visualization in Data Visualization followed by basic statistics in Graphics panel. The result of *Q* (what initially should be data similar to Table 1) changes the plot at the Data Visualization and Graphics panels each time the query is changed. The Data Visualization panel at Fig. 7 (right side), has the origin data in blue and destination data in red. It is possible to execute intra-regions queries, selecting the same region as origin and destination. The prototype also presents a Graphic panel (Fig. 8), with the following information: general drop-in in classification by age ("A"), general drop-in in classification by gender ("B"), drop-in classification by gender through the selected days("C"), drop-in classification by age through the selected days ("D"), and drop-in classification by age through the selected days ("E").

Note that the technologies presented in Fig. 3 uses data which is stored in a remote GIS database, a web-based engine (OpenStreetMap), along with several libraries in order to provide the results to a client.

```
select movimentacoes.cartao_sexo,
    movimentacoes.idade,
    movimentacoes.cartao_data,
    to_char(movimentacoes.cartao_datahora,
      'DD/MM/YYYY, HH24h') as datahora_formatada,
    movimentacoes.origem_lat, movimentacoes.origem_lng,
    movimentacoes.destino_lat, movimentacoes.destino_lng
from
(
    select id, roi as contornoPg
    from transporte_dinamico.yp_rois
    where id in (2029)
) origens,
(
    select  id, roi as contornoPg
    from transporte_dinamico.yp_rois
    where id in (2013)
) destinos,
(
    select *
    from transporte_dinamico.np_movimentacao
   where cartao_data between '2017-10-01' and '2017-10-07'
        and cartao_hora between '0:00' and '23:59'
      and upper(cartao_sexo) = 'F'
) movimentacoes
where st_Contains(origens.contornoPg, movimentacoes.origem)
    and st_Contains(destinos.contornoPg, movimentacoes.destino)
order by 1 asc
```

Figure 5: SQL query for showing the OD data in the prototype.

3.5 Preliminar Evaluation.

In order to evaluate the usability of prototype, an experiment was conducted in November 12th, 2019, with five males and two females, from 22 to 32 years old. Three of them were not from computer science (Sociology, Mathematics and Geography). The participants had to perform three activities: 1) an "easy one", using a gender filter with data for one day (with an average execution time of 12,25 secs); 2) a "medium one", using age classification for ten days of data (with an average execution time of 26,90 secs); and a 3) a "difficult one", using 25 days of data (with an average execution time of 36 secs).

The tests ran in home PCs, in parallel. All the participants were able to perform the three activities, without any background in the application. A questionnaire (with four specific questions and one descriptive question) was submitted to them, with the following results: 1) the prototype was intuitive to use; 2) the filters evaluation were fine; and 3) the results were easy to understand using the filters. In parallel, a specialist which used the city report (Pichiliani, 2017) stated that the prototype enhanced the overall analysis. Further details are available at Parcianello (2020).

4 Case Studies.

In this section we present case studies in order to illustrate the usability of the prototype and the basic patterns exploring subjects such as clusterization.

Figure 6: *ROIS* from geometric definitions of K-Means (5, 10, 20, 30, 40, 50, 100, 250, 500, 1000) followed by regionals and neighborhood.

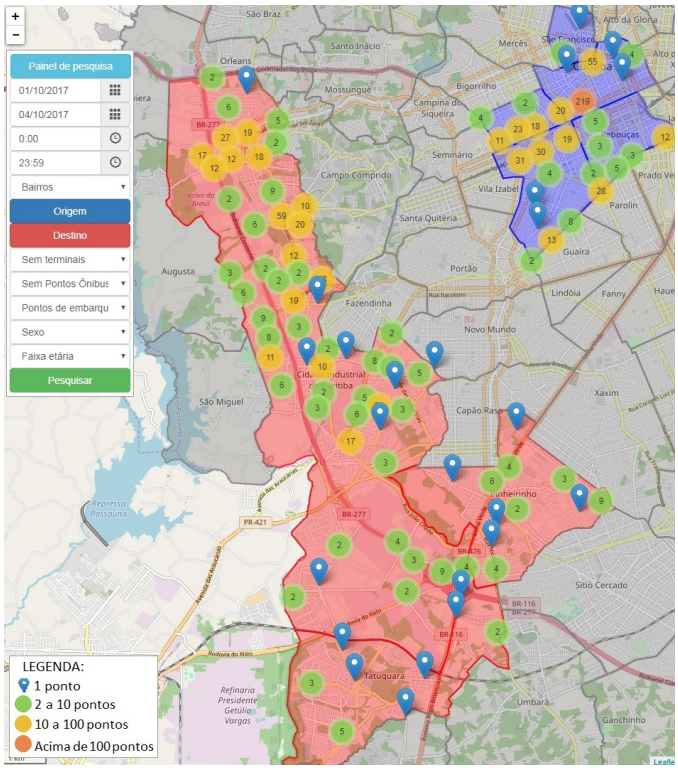

Figure 7: Filter and Data Visualization Panel.

4.1 Classification drop-ins across day shifts in a month.

If we analyze the drop-ins by shift, we note that the majority is concentrated at morning shifts (as shown in Fig. 9). Note that there is a higher demand on weekdays, and a decrease on weekends. Exceptions are noted in Saturday 30/09 and Thursday 12/10 (Children's Day holiday). Note also that the demand is higher during the morning shifts, followed by afternoon, night and dawn.

4.2 Neighborhood drop-ins.

In order to analyze the drop-ins (origin) aggregated by neighborhood, Eq. (1) was used.

$$IO_{co} = \frac{\text{Number of Events in the Neighborhood}}{\text{Demographic Density of the Neighborhood}} \quad (1)$$

Fig. 10-A presents the results using IO_{co} for three weekdays in Oct. 2017. Note that CIC, Curitiba's most populous neighborhood (more than 170,000 according to IBGE) and with the largest territorial area (about 44km 2) is also the one with the highest IO_{co}, indicating that there is a considerably greater movement of people. In particular, if we just analyze the drop-ins in CIC, we note that these citizens have as drop-offs the CIC itself, along with the neighborhoods in the region, as shown in Fig. 11.

The prototype analysis of people who has the drop-ins in CIC during the morning shifts (Fig. 12) for thirty days shows that the majority of the citizens are female, with the highest amount of ages concentrated among 23 and 33 years. The map also shows that bus stop drop-ins are equally distributed in the CIC neighborhood.

4.3 Intra-Region Movements.

If we consider intra-region analysis (the movements where origin neighborhood are the same as the destination one), we can also notice that the CIC neighborhood and the downtown area have the highest rates (Fig. 10-B).

The prototype analysis using K-Means equals 100 (for data in Oct. 04th, 2017 - Fig. 13) shows that CIC has some concentration of bus stops for drop-ins and drop-offs.

4.4 Regional Movements.

If we consider intra-region analysis (the movements where origin neighborhood is the same as the destination one), we can also notice that the CIC neighborhood and the downtown area have the highest rates (Fig. 10).

The prototype analysis of regionals from CIC and Downtown area (using them as drop-ins and drop-offs, with one month of data) shows that not only the main neighborhoods have the highest concentration of citizens, but also the districts around them. Here, the majority is still female, ranging from 19 to 32 years. Note that the Downtown regional has a highest concentration compared to CIC regional.

4.5 K-Means for K=500

In order to understand smaller regions (compared to neighborhood and regionals), we included a test using for k-means for K=500. Note that this type of analysis is not available at Pichiliani (2017).

The prototype analysis using one month of data, using drop-ins in CIC and drop-offs in Downtown area (Fig. 14) with K-Means for K=500 shows that the OD distribution among the regions are equal, with a highest concentration in the downtown area. That shows that not only the regions have a higher concentration of passengers, but also that the concentration is across their territories.

4.6 Lessons Learned.

Along with the data presentation, several factors should be considered in order to design and impact OD prototypes: 1) the filters and their classification (which ones are relevant to the user, how easy are their use and implementation, and which results are better understood). Although this paper mainly focused on filters based on data (with twelve categories and 32 subcategories) using as base the local reports, several others could be included (such as different visualization techniques, events, cause/effect), which may lead to too many filter options; 2) the final user and restrictions/technology background: the objective was to use as a starting point an interface which was familiar to the end user, and enhance it, to maximize the probability

Figure 8: Graphic Panel: results obtained querying multiple neighborhoods within day 07/02/2018.

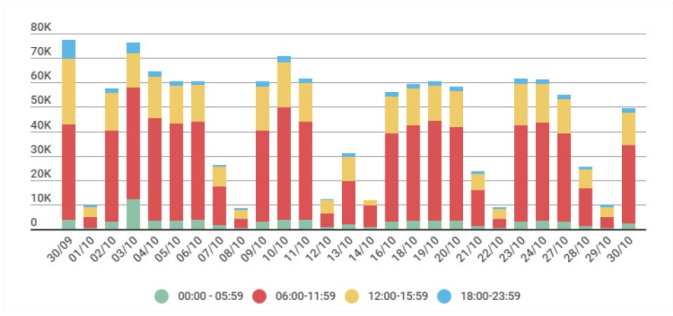

Figure 9: Number of Drop-ins by shift in a month.

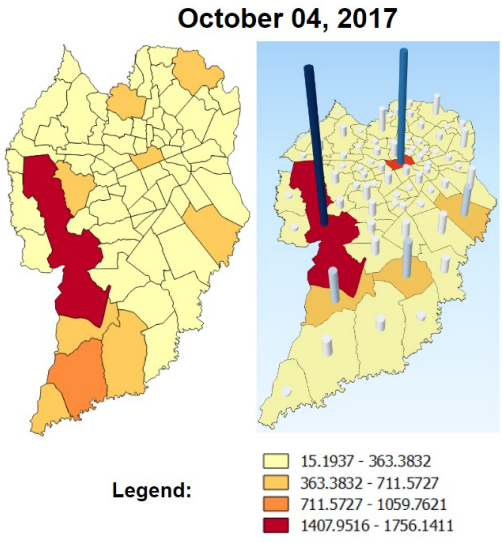

Figure 10: Neighborhood drop-ins (A) and Intra-Region Analysis (B).

of its use; 3) the overall data visualization and which basic statistics impact the user (what general overview the final user is interested); 4) movement and region variations (intra-region, regional, smaller variations (such as K-Means); 5) the analysis of the architecture, data, database structure and queries in order to improve optimization (techniques such as query optimization, indexes, table partitions); 6) the test and integration of different libraries (sometimes not freely available, or hard to integrate, or non available to different OS). In the prototype, for example, some Leafjet plugins were not compatible with others that we intended to add to the project.

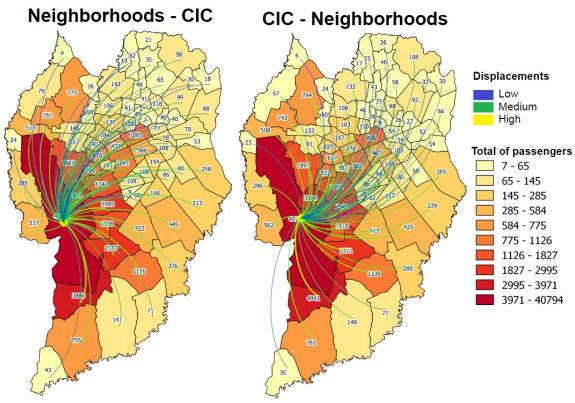

Figure 11: Movements having CIC as destination (left) and CIC as origin (right).

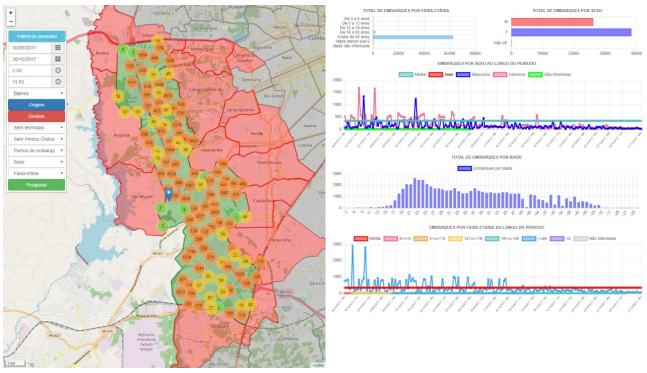

Figure 12: Prototype Analysis for drop-ins in CIC, in the morning shift, for one month of data.

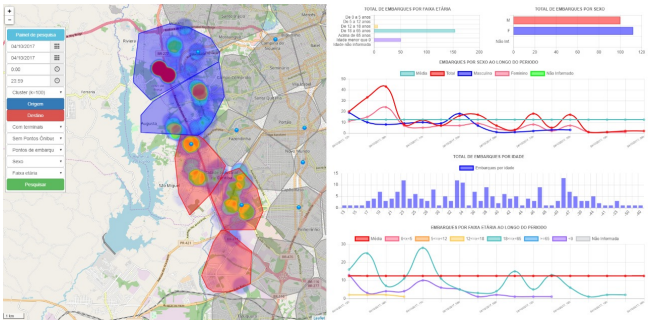

Figure 13: Prototype Analysis for 04/10/2019 with K-means as 100, using the data from Oct. 04th, 2017.

The prototype is available through video[36] and source code [37].

[36] https://youtu.be/KOzFHRc7lXA
[37] https://github.com/yussefparcianello/OrigemDestinoStpCuritiba

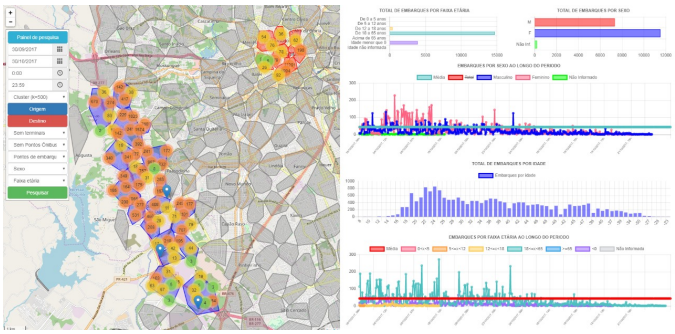

Figure 14: Prototype Analysis with K-means as 500, using one month of data, with drop-ins in CIC neighborhood and drop-offs in Downtown area.

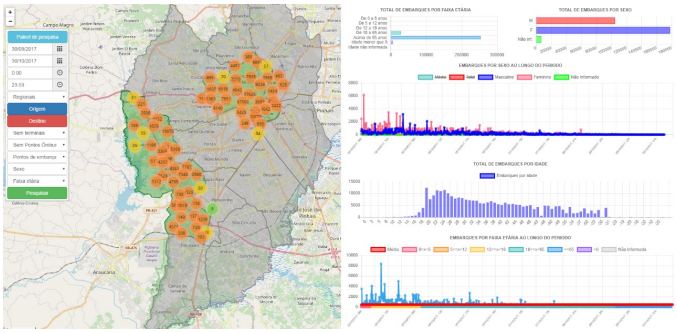

Figure 15: Prototype Analysis for drop-ins in CIC and Downtown Regionals, for one month of data.

5 Conclusion

There are several challenges related to urban mobility, mainly faced by large urban centers. Public transport supply and demand, traffic jams, travel times, and the distribution and patterns of mobility are some of the problems that impact the development and planning of a city.

In this sense, this paper presented a prototype for visualization of OD data. The novelty relies on visualization through clustering of georreferenced data, allowing the analysis of different regions of interests (neighborhood, regionals or mathematical regions using K-means algorithm). The prototype was based on: 1) queries associated with a set of trips (and related attributes) for non experts (Ferreira et al., 2013), 2) analysis of trips and services for Curitiba mentioned in Diniz Jr. (2017), 3) scalability for visualization, using clustering of georreferenced data (Vila et al., 2016), and 4) local study of Origin-Destination from the Public Transportation of Curitiba (Pichiliani, 2017), along with interviews of local citizens. The filter options offered were established based on the analysis from data collected via a questionnaire applied to the academic community. The data was stored in a GIS database, using indexes and table partitioning in order to improve performance. The results are presented in two panels: the OD visualization in Data Visualization followed by basic statistics in Graphics panel.

A series of case studies were used in order to understand

data patterns, using one month of data from Curitiba, with an average of 8.710.082 smart card entries. A preliminar evaluation with 7 users has shown that filters are helpful to understand the data, and the results were easy to understand through the interface. Within the lessons learned, we listed issues which impact not only the interface, but also the final user, technologies, architecture and database optimization.

For future work, we can mention the inclusion of complex queries, the expansion of filters (vehicle occupancy rate, average travel speed, etc.), the automation of data insertion, along with other visualizations (flowlines, displacements, speed alerts, etc.) and data (velocity alerts, route deviation, weather, points of interests, etc.).

Acknowledgments

We would like to thank IPPUC, URBS and the Municipality of Curitiba.

References

Añez, J., Barra, T. D. L. and Pérez, B. (1996). Dual graph representation of transport networks, *Transportation Research Part B: Methodological* **30**(3): 209 – 216.

Andrienko, N. and Andrienko, G. (2013). Visual analytics of movement: An overview of methods, tools and procedures, *Information Visualization* **12**(1): 3–24. https://doi.org/10.1177%2F1473871612457601.

Andrienko, N., Andrienko, G., Patterson, F. and Stange, H. (2020). Visual analysis of place connectedness by public transport, *IEEE Transactions on Intelligent Transportation Systems* **21**(8): 3196–3208. https://doi.org/10.1109/TITS.2019.2924796.

Braz, T., Maciel, M., Mestre, D. G., Andrade, N., Pires, C. E., Queiroz, A. R. and Santos, V. B. (2018). Estimating inefficiency in bus trip choices from a user perspective with schedule, positioning, and ticketing data, *IEEE Transactions on Intelligent Transportation Systems* **19**(11): 3630–3641. https://doi.org/10.1109/TITS.2018.2846036.

Cassiano, K. M. (2014). *Análise de séries temporais usando análise espectral singular (ssa) e clusterização de suas componentes baseada em densidade*, Master's thesis, Pontifícia Universidade Católica.

Chapleau, R. and Morency, C. (2005). Dynamic spatial analysis of urban travel survey data using GIS, *25th Annual ESRI International User Conference, San Diego, California*, pp. 1–14. Available at https://proceedings.esri.com/library/userconf/proc05/papers/pap1232.pdf.

Couto, E. A. (2018). *Aplicação dos indicadores de desenvolvimento sustentável da norma ABNT NBR ISO 37120:2017 para a cidade do Rio de Janeiro e análise comparativa com cidades da América Latina* , B.S. thesis, Universidade Federal do Rio de Janeiro.

Cruz, A., Ferreira, J., Carvalho, D., Mendes, E., Pacitti, E., Coutinho, R., Porto, F. and Ogasawara, E. (2018). Detecção de anomalias frequentes no transporte rodoviário urbano, *SBBD: Brazilian Symposium on Databases*, SBC, pp. 271–276. Available at http://sbbd.org.br/2018/wp-content/uploads/sites/3/2018/02/p240-245.pdf.

Diniz Jr., P. C. (2017). *Serviços telemáticos em uma rede de transporte público baseados em veículos conectados e dados abertos*, Master's thesis, PPGEI, Universidade Tecnológica Federal do Paraná, Campus Curitiba.

Ferreira, N., Poco, J., Vo, H. T., Freire, J. and Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips, *IEEE Transactions on Visualization and Computer Graphics* **19**(12): 2149–2158. https://doi.org/10.1109/TVCG.2013.226.

Guerra, A. L. (2011). *Determinação de matriz origem destino utilizando dados do sistema de bilhetagem eletrônica*, Master's thesis, Universidade Federal de Minas Gerais.

Itoh, M., Yokoyama, D., Toyoda, M., Tomita, Y., Kawamura, S. and Kitsuregawa, M. (2016). Visual exploration of changes in passenger flows and tweets on mega-city metro network, *IEEE Transactions on Big Data* **2**(1): 85–99. https://doi.org/10.1109/TBDATA.2016.2546301.

Osama, D., Ghoneim, A. and Manjaunath, B. (2015). Air pollution clustering using k means algorithm in smart city, *International Journal of Innovative Research in Computer and Communication Engineering* **3**(7): 51–57. https://doi.org/10.3906/elk-1707-99.

Palomo, C., Guo, Z., Silva, C. T. and Freire, J. (2016). Visually exploring transportation schedules, *IEEE Transactions on Visualization and Computer Graphics* **22**(1): 170–179. https://doi.org/10.1109/tvcg.2015.2467592.

Parcianello, Y. (2020). *Análise de origem-destino do uso de sistema de transporte coletivo de curitiba sob o ponto de vista de regions of interest*, Master's thesis, Universidade Tecnológica Federal do Paraná.

Parcianello, Y., Kozievitch, N. P., Fonseca, K. V. O., Rosa, M. d. O., Gadda, T. M. C. and Malucelli, F. C. (2018). Transportation: An overview from open data approach, *2018 IEEE International Smart Cities Conference (ISC2)*, pp. 1–8. https://doi.org/10.1109/ISC2.2018.8656937.

Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems, *Annals of the Association of American Geographers* **84**(3): 441–461. https://doi.org/10.1111/j.1467-8306.1994.tb01869.x.

Pichiliani, M. (2017). Consolidação de dados de oferta, demanda, sistema viário e zoneamento. Relatório 5: Pesquisa de Origem-Destino domiciliar (Consolidation of supply, demand, road system and zoning data. Report 5: Household Origin-Destination Survey). Available

at http://www.ippuc.org.br/visualizar.php?doc=http://admsite2013.ippuc.org.br/arquivos/documentos/D536/D536_002_BR.pdf.

Saberi, M., Mahmassani, H. S., Brockmann, D. and Hosseini, A. (2017). A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks, *Transportation* **44**(6): 1383–1402. https://doi.org/10.1007/s11116-016-9706-6.

Silva, E. L. C., Rosa, M. O., Fonseca, K. V. O., Luders, R. and Kozievitch, N. P. (2016). Combining k-means method and complex network analysis to evaluate city mobility, *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, IEEE, pp. 1666–1671. https://doi.org/10.1109/ITSC.2016.7795782.

Spadon, G., Scabora, L. C., Nesso-Jr, M. R., Traina Junior, C. and Rodrigues Junior, J. F. (2018). Caracterização topológica de redes viárias por meio da análise de vetores de características e técnicas de agrupamento, *SBBD: Brazilian Symposium on Databases*, SBC, pp. 157–168. Available at https://sbbd.org.br/2018/wp-content/uploads/sites/5/2018/08/157-sbbd_2018-fp.pdf.

Stolfi, D. H., Alba, E. and Yao, X. (2017). Predicting car park occupancy rates in smart cities, *International Conference on Smart Cities*, Springer, pp. 107–117. https://doi.org/10.1007/978-3-319-59513-9_11.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society* **63**(2): 411–423. https://doi.org/10.1111/1467-9868.00293.

Vila, J. J. F. R. (2016). *Clusterização e visualização espaço-temporal de dados georreferenciados adaptando o algoritmo marker clusterer: um caso de uso em Curitiba*, Master's thesis, Universidade Tecnológica Federal do Paraná.

Vila, J. R., Kozievitch, N. P., Fonseca, K. V. O., Gadda, T., Rosa, M., Gomes-jr, L. C. and Akbar, M. (2016). Urban Mobility Challenges – An Exploratory Analysis of Public Transportation Data in Curitiba, *Revista de Informática Aplicada* **12**: 1. https://doi.org/10.13037/ras.vol12n1.145.

Wood, J., Dykes, J. and Slingsby, A. (2010). Visualisation of origins, destinations and flows with OD maps, *The Cartographic Journal* **47**(2): 117–129. https://doi.org/10.1179/000870410X12658023467367.

Zaiane, O. R., Foss, A., Lee, C. H. and Wang, W. (2002). On data clustering analysis: Scalability, constraints, and validation, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, pp. 28–39. https://doi.org/10.1007/3-540-47887-6_4.

Zhang, H., Zhao, P., Wang, Y., Yao, X. and Zhuge, C. (2015). Evaluation of bus networks in china: from topology and transfer perspectives, *Discrete Dynamics in Nature and Society* **2015**. https://doi.org/10.1155/2015/328320.