Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Topic Models to Infer Socio-Economic Maps

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

Socio-economic maps contain important information regarding the population of a country. Computing these maps is critical given that policy makers often times make important decisions based upon such information. However, the compilation of socio-economic maps requires extensive resources and becomes highly expensive. On the other hand, the ubiquitous presence of cell phones, is generating large amounts of spatio-temporal data that can reveal human behavioral traits related to specific socio-economic characteristics. Traditional inference approaches have taken advantage of these datasets to infer regional socio-economic characteristics. In this paper, we propose a novel approach whereby topic models are used to infer socio-economic levels from large-scale spatio-temporal data. Instead of using a pre-determined set of features, we use latent Dirichlet Allocation (LDA) to extract latent recurring patterns of co-occurring behaviors across regions, which are then used in the prediction of socio-economic levels. We show that our approach improves state of the art prediction results by $\approx 9\%$.

Introduction

Socio-economic maps gather large amounts of information regarding the status of households at a national scale. These maps contain information that characterizes various social and economic aspects like the educational level of the citizens or the access to electricity. Such information is aggregated and reported at various granularity levels, from a national scale, to states, all the way down to urban geographic areas of a few square kilometers. 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. To reduce costs, countries have made cuts both in the number of interview questions and in the number of citizens interviewed, which unfortunately impacts the quality of the final census information.

On the other hand, 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). For example, higher socio-economic levels have been associated to stronger social networks or longer distances traveled (Blumenstock and Eagle 2010). Furthermore, previous work has shown that regional socio-economic levels can be inferred from cell phone-based behavioral features with acceptable accuracies using various regression and classification techniques over a set of behavioral features (Frias-Martinez et al. 2012). 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. Such features are usually defined based on ad-hoc hypothesis about human behaviors and socio-economic values often neglecting underlying spatio-temporal relationships. 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.

In this paper, we propose a novel approach to the problem of inferring regional socio-economic levels from spatio-temporal data. Specifically, we propose a topic modeling framework based on Latent Dirichlet Allocation (LDA). LDA models are typically used in research involving documents whereby a document is represented as a set of words and distributions are drawn to identify topics across documents in an unsupervised manner. However, they have also been successfully applied in a variety of different scenarios for model estimation and inference including computer vision (Fei-Fei and Perona 2005) or climate modeling (Tand and Monteleoni 2014). We investigate the hypothesis that using individual behaviors (words) collected from cell phone data to identify probabilistic large-scale population behaviors (topics) across regions might allow to improve the cur-

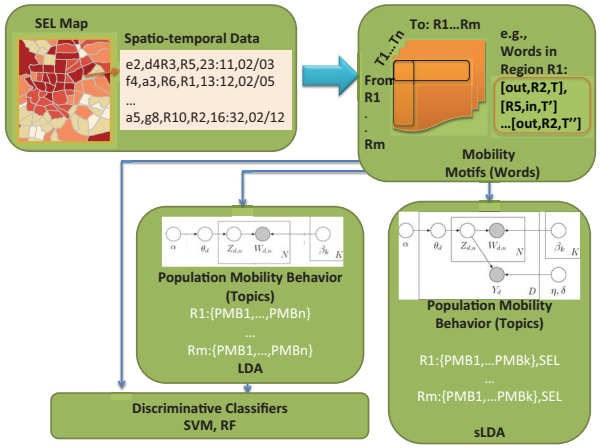

Figure 1: General Approach. sLDA and LDA plate notation from (Mcauliffe and Blei 2008).

rent state of the art in the prediction of socio-economic levels using spatio-temporal data. The rest of the paper is organized as follows: we first present our general approach, followed by our proposed method and results. We finalize with related work and conclusions.

Approach Overview

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 $(SEL, Spatio-temporal\ data)$ where SEL (socio-economic level) is a label manually gathered through surveys. These regions might represent states, cities, neighborhoods or *geographical units (GUs)*, the smallest geographical division, which divides cities into small areas of up to a few square kilometers (approximately blocks). Although the SEL is a continuous variable, it is often times expressed as a discrete value through a letter (A, B, C, etc.). The granularity of the SELs *i.e.,* the number of SEL classes in which the continuous values are divided into, varies a lot across studies. Some researchers differentiate three socio-economic levels (Wyatt and Mattern 2011) while others prefer to use a larger range of values in their analyses (Worrall, Basu, and Hanson 2003).

On the other hand, the spatio-temporal data of a given region contains all individual calls from or to a person in that region for a given period of time. Specifically, each call in the dataset contains information regarding origin and destination encrypted phone numbers, origin and destination regions where caller and callee where when the phone call was made, as well as time and date at which a given individual called another one. Using the information about origin and destination regions, we can build individual mobility patterns that indicate the continuous locations visited by an individual at different times of the day. 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. Previous work has shown that mobility patterns are predictive of socio-economic levels (Smith-Clarke, Mashhadi, and Capra 2014). Here, we explore whether using a generative approach to extract latent, more complex behavioral features which are then used for discriminative prediction can enhance previous state of the art approaches. Specifically, we consider each region as a document which we characterize by a set of individual mobility motifs (words) that have the region as origin or destination. We assume that the mobility motifs arise from a set of latent topics, that is, a set of unknown distributions over the patterns. In this scenario, the topics can be interpreted as *population mobility behaviors* at large scale. For example, commercial regions in a city could have as a popular motif incoming and outgoing trips from/to residential regions during the weekends. The topic would be *weekend shopping* and the words could be the two mobility patterns described.

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 to understand the impact of using latent topics versus pre-defined features in the prediction of regional socio-economic levels. The first approach, proposes 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). In this scenario, both the motifs in the region as well as the socio-economic label (response variable) are used to model the latent population mobility behaviors (topics) and predict the SELs. For the second approach, *PMB-LDA*, we propose to 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. Finally, the last approach, Pre-determined Features (PF), represents each region as a vector of pre-determined mobility motifs where each component shows the number of times a given motif happens. The vectors are then used as input to regression or classification techniques. 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. Next, we explain each approach in detail.

Method

We first explain how we compute the mobility motifs from the spatio-temporal data. Next, we cover each of the pre-

dictive approaches. At this point, it is important to clarify that the predictive approaches we propose will be adapted to work for either continuous or categorical values since the socio-economic levels (SEL) can sometimes be expressed as one or the other. Categorical SELs are typically represented by letters *e.g.,* letter A represents high socio-economic levels, B medium and C lower socio-economic levels where A covers the range $[1 - .66)$, B $[.66 - .33)$ and C $[.33 - 0]$.

Mobility Motifs

The proposed approach uses large-scale spatio-temporal data collected from cell phones to model individual mobility. Specifically, 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. Every time a phone call is made, a record of such type is collected thus building data collections of millions of interactions. 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.

Formally, 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'. Since the mobility motifs are based on calling data, not GPS, we do not have a large number of daily data points (regions visited) per individual. 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 $R_i = \bigcup_{j \in 1...6*R^2} (out, R_j, T) \vee (in, R_j, T)$ 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. For example, one such PMB might be *commuting* and it will be prominent in residential regions where motifs reaching office regions in the morning and leaving them in the afternoon will be common place. 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. Given that SELs can be continuous or categorical values, the algorithm (see 1) regresses the SEL labels on the topic frequencies when the values are continuous. When SELs are expressed as categories, it is considered to be drawn from a softmax regression for classification allowing to do a multi-class sLDA as explained in (Wang, Blei, and Li 2009).

Algorithm 1 PMBSEL-sLDA

```
Draw PMB proportions θ|α ≈ Dir(α).
for each mobility motif do
    Draw PMB assignment z_n|θ ≈ Mult(θ).
    Draw mobility motif w_n|z_n, β_{1:P} ≈ Mult(θ)).
end for
if SEL ∈ ℜ then
    Draw SEL y|z_{1:N}, η, σ² ≈ N(η^T ẑ, σ²).
else
    Draw SEL y|z_{1:N} ≈ softmax(ẑ, η).
end if
```

In the algorithm, P is the number of topics, $\beta_{1:P}$ a vector with probabilities distribution, $\hat{z} = (1/N) \sum_{n=1}^{N} z_n$, and the softmax function provides the distribution $p(c|\hat{z}, \eta) = exp(\eta_c^T \hat{z}) / \sum_{l=1}^{C} exp(\eta_l^T \hat{z}$ to be able to run classification on C classes instead of inferring a continuous response. Parameters $\beta_{1:P}$, η and σ^2 are estimated with maximum-likelihood estimation using a variational EM procedure.

To test the accuracy of this approach, we randomly divide the set of regions in the area under study into a training and a testing set $(75\% - 25\%)$ and repeat it 100 times. We use the training set for PMBSEL-sLDA model estimation and the testing set for the inference of SELs from mobility motifs and topics, and report average accuracy values across all runs. For continuous SEL values, we report R^2 and $RMSE$. For SEL classes, we report precision and recall per class, accuracy and average and per-class F1 score since the accuracy measures tend to be biased towards the majority class if the class distribution is not homogeneous. The F1 score is computed as $F1 = 2 * \frac{precision * recall}{precision + recall}$ where precision is defined as the fraction of correct results with respect to a given class and recall as the fraction of correct results with respect to all the correct results that should have been returned by the classifier for that class.

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 *i.e.,* instead of working with all possible mobility motifs (words) as features, the LDA extracts the distributions of the latent population mobility behaviors for each

region which are in turn used for SEL prediction. The process, as shown in algorithm 2, first generates a set of PMB features over all regions and assigns the mobility motifs to each PMB. Next, for each region R_i, it adds the vector containing its PMB frequencies to a dataset (TS) that will be used for training and testing. The regional SELs are then predicted using the regional PMB frequencies as features. Since mixed generative-discriminative classifiers have been reported to improve accuracy results, we evaluate the performance of Support Vector Machines (SVM) and Random Forest (RF) (Jaakkola and Haussler 1998) (Joachims 1999) (Breiman 2001). Finally, depending on the nature of the socio-economic data, the algorithm executes regressions (SVR,RFR) or classifications (SVM,RF).

Algorithm 2 PMB-LDA

```
Draw PMB proportions θ|α ≈ Dir(α).
for each mobility motif do
    Draw PMB assignment z_n|θ ≈ Mult(θ).
    Draw mobility motif w_n|z_n, β_{1:P} ≈ Mult(θ)).
end for
for each region R_i do
    Extract PMB frequencies R_i = [PMB_1..PMB_P].
    TTS = ⋃_{i=1}^{R} (R_i, SEL)
end for
if SEL ∈ ℜ then
    Train and test with TTS using SVR and RFR
else
    Train and test with TTS using SVM and RF
end if
```

As shown, P is the number of topics and $\beta_{1:P}$ a vector with probabilities distribution. All parameters are estimated with maximum-likelihood estimation using a variational EM procedure. We compute the accuracy by randomly dividing the set of regions in the area under study into a training and a testing set ($75\% - 25\%$) and repeating it 100 times. We use the training set to estimate the LDA model, ignoring the regional SEL labels; and the testing set to infer SELs using SVM or RF over the topic proportions. We report average accuracies across runs for both classifications and regressions.

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. As shown in the algorithm 3, each region is represented as a vector containing for each of the $6 * R^2$ fields the normalized number of times a mobility motif mf_i is observed. Depending on the vector's size, for large R values Laplacian smoothing (Field 1988) is applied to account for cases that, although not observed in the dataset, might have happened but not been collected. To be able to compare results against topic modeling approaches (which exert some type of loosely defined feature selection), we also apply feature selection techniques (mRMR) on the hard-coded elements of the vector (Ding and Peng 2005; Chang and Lin 2001). Then, each regional vector –after feature selection is applied– is paired up with its SEL label and all vectors are used as input to a discriminative classifier. As in previous approaches, both classification and regression are considered depending on the nature of the SELs. The accuracy for this approach is computed by randomly dividing the set of regions into a training and a testing set ($75\% - 25\%$) and repeating it 100 times. We use the training set to estimate the SVM/RF model; and the testing set to infer SELs. We report average accuracies across runs for both classifications and regressions.

The mobility motifs presented in this paper were designed such that they would be amenable to topic models. However, related work has used different types of pre-determined mobility-related features to predict socio-economic levels. Most of them tend to be continuous variables modeling factors such as radius of gyration, distances traveled, etc. For completeness, and in an attempt to compare pre-determined feature approaches, we use 26 different mobility features described in recent related literature (Blumenstock and Eagle 2010; Smith-Clarke, Mashhadi, and Capra 2014) to characterize each region and we run algorithm 3 over them. Specifically, we consider the following features: number of incoming and outgoing transitions, average traveled distance for trips to or from a given region, radius of gyration, population and number of visitors to/from a region and entropy considering separate weekday and weekend values as well as several ratios among them. We will refer to this approach as $PF2$ and discuss results in the next section.

Algorithm 3 PF

```
for each region R_i do
    for each mobility motif in region R_i do
        Update R_i = [mf_1, ..., mf_{6*R^2}]
    end for
    Apply smoothing technique to R_i if necessary
    TTS = ⋃_{i=1}^{R} (R_i, SEL)
end for
if SEL ∈ ℜ then
    FS, Train, test with TTS using SVR and RFR
else
    FS, Train and test with TTS using SVM and RF
end if
```

Results

Dataset

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. For privacy reasons involving the use of cell phone data at large-scale, we

cannot reveal the name of the country. 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. Each region has a SEL value expressed as a real number. For completeness, we also report results for discrete SELs. For that purpose, we discretize SEL values into a range of two to six different categories and report results for the best approach which was three *i.e.,* level A, or high socio-economic level, covers the range $[1 - .66)$, level B $[.66 - .33)$ and level C represents the lower socio-economic level $[.33 - 0]$. The distribution of regions across classes was 48 regions with level A, 71 regions with level B and 67 regions with level C.

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 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. Finally, when comparing pre-determined feature approaches we can see that the mobility motifs (PF) outperform the set of mobility variables in $PF2$ by $\approx 7\%$. Incidentally, the words defined to carry out the experiment, which model events co-occurring in space and time (connected regions, direction of the flow and time stamp) appear to represent the human behavior associated to socio-economic levels better than simpler mobility measures recently used in the literature. Somewhat similar findings were reported in (Yuan, Zheng, and Xie 2012) with respect to identifying urban land uses from GPS-based human behavior, although the task was fully unsupervised.

On the other hand, Table 2 shows the accuracies and F1

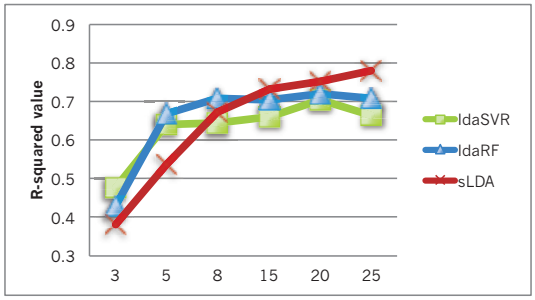

Figure 2: R^2 per number of topics for PMB-LDA (SVR and RF) and PMBSEL-sLDA approaches.

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.

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. Finally, the discretization of the SEL values was done by simply dividing the values into three equal-length ranges. However, future work will explore more sensitive discretizations with respect to socio-economic levels such as computing range widths based on density estimations of SELs which might model better how poverty is distributed (Schmidberger and Frank 2005).

Exploring SEL-based Topics

The mobility motifs we have defined are not as intuitive as natural words and as a result, it makes it challenging to show

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.

what motifs are associated to different topics so as to carry out qualitative explorations of the type natural language processing researchers do. For that reason, we propose a qualitative evaluation of the topics and words based on socio-economic levels since these are the most relevant variables for policy makers. For each PMB (topic), we take its mobility motifs (regardless of their frequency), extract the socio-economic category of the regions involved in those words, and plot a colored circle for each observed category in each region. Figure 3 shows an example for one of the topics detected in one of the three cities under study. Each region is color-coded according to its SEL: A is orange, B dark yellow and C light yellow; and the circles follow the same color code. Interestingly, we observe that most of the regions only have mobility motifs to/from regions with the same socio-economic level (see #1 in the Figure) thus showing that people from similar SELs tend to cluster together and revealing potential urban pockets based on SELs. There exist a few mid- and high-SEL regions that show transitions between them, probably because they share borders (see #2) although, as can be seen, there exist other regions that share borders without sharing flows of different SELs (see #3). Additionally, there exists a unique region that has flows to/from all socio-economic levels (see #4 in Figure) which contains a very important touristic attraction that probably attracts frequent trips from all regions. We also observe three rare cases of *sink regions* where flows to/from the region are only from a different SEL than its own (see #5). One could argue that these findings simply reflect the temporal limitations that individuals have to reach different regions and that in general they visit the ones that are nearby. However, recall that the mobility motifs are computed for $4 - hour$ periods (avg. transition length of $3.2h$) which is sufficient time to reach any region of the city from any other. Thus, it appears that individuals choose to focus on specific SEL-influenced flows among all the possibilities. By repeating this analysis for other topics, and adding word frequencies or temporal ranges, we might be able to extract urban dynamics that could inform policy makers about inequalities or abnormalities that should be solved to improve the cities we live in. The topic modeling framework proposed in this paper is amenable to that type of analysis.

Related Work

(Smith-Clarke, Mashhadi, and Capra 2014) used cell phone data from two countries to extract a set of mobility and calling features including call volumes, regional distances or

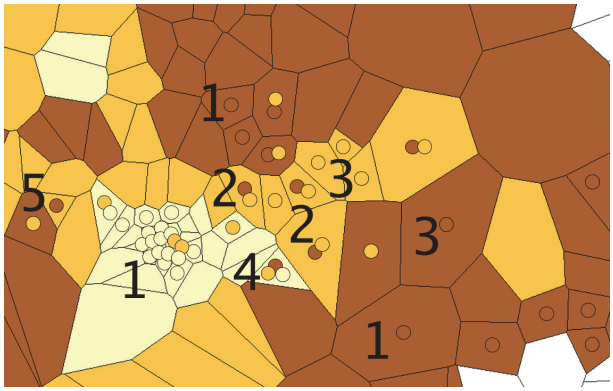

Figure 3: SEL-based mobility motifs. Darker colors represent higher SELs. Circles represent to/from transitions.

entropy. Such features were then used to analyze various types of correlations as well as to run a simple linear regression to approximate poverty information from cell phone data. Other papers have presented similar approaches using other machine learning techniques including SVMs, RF or EM clustering to predict socio-economic maps (Rubio et al. 2010; Eagle, Macy, and Claxton 2010; Blumenstock and Eagle 2010). Other works not specifically focused on SEL prediction but on potential proxies include (Quercia et al. 2012; Quercia, Seaghdha, and Crowcroft 2012) who found that deprivation values were related to pre-determined Twitter features such as conversations or sentiment; or (Doll, Muller, and Elvidge 2000; Ebener et al. 2005; Elvidge et al. 1997) who showed that a country's GDP was correlated to night time light (NTL) measured from satellite imagery. Topic models have been used outside natural language processing in various capacities (Barnard et al. 2003; Blei and Jordan 2003; Bosch, Zisserman, and Munoz 2006; Duygulu et al. 2002). In this paper, we have explored the use of topic models to enhance the existing approaches to SEL prediction.

Conclusions

Computing socio-economic maps is critical for countries. However, its compilation requires extensive resources and becomes highly expensive. Traditional approaches have used prediction techniques based on a set of pre-determined features computed from spatio-temporal data. In this paper, we have presented a novel approach that uses topic modeling techniques to extract a set of latent features (the population mobility behaviors) that are used to infer socio-economic regional labels. We have shown that our approach improves state of the art techniques by $\approx 9\%$. Future work will explore the use of social media sources to explore improvements in socio-economic level predictions through multimodal systems.

References

Barnard, K.; Duygulu, P.; de Freitas, F.; Forsyth, D.; Blei, D.; and Jordan, M. 2003. Matching words and pictures.

Journal of Machine Learning 3:1107–1135.

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*.

Blei, D., and Jordan, I. 2003. Modeling annotated data. *SIGIR*.

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

Bosch, A.; Zisserman, A.; and Munoz, X. 2006. Scene classification via plsa. *European Conference on Computer Vision*.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Chang, C., and Lin, C. 2001. Libsvm: a library support for vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Ding, C., and Peng, H. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* 3(2):185–206.

Doll, C.; Muller, J.; and Elvidge, C. 2000. Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. *AMBIO:Journal of the Human Environment* 29(3):157–162.

Duygulu, P.; Barnard, J.; de Freitas, F.; and Forsyth, D. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *European Conference on Computer Vision*.

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

Ebener, S.; Murray, C.; Tandon, A.; and Elvidge, C. 2005. From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. *International Journal of Health Geographics* 4(1).

Elvidge, C.; Baugh, K.; Kihn, E.; and Kroehl, H. 1997. Mapping city lights with night time data from the dmsp operational linescan system. *Photogrammetric Engineering and Remote Sensing* 63(6):727–734.

Fei-Fei, L., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, 524–531.

Field, D. A. 1988. Laplacian smoothing and delaunay triangulations. *Communications in applied numerical methods* 4(6):709–712.

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*.

Jaakkola, T., and Haussler, D. 1998. Exploiting generative models in discriminative classifiers. In *In Advances in Neural Information Processing Systems 11*.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*.

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

Quercia, D.; Ellis, J.; Capra, L.; and Crowcroft, J. 2012. Tracking gross community happiness from tweets. *In Proceedings of the ACM Computer-Supported Cooperative Work and Social Computing*.

Quercia, D.; Seaghdha, D.; and Crowcroft, J. 2012. Talk of the city: Our tweets, our community happiness. *In Proceedings of the AAAI International Conference on Web and Social Media*.

Rubio, A.; Frias-Martinez, V.; E., F.-M.; and Oliver, N. 2010. Human mobility in advanced and developing economies: A comparative analysis. *AAAI Spring Symposium: Artificial Intelligence for Development*.

Schmidberger, G., and Frank, E. 2005. Unsupervised discretization using tree-based density estimation. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD'05.

Smith-Clarke, C.; Mashhadi, A.; and Capra, L. 2014. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *CHI*.

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*.

Tand, C., and Monteleoni, C. 2014. Detecting extreme events from climate time series via topic modeling. In *International Workshop on Climate Informatics*.

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*.

Wang, C.; Blei, D.; and Li, F.-F. 2009. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1903–1910. IEEE.

Worrall, E.; Basu, S.; and Hanson, K. 2003. The relationship between socio-economic status and malaria: a review of the literature. *Health Economics for Developing Countries*.

Wyatt, J., and Mattern, K. 2011. Low-ses students and college outcomes: the role of ap fee reductions. *College Board: AP Data and Records*.

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*.