Business Activity Clustering: A Use Case in Curitiba

Yuri S. Bichibichi[1], **Nádia P. Kozievitch**[1], **Ricardo da S. Dutra**[1], **Artur Ziviani**[2]

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

[2]National Laboratory for Scientific Computing (LNCC), Petrópolis, RJ – Brazil

yuribichibichi@alunos.utfpr.edu.br, {nadiap, rdsilva}@utfpr.edu.br,

ziviani@lncc.br

Abstract. *In the context of smart cities, the information of businesses licenses has the potential to discriminate economics characteristics of the observed urban environment. This work performs an initial analysis on business activity clustering using the k-means algorithm with data from the granting of business licenses (from 1980 to 2016) in the city of Curitiba - Brazil.*

1. Introduction

Nowadays large-scale data analytics enables analyzing socio-economic factors in metropolitan areas with less resources than by traditional surveys, such as censuses and questionnaires. The challenge typically resides, however, in how to explore the available data to achieve relevant results [Silva and Loureiro 2016]. In this context, business licenses, which are granted by the municipality to entities that intend to exert commercial activity, can be used to estimate the development and distribution of commercial agglomerates [Carr et al. 2003]. This type of analysis can be used for the benefit of the community, by influencing public strategies, for example. In addition, the study of the development of a commercial area offers possible indicators that should help identifying potential new commercial areas.

In particular, the city of Curitiba in southern Brazil was already considered one of the world's smartest cities [1] and also belongs to a group of cities, the *C40 cities* [2], which set ambitious targets to improve urban life quality and protect their environment. Curitiba has 1.8 million people inside a total area of 430,9 km^2, Human Development Index (HDI) of 0,823, according to the Brazilian Institute of Geography and Statistics (IBGE) [3]. This area encompasses 75 neighborhood districts. In particular, the Batel neighborhood (considered the second Downtown of the city), is a former traditional residential area of Curitiba, with a population of about 12,000 people. It is currently full of restaurants and options for nightlife. In the early 20th century, the region already had two breweries, two tea processing plants, one soap factory, and some small shops. The current economy of this neighborhood is mainly composed of services (61%) and trade (29%). Batel has a density of 66.92 pop./km^2 and it has 5,243 households, 90% composed of apartments. The

[1] https://www.forbes.com/2009/12/03/infrastructure-economy-urban-opinions-columnists-smart-cities-09-joel-kotkin.html – Last visited on May 24th, 2018.

[2] http://www.c40.org – Last visited on May 24th, 2018.

[3] http://www.ibge.gov.br – Last visited on May 24th, 2018.

neighborhood has old mansions of the "Mate Barons", and its larger range of households (28.5%) has incomes between 5 and 10 times the Brazilian minimum wages.

This paper presents a business license analysis using the *k-means* algorithm, with a use case in Curitiba. Several values for k are tested using the *Elbow Method*. The objective was to identify correlations between category, localization, and the creation date in business licenses using clusterization. The contribution here is search a relation between theses types of data. The results shows that a strong relation couldn't be find out, probably due to the characteristics of the data.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 presents the studied dataset as well as data processing details. Section 4 shows the study about the business activity clustering. Finally, Section 5 presents the conclusion of the paper and future work.

2. Related work

Curitiba open data has been explored for urban mobility research projects aiming at modeling the public transportation system operation [da Silva et al. 2016b, Bona et al. 2016, Stenneth et al. 2011], getting insights of its quality of service [da Silva et al. 2016a], suggesting data exploitation for new services [Diniz Junior 2017] or technologies [Sebastiani et al. 2016], and environmental impacts [Dreier and Silveira 2016, Dreier et al. 2015]. [Gay et al. 2016] presented a study involving the data obtained from the GeoSampa portal of the city of São Paulo-SP. Using GIS, the researchers identified the most vulnerable places of São Paulo, as those with the highest levels of flood and with the lowest levels of accessibility. Geospatial open data is also used to risk assessment in Curitiba[4] through special applications (vicom saga[5]). Many of these application systems are based on global positioning systems (GPS) [Stenneth et al. 2011], using specific data (such as real time bus locations, spatial rail and spatial bus stop information) and specific techniques (such as spatial data mining [Mennis and Guo 2009]).

In particular, if we consider business licenses, [Rosa et al. 2016] analyzed the entropy from the districts Center, Batel, and Tatuquara, concluding that it tends to reduce when are a large number of business licenses, *i.e.*, the business categories tends to be dispersed. [Bichibichi et al. 2018] presented an study for the surrounding areas for two shoppings within the district Batel. The same business license data was used for general clusterization, using heatmaps in [Vila et al. 2016]. Finally, [Kozievitch et al. 2017] at the other hand, showed the challenges related to the data.

3. Dataset and Pre-Processing Steps

A dataset of historical data on business licenses was provided by the Curitiba City Hall (*Prefeitura Municipal de Curitiba* – PMC). This dataset covered granted business licenses from January 1st, 1980 to December 31st, 2015, in a total of 172,173 records. For this analysis, initially only the data from the Batel district is used, in a total of 7,268 records.

Approximately 0.05% of the records presented problems in the address geocoding process, requiring manual intervention to correct the latitude and longitude information.

[4] `https://www.viconsaga.com.br/grrd` – Last accessed on May 24th, 2018.

[5] `https://www.viconsaga.com.br/site/language=us` – Last accessed on May 24th, 2018.

The dataset has only the creation dates for the entries and it was not possible to infer if the business eventually stopped operating. Economic activities, initially cataloged in 2,977 detailed types (such as dancing restaurant, pizza restaurant, etc.) were reassembled into 73 types of *aggregated activities* (such as office, restaurant, banking, construction, etc.), following the recently presented steps in [Kono 2016]. Table 1 shows the attributes present within the data along with their description.

The Curitiba Municipality Data was imported into a PostGIS server, where tablespaces, indexes, and schemes were created. Subsequently, the QGIS[6] software was used to integrate the data with other tools: *GoogleMaps*[7] and *OpenStreetMaps*.[8] The k-means algorithm and Elbow Distortion were implemented in python (along with scikit, pandas and matplotlib libraries).

Table 1. Attributes available in the dataset for business activities.

Attribute	Description	Attribute	Description
NOME_EMPRESARIAL	Company name	CEP	Zip code
NUMERO_DO_ALVARA	Code of the business license	DATA_EMISSAO	Issue date of the business license
DATA_EXPIRACAO	Expiration date of the business license	UNIDADE	Detail related to address
ATIVIDADE_SECUNDARIA1	Description of the first activity	ENDERECO	Address
ATIVIDADE_SECUNDARIA2	Description of the second activity	NUMERO	Number detail from address
ATIVIDADE_PRINCIPAL	Description of main activity	ANDAR	Identification of floor number
COMPLEMENTO	Address complement	BAIRRO	Address district
INICIO_ATIVIDADE	Date of business activity started up	ATIVIDADE_AGREGADA	General category for activity

4. Business activity analysis

From the attributes listed in Table 1, only location, time, and aggregated business license types were used. The underlying hypothesis is that there are clusters (within a combination of location, time and type) which present more affinity (for example, given the location of a restaurant, we might have a drugstore nearby).

The attributes for location (latitude and longitude) were both initially normalized to the interval $[0, 1]$ according to the minimum and maximum values for latitude and longitude registered for Curitiba. The accuracy of the data remains similar, since the curvature effect of the globe captured by latitude and longitude is negligible within a single city district. The attribute *INICIO_ATIVIDADE* was also normalized, so zero represents the oldest date and one represents the newest date. For each value on the attribute that represents the aggregation of the business activity, a matrix was created such that only the

[6] `http://www.qgis.org/en/site/` – Last visited on May 24th, 2018.
[7] `https://www.google.com.br/maps` – Last visited on May 24th, 2018.
[8] `https://www.openstreetmap.org/` – Last visited on May 24th, 2018.

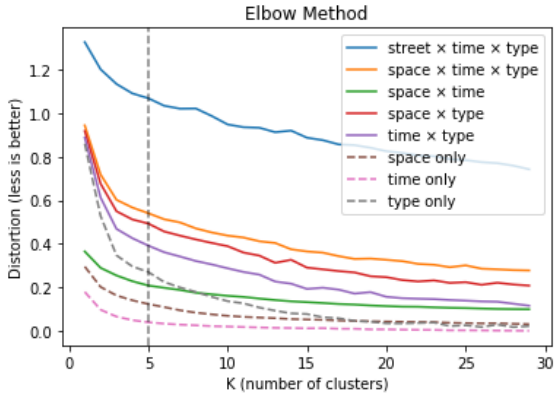

Figure 1. Elbow Distortion.

actual activity is marked as one, while the others are marked as zero. Table 2 shows an example of that procedure. Note that all normalized values are between 0 and 1 and the columns referring to the business categories are sparse (is this example, Gym, Butcher shop, and Agency).

Table 2. Example of data normalization.

number	normalized_lat	normalized_lng	normalized_date	normalized_Gym	normalized_Butcher_Shop	normalized_Agency
1121901	0.861496	0.672512	0.916667	0	1	0
1055724	0.567241	0.788929	0.861111	0	0	1
1128151	0.130912	0.416030	0.916667	1	0	0

For the clustering analysis the *k-means* algorithm was used. The choice of the cluster number k is determined using the *Elbow Method*, that shows the sum of squared errors (SSE) for different values of k. Figure 1 shows the Elbow Method for the following combinations of the clustering parameters: space, time, type (of business license), time $\times$ type; space $\times$ type; space and time; space $\times$ type $\times$ time; and street $\times$ time $\times$ type. Note that "space $\times$ type $\times$ time", "space $\times$ type", and "time $\times$ type" suggest $k \geq 3$ while "space $\times$ time" suggests $k \geq 5$.

Figure 2 (for Batel) and Figure 3 (for Curitiba) present more details about all the aggregated business licenses over time: light green presents the most common types and blue lines represent the cluster *others*. Note that (i) not all aggregated license types are present at the Batel District (only 54 out of 73) and that (ii) the office and retail trade types

Table 3. Clusters created when $k = 5$.

Cluster Nr.	Nr. of Records	Main Activity
0	2489	Office
1	1503	Retail trade
2	334	Hospital Service
3	204	Hairdresser
4	1718	(All other types)

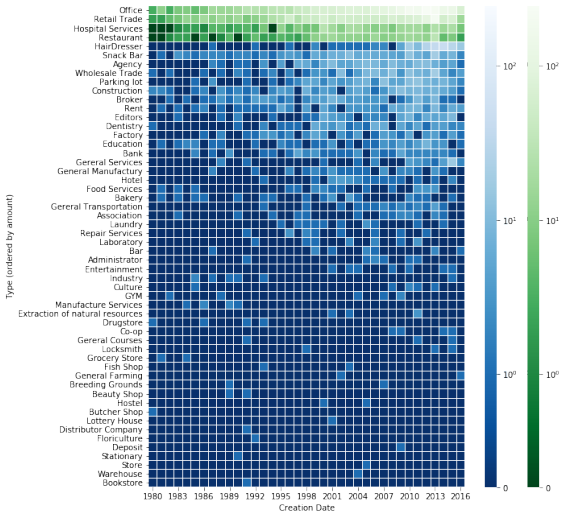

Figure 2. Business Licenses types over time for Batel: light blue or green present the most common types.

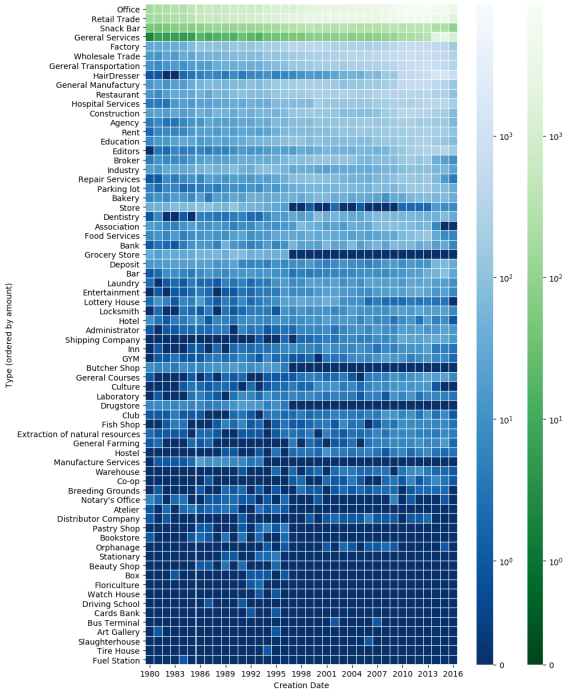

Figure 3. Business Licenses types over time for Curitiba: light blue or green present the most common types.

are the most common types for both the city of Curitiba and for the Batel District.

The data clustering suggested that the attributes space, time and type produce clusters which are divided only by the type of business license (as shown in Table 3, which uses $k = 5$). Figure 4 shows the five clusters for Batel. Note that clusters such as office and retail trade ignore the time factor. Darker colors indicate a higher amount of business licenses of that type in a specific date.

The experiments using several k indicated that space, time and business license type are disjoint when clustered: in summary, the three dimensions are unrelated. The most commons types (*Office* and *Retail Trade* - representing half of the data), are distributed equally in time and space (Figures 2 and 3). Since there are different numbers of records for each type of business license (as shown in Figure 6), the less common business license types end up not influencing in the larger clusters. This presented a major impact during the analysis, since the less common business license types are the ones responsible for characterizing districts. Batel for example, is a district which has the majority of hospital services in Curitiba. That information can be noticed in the third cluster in Table 3.

In other words, the most common business license types do not present correlation, since they are present over all the time and over all the space. Probably this occurs because the most common business licenses are spread homogeneously and have more influence than the smaller clusters. Future studies include the analysis of smaller clusters (such as

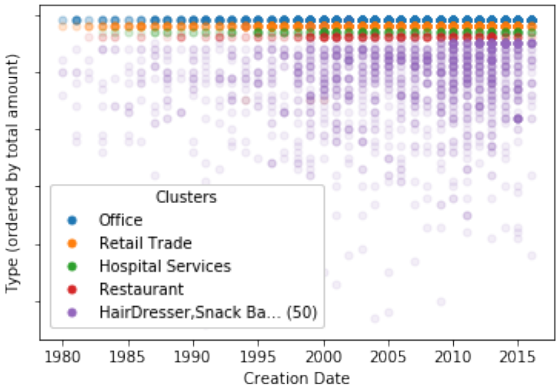

Figure 4. Business License Type (y axis) over time (x axis) in Batel. Colors appoint clusters (Note that clusters ignore time factor).

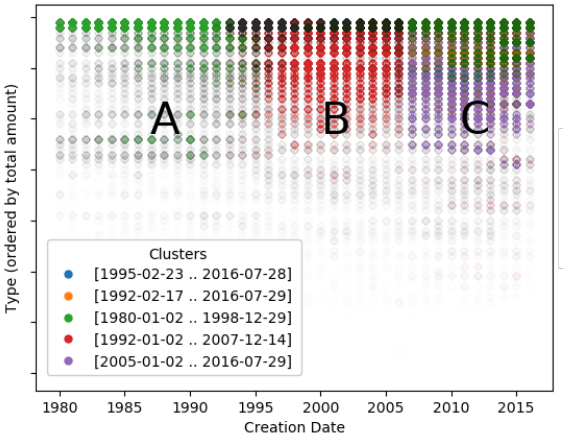

Figure 5. Business License Type (y axis) over time (x axis) in Curitiba. Note that clusters ignore type factor and are grouped in A, B and C).

Hospital Services in Batel - Figure 2) so that business licenses characterizing the districts of the city can be better understood.

5. Conclusion

This paper presented a business license analysis using the *k-means* algorithm, with the *Elbow Method*. The objective was to identify the affinity of data, considering the different types of business, and their distribution along time and space, in a use case of three decades of data in Curitiba.

The results for Batel indicated that offices, retail trade, hospital services and restaurants are the activities which identify the first four clusters, followed by the rest of business license types.

The studies presented no correlation among location, time and business license type, indicating that: (i) these type of data is unrelated, i.e, opening a restaurant does not indicate a new pharmacy, for example, at the same location and year; and (ii) the most commons business licenses (*Office*, *Retail Trade* and others) are distributed equally in time and space and the smalls clusters, which show more interesting relations, are overshadowed by the biggest clusters.

Within future work, we can mention the integration using other techniques, along with the analysis of smaller clusters, present only on specific districts.

Acknowledgments. We would like to thank the Municipality of Curitiba, IPPUC, CAPES, CNPq, Fapemig, FAPERJ, FAPESP, EUBra-BIGSEA project (EC/MCTIC 3rd Coordinated Call), and INCT em Ciência de Dados (INCT-CiD).

References

[Bichibichi et al. 2018] Bichibichi, Y. S., Kozievitch, N. P., and Carvalho, R. A. M. (2018). Análise de evolução de emissão de alvarás. In *Anais da XIV Escola Regional de Banco de Dados*.

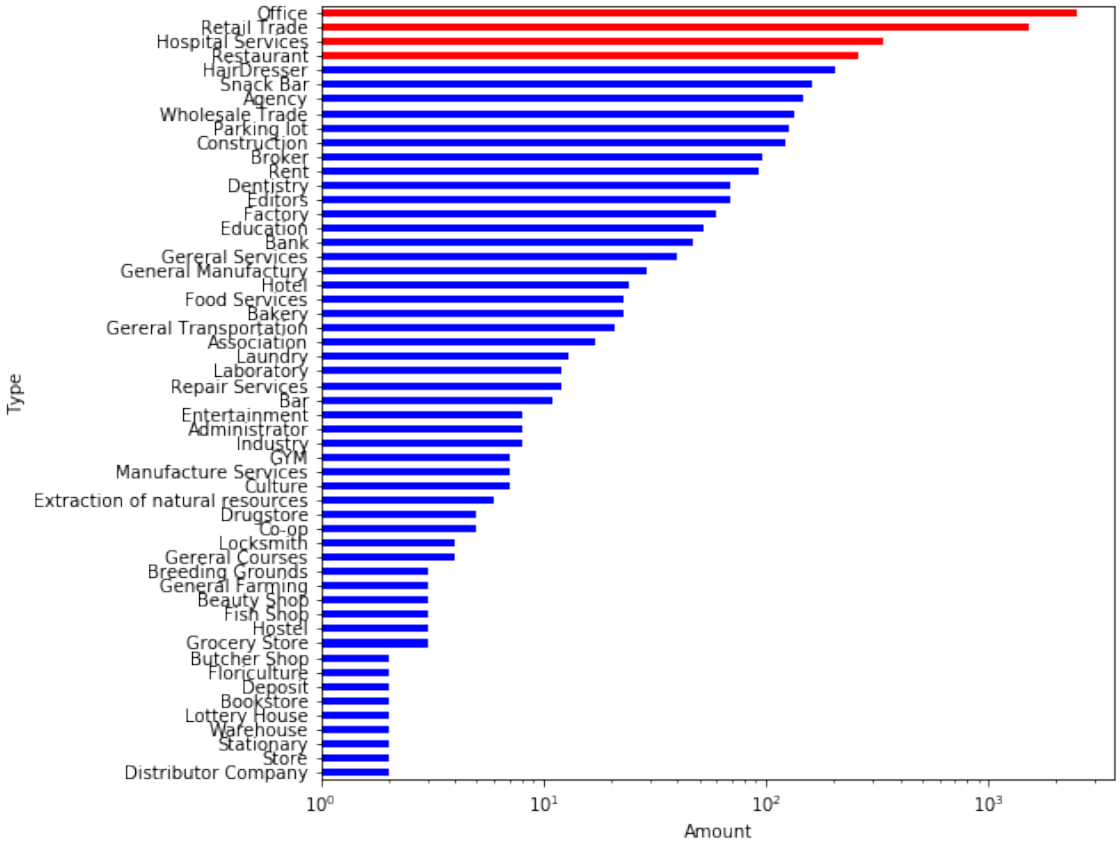

Figure 6. Number of Business License Types for Curitiba.

[Bona et al. 2016] Bona, A. A. D., Fonseca, K. V. O., Rosa, M. O., Luders, R., and Delgado, M. R. B. S. (2016). Analysis of public bus transportation of a brazilian city based on the theory of complex networks using the p-space. *Mathematical Problems in Engineering*, 2016:1–12.

[Carr et al. 2003] Carr, D., Education, D. R. E., Lawson, J., Lawson, J., and Schultz, J. (2003). *Mastering Real Estate Appraisal*. Kaplan Financial Series. Kaplan.

[da Silva et al. 2016a] da Silva, E. L. C., de Oliveira Rosa, M., Fonseca, K. V. O., Luders, R., and Kozievitch, N. P. (2016a). Combining k-means method and complex network analysis to evaluate city mobility. In *ITSC'2016*, pages 1666–1671. IEEE.

[da Silva et al. 2016b] da Silva, E. L. C., Fonseca, K. V. O., de Oliveira Rosa, M., and Munaretto, A. (2016b). Analysis of curitiba's public transport system as a complex network. In *ISPE TE*, pages 267–276.

[Diniz Junior 2017] Diniz Junior, P. C. (2017). Serviços telemáticos em uma rede de transporte público baseados em veículos conectados e dados abertos. Msc. thesis, UTFPR.

[Dreier and Silveira 2016] Dreier, D. and Silveira, S. (2016). Smart City Concepts in Curitiba-innovation for sustainable mobility and energy efficiency: Project NEWSLETTER, January 2016.

[Dreier et al. 2015] Dreier, D., Silveira, S., Khatiwada, D., Fonseca, K. V. O., Nieweglowski, R., and Schepanski, R. (2015). Energy use and co2 emissions of city buses in

curitiba, brazil. In *Systems Analysis 2015, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, 11.-13. November 2015*. KTH Royal Institute of Technology.

[Gay et al. 2016] Gay, J. S., Giannotti, M. A., and Tomasiello, D. B. (2016). Accessibility and flood risk spatial indicators as measures of vulnerability. *XVII GEOINFO*, (17):93–104.

[Kono 2016] Kono, F. (2016). Um modelo de representação computacional baseado em conceitos de crescimento urbano associados a alvarás e primitivas em banco de dados espacial. Master's thesis, Department of Informatics, UTFPR, Brazil.

[Kozievitch et al. 2017] Kozievitch, N. P., Silva, T. H., Ziviani, A., Costa, G., and Lugo, G. (2017). Three decades of business activity evolution in curitiba: A case study. *Annals of Data Science*, 4(3):307–327.

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

[Rosa et al. 2016] Rosa, J., Silva, T. H., Kozievitch, N. P., and Ziviani, A. (2016). Ciência de dados: Explorando três décadas de evolução da atividade econômica em curitiba. In *Anais da XII Escola Regional de Banco de Dados*, pages 139–142.

[Sebastiani et al. 2016] Sebastiani, M. T., Lüders, R., and Fonseca, K. V. O. (2016). Evaluating electric bus operation for a real-world brt public transportation using simulation ptimization. *ITSC'2016*, 17(10):2777–2786.

[Silva and Loureiro 2016] Silva, T. H. and Loureiro, A. A. (2016). Users in the urban sensing process: Challenges and research opportunities. In *Pervasive Computing: Next Generation Platforms for Intelligent Data Collection*, pages 45–95. Academic Press.

[Stenneth et al. 2011] 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.

[Vila et al. 2016] Vila, J. J. R., Kozievitch, N. P., Gadda, T. M., Fonseca, K., Rosa, M. O., 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).