LDA Mapping of Regional Socioeconomic Status

Lingzi Hong
College of Information Studies
University of Maryland
lzhong@umd.edu

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
enrique.friasmartinez@telefonica.com

Vanessa Frias-Martinez
College of Information Studies
University of Maryland
vfrias@umd.edu

Abstract

The socioeconomic status of a region provides an understanding of the access of its citizens to basic services. While socieconomic maps are key for policy makers, its compilation requires extensive resources and becomes highly expensive. As a result, traditional methods are now using pervasive datasets (such as cell phone traces for example) to infer regional socio-economic characteristics in a cost efficient manner. In this paper we use mobility information derived from cell phone records to identify socioeconomic levels by using a Latent Dirichlet Allocation that extracts recurring patterns of co-occurring behaviors across regions.

Introduction

Socio-economic maps contain information that characterizes various social and economic aspects like the educational level of the citizens or the access to electricity. The accuracy of these maps is critical given that many policy decisions made by governments and international organizations are based upon such information. National Statistical Institutes (NSIs) compute these maps every five to ten years, and typically require a large number of enumerators that carry out interviews gathering information pertaining the main socio-economic characteristics of each household. All these prerequisites make the computation highly expensive, especially for budget-constraint emerging economies. The ubiquitous presence of cell phones worldwide is generating datasets of spatio-temporal data across large groups of individuals. As previous research has shown, cell phone data can offer a detailed picture of how humans move and interact with each other (Becker et al. 2013). Recent results found that cell phone-based behavioral patterns might be correlated to specific socio-economic characteristics (Eagle, Macy, and Claxton 2010; Soto et al. 2011; Frias-Martinez et al. 2013; 2012). For example, higher socio-economic levels have been associated to stronger social networks or longer distances traveled (Blumenstock and Eagle 2010). Framing the problem as a supervised learning setting, these approaches use the spatio-temporal data to compute a set of pre-determined behavioral features per region and attempt to predict the regional socio-economic levels manually collected by the NSIs. Rather than pre-determined features, regions might be better characterized by probabilistic models of latent behaviors not obvious through observation. Such latent recurring patterns might best reflect the complex nature of human behaviors and the impact that geography and time might have on that behavior. We propose a novel approach to the problem of inferring regional socio-economic levels from spatio-temporal data by using a topic modeling framework based on Latent Dirichlet Allocation (LDA).

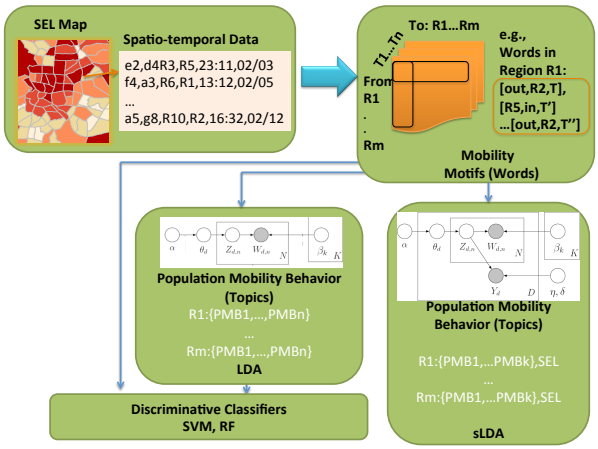

Figure 1: General Approach. sLDA and LDA plate notation from (Mcauliffe and Blei 2008).

Learning Architecture

Figure 1 shows the general approach proposed in this paper. We start with a socio-economic map where each region is represented by a pair $(SEL, Spatio-temporal;;data)$. Although the SEL is a continuous variable, it is often times expressed as a discrete value through a letter (A, B, C, etc.). On the other hand, the spatio-temporal data of a given region contains all calls in that region for a given period of time. Since this information is collected by telecommunication companies for billing purposes, it contains behavioral information about millions of users. As a result, and as shown in previous research (Vieira et al. 2010), the mobility patterns extracted from such data can be representative of the regional population at large, and thus of the underlying socio-economic level.

In this paper, we use the spatio-temporal data as a proxy of the mobility across regions, which will in turn be used as a predictor of regional socio-economic levels. With this approach in mind, we organize the spatio-temporal data into geographic data structures amenable to represent mobility as words. We define the *mobility motifs* as individual transitions containing origin and destination regions together with the time range at which that event happens. As a result, each region will have a set of such motifs every time a transition is observed for the time period under study.

With these regions and their motifs, we propose three approaches: (1) the use of supervised Latent Dirichlet Allocation to do both latent population mobility behaviors (PMB) extraction and SEL prediction over the mobility motifs, ($PMBSEL - sLDA$) (Mcauliffe and Blei 2008); (2) use unsupervised LDA to reveal the population mobility behavior (topic) proportions across regions which are in turn used as input to discriminative regression and classification algorithms to predict the SELs, *PMB-LDA*; and (3) Predetermined Features (PF), which represents each region as a vector of pre-determined mobility motifs where each component shows the number of times a given motif happens. By comparing the accuracy of the three approaches, PMBSEL-sLDA, PMB-LDA and PF, we expect to quantify the impact of using latent topics to predict socio-economic levels from spatio-temporal data.

Mobility Motifs and LDA approaches

The proposed approach uses large-scale spatio-temporal data collected from cell phones to model mobility. Each record collected is of the type $(i, j, R_i, R_j, T, D$ where i, j are encrypted phone numbers, R_z the regions where the individuals were when the phone call was made and T, D are time and date of the call. We use such records to compute the *mobility motifs* of a region R_i as the set of individual continuous transitions that depart or reach that region for a given period of time.

Given two call records from the same individual i, we can build a transition as follows. If i was in region R_i and called individual j in region R_j at time T and date D *i.e.,* (i, j, R_i, R_j, T, D) and next the individual i moved to region R'_i and called to k in region R'_k at time T' and date $D' \geq D$ *i.e.,* $(i, k, R'_i, R'_k, T', D')$ we can extract a mobility motif for region R_i as the tuple $mm = (out, R'_i, T)$ meaning that we observe an individual outgoing transition from region R_i to region R'_i at time T; and a mobility motif for region R'_i as the tuple $mm = (in, R_i, T')$ meaning that we observe an individual incoming transition from region R_i to region R'_i at time T'. The average time between visited regions is $3.2h$, thus, we discretize the time into six four-hour ranges *i.e.,* $T \in \{[0-4), [4-8), ...[20, 24)\}$. Repeating the process for all transitions observed, we can build a collection of regions (documents) each containing a specific set of mobility motifs modeled from the observed calling data as $R_i = \bigcup_{j=1 \in 1...6*R^2} (out, R_j, T) \vee (R_j, in, T)$ where $6 * R^2$ is the size of the vocabulary accounting for all possible bidirectional transitions between any two given regions in the area under study at any four-hour time range.

PMBSEL-sLDA

In this approach, we assume that the mobility motifs in each region arise from a set of latent topics or population mobility behaviors (PMB) at large scale *i.e.,* a set of unknown distributions over the mobility motifs. The set of PMBs is common to all regions, but each region will have a different combination of them.

We propose to use a supervised latent Dirichlet allocation (sLDA) in such a way that the generative process also includes the socio-economic label for each region as part of the model (Mcauliffe and Blei 2008). As a result, the inference is based on model estimates that take into account the socio-economic labels i.e., the empirical PMB frequencies put together and non-exchangeably both mobility motifs and SELs.

PMB-LDA

This second approach focuses on using topic modeling to extract the population mobility behaviors (PMB) in an unsupervised manner *i.e.,* topics are identified with the regions treated as unlabelled. Next, we use the PMBs as features to predict SELs either as continuous values or as classes. In this scenario, the LDA is used for dimensionality reduction.

Pre-determined Features

In this approach, each region R_i is represented by a vector containing all possible mobility motifs as features. We refer to the features as pre-determined because they are defined from behavioral hypothesis about human behavior and socio-economic levels rather than from latent topics directly extracted from the features which add the possibility of finding more complex behaviors. Instead of using population mobility behaviors extracted with LDA, here the regions have hard-coded all possible mobility motifs.

Results

To evaluate the accuracy of the approaches proposed, we use two datasets: a large-scale spatio-temporal dataset containing one month of calling activity for three cities from the same country, and the socio-economic map for those three cities containing regional SEL information. The spatio-temporal dataset contains a total of $134M$ calls and $1.8M$ individuals; while the SEL map contains a total of 186 regions distributed across the three cities.

SEL Inference

We first compute the mobility motifs from the spatio-temporal dataset. We obtain a total of $\approx 4.4M$ mobility motifs across all 186 regions, with an average of $\approx 24K$ motifs per region ($\sigma = \approx 32K$).

Table 1 shows the results for the four approaches using regression to infer SELs as continuous values. The results reported are for 20 topics for PMB-LDA (RF and SVR) and 25 for PMBSEL-sLDA, which turned out to be the number of topics that had the best results in terms of accuracy (R^2) as shown in Figure 2. For Support Vector Regression, we used a Gaussian RBF kernel and the parameters (C, γ, ϵ) were selected using 5-fold cross validation to minimize the

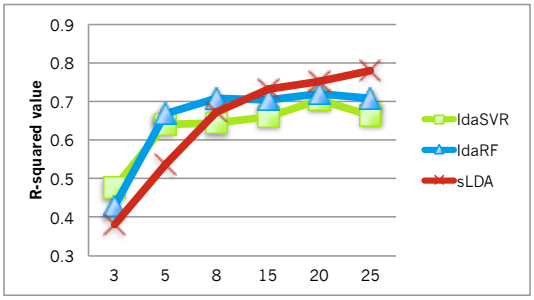

Figure 2: R^2 per number of topics for PMB-LDA (SVR and RF) and PMBSEL-sLDA approaches.

REGRESSION		R^2	**RMSE**
PMBSEL-sLDA		**0.7802**	0.0902
PMB-LDA	SVR	0.7050	0.1088
	RF	**0.7188**	0.1058
PF	SVR	0.2573	0.1731
	RF	**0.6927**	0.1156
PF2	SVR	0.5721	0.1290
	RF	**0.6288**	0.1195

Table 1: Accuracies for Regression with topic models and pre-determined features.

mean squared error. For Random Forest, the results are reported for 8 random trees in PMB-SEL; 146 trees in PF and 14 trees in $PF2$.

We can observe that both topic model approaches, PMBSEL-sLDA and PMB-LDA, have the best R^2 values together with the lowest error. In the case of PMB-LDA the best results are obtained using RF although SVR gave results with R^2 only $\approx$ 1% worse. As hypothesized, the topic model approaches outperform the pre-determined feature approaches by $\approx$ 9% in the best case. In fact, $R^2 = 0.7802$ for PMBSEL-sLDA while $R^2 = 0.6927$ for PF. These results show that the topic models reveal latent population mobility behaviors (PMB) that appear to characterize SELs (and the complex behaviors associated to them) better than the mobilility motifs in which the PMBs are based on. Comparing both topic model approaches, the supervised approach gave $\approx$ 6% better R^2 than the unsupervised approach combined with RF. A similar result was also reported by (Mcauliffe and Blei 2008) in an experiment inferring movie ratings with sLDA.

Table 2 shows the accuracies and F1 scores for all four approaches when SELs are defined as three discrete classes: A, B and C (from high to low socio-economic level). The results are reported for 25 topics for PMBSEL-sLDA and 15 topics for PMB-LDA, which are the topics that gave the highest F1 scores. For SVM, we used an RBF Gaussian Kernel and for RF the number of trees were 8, 146 and 2 for PMB-LDA, PF and PF2, respectively.

In general, the findings and trends are similar to the ones already discussed for the regression results. Here again, we observe that both topic models appear to improve the average F1 score obtained with the pre-determined features approach by $\approx$ 4% in the best case scenario (PMBSEL-sLDA vs. PF). It seems that LDA-based approaches might be doing a better job at extracting more complex population mobility behaviors than just the mobility motifs. Similarly, the pre-determined mobility motifs approach is slightly better than the simpler features of PF2 when RF are used. Moving on to the per-class F1 scores, we observe that the results across classes are quite balanced, specially when topic models are used. Interestingly, this fact reveals that regions are not simply being classified as the *most frequent class* which would be B in this case.

CLASSIFICATION		**ACC**	**AVG.F1**	F1		
				A	B	C
PMBSEL-sLDA		0.7565	**0.7526**	0.7273	0.7283	0.8023
PMB-LDA	SVM	0.6237	0.6302	0.6609	0.5519	0.6777
	RF	0.7130	**0.7212**	0.7786	0.6572	0.7276
PF	SVM	0.4522	0.4510	0.7856	0.6283	0.7160
	RF	0.7004	**0.7100**	0.7856	0.6283	0.7160
PF2	SVM	0.6200	0.6374	0.7409	0.5586	0.6128
	RF	0.6440	**0.6567**	0.7468	0.5847	0.6387

Table 2: Accuracy (ACC), average F1 and per-class F1 score with topic models and pre-determined features.

References

Becker, R.; Caceres, R.; Hanson, K.; Isaacman, S.; Loh, J.; Martonosi, M.; Rowland, J.; Urbanek, S.; Varshavsky, A.; and Volinsky, C. 2013. Human Mobility Characterization from Cellular Network Data. In *CACM*.

Blumenstock, J., and Eagle, N. 2010. Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda.

Eagle, N.; Macy, M.; and Claxton, R. 2010. Network diversity and econmic development. *Science* 328.

Frias-Martinez, V.; Soto, V.; Virseda, J.; and Frias-Martinez, E. 2012. Computing cost-effective census maps from cell phone traces. *Workshop on Pervasive Urban Applications, PURBA*.

Frias-Martinez, V.; Soguero-Ruiz, C.; Frias-Martinez, E.; and Josephidou, M. 2013. Forecasting socioeconomic trends with cell phone records. *ACM Symposium on Computing for Development*.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *Advances in neural information processing systems*, 121–128.

Soto, V.; Frias-Martinez, V.; Virseda, J.; and Frias-Martinez, E. 2011. Prediction of socioeconomic levels using cell phone records. In *Proceedings of the Int. Conference on User Modeling, Adaption, and Personalization*.

Vieira, M.; Frias-Martinez, E.; Bakalov, P.; Frias-Martinez, V.; and Tsotras, V. 2010. Querying spatio-temporal patterns in mobile phone-call datasets. *International Conference of Mobile Data Management*.