Computers, Environment and Urban Systems 61 (2017) 39–48

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Land Use detection with cell phone data using topic models: Case Santiago, Chile

Sebastián A. Ríos*,[1], Ricardo Muñoz

Business Intelligence Research Center, Department of Industrial Engineering, Universidad de Chile, Beauchef 851-OF. 615,P.O. Box: 8370456 Santiago, Chile

ARTICLE INFO

Article history:
Received 3 February 2016
Received in revised form 22 August 2016
Accepted 23 August 2016
Available online 15 October 2016

Keywords:
Land Use detection
Topic models
Mobile phone data
Human mobility
Latent Dirichlet allocation
Big data

ABSTRACT

Today we have the opportunity without precedents to analyze human land use or mobility behavior in a city, country or even the globe. Some studies have analyzed existing data generated daily by mobile networks, mostly using geo-localization in Twitter, Foursquare or cell phone records. Most of these studies use a small portion of data (a few days or a couple million records). This time we will show a novel way to apply latent semantic topic models to detect Land Use Patterns in a real big dataset of 880,000,000 calls made in Santiago City (Chile) over 77days by about 3 million customers of a major telecommunications company. We proposed to use a latent variables clustering technique which allow us to detect four interesting clusters. We found out that the application of LDA allow us to discover two well known clusters (residential and office area clusters) but also we discover two new clusters: Leisure-Commerce and Rush Hour patterns.

1. Introduction

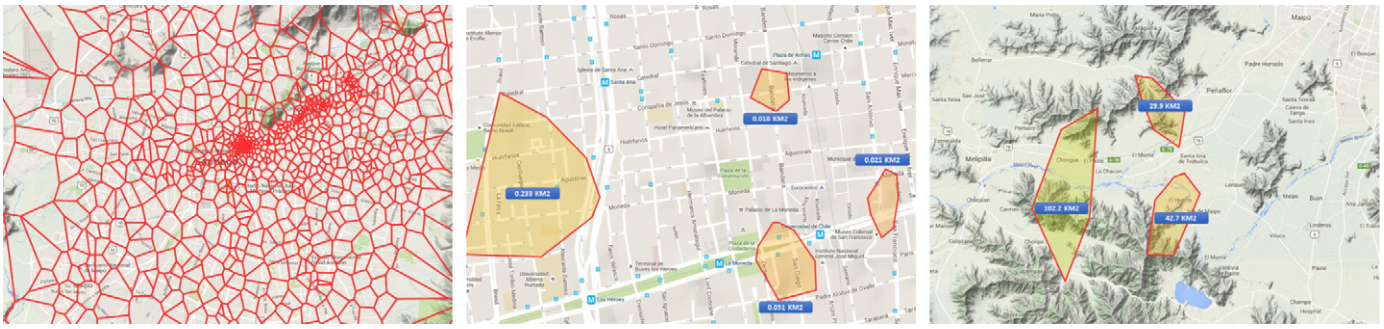
Collecting data on human behavior used to be done by laborious methods such as surveys, which are applied to small samples of test subjects. Such results are difficult to update. Moreover, these methods are expensive in time and money. Over the past few years multiple channels have arisen where people are willing to disclose personal information that is useful. This facilitates this type of analysis. The best examples are social networks such as Facebook or Twitter. Every minute Twitter users send over 100,000 tweets and 2% of all tweets include geographic metadata (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013). Additionally, since smartphones and data plans are more affordable, cell phones have become one of the main sensors of human activities, thanks to their growing market penetration, the wealth of applications to the end user (Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2012) and its simplicity to share information anytime and anyplace.

This vast amount of data, generated every second, has been used for social network analysis (Baruah & Angelov, 2012; Catanese, Ferrara, & Fiumara, 2013; Xu, Cui, Tie, & Zhang, 2012), urban dynamics (Calabrese, Colonna, Lovisolo, Parata, & Ratti, 2011) and the understanding of customer behavior (Dragana & Becejski-Vujaklija Dragana, 2009); and presents an opportunity without precedents to analyze human behavior on a city, country or even the globe. In fact, cell phones have also the advantage of being carried by the same individual during his/her daily routine. This offers the best proxy to capture individual human trajectories (Gonzalez, Hidalgo, & Barabasi, 2008) and their geo-localization (by the serving antenna geographical position) providing insight into the spatial organization of individual and human networks (Chi, Thill, Tong, Li, & Yu, 2014; Phithakkitnukoon, Smoreda, & Olivier, 2012). But, to do so requires facing the big data processing challenge, which affects actual methods and software architectures employed to research.

There are some studies on data generated daily by mobile networks, mostly using geo-localization in Twitter (Becker et al., 2011; Frias-Martinez et al., 2012; Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2014; Fujisaka, Lee, & Sumiya, 2010; Wakamiya, Lee, & Sumiya, 2011) or cell phone records (Gonzalez et al., 2008; Phithakkitnukoon et al., 2012; Reades, Calabrese, Sevtsuk, & Ratti, 2007). All of those studies have been done using K-means, Self-Organizing Map (SOM) or other clustering methods based on distance. But in this paper we proposed to use a latent topic modeling in cell phones data for automatic identification of Land Use Patterns on a very large real database of 880,000,000 calls made in Santiago over 77 days. Nearly 6,300,000 people live in Santiago City, therefore we have data of 50% of the population in this city.

* Corresponding author.
E-mail addresses: srios@dii.uchile.cl (S. Ríos), rimunoz@ing.uchile.cl (R. Muñoz).
[1] http://www.ceine.cl

http://dx.doi.org/10.1016/j.compenvurbsys.2016.08.007
0198-9715/



(a) Santiago Tessellation (b) Urban Areas (c) Rural Areas

Fig. 1. Voronoi Tessellation.

The main idea is to model the antenna's activity (BTS) discovering latent variables, which are not directly observed on the data, assuming a mixture of probabilistic distributions over it and then reducing their dimensionality. Our main contribution is to innovate in pattern recognition method application, proving that topic models can be used to detect Land Use Patterns. This idea and our background on text mining on social networks allowed us to hypothesize that topic models could be applied to unveil Land Use Patterns. Topic models are used to discover topics on free texts by assuming that topics cannot be observed directly on the text of a web page or a social network comment, but they are expressed by a set of terms on the text. Finally, we adapted, calibrated and evaluated the quality of results by expert knowledge of Santiago.

The paper is organized as follows: Section 2 presents related work in the characterization of urban Land Use. In Section 3, we introduce our adaptation of LDA in Land Use Pattern identification and the experiments with its results in Section 4. Finally, Section 5 presents the conclusions of this work.

2. Related work

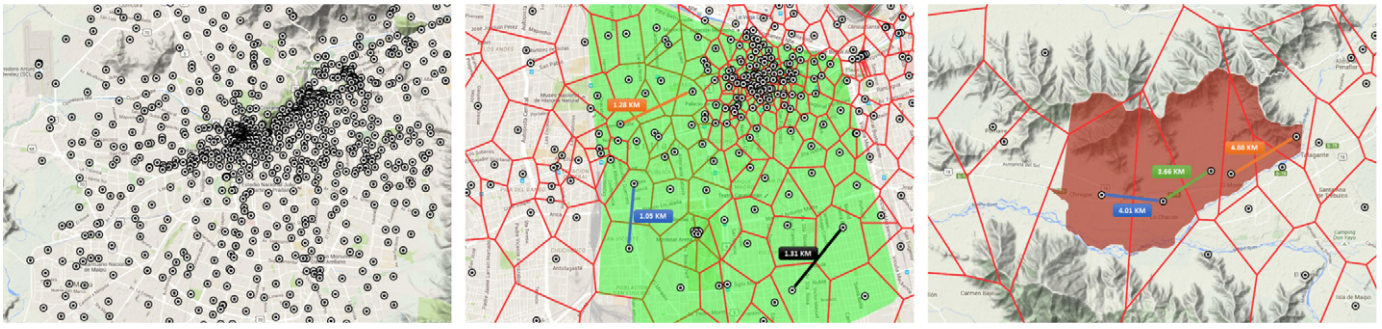
Much research has been done on characterizing patterns in urban areas using social crowd-based resources like geo-tagged tweets or cell phone records. Fujisaka et al. (2010) discovered regional characteristic patterns from movement histories using aggregation and dispersion models in order to understand the nature of human mobility. Similar work was developed by Wakamiya et al. (2011), where they defined the geographic regularity of an urban area using daily crowd activity patterns and analyzing their changes over time. Also, Noulas, Scellato, Mascolo, and Pontil (2011) applying spectral clustering, modeled crowd activity patterns in two cities using geolocated information provided by Foursquare. Crandall, Backstrom, Huttenlocher, and Kleinberg (2009) performed landmark location using data from geo-tagged photos on Flickr with the mean-shift algorithm. Additionally, Frias-Martinez et al. (2012) evaluated the use of geo-located tweets as a complementary source of information for urban planning applications using SOM, Voronoi Tessellation and K-means algorithm. Those authors Frias-Martinez et al. (2014) also proposed a technique that automatically determines land uses in urban areas by clustering geographical regions with similar tweeting activity patterns.

Related to cell phone data, Soto and Frias-Martinez (2011) presented a technique to automatically identify the uses that citizens give to different parts of a city using the information contained in cell phone records, applying fuzzy clustering techniques. Reades et al. (2007) monitorized the dynamics of Rome and obtained clusters of geographical areas measuring cell phone tower activity using Earlangs.

All these works use clustering methods that calculate distances either as *inter-group* or *intra-group* data. We emphasize the advantages of using latent variables over traditional clustering techniques and we validated the results of this topic model as an excellent model to characterize Land Use in urban areas.

3. Proposed model

In this work, Land Use Patterns are detected using an generative statistical method called Latent Dirichlet Allocation (LDA), which is often used in the text mining and natural language processing research to discover underlying free text topics in web pages, social network comments, news, etc. for example in Ríos, Aguilera, and Guerrero (2009), L'Huillier et al., (2010), and Ríos and Muñoz (2016). In the case of text applications, we commonly use the concepts of



(a) Santiago BTS Positions (b) Urban Areas (c) Rural Areas

Fig. 2. Voronoi Tessellation showing antenna positions.

text corpus, documents, words and topics; therefore, we adapted these concepts in order to apply LDA to discover Land Use Patterns using the cell phone network data, which to the best of our knowledge has not been performed before. In the context of this work *words* refers to the activity (calls) over some time window (see below Activity Block). A *document* refers to a complete week of words (Activity Blocks) for a specific BTS (see below BTS Activity Pattern). The *corpus* is the full data obtained from BTSs that is the set of all documents. Finally, *topics* are equivalent to the Land Use Patterns and the main goal of this work is to detect them. Therefore, the aim is to learn both what the topics are and which documents employ them in which proportion.

We found a recent research where LDA is applied to text on geolocated tweets like (Steiger, Westerholt, Resch, & Zipf, 2015). But our research is very different since we do not apply LDA to a text, but to antennas activities. We are able to obtain the same clusters they found (office and home areas clusters) but we discover two totally new clusters: Rush hour areas and Shopping areas.

To the best of our knowledge, the method commonly used to process activity data on a city are K-means, fuzzy c-means, SOM, hierarchical clustering (which are very simple methods and very fast to compute on big data) but they are all distance-based methods (like Euclidean distance) which is used to compute proximity between examples. This is the main difference with LDA (and other generative algorithms) which assume a probability distribution over the BTS usage that form a hidden cluster. When we use LDA, we are assuming that the land usages clusters follow a Dirichlet distribution (similar to topics in a text of a web page or a news paper). Therefore, this mechanism allow us to discover hidden usage patterns; since it is not only dependant just from the data itself or the distance from one observation to the rest of them, but in their co-occurrence over time following the Dirichlet distribution.

3.1. Basic notation

A *Base Transceiver Station* (BTS) is the equipment that facilitates the communication between cell phone devices and the cell phone network. Each BTS has one or more antennas distributed over a given area in order to provide the best possible radio coverage through regions called cells. Each time a user makes or receives a phone call, the call is processed by the closest BTS and it is geographically tagged with the latitude and longitude of this BTS.

A *BTS Activity Pattern* (BAP) related to the BTS $b \in B$ (B is the set of BTSs) is a list of consecutive time frames (time windows) A. Each time frame is, in fact, a single *Activity Block* $A^b(t_i, t_{i+1})$; where t_i is the start time of the period of interest and t_{i+1} the end time of the period of interest.

$$BAP_b = \langle A_1^b, A_2^b, \ldots, A_m^b \rangle, m \in N$$
$$A_i^b = Norm.calls.number(b, t_i, t_{i+1}), i = 1, 2, \ldots, m$$

where $Norm.calls.number(b, t_i, t_{i+1})$ is a function which returns the proportion of calls made in the period $[t_i, t_{i+1}]$ for BTS b. In this work, disjoint intervals will be used, it means one interval starts when the previous interval ends.

For example, in order to study the behavior of a BTS j every hour during the week; we define its Activity Pattern BAP_j, which is a vector with 168 components (24 h, seven days per week). Components are: $A_1^j \ldots A_{168}^j$, where A_i^j is the proportion of calls made in an hour in the antenna BTS_j.

3.2. Applying a topic modeling on cell phone data

A topic model, e.g. Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), can be considered to be a probabilistic model that relates documents and words through latent variables which represent the main topics inferred from the text itself. In this work, this idea has been adapted to understand activities in the city using cell phone data. The rationale of using LDA in this problem is to model *BTS activity patterns* as arising from multiple latent variables (topics) of Land Use Patterns, where a land usages are defined to be a distribution over a fixed *Activity Blocks* set. Specifically, we assume that K Land Use Patterns (topics) are associated with the set of BTS, and each BTS exhibits these usage patterns with different proportions. In this context, a *BTS activity pattern* can be considered as a mixture

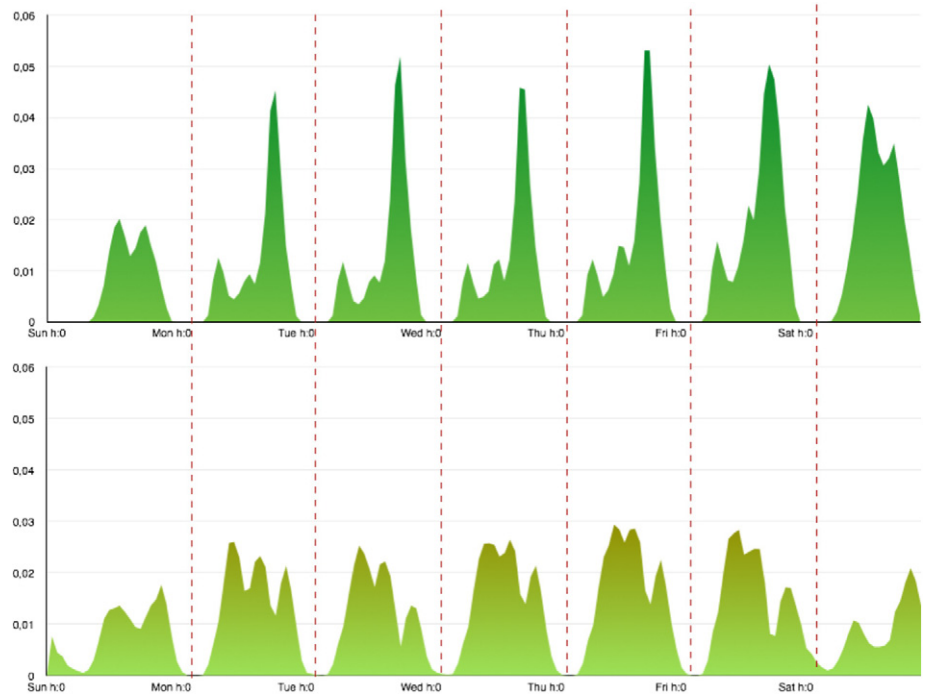


Fig. 3. Examples of BPS activity patterns from two different BTS antennas.

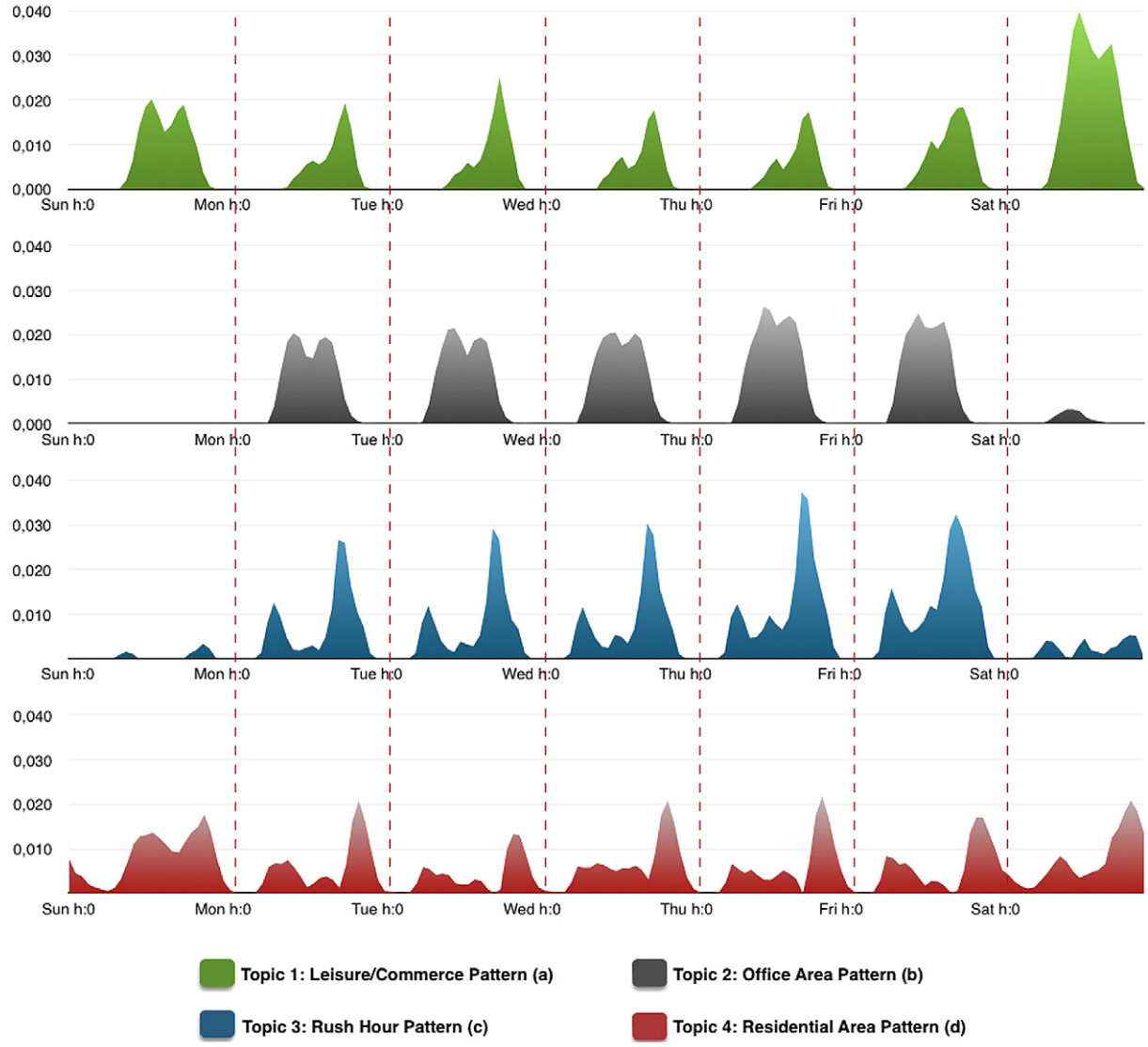


Fig. 4. Latent Land Use Patterns detected after applying LDA.

of usage patterns, represented by probability distributions, which can generate the *Activity Blocks* in a *BTS activity pattern* given these usages (topics). The inferring process of the latent variables is the key component of this model, whose main objective is to learn from cell phone data the distribution of the underlying usage patterns in a given dataset of *BTS activity patterns*.

With LDA, given the smoothing parameters β and α and a joint distribution of a topic (Land Use Pattern) mixture θ, the idea is to determine the probability distribution to generate – from a set of topics $\mathcal{K}$ – a *BTS activity pattern* composed by a set of *S activity blocks* $\mathbf{a}$ $(\mathbf{a} = (a^1, \ldots, a^S))$,

$$p(\theta, z, \mathbf{a}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^{S} p(z_s|\theta) p\left(a^s|z_s, \beta\right) \quad (1)$$

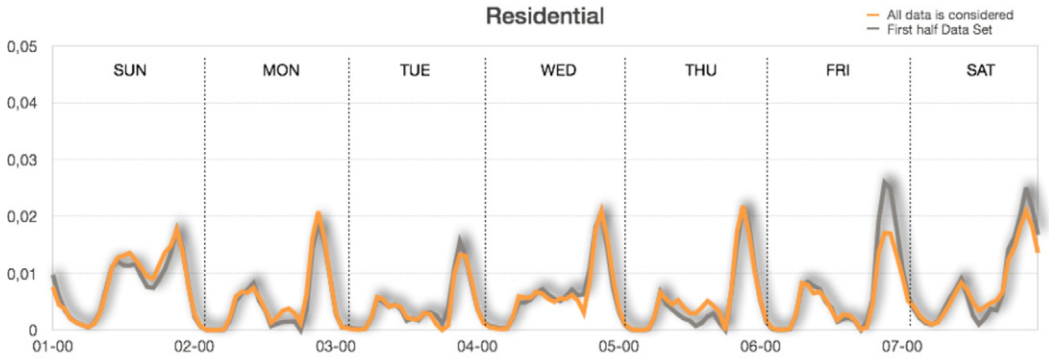
where $p(z_s|\theta)$ can be represented by the random variable θ_b; this topic z_s is present in BTS b $\left(z_s^b = 1\right)$. A final expression can be deduced by integrating Eq. (1) over the random variable θ and summing topics $z \in \mathcal{K}$.

3.2.1. Generative model

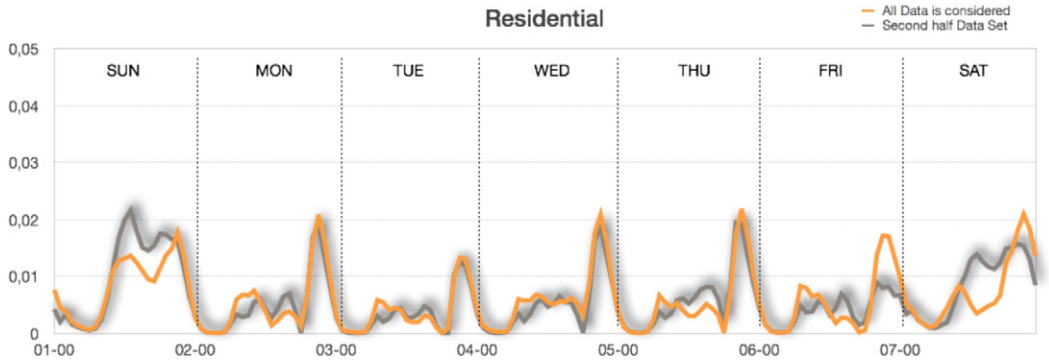
Let K be a specified number of Land Use Patterns (topics), B the number of BTS, $\vec{\alpha}$ a positive K-vector and η a scalar. We let $Dir_B(\vec{\alpha})$ denote a B-dimensional Dirichlet with vector parameter $\vec{\alpha}$ and $Dir_K(\eta)$ denote a K dimensional symmetric Dirichlet with scalar parameter η.

1. For each Land Use Pattern $k \in [1, K]$,
   (a) Draw a distribution over Activity Blocks $\vec{\beta}_k \sim Dir_K(\eta)$
2. For each BTS $b \in [1, B]$,
   (a) Draw a vector of Land Use Pattern proportions $\vec{\theta}_b \sim Dir_B(\vec{\alpha})$
   (b) For each Activity Block $a \in [1, A_b]$ in BTS b,
      (i) Draw a Land Use Pattern assignment $Z_{b,a} \sim Mult\left(\vec{\theta}_b\right)$, $Z_{b,a} \in \{1, \ldots, K\}$
      (ii) Draw an Act. Block $W_{b,a} \sim Mult\left(\vec{\beta_{Z_{b,a}}}\right), W_{b,a} \in \{1, \ldots, B\}$

The hidden topical structure of a collection is represented in the hidden random variables: the Land Use Patterns $\vec{\beta}_{1:K}$, the per-BTS Land Use Pattern proportions $\vec{\theta}_{1:B}$, and the per-activity block Land Use Pattern assignments $Z_{1:B,1:A}$. With these variables, LDA is a type of mixed-membership model.



(a) First Half Dataset



(b) Second Half Dataset

Fig. 5. Residential Pattern stability.

4. Experimental setup and results

The study area of this research is the Santiago Metropolitan Region, one of Chile's fifteen administrative divisions. The region is divided into 52 communes, covering an area of 15,403.2 km^2. The total population of the city is about 6 million. For the purpose of this work, the dataset was originally collected by the major telecommunications company in Chile, the dataset consists of 880 million phone calls recorded over a 77 day period for approximately 3 million anonymized mobile phone users. It contains the information about the phone call: date, time, duration and coordinates (latitude and longitude) of the BTS routing the communication for each phone call. Furthermore, we only know the coordinates of the BTS routing the communication, hence exact location of users are not known within a tower's service area.

In our dataset, each BTS serves an area assigned by a Voronoi Tessellation (see Fig. 1 (a)), in urban areas (see Fig. 1 (b)) each BTS server an area of approximately 0.021 km^2 and 74.6 km^2 in rural areas (see Fig. 1 (c)). There are 1183 BTS towers routing the communication in Santiago (see Fig. 2 (a)), the distance between BTS's can be a few meters (see Fig. 2 (b)) in areas up to several kilometers in rural areas (see Fig. 2 (c)).

In our study, we set every BAP related to the BTS b as the number of calls that are managed by that BTS every hour in a seven-day week. Therefore, each BAP is a vector with $N = 168$ components (24 Activity Blocks per day, seven days per week), where every component reveals the activity of b during 1 h. This is achieved by holding the proportion of calls during that hour compared with the amount of calls made in the whole week.

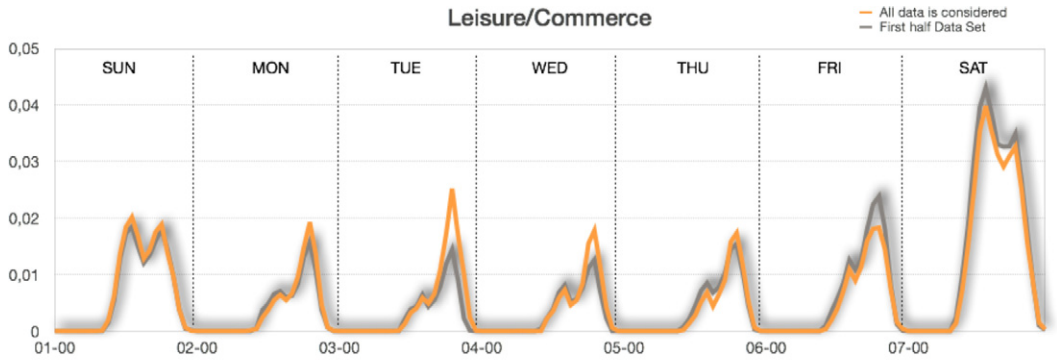
We would like to notice that the time needed to generate the BAP table based on our big dataset took more than a day. However, LDA algorithm (or any other) run over the BAP table which has 77 records with 168 components for each BTS, which give us about 91,000 records of 168 components. Therefore, the performance of the methods is not an issue once we have this summarized data.

Fig. 3 illustrates two BAP related to different BTSs. Every BAP starts on Sunday and finishes on Saturdays.

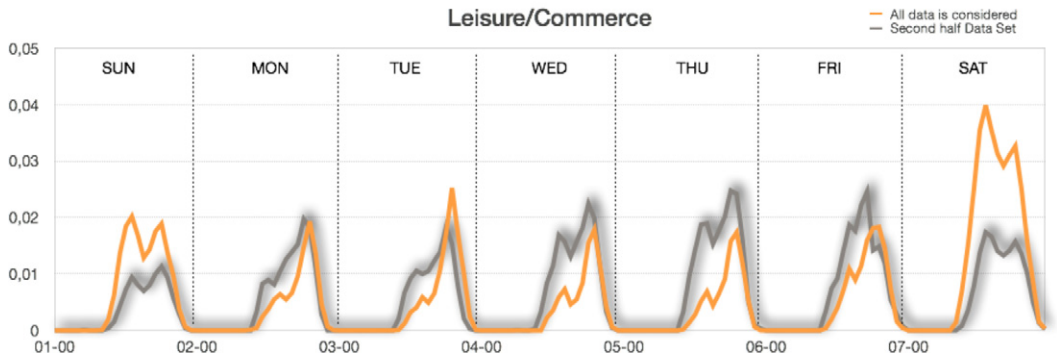
4.1. Land Use Pattern identification

We have used Latent Dirichlet Allocation (LDA) to detect latent Land Use Patterns which define land uses. LDA needs as input the number of topics K (land uses) which represents activities in the city. In order to validate the optimal number of topics we executed LDA for each value $K = 2, \ldots, 10$ and selected, using expert knowledge, the value of K which provides the maximal information and the minimal dimensionality.

It is important to notice that as mentioned above, we compute perplexity for $K = 2, \ldots, 10$; a common evaluation metric for generative models. The better performance was obtained for 10 clusters (topics). However, when we analyzed the resulting clusters we were able to find out the two well known clusters of working areas and residential areas; plus two clusters that we were able to interpret: leisure-commerce and rush hour clusters with objective information from Google maps. Unfortunately, we were not able to interpret the rest of the clusters. This is why we only reported four new clusters obtained with our method. However, it is possible that with better sources of information we could derive interpretations for those clusters.



(a) First Half Dataset



(b) Second Half Dataset

Fig. 6. Leisure/Commerce Pattern stability.

A final remark on evaluation is that, several authors have shown that perplexity is not strongly related with human judgment. A good example is Chang, Gerrish, Wang, Boyd-Graber, and Blei (2009) and more recently Contreras-Piña and Ríos (2016), where they have run experiments in big datasets of text and then have conducted experiments on humans to measure the quality of hidden topics discovered. Thus, this is an argument to explain why if we obtain better perplexity for $K = 10$ in the end, we only could interpret four clusters.

Fig. 4 shows the Land Use representatives (topics) obtained after applying LDA with four Land Use Patterns. An analysis of these topics allowed to hypothesize about the land uses. Fig. 4 (a) describes a land use characterized by high activity during weekends specially on Saturdays. During weekdays the behavior is regular through days showing an increasing activity with peaks in afternoons (19:00 PM). This behavior seems to belong to leisure or commercial areas.

Fig. 4 (b) shows a land use characterized by a high and regular activity during weekdays and almost non-existing activity during weekends. During the day there is a decreasing activity at lunchtime (13:00 PM), indicating probably office areas activity. A similar behavior to Fig. 4 (b) is presented in Fig. 4 (c), but in this case the activity during the day shows a different pattern. Each day is characterized by three peaks which might be associated to the times at which people typically get to work, go for lunch, and leave work. The first one is in the morning at 09:00 AM, the second one and considerably less than the others occurs at lunchtime and the last one is the highest peak during the day and this occurs in the afternoon at 19:00 PM. These activity patterns seem to belong to areas with high human displacement and traffic jams because every peak occurs at rush hour. Moreover, the lunchtime peak almost dissipates on Fridays in order to contribute to the afternoon peak; this phenomena could be explained because people leave work early on Fridays.

Finally, Fig. 4 (d) presents a land use with activity during all week, but the behavior during weekdays and weekends is different. During weekdays the activity starts at 06:00 AM, decreases along morning and increases again after 19:00 PM with a peak at 21:00 PM. Moreover, the activity is higher on weekends, especially on Sundays. This behavior is typical of residential areas in which individuals come from work in the afternoon during weekdays and during weekends stay at home.

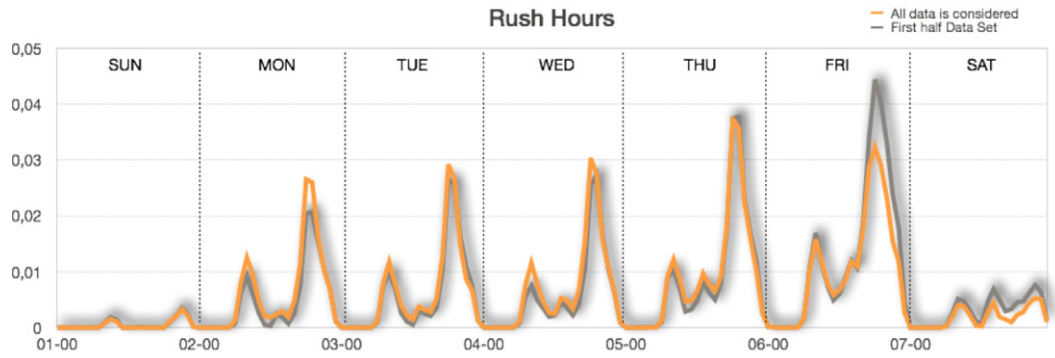
4.2. Land Use Pattern stability

In order to analyze the stability of the patterns discovered using the methodology presented in this work, we divided our dataset (S) in two equal data subsets. We applied our methodology to discover Land Use Patterns using each subset. The first dataset ($S1$) contains calls made between 04/18/2013 and 05/26/2013, and the second one ($S2$) between 05/27/2013 and 07/04/2013.

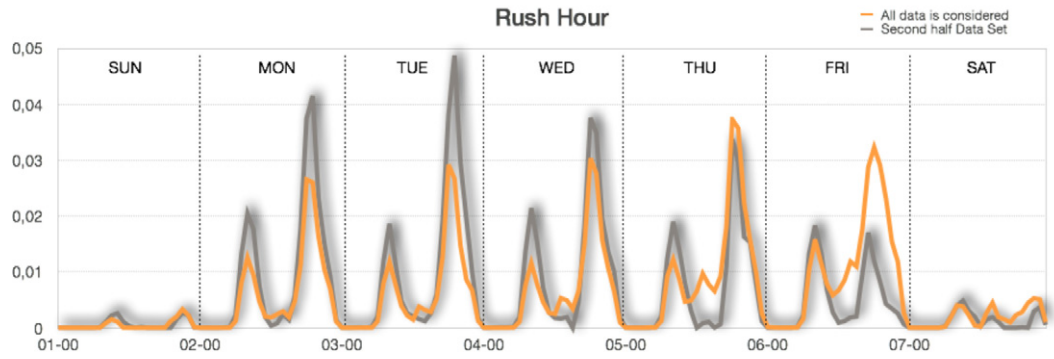
To quantify the pattern stability under different datasets, we used Cosine Similarity. This measure is defined as follow:

$$COS(BAP_b, BAP_c) = \frac{\sum_{i=1..N} A_i^b \cdot A_i^c}{\sqrt{\sum_{i=1..N} \left(A_i^b\right)^2} \cdot \sqrt{\sum_{i=1..N} \left(A_i^c\right)^2}}$$

This variable is in the range $[-1, 1]$ and equals 1 only when the two BAP, BAP_b and BAP_c are exactly coincident. We examined how the discovered patterns change as the dataset varies from the first ($S1$) and the second half ($S2$) to the whole dataset (S). The results for the comparison between the patterns discovered using the first half and the whole dataset are showed in Figs. 5 (a), 6(a), 7(a), and 8(a).



(a) First Half Dataset



(b) Second Half Dataset

Fig. 7. Rush Hour Pattern stability.

In general, results show that most discovered Land Use Patterns are very stable when dataset is reduced. Indeed, cosine similarity between S and $S1$ are 0.978, 0.982, 0.983 and 0.988 for Rush Hour, Residential, Leisure/Commerce and Office Areas respectively. Similarly, the comparison between S and $S2$ – presented in Figs. 5 (b), 6(b), 7(b), and 8(b) – also exhibit high rates of similarity: 0.885, 0.936, 0.806, and 0.9886 for Rush Hour, Residential, Leisure/Commerce and Office Areas respectively. Patterns in this subsets are lesser stable than patterns from the first half. The largest differences occur in Rush Hour and Leisure/Commerce patterns. Some of these differences are explained because within this period are the Chilean winter holidays (June to July). This causes there are fewer people circulating through the city in rush hour and displaces some of the recreational/commercial activities to weekdays.

As a final remark, it is possible to say that discovered patterns are very stable over time. Also, our methodology continues to find out the same patterns although a great mobility pattern change was present in $S2$ dataset (winter vacations). Of course, as a subject for a future work it would be interesting to discover which is the minimum dataset needed to avoid Land Use Pattern change or vice versa.

4.3. Land Use Pattern validation

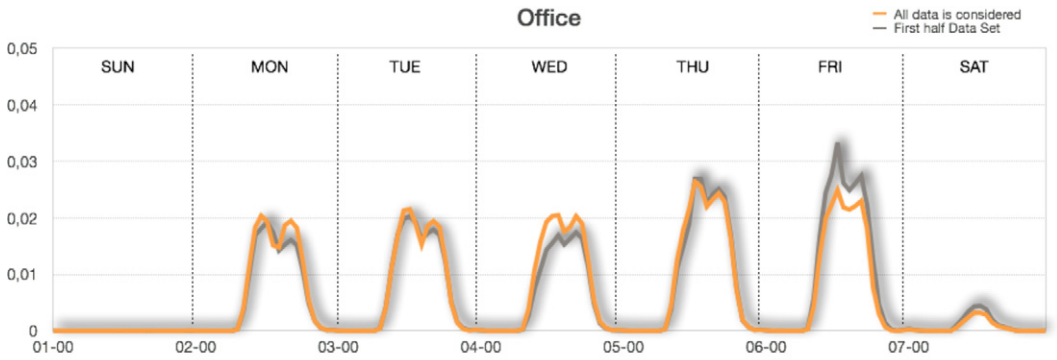
In order to validate our land use hypothesis, we used our results as a layer over the city map to have a geographical representation of the areas where we have its real land use. The use of LDA allows the capturing of the degree to which each Land Use is present for each BTS b. LDA returns a Land Use score g_{kb} for Land Use Pattern (topic) k and BTS b, $\sum_k g_{kb} = 1\ \forall b$. Fig. 9 illustrates each Land Use Pattern (topic) over the city, where we see how different every Land Use Pattern distribution is. In order to identify Land Use easily we discard all BTS towers with a Land Use score lower than a given threshold θ.

The validation consists in checking if the interpretation of the Land Use Pattern given in Section 4.1 correlates the BTS infrastructure located in a geographical area and its vicinity. Since a database detailing the actual land use of the city and the different uses we have identified is not available, we use our expert knowledge of the City of Santiago.

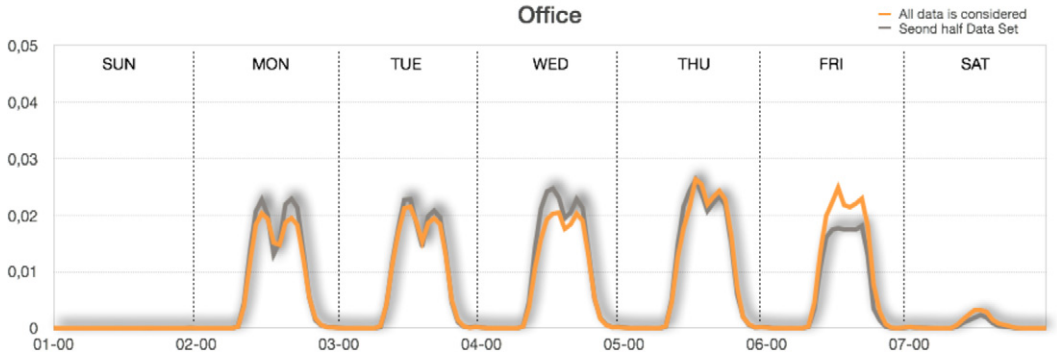
The residential Land Use Pattern represented in Fig. 9 (a) shows higher Land Use score in the periphery of the city center, where the business and commercial center is located. The periphery of the city center contains the biggest residential zones in Santiago. In this pattern there is no presence of zones with scores considerably higher than others, but just one area which covers the *Movistar Arena*. It is one of the largest multi-purpose arenas in South America behind some Brazilians arenas like Ginásio Ibirapuera, HSBC Arena (Rio de Janeiro) and Maracana Arena. The high score in this area is due to the events presented in this location being 1cheduled at a time when the residential pattern presents high activity.

Fig. 10 (a) presents the Leisure-Commerce Land Use pattern. This Land Use Pattern presents a high intensity in some points of the city. In order to validate this pattern, we have highlighted these points (see Fig. 10 (b)) where the blue circles contain the principal Shopping Malls in Santiago and the green circle contains the largest stadium in Chile, where are located some tennis courts, an aquatic center, a gymnasium, a velodrome and a BMX circuit.

The Rush Hour Land Use Pattern is presented in Fig. 11(a). This pattern has a high Land Use score in two main areas, the first runs horizontally through Santiago and the second one runs vertically. These patterns are highly correlated to the subway network of Santiago (see Fig. 11 (b)).



(a) First Half Dataset

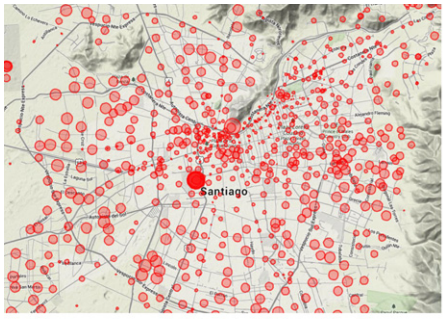


(b) Second Half Dataset

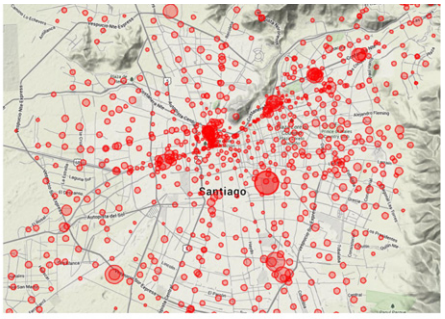
Fig. 8. Office areas pattern stability.

4.4. Discussion

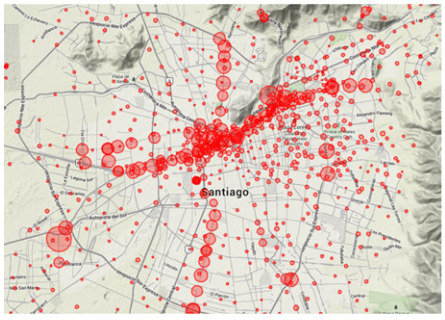
It is very important to remark that the cell phone network data cannot provide the exact location of a cell phone. In fact, the position of cell phones varies within a radius r from a BTS. This radius depends on the density of BTSs that are in a specific area. In a very BTS dense area – such as a city downtown – the radius r is just a few hundred meters but in areas where there are few BTSs the radius can be



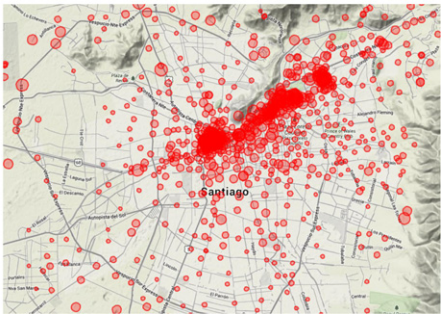
(a) Residential



(b) Leisure-Commerce

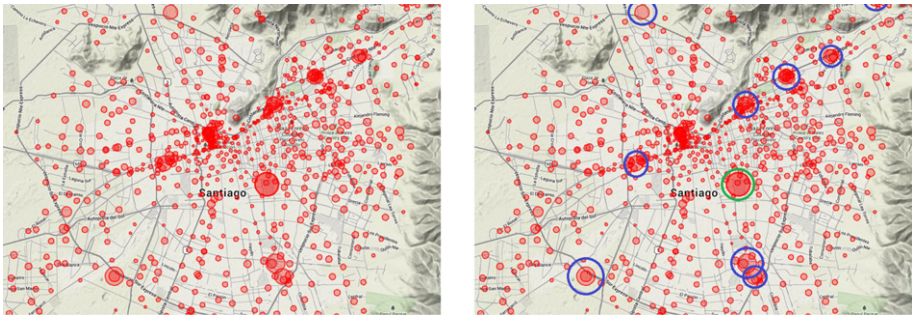


(c) Rush Hour



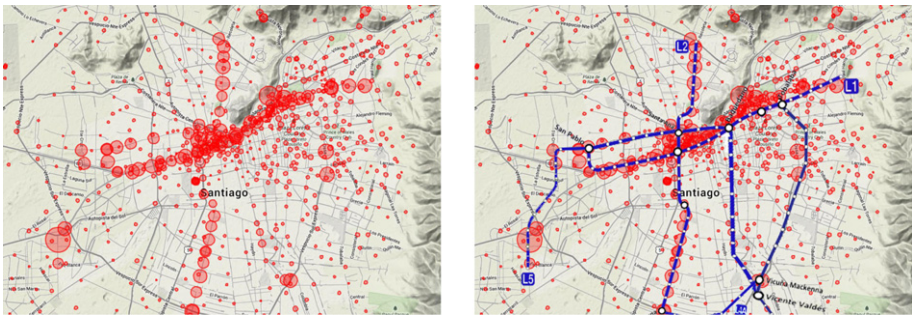
(d) Offices Areas

Fig. 9. Geographical representation of Land Use Patterns.



(a) Leisure-Commerce (b) Shopping Malls of Santiago

Fig. 10. Leisure-Commerce Pattern validation.



(a) Rush Hour (b) Rush Hour & Subway Network

Fig. 11. Rush Hour Pattern validation.

several kilometers. Thus, in Santiago City downtown (see Fig. 2) the average radius is about 250 m or about 3 blocks; but in the residential areas, we can observe areas of radius of 350 m to 750 m (from 3 to 7 blocks). Thus, our clusters are approximate behavior from cell phones in that radius.

Because the amount of data used is huge (more than 3 million cell phones over 77 days); the amount of clusters that can be obtained is quite high, but most of them will probably be noise. Therefore, we think new studies can be produced using this data and some criteria to study a specify area of the city. However, to do so, we need to improve cell phone location resolution.

5. Conclusions & future work

We have shown a methodology to understand the behavior of a city by discovering Land Use Patterns. The novelty of our approach is the use of latent variables over big data from cell phones network (more than 3 million cell phone data). Inferring these variables – which are not directly observed in data – have proved to be highly satisfactory.

We discovered four Land Use Patterns, two of these patterns are very well known by the community (office areas and residential areas); and two patterns are new information: Leisure-Commerce Pattern and Rush Hour Pattern. Leisure-Commerce Pattern is related to where people can spend their free time, this pattern correlates with shopping malls, cinemas, parks, etc. Rush Hour Pattern appears at certain time of the day and over certain streets, avenues, and highways.

This way, we successfully showed that a probabilistic generative algorithm is able to discover new information. This information could be important for several uses like urban planning.

For future work, we are focusing on finding new sources of information about land usage that will allow us to validate our approach with a ground truth. At the same time, we are accumulating data in order to develop a whole year study which probably generate interesting information. A very important point is to enhance cell phone location resolution to be able to develop in detail land usage studies for areas where we have very good information regarding its uses.

We also propose to study the evolution of a city along some years, find ways to set a number of Land Use Patterns, and cross this information with other sources, such as geo-referenced data from Twitter.

Acknowledgments

This work was made possible by the Business Intelligence Research Center (CEINE) from The University of Chile and the Complex Engineering Systems Institute (ISCI) (ICM-FIC: P05-004-F,CONICYT:FB0816).

References

Baruah, R.D., & Angelov, P. (2012). Evolving social network analysis: A case study on mobile phone data. *Evolving and Adaptive Intelligent Systems (EAIS), 2012 IEEE Conference on. IEEE.* (pp. 114–120).

Becker, R.A., Cáceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. *Pervasive, Computing, IEEE*, *10*(4), 18–26.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993—102 2.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 141–151.

Catanese, S., Ferrara, E., & Fiumara, G. (2013). Forensic analysis of phone call networks. *Social Network Analysis and Mining*, *3*(1), 15–33.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., & Blei, D.L. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296.

Chi, G., Thill, J.-C., Tong, D., Li, S., & Yu, L. (2014). Uncovering regional characteristics from mobile phone data: A network science approach. *Papers in Regional Science*, *95*(3), 613–631.

Contreras-Piña, C., & Ríos, S.A. (2016). An empirical comparison of latent sematic models for applications in industry. *Neurocomputing*, *17 9*, 176–185.

Crandall, D.J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. *Proceedings of the 18th International Conference on World Wide Web. New York, NY, USA*. (pp. 761–770).

Dragana, C., & Becejski-Vujaklija Dragana, G.N. (2009). A call detail records data mart: Data modelling and olap analysis. *Computer Science and Information Systems*, *6*(2), 87–110.

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. *Socialcom/PASSAT. IEEE*. (pp. 239–248).

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2014). Sensing urban land use with Twitter activity. *International Conference on Social Computing (SocialCom)*, 239–248.

Fujisaka, T., Lee, R., & Sumiya, K. (2010). Exploring urban characteristics using movement history of mass mobile microbloggers. *Proceedings of the Eleventh Workshop on Mobile Computing Systems and Applications. New York, USA*. (pp. 13–18).

Gonzalez, M.C., Hidalgo, C.A., & Barabasi, A.L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, *18*(5),

L'Huillier, G., Ríos, S.A., Alvarez, H., & Aguilera, F. (2010). Topic-based social network analysis for virtual communities of interests in the Dark Web. *ACM SIGKDD Workshop. New York, New York, USA*. 9–9.

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *The Social Mobile Web. Vol. 11*. (pp. p. 02.).

Phithakkitnukoon, S., Smoreda, Z., & Olivier, P. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS One*, *7*(6),

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, *6*(3), 30–38.

Ríos, S.A., & Muñoz, R. (2016). Content patterns in topic-based overlapping communities. *The Scientific World Journal*, *2014*,

Ríos, S.A., Aguilera, F., & Guerrero, L.A. (2009). Virtual communities of practice's purpose evolution analysis using a concept-based mining approach. *Knowledge-Based and Intelligent Information and Engineering Systems. vol. 5712*. (pp. 480–489). Berlin Heidelberg: Springer.

Soto, V., & Frias-Martinez, E. (2011). Robust land use characterization of urban landscapes using cell phone data. *Proceedings of the 1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing.*.

Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, *54*, 255–265. November 2015.

Wakamiya, S., Lee, R., & Sumiya, K. (2011). Urban area characterization based on semantics of crowd activities in Twitter. *Proceedings of the 4th International Conference on Geospatial Semantics, Geos'11*. (pp. 108–123). Berlin, Heidelberg: Springer-Verlag.

Xu, K., Cui, W., Tie, J., & Zhang, X. (2012). An algorithm for detecting groups in mobile social network. *Journal of Networks*, *7*(10), 1584–1591.