Computers, Environment and Urban Systems 61 (2016) 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 xxxx

*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.

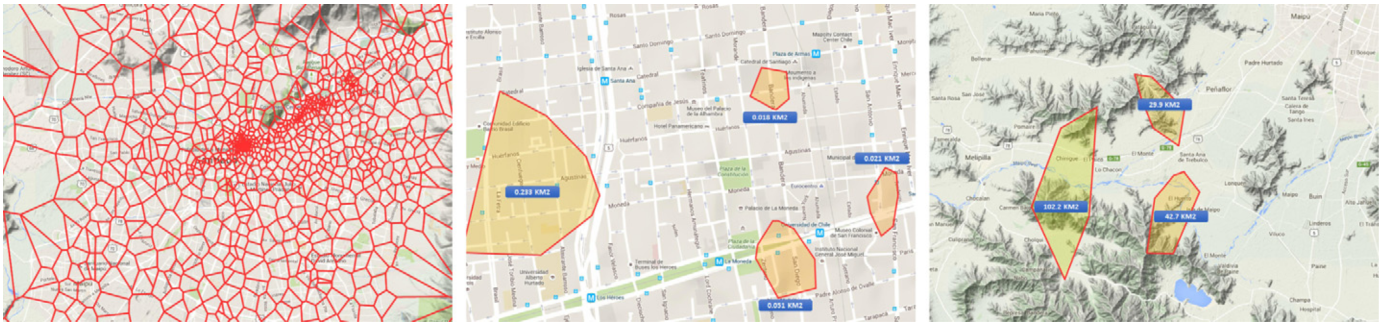
This vast amount of data, generated every second, has been used for social network analysis (Baruah & Angelov, 2012; Catanese, Ferrara, & Fiumara, 2013; Xu, Cui, Tie, & Zhang, 2012), urban dynamics (Calabrese, Colonna, Lovisolo, Parata, & Ratti, 2011) and the understanding of customer behavior (Dragana & 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/© 2016 Elsevier Ltd. All rights reserved.



(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

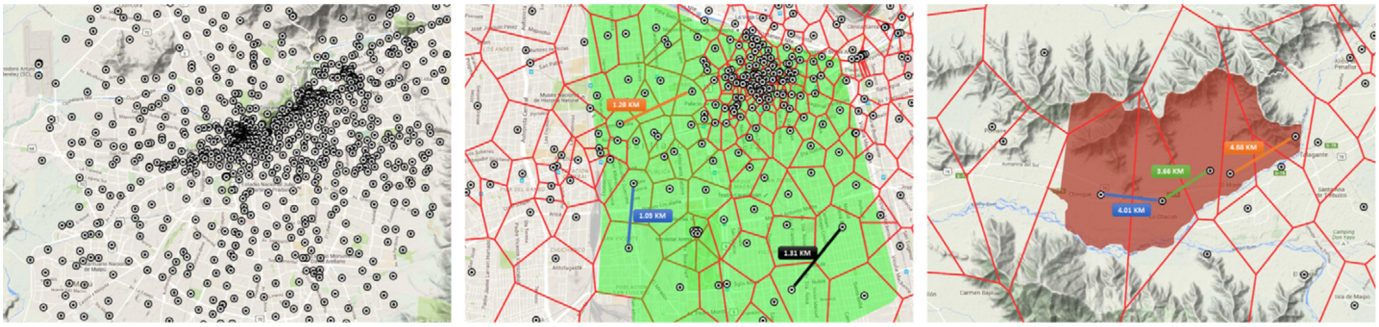
Much research has been done on characterizing patterns in urban areas using social crowd-based resources like geo-tagged tweets or cell phone records. Fujisaka et al. (2010) discovered regional characteristic patterns from movement histories using aggregation and dispersion models in order to understand 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.