Challenges for matching spatial data on economic activities from official and alternative sources

Rodrigo Wenceslau[1], Clodoveu A. Davis Jr.[1], Rodrigo Smarzaro[1]

[1]Depto. de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Av. Presidente Antônio Carlos, 6627 – 31270-901 – Belo Horizonte – MG – Brasil

{rwtorres,clodoveu,smarzaro}@dcc.ufmg.br

Abstract. *One of the most interesting challenges for urban geographic applications is the integration of multiple heterogeneous data sources. Given current limitations in the access to official data, in spite of modern Open Governmental Data policies, it is necessary to assess whether unofficial data sources can be used to replace official ones, or used along with them, in a complementary fashion. This work proposes a methodology for matching and comparing official governmental and alternative data on economic activities in an urban area. Applied to data from Belo Horizonte, Brazil, the proposed approach led to the accurate matching of up to 75% of Google Places entries to official municipal tax records in some categories. Results show that alternative data sources on businesses can be more accessible and dynamic than official datasets, especially when such businesses benefit from the online exposure provided by freely accessible Web applications.*

1. Introduction

Many demands from urban dwellers are based on the ready availability of data on various aspects of contemporary life. Governmental sources account for several important data categories, particularly in urban environments. Such data includes street traffic, public transportation, health services and safety.

However, solving complex urban problems requires more than governmental actions. Private initiatives that endeavor to help citizens with information technologies for solving daily problems often need to access public data. Open Governmental Data (OGD)[1] initiatives are coming up as a state policy around the world. For example, Canada[2], Switzerland[3], Brazil[4], United States[5] and many other countries maintain online platforms to make data available to everyone. OGD are based on eight principles that establish that data should be complete, primary, timely, constantly accessible, machine processable, for everyone, for any use [Perritt Jr 1997, McDermott 2010, Bertot et al. 2012, Gov.UK 2013]. In reality, however, open data portals can make existing data available and easily accessible, but not necessarily up-to-date or completely reliable, or even covering all aspects of modern urban life [Chakraborty et al. 2015, Lourenço 2015].

[1] http://opengovernmentdata.org [Accessed on July, 12, 2017]

[2] http://open.canada.ca/en/open-data [Accessed on July, 12, 2017]

[3] http://opendata.swiss/en/ [Accessed on July, 12, 2017]

[4] http://dados.gov.br/ [Accessed on July, 12, 2017]

[5] http://data.gov [Accessed on July, 12, 2017]

On the other hand, technologies such as GPS-enabled smartphones allow citizens to become geographic data producers. Crowdsourcing and Web 2.0 services are increasingly used as sources of data that are updated frequently by its own users. Services like OpenStreetMaps, Google Places, Foursquare and Waze allow users to contribute valuable georeferenced data on businesses, transit and events in near real time. Although crowdsourced data can be updated continuously, it often has problems in aspects such as coverage, reliability and positional accuracy.

The objective of this work is twofold: first, we propose to integrate government and crowdsourced data to produce new datasets with a complementary coverage or complementary aspects of a given subject. Second, we propose to assess the usefulness of data from alternative sources by comparing them with official governmental data, so that official data sources can be reasonably replaced when unavailable or outdated. We present a case study with data on economic activities from official and collaboratively-built alternative sources.

This article is organized as follows. Section 2 discusses related work and presents a general comparison of official and alternative data sources. Section 3 introduces a methodology for integrating data sources on economic activities. Section 4 presents a case study, involving data from economic activities extracted from a municipal tax collection system, and data from Google Places. Section 5 presents and discusses the results. Finally, Section 6 presents conclusions and provides indications for future work.

2. Related Work

The possibility of using online data sources to infer characteristics of specific aspects of an urban environment have been the focus of recent attention. Yuan et al. (2012) propose a framework capable of dividing an urban center into different regions based on its economical functions (commercial, leisure, residential, entertainment, etc.). The work uses data extracted from taxi routes of the city of Beijing, China. However, an approach using only taxi data can be too simplistic to provide a represent the complexities of urban mobility.

Quercia and Saez (2014) use data from social media for studying the relationship between resources and neighborhood deprivation. Authors gather data from Foursquare users in the city of London and use classification algorithms to infer land-use information based those users' locations. Data from Foursquare can only provide a basic view of deprivation for classification, since it focuses only on the users of that specific platform. Integrating socioeconomic data from official sources could enhance the analysis.

In a recent study, Shelton et al. (2014) perform an extensive analysis of the data gathered from geolocated tweets in the city of Louisville, KY, USA for mapping and inferring issues on various urban topics, such as neighborhood segregation, mobility and inequality within the city. The work combines GIS and socio-spatial analysis to support its methodology and conclusions. Although this is a very interesting interdisciplinary work, data from a single source (in this case, Twitter) can be very biased. Furthermore, this work covers a city with specific features, which raises questions if the same methodology could be directly applicable to larger urban centers.

Fonte et al. (2015) addresses the validation of alternative data sources as compared to official data. The work gives examples of projects which use data from volunteered geographic information (VGI) platforms, such as OpenStreetMaps and Panoramio, in order to validate previously mapped land cover areas. Our work proposes a similar validation approach, although in a different context, and also proposes a method for validating alternative data sources in the urban economic activities landscape.

In this work, we assess the feasibility of integrating official data sources to Web-based ones, in order to promote either the expansion of existing data or the replacement of governmental data with current data from alternative sources. In the next subsections, we discuss the characteristics of official and alternative data sources, in preparation for discussing an integration methodology, which focuses on urban economic activities.

2.1. Official Sources

For this work, any data source backed by a governmental entity is considered an **official source**. We exemplify with data on economic activities in a city. Local government entities need to maintain a registry of each business within its jurisdiction, in order to be able to collect business-related taxes and to determine whether a business is formally authorized to operate. Such activities are classified using a standardized list of categories, so that the branch of activities undertaken by any business can be grouped with similar ones for analysis purposes and to support the application of any specific legislation.

Due to these regulations, official sources of data are usually accurate, reliable and well structured [Kalampokis et al. 2011]. However, these same regulations that enforce the completeness of the data may postpone its availability, since many governmental agencies may be involved and their data integration routines are often poorly organized. For example, the quality of urban life index (IQVU) [Nahas 2002] for the city of Belo Horizonte has as one of its principles to be based on indicators that are easily available from government agencies. However, it takes a long time to release results. The IQVU for 2014 was only available on June 2016 [GPDS 2016]. With such delays in the dissemination of official data it is difficult to capture the dynamic behavior of urban activities from this kind of source. Furthermore, such data can be hard to obtain, as local governments are reluctant to provide information on private businesses, even within the scope of open data policies.

2.2. Alternative Sources

Alternative (or Non-Official) sources in this work can be described as any online resource containing data that can be easily obtained through APIs or Web services. Typically, such data are provided and maintained by users, or, in the case of economic activities, directly by entrepreneurs, who are interested in promoting their businesses within the scope of a given application. Collaborative platforms such as Google Places[6], Yelp, Foursquare and Facebook Business are examples of this kind of source.

As in the case of official sources, alternative sources have their own advantages and setbacks. Data can be easily gathered, using crawlers or APIs on the Web. The refresh period is considerably lower, since data collection APIs access the current version of the data, not a copy extracted from a conventional information system. On the other hand, collaborative data can be less reliable, since at first there is nothing to prevent someone from entering false information. In the long run, users are able to filter out

[6] http://developers.google.com/places/web-service/ [Accessed on August 15, 2017]

spurious contributions and help in the curation of the data. In addition, data on small businesses or activities may be missing due to the lack of users with motivation to enter their information. Due to the more flexible Web-based platforms and the technologies they use, data from Web sources tends to be less structured than official data.

3. Methods

A proposed method for integrating official and alternative sources of geographic data on economic activities is depicted in Figure 1. Each step of the method is described next.

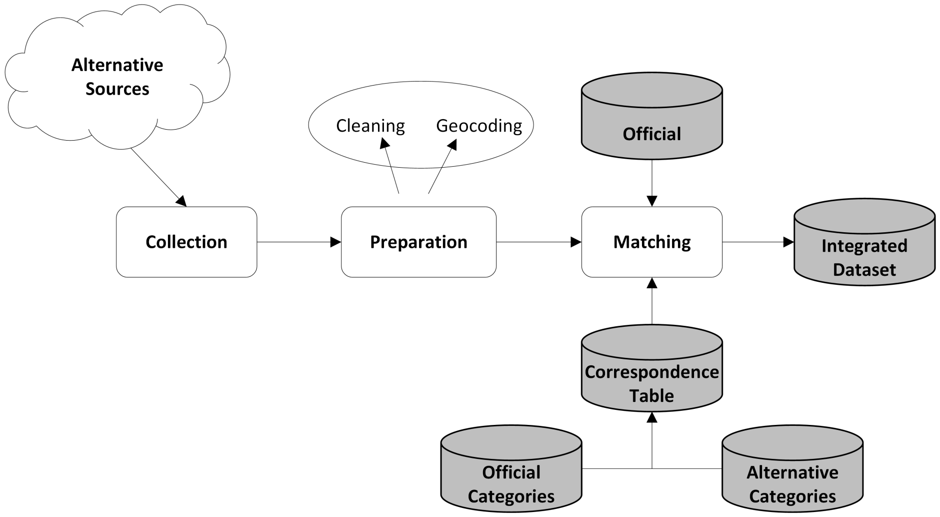

Figure 1. Integration steps

The first step regards data acquisition from both official and alternative sources, comprising active businesses in the city. One reliable source of official data for that purpose is the list of all licensed businesses and services, maintained by local governments for tax collection purposes. Such data can be geographically located using street addresses, but in many local governments such data are already georeferenced.

Any online collaborative platform that contains information on businesses of a region can be used as an alternative data source. Such data should undergo additional preparation, since crowdsourcing often introduces mistakes and irrelevant information. Once cleaning is done, if the Web source provides coordinates, a reverse geocoding step can be used to obtain a street address for each business, if this information is not available in attributes.

At this point, there are two datasets with geolocated business data. Integration is then done by matching attributes of each business, found in both sources. Initially, the business name, address and segment of activity are the first candidate attributes for integration, but, unfortunately, not all of them may be present in both datasets.

In the official dataset used in our case study a “formal” name of each business was not provided. We proceed on using only the coordinates and segment of activity information of each business on both datasets on our matching algorithm. Furthermore, activity classification schemes are rather different. While official sources often use a tax-related classification, alternative sources group businesses as needed to fit the application's purposes. The classification (or *translation*) of each category is done using a correspondence table, which was created by manually matching the closest segment that describes the category of both sources.

This method was instantiated and used on a case study with data from the city of Belo Horizonte, Minas Gerais, Brazil, presented in detail in Section 4.

4. Case Study

For this case study, we used official data supplied by Belo Horizonte's municipal administration, comprising business data from the city's service tax cadastre and data from Google Places.

4.1. Data acquisition

All licensed businesses and/or service providers in Brazil must pay municipal taxes, which vary according to the category of business. In Belo Horizonte, service tax ISSQN (*Imposto Sobre Serviços de Qualquer Natureza*) rates vary between 2.5% and 5%. Data are very detailed, with various aspects on the businesses recorded in attributes. However, there is no record of the "popular" name of the business, as displayed at its site or building. All records are georeferenced and also include postal addresses. Each business is associated to one or more categories, according to a national economic activities classification (see Section 4.3). One of these categories is considered the business' main class of activities.

To gather data from Google Places we developed a crawler to use the available API[7] provided by the application's developers. The API does not have a function that allows collecting business data within a given polygon. We used an API function that returns a JSON response containing data on businesses at a given distance from a given geographic point. This function was repeatedly called with a geographic coordinate and an integer representing a distance radius (in meters).

The geographic points which fed each API call were generated using the intersection between a regular grid and a polygon representing the boundaries of Belo Horizonte. We generated 530,044 distinct points, with a distance of 25m to each other. The radius was also set to 25m, in order to ensure total coverage of the city's territory.

4.2. Data Preparation

ISSQN data comes directly from a conventional information system, enhanced with geographic features, so no cleaning or geocoding was necessary for the official data. Each economic activity address is recorded using a thoroughfare code and the street number, which translate into a pair of coordinates by checking the city's georeferenced addresses table. Google Places data, however, needed to go through cleaning and reverse geocoding steps.

Due to the grid-based data collection method, there were many duplicate entries in the dataset created from Google Places data. These duplicates came from many intersecting regions, searched multiple times by API calls. Also, the original data from Google

[7] https://developers.google.com/places/documentation/ [Accessed on August 15, 2017]

Places contains many duplicates itself, created by its users. We considered as duplicates those records in which the business name was similar and the geographic position was closer than 25 meters. The string similarity[8] function used compares trigrams, groups of three consecutive characters taken from a string, to test if two strings are similar by counting the number of trigrams they share. We experimented with the threshold and visually inspected the results of string similarity, and found that 65% was suitable for the purposes of this work. For instance, the threshold allows detecting similar names such as "Lar Idosos St Antônio Pádua Ve" and "Lar dos Idosos Santo Antônio de Pádua" (67% string similarity), as well as other cases in which place names contain abbreviations.

Besides duplicates, many others records presented inconsistencies. The most common were entries with incomplete information (*e.g.* business category and/or address missing). Such entries were eliminated as well as irrelevant data. We considered irrelevant those records describing simple urban locations or points of reference (e.g., squares, streets, avenues, corners).

All entries in the official and alternative datasets are georeferenced, so a geocoding step was not necessary. However, Google Places provides a coordinate for each business, but not a street address. We executed a reverse geocoding step to obtain address information that could be used in the matching step. Reverse geocoding is the process of obtaining textual information (place names or addresses) from geographic coordinates [Kounadi et al. 2013]. We used functions in the Google Maps API for reverse geocoding and received the street address that is closest to the given coordinates.

After data preparation, the ISSQN dataset includes 270,152 businesses. Google Places dataset contains data on 76,864 businesses.

4.3. Business Segment Classification

Entries from the ISSQN dataset are classified according to business segment for tax collection purposes. Each business can be associated to one or more categories, depending on its range of activities. Classification codes are assigned according to CNAE, a national classification of economic activities (*Classificação Nacional de Atividades Econômicas*, in portuguese)[9]. CNAE is currently in its 2.0 version, which derives from version 4 of the International Standard Industrial Classification of All Economic Activities (ISIC 4), managed by the United Nations Statistics Division [United Nations 2008]. In Brazil, CNAE is officially adopted by the national statistical system and all federal organizations in charge of administrative records. Its adoption in local governments is ongoing. CNAE codes are 7-digit numbers that are hierarchically structured into 21 sections, 87 divisions, 285 groups, 673 classes, and 1301 subclasses.

Google Places, on the other hand, uses a flat classification with 96 distinct categories. We manually classified each of these categories to the closest CNAE code to represent that business segment. Manual classification was preferred, since terminological differences and category naming subtleties precluded using automated methods, and the number of categories is not too large. Table 1 shows examples of the correspondence between CNAE and Google Places classifications.

[8] Available on the pg_trgm module of PostgreSQL

[9] http://cnae.ibge.gov.br/classificacoes/por-tema/atividades-economicas/classificacao-nacional-de-atividades-economicas, [Accessed on July 29, 2017]

4.4. Matching Algorithm

We implemented a matching algorithm to operate over two attributes that are present in both official and alternative datasets: coordinates and the economic activity category.

Our matching algorithm uses the geographic point contained in each entry from Google Places dataset to search any ISSQN record located at most of 150m (which is about the typical size of a city block in Belo Horizonte). The algorithm then checks which ISSQN entries have the same CNAE code as the Google Places entry to determine a match. If there are multiple ISSQN entries with a matching CNAE code in the vicinity, *the algorithm picks the closest of them.* Notice that a simple business name-based matching is not feasible, since official sources in Brazil record the contractual or corporate name of the business, while unofficial sources use the name by which the business is exernally known, which is displayed at its front entrance. For instance, a restaurant officially known as "Ferreira Comércio de Alimentos" is publicly recognized as "Le Petit Gateau".

We use two different parameter setups for the matching algorithm when it comes to comparing economic activity category codes. In the first setup, we only match entries when both CNAE codes are *exactly* the same. In the second one, we considered matches on the first three digits of CNAE in order to judge if both entries were the same or not. The first three digits of the CNAE code indicate a hierarchically coarser classification, with 87 distinct categories. This number of categories is closer to the dimension of Google Places' classification scheme, with 96 categories.

Google Place and ISSQN entries can have more than one CNAE code associated to them (ISSQN consider one as the main CNAE code). We considered two matching situations: first, the Google Places code matches the *main* category of a ISSQN entry; second, *any* of the various categories that can be listed for a Google Places entry matches with *any* CNAE code from the ISSQN entry. Therefore, there are four types of matches, combining 7-digit and 3-digit CNAE codes to compare the *main* or *any* business category present at Google Places and ISSQN data.

Table 1. Correspondence example between Google Places and CNAE categories

Google Places Category	CNAE Code	CNAE Name
Airport	5240101	Airport Operation and Landing Field
Bank	6421200	Commercial Bank
Gym	9313100	Physical Conditioning Activities
Hospital	8610101	Human Health Care Activities (except Emergency Unit)
Lawyer	6911701	Attorney Services
University	8531700	Higher Education - Graduation Only

5. Results

Matching ISSQN records to Google Places entries based on the full 7-digit CNAE category (1,301 classes) resulted in a rate of success of 18%: 13,622 businesses from the 76,864 entries collected from Google Places were matched to ISSQN records. The second match, considering the three first digits of the CNAE code, which correspond to a set of 87 classes, achieved a considerably better success rate: 25%, with 18,829 matches.

We also ran analysis comparing the list of available CNAE for each ISSQN entry (up to ten codes) with the CNAE list from Google Places' entry. The results from exact comparison were 19,976 matched entries (26% from Google Places) and using just the first three digits we were able to reach 30% of matching ratio (representing 22,733 businesses from Google's dataset). The results of the matching runs are summarized in Table 2.

Table 2. Number of matching entries and ratio for different CNAE length

CNAE Code	Main Category	Any Category
7-digit	13,622 **(18%)**	18,829 **(25%)**
3-digit	19,976 **(26%)**	22,733 **(30%)**

After calculating the matching ratio of our algorithm we tried to understand the reasons involving the results observed. Some early insights suggested that matching in our study could be highly affected by the segment in which a business belongs so, we aimed on exploring this attribute further in our analysis.

5.1. Matched Entries Analysis

Going in details around the results obtained in our matching algorithm, we now explore the categories from Google Places which had the most success in terms of matching. All the following analysis were done using the results of the algorithm considering three first digits from CNAE code and comparing it with any of the categories from a Google Places entry.

Table 3. Top 5 Matching Business Groups (3-digits CNAE)

Segment	**CNAE**	**Matching Ratio (%)**
- Computer and Accessories Retail Stores	475	75%
- General Retail Stores	478	74%
- Electric and Hydraulic Services	432	70%
- Personal Care Services	960	68%
- Pharmaceutical, Cosmetic or Orthopedic Retail Store	477	67%

Picking our top segments in Table 3 we see that *Non-specialized Retail Stores* are the leading business segment. That gives us an important insight, suggesting that a considerable portion of the entries listed on Google Places is composed by businesses which consider themselves as being a general commercial store in some degree. Analyzing the following groups in our table we notice they consist of stores which deal directly with the general public (called B2C[10] businesses), having a genuine interest on exposing themselves on Google's platform to attract more customers and finally suggesting that these activities are, in general, well mapped by Google's platform.

[10] Business-to-Customer

5.2. Non-matching Entries Analysis

We mapped all businesses in ISSQN dataset which our algorithm couldn't find a match in Google Places list of entries. The analysis shown in this subsection is split into two different contexts. First, we have a ranking of the categories from *Google Places* dataset with the lowest matching ratio. Then, we do the same approach, but using the total number of occurrences of *ISSQN* because the entire matching ratio has been the same for the worst cases.

5.2.1. Non-matching Google Places entries

Table 4 shows categories with the lowest matching ratio in Google Places dataset. These are categories found on Google Places but, for some reason couldn't find a matching pair on ISSQN dataset.

Table 4. Top 5 Non-matching categories from Google Places dataset

Segment	CNAE	Occurrences	Ratio
- Agricultural Equipment and Livestock Retail Stores	462	231	12%
- Insurance Services (Life and Properties)	651	403	15%
- General Retail Stores	471	43,772	17%
- Accommodation Services (excluding Hotels and Similar)	559	403	18%
- Domestic Services	970	1084	21%

The categories shown in Table 4 presented as being too generic by themselves. The poor matching ratio of the entries suggest that the correspondence between their categories informed in Google Places and the official CNAE code wasn't specific enough in the correspondence table, leading to many cases of missed entries.

5.2.2. Non-matching ISSQN entries

The categories shown in Table 5 didn't have any reported match by our algorithm but were present in the official ISSQN dataset.

Table 5. Top 5 Non-Matching Business Groups from ISSQN (3-digits CNAE)

Segment	CNAE	Occurrences	Ratio
- Other Financial Services (Factoring, Leasing, etc.)	649	318	0%
- Demolition and Land Preparation	431	266	0%
- Electrical Energy Generation and Distribution	351	164	0%
- Investment Fund Administration	663	118	0%
- Market and Public Survey Research Offices	732	110	0%

As seen, the segments without any encountered match are composed by activities which obviously don't rely on Google Places in order to acquire customers. Those are services which don't deal with the general public, having little or no interest on exposing their brand in the platform.

6. Conclusions and Future Work

Dealing with alternative sources of data proves to be a challenge. At the same time, there are several opportunities in integrating unofficial data to governmental sources in order to create a wider picture of business offerings within an urban center.

We found that the reliability of alternative data is more linked with business segmentation than expected, proving to be a good representation of the real business landscape of a city, specially for business-to-customer segments, and a reliable alternative to official data for mapping those categories. When it comes to categories of business where the target market isn't the general customer (business-to-business companies), Google Places clearly didn't have enough data coverage, resulting in a reduced number of matches. The results found also served as a motivation to continue working on integrating different sources and paving the way for the creation of decision-support tools for businesses in the near future.

In terms of future work, the correspondence between the CNAE code and the segment classification used by unofficial sources should be improved, as CNAE has numerous subdivisions and a full mapping has not yet been achieved. The use of matched records to help in this respect must be considered. Machine learning algorithms such as KNN [Zhang et al. 2006], for example, can provide a more refined solution for classification, so that matching algorithms can improve both in processing time and in terms of matching accuracy. Another improvement could be the use of clustered segments in the correspondence table, so each business category from alternative sources can be translated into many CNAE codes (many-to-many relationship).

7. Acknowledgements

The authors wish to thank the Prefeitura de Belo Horizonte (specially the Secretaria Municipal de Finanças) for providing official data, and CNPq, CAPES and FAPEMIG, Brazilian agencies in charge of fostering research and development.

References

Bertot, J. C., McDermott, P., and Smith, T. (2012). Measurement of Open Government: Metrics and Process. In *2012 45th Hawaii International Conference on System Sciences*, pages 2491–2499. IEEE.

Chakraborty, A., Wilson, B., Sarraf, S., and Jana, A. (2015). Open data for informal settlements: Toward a user's guide for urban managers and planners. *Journal of Urban Management*, 4(2):74–91.

Fonte, C. C., Bastin, L., See, L., Foody, G., and Lupia, F. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science*, 29(7):1269–1291.

Gov.UK (2013). G8 Open Data Charter. Available: https://www.gov.uk/government/publications/g8-open-data-charter-national-action-plan [Accessed on September 28, 2017].

GPDS (2016). Relatório geral sobre o cálculo do Índice de qualidade de vida urbana de belo horizonte (iqvu-bh) para 2014. online. Available: https://monitorabh.pbh.gov.br/

sites/monitorabh.pbh.gov.br/files/IQVU/reliqvu14_sitecor.pdf [Accessed on September 28, 2017].

Kalampokis, E., Tambouris, E., and Tarabanis, K. (2011). Open Government Data: A Stage Model. In Janssen, M., Scholl, H. J., Wimmer, M. A., and Tan, Y.-h., editors, *Electronic Government: 10th IFIP WG 8.5 International Conference, EGOV 2011, Delft, The Netherlands, August 28 – September 2, 2011. Proceedings*, pages 235–246. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kounadi, O., Lampoltshammer, T. J., Leitner, M., Heistracher, T., and Francis, T. (2013). Accuracy and privacy aspects in free online reverse geocoding services. *Cartography and Geographic Information Science (CaGIS)*, 40(2):140–153.

Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly*, 32(3):323–332.

McDermott, P. (2010). Building open government. *Government Information Quarterly*, 27(4):401–413.

Nahas, M. I. P. (2002). *Bases teóricas, metodologia de elaboração e aplicabilidade de indicadores intra-urbanos na gestão municipal da qualidade de vida urbana em grandes cidades : o caso de Belo Horizonte*. Phd, Universidade Federal de São Carlos.

Perritt Jr, H. H. (1997). Open Government. *Government Information Quarterly*, 14(4):397–406.

Quercia, D. and Saez, D. (2014). Mining Urban Deprivation from Foursquare: Implicit Crowdsourcing of City Land Use. *IEEE Pervasive Computing*, 13(2):30–36.

Shelton, T., Poorthuis, A., and Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, 142:198–211.

United Nations (2008). International Standard Industrial Classification of All Economic Acitivities (ISIC), rev.4. online. Available: http://unstats.un.org/unsd/cr/registry/regdntransfer.asp?f=135 [Accessed on September 28, 2017].

Yuan, J., Zheng, Y., and Xie, X. (2012). Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, pages 186–194, New York, New York, USA. ACM Press.

Zhang, H., Berg, A. C., Maire, M., and Malik, J. (2006). SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2126–2136.