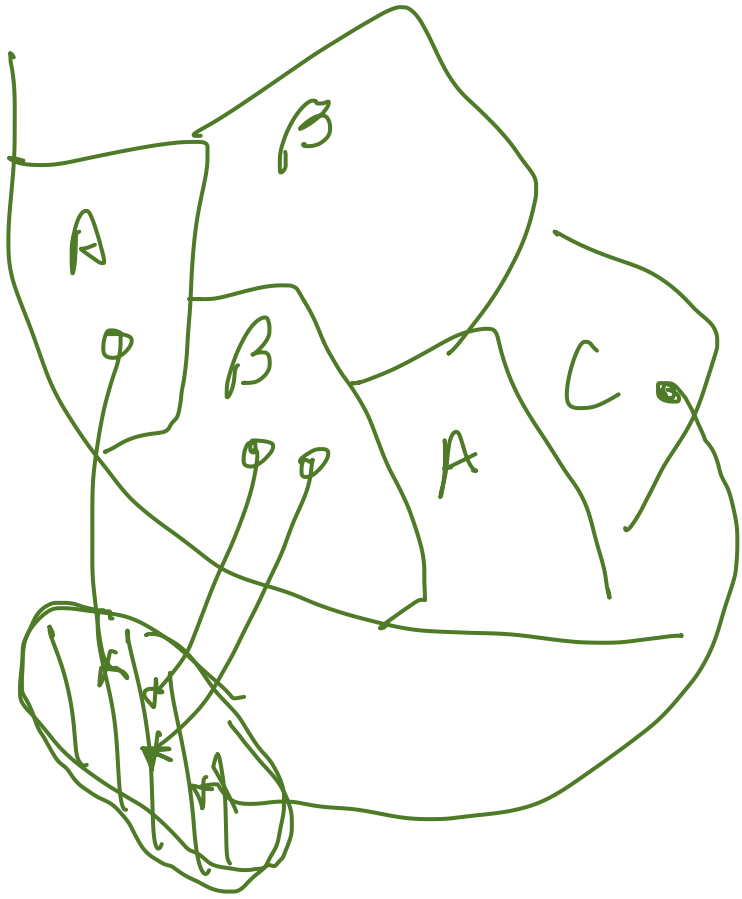
Project: A - Characterization of Urban Areas Week: W06

Team: DINIS CAQUALHO ✓
JOÃO TEIXEIRA ✓

Progress (0..5):

This week:

- CENSUS
- POIS
- CAOS



Next Tasks:

CENSUS
+
POIS
+
etc

CLAUDIA
... ideias de como classificar zonas (de grelha)

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Human activity recognition from spatial data sources

Zolzaya Dashdorj[1,2,3], Stanislav Sobolevsky[2], Luciano Serafini[3], and Carlo Ratti[2]

[1]University of Trento and Telecom Italia, Italy Via Sommarive, 9 Povo, TN, Italy

{*dashdorj*}*@disi.unitn.it*

[2]Massachusetts Institute of Technology, MIT 77 Massachusetts Avenue Cambridge, MA, USA

{*stanly,ratti*}*@mit.edu*

[3]Fondazione Bruno Kessler, Via Sommarive 18 Povo, TN, Italy

{*serafini*}*@fbk.eu*

ABSTRACT

Recent availability of big data of digital traces of human activity boosted research on human behavior. However, in most of the datasets such as mobile phone data or GPS traces, an important layer of information is typically missing: providing an extensive information of when and where people go typically does not allow understanding of what they do there. Predicting the context of human behavior in such cases where such information is not directly available from the data is a complex task that addresses context recognition problems. To fill in the contextual information for such data, we developed an ontological and stochastic model (HRBModel) that interprets semantic (high-level) human behaviors from geographical maps like OpenStreetMap, analyzing the distribution of Points of Interest(POIs), in a given region and time period. The semantic human behaviors are human activities that are accompanied by their likelihood, depending on their location and time. In this paper, we perform an extended evaluation of this model based on other qualitative data source, namely a country-wide anonymized bank card transaction data in Spain, which contains contextual information about the locations and the types of business categories where transactions occurred. This allows us to validate the model, by matching our predicted activity patterns with the actually observed ones, so that it can be later applied to the cases where such information is unavailable. This extended evaluation aimed to define the applicability of the predictive model, HRBModel, taking various type of spatial and temporal factors into account.

Ground-thruth data

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; H.1 [**Models and Principles**]: Information theory, Human factors

SIGSPATIAL'14, November 04-07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2675316.2675321

General Terms

Algorithms, Human Factors

Keywords

Geo-spatial data and knowledge, Urban and Environmental Planning, Spatial Data Quality and Uncertainty, Statistical matching, Human activity recognition, Context recognition, bank card transactions, big data

1. INTRODUCTION

The availability of massive mobile phone Call Data Records (CDR) and other data created by different aspects of human activity has opened vast opportunities for the analysis and better understanding of real-life social phenomena and human dynamics providing valuable insights for interested parties. Such analysis includes discovering patterns of human mobility [7, 5, 17, 1], communication [19, 13] and urban activities [6, 10] through cell-phone usage at the regional scale and bank card transactions [18]. An important addition to the general analysis based on raw CDR data or any other data of human activity over time and space might be obtained from studying the contextual information about the circumstances under which the activity happened. Such contextual information might include weather conditions, location features (e.g., points of interest), public and private events (e.g., concerts, sports matches, etc) and emergency events (e.g., accidents, strikes etc), providing a unique platform to better understand heterogeneous human behaviors including their potential motifs.

Our objective is to use all of this informatio

In particular, semantic meaning of places where such events happen is an emerging topic for understanding spatial dependency of events on human activities. Many different approaches are studied to interpret the semantics of places: collection of user survey data, classification of data or analysis of the geographical area features [3, 20, 10]. For example, Krumm et al [8] used diary surveys for predicting semantic places using a multiple classifier based on a forest of boosted decision trees taking individuals' demographics, times they visited the places and nearby businesses into account. On top of this approach, the learning process such as a conventional decision tree followed by re-learning misclassification improves the prediction accuracy. Ye et al [22]

used a binary support vector machine (SVM) classifier like multi-label classification for learning approach that annotates all places with category tags.

Geographical area features provide a more meaningful way to understand the semantics of places [23, 15, 14], as well as of the dependency of human activities and urban city structure. Phithakkitnukoon et al [11] used geographical area features like POIs from Yahoo Maps and analyzed the correlation of geographic areas and human activity patterns (i.e., sequence of daily activities) by annotating the points of interest to the human activities like eating, recreation, shopping and entertainment, and then the Bayes theorem is used to classify the area into crisp distribution map of activities. Identifying the mobility choices of users for daily activity patterns, the study shows that the people who have the same work profiles have strongly similar daily activity patterns. But this similarity is reduced if the distance between people's work profile location are increased. However, spatial data distribution is very uncertain and heterogeneous and extracting meaningful information from it is a very challenging task. The data can be imprecise, noisy, and not available in some areas. The heterogeneity of the data directly influences the confidence of how the data characterized. Sengstock et el [16] demonstrated the functionality of a probabilistic model developed for extracting spatio-temporal data about semantic information of real-world phenomena from social media based on Bayesian Networks. Twitter data is used to validate the model comparing the probabilistic distributions generated from those two spatial data-sources.

In this paper, we consider a validation of the model developed by Dashdorj et al in [4] that extracts human activities from geographical map features, using another type of data source providing precise information about customer actual activity, such as bank card information. The model, High level Representation of Behavioral Model (HRBModel) extracts a wide range of human activities (e.g., working, attending a concert, shopping, etc) typically associated with the considered location and calculates the probabilities of such activities to be performed in the area during different times of the day/week by analyzing the importance of POIs in a given location or nearby location of people. The model is aimed at providing qualitative information about human activity depending on geographical location for giving a great impact on human behavior classification task from raw data set.

validation

Therefore, we estimate the matching error percentage between HRBModel and bank card transaction data in the areas with different land-use categories in a city. We further compare the applicability of the model in different cities(first focusing on major cities, such as Barcelona and Madrid). Such a validation would justify the model applicability for predicting the context of human behavior in the cases when the context information is not directly available from the data (e.g., raw CDR or taxi data).

The validation of the model is very important to measure the noise,scalability and accuracy of spatial data-sources, so that a bank card transaction data can provide enough business activities for a macro level analysis like Street centrality and the Location of Economic Activities by Sergio et al [12].

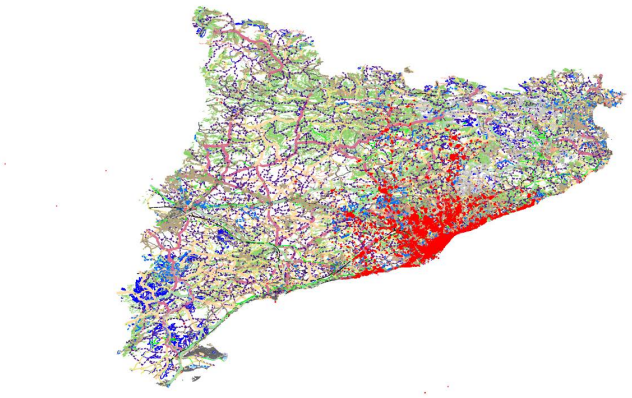


Figure 1: Points of Sale distribution in Barcelona Province

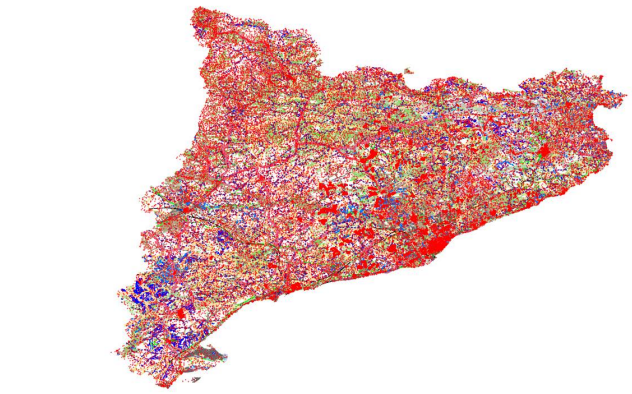


Figure 2: Points of Interest distribution in Barcelona Province

The remainder of the paper is structured as follows. In the following section, we introduce the used data sources. In Section 4, we describe the basic concepts of our problem statement and define our proposed methodology for the extended validation. We present experiments and discuss the results and inference in Section 5. We conclude the discussion in Section 6.

2. PROBLEM STATEMENT

Human activity recognition from big data is an emerging research area in smart environment for understanding a pulse of city, city dynamic as well as human dynamics. There is a need to understand activity of interest from geographical objects distribution in urban and rural areas under normal situation. The surrounding geographic object distribution - shop, hospital and traffic road, etc - in an area is a key factor for identifying human activity of interests. Dashdorj et al in [4] proposed a model, High-level Representation Behavioral Model (HRBModel) which uses an ontology for analyzing geographical objects collected in OpenStreetMap to interpret human activity of interests. Through an user survey, the model is evaluated and the result shows that human activity of interest can be identified from the geographical object

distribution. However from the user survey, the actual activities are predicted 70.89% and the high level activities are predicted 80.2%. But the top-5 activities are predicted 62%. In this paper, we extend this evaluation using other spatial data source; bank card transaction data, which contains contextual information about the locations and the types of business categories where transactions occurred. So we validate the HRBModel by matching their predictive activity patterns with the actually observed ones (from bank card transaction data).

3. DATASET SOURCES

We used a dataset of 1.6 million bank card transaction provided for research purposes by one of the largest Spanish banks. The data was completely anonymized on the bank's side preventing access to any personal information of the customers in accordance with all applicable privacy protection regulations. Each transaction is labeled with one of the 76 business activity categories, such as groceries, fashion, bars, restaurants and sport shops etc,. We also used OpenStreetMap [1] for the extraction of POIs and collected around 1.7 million POIs in Barcelona city.

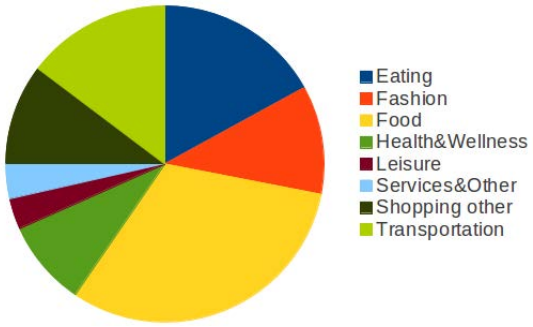


Figure 3: Business activity distribution from bank card transaction data in Barcelona City

In order to match the bank card transaction data with HRBModel, firstly, we need to transform the semantic taxonomies of those two datasource into same categorical taxonomies. To avoid a semantic gap between these two data categories, taxonomies of business categories (Figure 3) are manually mapped with the possible taxonomies of POI. We used the same methodology to relate human activities to the POIs as described in the ontology of HRBModel in [4]. In the ontology, we described 8 types of business activities. In Table 1 they are grouped in terms of the similarity.

4. METHODOLOGY

Here, we describe a methodology for an extended validation task for human activity extraction from HRBModel if the model is applicable to provide contextual information to raw datasets like CDR, and taxi data or any other data containing customer mobility in space and time for extracting human behaviors. In order to compare the model predicted activities to the actually observed type of customer activity, we use bank card transaction data source concentrating on the business activities as described in Section 3.

[1]http://www.openstreetmap.org

We present a general model for comparing the probabilistic business activity distributions generated from both datasources: 1) HRBModel based on the geographical map datasource 2) bank card transaction data. Let $P_{POI}(L)$ be the distribution of business activity in a given location L extracted from the HRBModel and let $P_{POS}(L)$ be the distribution extracted from the bank card transaction data. L is a discrete random variable representing geographical location point. Each record is an individual location randomly selected for extracting some spatio-temporal and semantic knowledge. The value domain of L is a set of random locations in our data set, i.e., $val(L) = l_1, ..., l_{50}$.

In order to present the results of the comparison for a region (Barcelona province) we take locations L randomly within a city shape or randomly among the POI locations to reflect that we are more interested in the dense areas of the city, while distribution of POI locations is a good proxy for density of business activity. The coverage radius for a given location is demonstrated in random radiuses, R = [300,..,1000] each increased by 50m increments.

For each location L (just an arbitrary point within a city), we define a probability of each action type based on HRBModel by analyzing the POI distribution in the coverage area of that location:

$$\mathrm{P_{POI}}(a|L) = \frac{\sum_{type(POI)=a} N(a, U(L))}{\sum N(U(L))} \quad (1)$$

where N(a,U(L)) denotes the number of POI of type "a" in the area U, while U(L) is the coverage area of location (radius 300m) that can be increased up to 1,000m by 50m increments. The activities are generated using the methodology described in [4].

We estimate the probability of making an action of each type in bank card transaction data for the location L:

$$\mathrm{P_{POS}}(a|L) = \frac{\sum_{pos \in U(L), type(pos)=a} W(pos)}{\sum_{pos \in U(L)} W(pos)} \quad (2)$$

where W(pos) denotes the number of transactions in the points of sale in the area U.

In order to quantify a deviation (error) between them we can use the following formulae, for instance, the sum of absolute distance defines the total distance of the common probabilities in both distributions based on transactions and POIs.

Then, what we do is simply compare distributions of values of $P_{POI}(a|L)$ and $P_{POS}(a|L)$ for each location L to see if the match is rather good or not, thus defining evaluating model applicability this way. There are different possibilities [21, 2] to quantify the degree of similarity/dissimilarity for two different activity distributions in a given location. The first and simplest one is to compute, in the two data sources involved, Jaccard similarity coefficient, which ranges between 0 and 1 and measures common variables in two sets, with the size of intersection divided by the union of variables in two data sources:

Table 1: Business categories

Category	Type of businesses
Eating	Bar/Restaurants
Fashion	Apparel/Accessories
Health/ Wellness	Hospitals, Optician's shop, Pharmacy
Food	Food, Hypermarkets, Supermarkets
Leisure	Bookstore, Sports, Entertainment
Shopping other	Tech, toy shops
Transportation	Travel, Tourism
Services	Service shops

$$Jaccard(L) = \frac{\sum_a (P_{POS}(a|L) \cap P_{POI}(a|L))}{\sum_a (P_{POS}(a|L) \cup P_{POI}(a|L))} \quad (3)$$

where, a frequency of variable is not taken into account. This coefficient is mostly used to measure the accuracy for common matching variables in two data-sources, with no missing data, errors and a high level of quality.

Another way is to compute the weighted frequency distributions for each variable of interest and to calculate the differences. The maximum value of these differences can be taken as a criterion for comparison. Often appearing in literature, a threshold for maximum difference is a simple measure for the coherence of the variable without much theoretical background.

However, a relative measure of differences in the distributions of various common variables at different levels measures distortion of distributions. A Hellinger distance [9] quantifies the similarity between two probability distributions and lies between 0 and 1. It is easy to interpret as the Euclidean norm but it does not take a sampling design into account. We consider two discrete probability distributions $P_{POI} = (p_1 \dots p_k)$ and $P_{POS} = (q_1 \dots q_k)$. Their Hellinger distance is defined as

$$HD(P_{POI}, P_{POS}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}, \quad (4)$$

where the two distributions indicate a perfect similarity if the Hellinger distance is 0. Otherwise any deviation creates a positive error, which is 1 if the estimations are completely different (like 100% dining from transactions and 100% fashion from POIs).

Many techniques use different sampling designs, such as Chi-square, Kolmogorov Smirnov, Rao Scott, etc.

5. EXPERIMENT

We now present experiments conducted with different models, which measure the accuracy, match error between the two activity probability distributions for HRBModel and bank card transaction data-source, depending on land-use classifications given by OpenStreetMap land-use classification features, as described in Table 2. We selected 50 locations randomly for each land-use classification and then,

Table 2: Land-use classifications

No	Land-use	Usage
1	Industrial	factories, warehouses, workshops
2	Recreational ground	college, companies, pitches
3	Commercial	offices, business parks
4	Railway	railway use
5	Retail	shops
6	Residential	houses, apartment buildings
7	Dense	downtown

for each location, we populated POIs and POSs in different radius coverage areas, with a threshold radius starting at 300m and going up to 1000m by 50m increments.

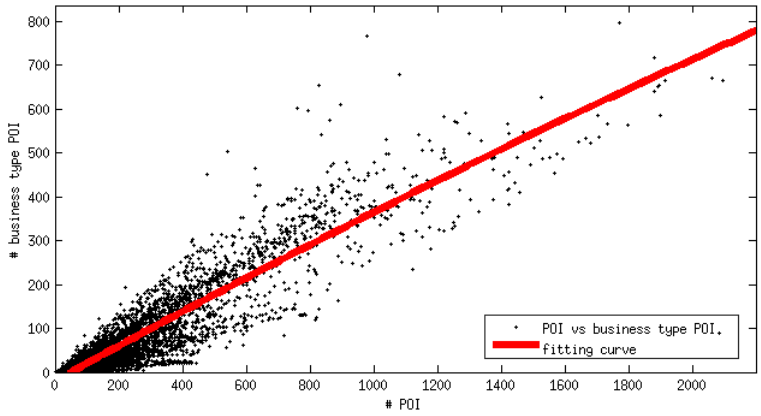


Figure 4: Correlation between POIs and its human activities for all land-use classifications, Polynomial fitting $R^2 = 0.85$

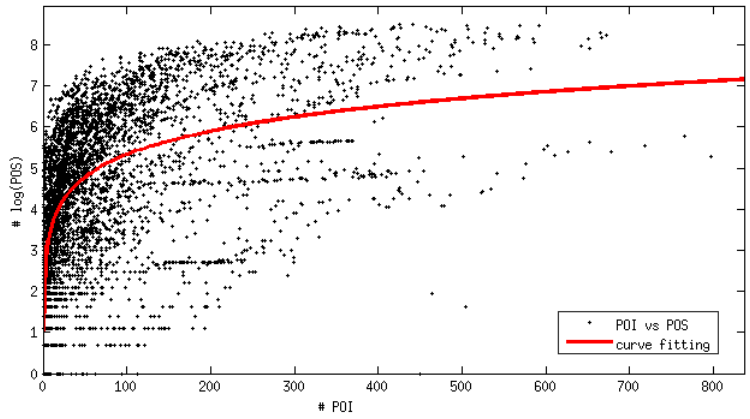


Figure 5: Correlation between business type POIs and POSs for all land-use classifications, Power fitting $R^2 = 0.34$

Some POIs where people do not perform a major business activity are ignored, (benches, power towers, farms etc). The POIs and business type of POIs for all land-use classifications within radius thresholds is polynominally increased

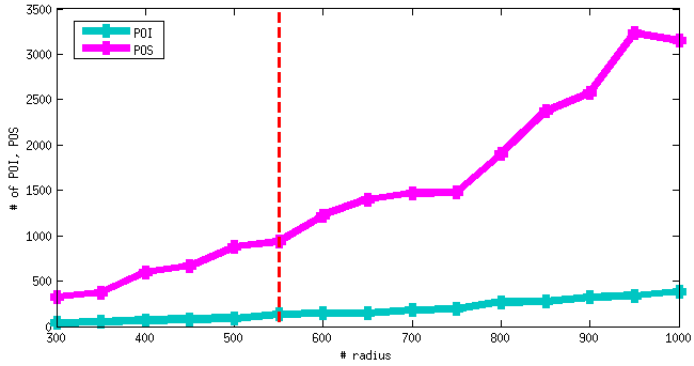


Figure 6: Dense land-use: Rate of POIs and POSs within the increasing radius

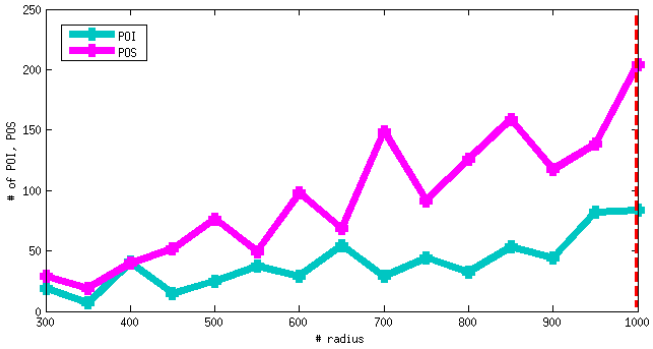


Figure 7: Residential land-use: Rate of POIs and POSs within the increasing radius

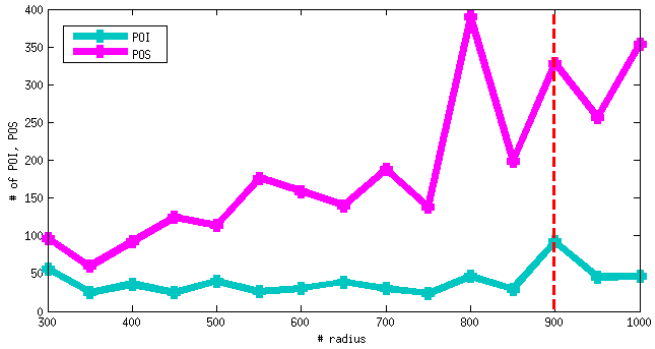


Figure 8: Retail land-use: Rate of POIs and POSs within the increasing radius

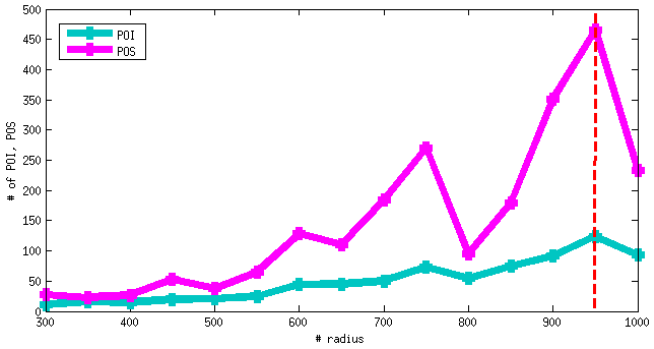


Figure 9: Railway land-use: Rate of POIs and POSs within the increasing radius

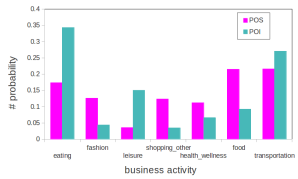
as described in Figure 5, with a strong correlation coefficient. But the correlation between business type of POIs and POSs is a weak as it is has random variables described in Figure 5.

First, in order to measure the data loss about activity categories in the both distribution, we estimated the matching accuracy regarding the existence of the activities in the two distributions using Jaccard similarity, which compares the similarity and diversity of the two activity distributions, with the size of the intersection divided by the size of the union of the activity distributions. Since POS distribution is hetergenous in every location, jaccard correlation coefficient is 50 percent, see Figure 13.

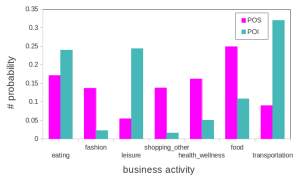
To avoid having fewer POIs or POSs in a location, we removed the 4 top, 4 bottom locations as outliers seperately from the cumulative distribution of the location by POIs and POS for each land-use classification and given radius thresholds. Figures 6 - 12 show the average rates of POIs and POSs when the threshold radius increases. We selected the optimal radius is shown in the figures as a dashed red line, considering the number of POIs are at least 100 at each location. For example, in a dense area are the optimal radius is 550m that collects 135 POIs, 935 POSs on average. Residential, recreational and industrial areas less populated with POIs. We use these optimal radiuses for each land-use for estimating the error between the two probability distributions.

The "service other" type of activity, see Table 2 is a group of non-classified, heterogeneous types of small businesses, such as insurance brokers, storage services and pet or video stores. We excluded this type of activity from the both probability distribution from bank card transaction data and HRBModel.

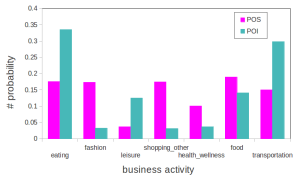
However, the actual probability distribution dissimilarity between the bank card transaction data and HRBModel expectation is more meaningful for estimating the data-loss, by allowing different types of land-use categories. The dissimilarity between the two probability distributions is a divergence of Hellinger distance(HD) and the divergence range for each land-use category is shown in Figure 5. In a dense area, the divergence is significantly lower than in other areas, with an average percentage of HD error is 33% (HD=0.33) there while recreational area divergence is around 41%. The difference of average probability distributions of the locations where the values are in a range of $\mu(HD, Land - use) \pm \sigma(HD, Land - use)$ for each land-use classification is depicted in Figures 16. Dense area has a small range of the HD error values between 0.2 and 0.5, it means at the maximum 80% of matching distribution can be found in the both data-source. Also recreation area is having a small range of HD error values between 0.32 to 0.52. Residential and industrial areas are having a big range of HD error values



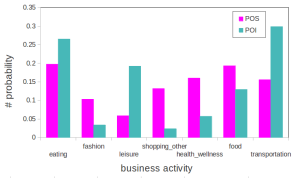
(a) Industrial area - activity distributions from POI and POS where the HD=0.28



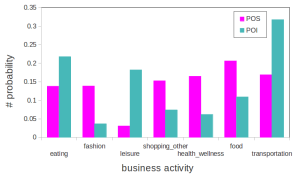
(b) Recreational ground area - activity distributions from POI and POS where the HD=0.4



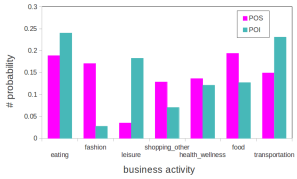
(c) Commercial area - activity distributions from POI and POS where the HD=0.33



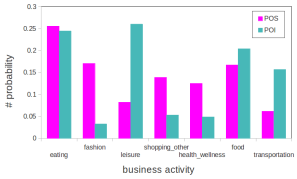
(d) Railway area - activity distributions from POI and POS where the HD=0.28



(e) Retail area - activity distributions from POI and POS where the HD=0.3



(f) Residential area - activity distributions from POI and POS where the HD=0.27



(g) Dense area - activity distributions from POI and POS where the HD=0.29

Figure 16: Activity distributions from POI and POS

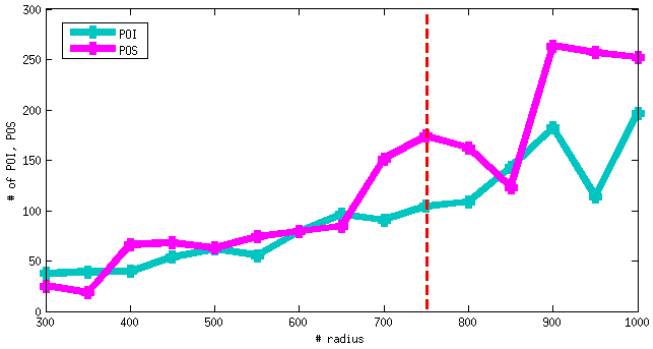


Figure 10: Commercial land-use: Rate of POIs and POSs within the increasing radius

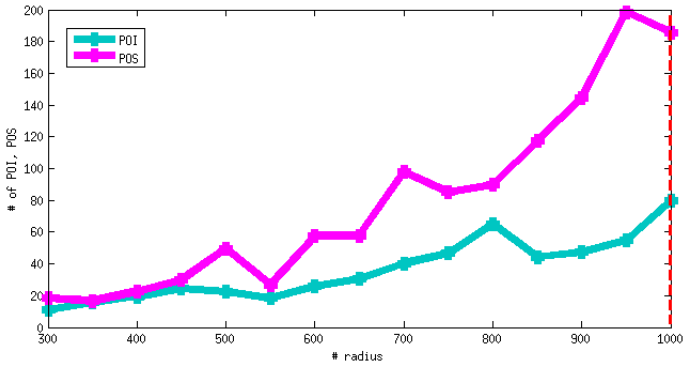


Figure 12: Industrial land-use: Rate of POIs and POSs within the increasing radius

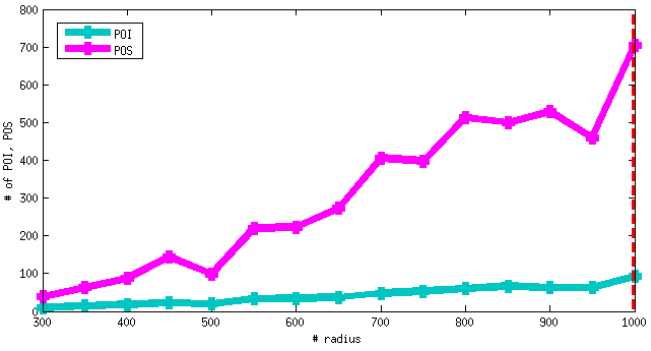


Figure 11: Recreational ground land-use: Rate of POIs and POSs within the increasing radius

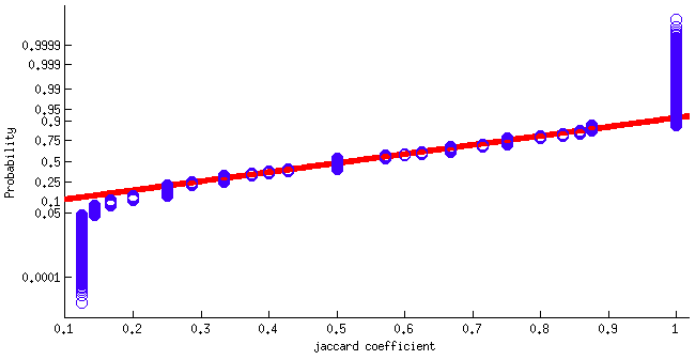


Figure 13: Normal probability distribution for Jaccard coefficient over all types of land-use within radius thresholds

and their activity distributions are more heterogeneous that could be a lack of dataset either POIs or POSs.

HD error distribution over all land-use classifications, has a peak at 0.35, this means 35% accuracy we get in average, see HD error normal distribution Figure 5. Depending on the land-use classifiction, the HD error values are varied in

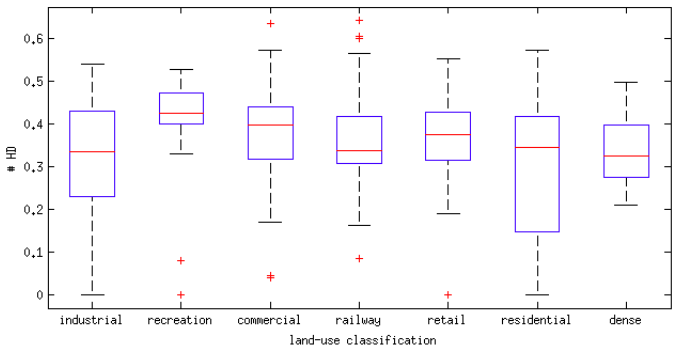


Figure 14: HD error range for all land-use classifications

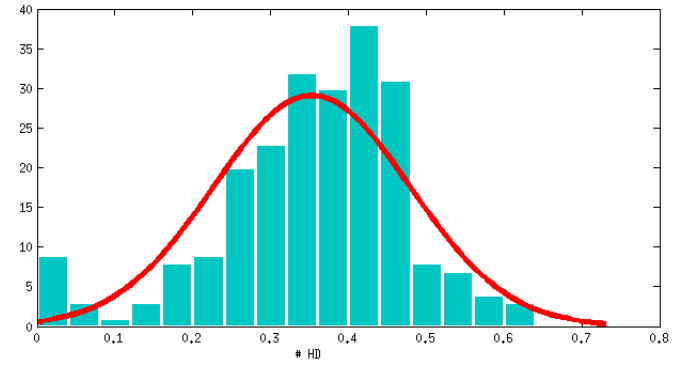


Figure 15: HD error distributions over all land-use classifications

normal distribution and the maximum error values are found in commercial and railway areas.

6. DISCUSSION

We described a straightforward methodology to perform an extended validation for HRBModel using bank card transaction data. We demonstrated a validation of the HRBModel comparing the two activity distribution between HRBModel and bank card transaction data in different land-use classifications. The overall accuracy was shown to be 65% (average HD is 35%) over all types of land-use categories.

Although some activities are randomly distributed regardless of land-use classification, the main representative activities are well described. In railway area, transportation activity is well described from POIs while transportation activity is not representative from POSs. In contrast, the retail area is well classified with eating, fashion, shopping other, health wellness and food activities from POSs. These activities are not well described from POIs in retail area. Interestingly, human activities in retail and commercial areas are similarly distributed from both POIs and POSs as those areas are representatives for commercial areas. In recreational area, leisure activity is a representative and well described from POIs but eating, shopping and fashion shopping activities are representatives from POSs in that area. That might be a reason of the lack of POSs information collected. The activities from POSs in dense and commercial areas are well aligned with the area characterizations than the activities from POIs. The matching error range, Hellinger distance between the human activity distributions from POSs and POIs is rather wide in residential and industrial areas and its activities are differently characterized in POSs and POIs. In the rest of land-use types, the human activity distributions from POIs and POSs are quite similar and its matching error range is around 10%. So the result shows that the significance of both data sources; POIs and POSs are around 60-70% and both relatively important to identify human activity of interests. But it suffers from the lack of POIs and POSs information collected.

The study presented in this paper showcases the performance of HRBModel evaluated with respect to the additional contextual data source (the bank card transaction data). We can further evaluate the HRBModel on various cities and countries to estimate the applicability and scalability of the model provided that the availability of OpenStreetMap resources (tags for POIs) in those regions is strong enough to support the use of our approach to improve the understanding of human behavior based on special data sources. As a further step one could also aim for determining how much and what kind of information/resources to consider to improve such understanding - e.g., the use of POS and POI data together. It is also possible to replace OpenStreetMap with some of the many static contextual data sources (e.g., credit transaction data, taxi data, and social events, etc.) that have a rich set of qualitative properties and can be utilized in our approach.

7. ACKNOWLEDGMENTS

The authors would like to thank the Banco Bilbao Vizcaya Argentaria (BBVA) for providing the dataset for this research. Special thanks to Assaf Biderman, Marco Bressan, Elena Alfaro Martinez, Juan Murillo Arias and Maria Hernandez Rubio for organizational support of the project and stimulating discussions. We further thank the BBVA, Ericsson, MIT SMART Program, the Center for Complex Engineering System (CCES) at KACST and MIT CCES program, the National Science Foundation, the MIT Portugal Program, the AT&T Foundation, Volkswagen Electronics Research Lab, The Coca Cola Company, Expo 2015, Ferrovial, The Regional Municipality of Wood Buffalo and all the members of the MIT Senseable City Lab Consortium for supporting this research.

8. REFERENCES

[1] A. Amini, K. Kung, C. Kang, S. Sobolevsky, and C. Ratti. The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1):6, 2014.

[2] M. A. Aura Leulescu. Statistical matching, a model based approach for data integration. *Eurostat - Methodologies and Working papers*, 2013.

[3] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: Analyzing cell-phone mobility and social events. In *Proceedings of the 8th International Conference on Pervasive Computing*, Pervasive'10, pages 22–37, Berlin, Heidelberg, 2010. Springer-Verlag.

[4] Z. Dashdorj, L. Serafini, F. Antonelli, and R. Larcher. Semantic enrichment of mobile phone data records. In *MUM*, page 35. ACM, 2013.

[5] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[6] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti. Towards a comparative science of cities: using mobile traffic records in new york, london and hong kong. *CoRR*, abs/1406.4400, 2014.

[7] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64:296–307, 2014.

[8] J. Krumm and D. Rouhana. Placer: semantic place labels from diary data. In F. Mattern, S. Santini, J. F. Canny, M. Langheinrich, and J. Rekimoto, editors, *UbiComp*, pages 163–172. ACM, 2013.

[9] M. S. Nikulin. Hellinger distance. *Encyclopaedia of Mathematics. Kluwer Academic Publishers*, 2002.

[10] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, (Volume 28, Issue 9):1–20, 2014.

[11] S. Phithakkitnukoon, T. Horanont, G. Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*, volume 6219 of *Lecture Notes in Computer Science*, pages 14–25. Springer Berlin Heidelberg, 2010.

[12] S. Porta, V. Latora, F. Wang, S. Rueda, E. Strano, S. Scellato, A. Cardillo, E. Belli, F. C. rdenas, B. Cormenzana, and L. Latora. Street centrality and the location of economic activities in barcelona. *Urban Studies*, 49(7):1471–1488, 2012.

[13] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.

[14] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *In Proceedings of the Nineteenth International WWW Conference (WWW2010). ACM*, 2010.

[15] C. Sengstock and M. Gertz. Exploration and comparison of geographic information sources using distance statistics. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 329–338, New York, NY, USA, 2011. ACM.

[16] C. Sengstock, M. Gertz, F. Flatow, and H. Abdelhaq. A probablistic model for spatio-temporal signal extraction from social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 274–283, New York, NY, USA, 2013. ACM.

[17] S. Sobolevsky, I. Sitko, R. T. D. Combes, B. Hawelka, J. M. Arias, and C. Ratti. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In *Big Data (BigData Congress), 2014 IEEE International Congress on*, pages 136–143. IEEE, 2014.

[18] S. Sobolevsky, I. Sitko, S. Grauwin, R. T. des Combes, B. Hawelka, J. M. Arias, and C. Ratti. Mining urban performance: Scale-independent classification of cities based on individual economic transactions. *CoRR*, abs/1405.4301, 2014.

[19] S. Sobolevsky, M. Szell, R. Campari, T. Couronné, Z. Smoreda, and C. Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PloS one*, 8(12):e81707, 2013.

[20] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, and E. de Lara. Mobility detection using everyday gsm traces. In *Proceedings of the 8th International Conference on Ubiquitous Computing*, UbiComp'06, pages 212–224, Berlin, Heidelberg, 2006. Springer-Verlag.

[21] B. Vantaggi. Statistical matching of multiple sources: A look through coherence. *International Journal of Approximate Reasoning*, 49(3):701 – 711, 2008.

[22] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 520–528, New York, NY, USA, 2011. ACM.

[23] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 247–256, New York, NY, USA, 2011. ACM.

Measuring Spatial Subdivisions in Urban Mobility with Mobile Phone Data

Eduardo Graells-Garrido
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Universidad del Desarrollo
Santiago, Chile
eduardo.graells@bsc.es

Irene Meta
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Universitat Internacional de Catalunya (UIC)
Barcelona, Spain
irene.meta@bsc.es

Feliu Serra-Burriel
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
feliu.serra@bsc.es

Patricio Reyes
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
patricio.reyes@bsc.es

Fernando M. Cucchietti
Barcelona Supercomputing Center (BSC)
Barcelona, Spain
fernando.cucchietti@bsc.es

ABSTRACT

Urban population grows constantly. By 2050 two thirds of the world population will reside in urban areas. This growth is faster and more complex than the ability of cities to measure and plan for their sustainability. To understand what makes a city inclusive for all, we define a methodology to identify and characterize spatial subdivisions: areas with over- and under-representation of specific population groups, named *hot* and *cold* spots respectively. Using aggregated mobile phone data, we apply this methodology to the city of Barcelona to assess the mobility of three groups of people: women, elders, and tourists. We find that, within the three groups, cold spots have a lower diversity of amenities and services than hot spots. Also, cold spots of women and tourists tend to have lower population income. These insights apply to the floating population of Barcelona, thus augmenting the scope of how inclusiveness can be analyzed in the city.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Urban Mobility, Mobile Phone Data, Spatial Analysis

ACM Reference Format:
Eduardo Graells-Garrido, Irene Meta, Feliu Serra-Burriel, Patricio Reyes, and Fernando M. Cucchietti. 2020. Measuring Spatial Subdivisions in Urban Mobility with Mobile Phone Data. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3366424.3384370

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3384370

1 INTRODUCTION

As of 2020, more than half of the population live in cities [25], and by 2050, two thirds of the population will reside in urban areas [28]. In this scenario of urban growth, the United Nations have declared a goal for sustainable development that aims to make cities "inclusive, safe, resilient, and sustainable" [1]. To reach these objectives, disciplines such as urbanism, architecture, ecology, sociology, and others provide frameworks to model the functioning of cities. Typically, the main data source for analysis is household and time-use surveys as well as travel diaries. Such instruments provide rich information that represents the general population of a city, which then informs urban design and policy making.

However, the goals of improving inclusiveness and safety are limited by the purpose of traditional methods, because typical data sources tend to under-represent specific sub-populations, including women [8] and elders [20]. For example, surveys fail to measure that women trips are shorter than those of men [6], and how these trips are chained to others, partially due to household and caretaking purposes [17]. Likewise, elders also move in shorter trips, but, in contrast to younger people, their trip purposes are focused mostly on feeling independent and interacting with others in social situations [30]. While it is known that traditional methods have these shortcomings, improving them to reach finer representation is expensive and impractical on a global scale. Even though specific methods can be designed for under-represented groups, such efforts may miss the global picture, which includes the relationship between mobility of different population groups. As a consequence, there is need of fine-grained city-scale data for the design of inclusive and safe urban spaces. Finally, data biases often occur due to underlying societal biases. Biased data means incorrect population statistics which can mislead city planning and design into amplifying the problems they aim to fix [26].

Recent technological advances and the availability of non-traditional data sets have allowed to study urban phenomena at spatio-temporal granularities that traditional methods cannot. Mobile phone data, for example, allows a cost-effective way to perform

studies about urban human mobility [7], as mobile operators already generate, store, and analyze the data for billing and marketing purposes. The aggregation of digital traces from mobile phone usage was used to uncover data gaps in mobility [11], a true seminal piece of work towards using this data source for inclusive cities. Inspired by this line of work, we extend the scope to understand mobility aspects of three groups of urban visitors: women, elders, and tourists. Under the assumption that all people access the city equally, our research questions are: *How to identify places with more (or less) presence of these groups than expected?* If these places can be pointed out, *what characterizes them?* Our proposed method is a pipeline that starts on the definition of visitor metrics related to these three groups; then, we perform spatial analysis on these metrics to identify whether there is spatial concentration of visitors (or the lack thereof). If so, we identify areas with over-representation of these groups, or *hot spots*, as well as the opposite, places with under-representation, or *cold spots*. We then proceed to use up-to-date information about income, amenities, and services in the city to characterize these areas based on their economic development and urban environment.

Different variables (semantics)

As a case study, we analyzed the city of Barcelona, the second largest city in Spain and one of the largest in Europe. The city is known for its urban planning tradition and was often ahead of its time compared to the Spanish general developments [16]. However, there are still challenges in improving the city for everyone from a sustainable perspective: overtourism [21], gender accessibility problems [15], and one fifth of its inhabitants are elderly people [9]. In an effort to augment its understanding of the city's urban dynamics, the local government (*Ajuntament de Barcelona*) acquired anonymized and aggregated mobile phone data from the mobile phone operator Vodafone. Access to this type of data for planning is a clear opportunity to compare and advance over other data sources focused only on census and residential populations, as it could help to understand the mobility patterns of residents and non-residents alike.

We find that the studied population subgroups behave differently. Cold spots for all groups are characterized by lower population income than hot spots, as well as less diversity of amenities and services. Hot spots for all groups are characterized for being less associated with public transport than the rest of the city. Urban infrastructure such as highways and the main streets of the city play a role when interpreting the locations of these areas of over- and under-representation, as cold spots tend to be outside of the area delimited by highways around the city whereas hot spots tend to be close to relevant primary streets. As such, our work contributes: (i) a methodology to identify and characterize hot and cold spots of the floating population in a city; (ii) a case study applying the methodology to Barcelona.

We conclude the paper with a discussion focused on the implications in public policy and the usage of non-traditional data to solve the complex problems that affect cities today.

2 RELATED WORK

Mobile phone data usually refers to the set of billing records from mobile phone networks, known as Data Detail Records (XDR) [7]. Other types of non-traditional data that has been used to understand mobility include micro-blogging platforms with geo-location [19] or inferred user attributes regarding mobility [29]; check-ins from location-based services [24]; and photos from photo-sharing services [5]. In comparison to these data sets, XDR allows a fine-grained analysis, not only in spatio-temporal aspects to, for instance, observe changes in mobility in short periods of time [12], but also in demographic ones, such as measuring the social diversity of visitors in shopping malls [4].

In terms of scope, our work is similar to recent efforts to uncover gender gaps in urban mobility [11]. The differences in our approach are two-fold. On the one hand, we use a data set that is aggregated from XDR in spatial and temporal aspects. As such, it does not include individualized information, allowing us to perform analysis at the area-level but not on the individual level. On the other hand, our methods rely primarily on established spatial analysis [3, 23], which brings a different perspective to the distance-based approach employed before. Then, our work contributes a different approach to an already identified problem, with extended coverage in terms of population groups, adding elders and tourists, and focusing on a different city, Barcelona.

3 CONTEXT AND DATA SETS

More than 1.6 million people reside in Barcelona. Its 100 Km^2 area is composed of 12 districts, split into 73 neighborhoods. Natural boundaries delimit the city: the Besos river limits the city at the north-west, and the Llobregat river does so at the south-west side. The Metropolitan Area is much wider, and it is impossible to distinguish the limit between Barcelona and the surrounding municipalities. The city extends on a mild slope from the sea (south-east) up to the edge of the Collserola mountain chain (north-west). The Collserola and the Montjuic (south) have limited the city expansion because of their relatively hard accessibility, and now are important areas of leisure and biodiversity within the city [16].

The social aspects of mobility that affect subpopulations of the city [9, 15] along with rising overtourism [21] and alarming pollution levels have urged urban planners to focus on sustainability [2]. In this context, we focus on one of the qualities of sustainability, inclusiveness.

City Data. The *Ajuntament* provides open access to socio-demographic attributes at the neighborhood level at a yearly frequency (some metrics are quarterly), including income and house pricing among other things. All these variables are scaled to the mean, which allows us to compare the different neighborhood areas in relative terms. We measure income through the mean family income (RF, or *Renda Familiar*),[1] which contains mean income at the neighborhood level (see Figure 1 (a) for its spatial distribution), normalized so that the whole city mean income equals 100.

Mobile Phone Data. The data obtained from the mobile phone operator consists of the number of visitors observed during the year 2018, at periods of four hours, grouped into 212 regions or cells (see Figure 1 (b)). The number of visitors is defined as the total number of mobile phones active inside each region during each period. *Active* means that the phone was initiated or received some activity (call, browse, text, etc) other than passive connections to the

[1] https://opendata-ajuntament.barcelona.cat/data/es/dataset/est-renda-familiar

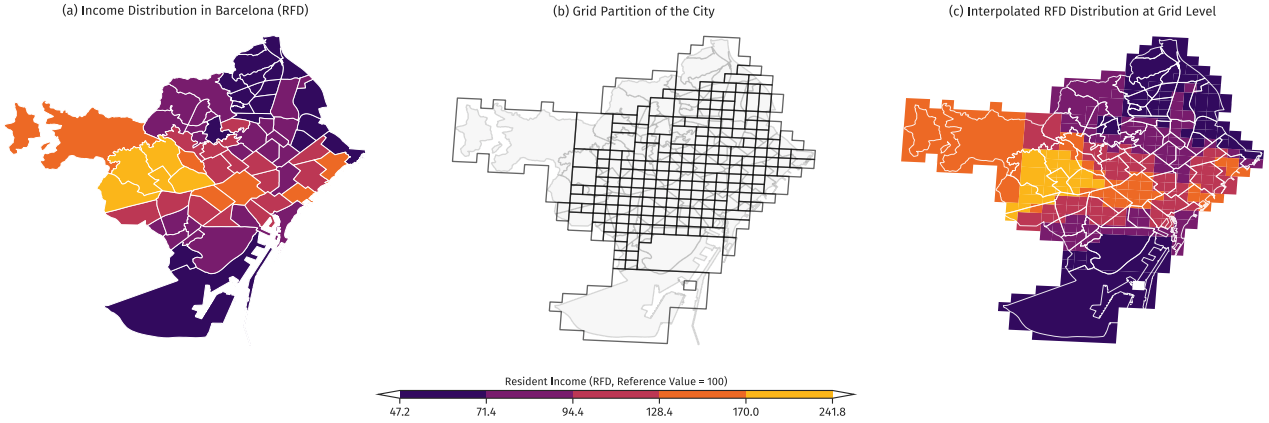


Figure 1: a) Map of neighborhoods with income data; b) Overlap of the grid areas with the neighborhoods; c) Spatial join of neighborhood income data with the grid.

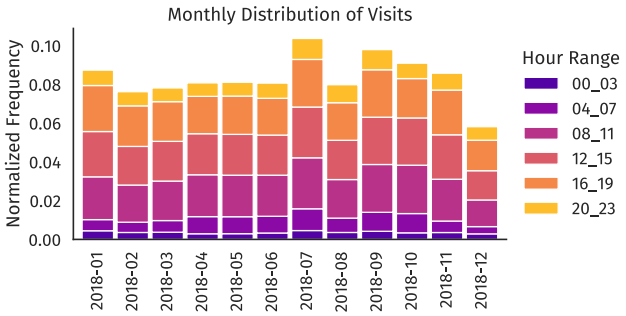


Figure 2: Normalized frequency of the sum of the monthly number of visitors captured in the dataset, provided by Vodafone to the Barcelona city hall. The colors stratify the sample into time-frames.

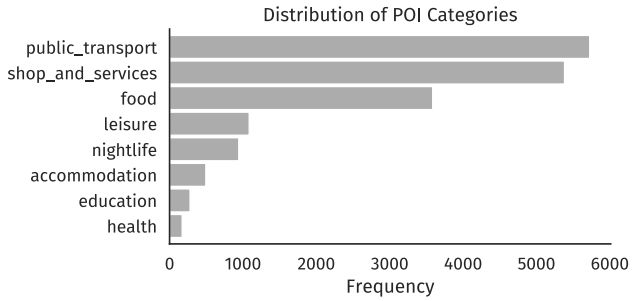


Figure 3: POI count per category. There are three main categories that represent more than 80% of all the POI.

network. This may introduce bias in the data, as people do not call while driving or at night, or they connect to their home or job WiFi falsely indicating less presence. In addition to the total number of visitors, the operator also provides the number of visitors according to specific demographic characteristics, including gender (binary, female and male), age cohorts, and tourists (national and foreign). The determination of these characteristics and its aggregation into a number of visitors was made directly by the mobile operator using activity criteria as well as billing and other information when available. In addition, cells with less than a given number of observations during each period were discarded from the data and cells that consistently exhibited few observations (below 500 visitors) were consolidated into grouped regions.

We estimated the total number of visits accounted per month and per hour range (see Figure 2, normalized to avoid revealing commercially sensitive data). The number of total visitors per month lies within the same order of magnitude, with fluctuations that could be explained by changes in the market share of the operator and seasonal factors such as tourism in July.

We estimated a mean income for each cell, defined as the weighted interpolation of the incomes in all areas that intersect with that cell (see Figure 1 (c)).

Open Street Map. To include aspects of the built environment and accessibility to amenities and services, we use data from OpenStreetMap (OSM).[2] OSM provides spatial and geographical data contributed freely and voluntarily by its members, and it has been identified as an accurate source of urban information [13].

From OSM we obtain Points of Interest (POIs) that allow us to understand part of the urban environment in our analysis. Points of interest (POIs) are geolocated attractors, such as shops, food spots, and tourist sights. We categorized most of POIs in the city as shown on Figure 3. These include (sorted by descending frequency): *public transport* (e.g., bus stops, metro stations), *shops and services* (e.g., convenience shops, government offices, professional services), *food* (e.g., cafés, restaurants), *leisure* (e.g., natural attractions, parks, stadiums), *nightlife* (e.g., bars), *accommodation* (e.g., hotels), *education* (e.g., universities, schools), and *health* (e.g., hospitals).

By analyzing the integration of these data sets, we aim to identify areas of the city with over- and under-representation of women,

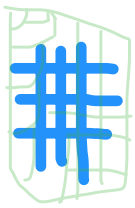
[2] http://openstreetmap.org

elders, and tourists, and characterize these areas according to their economic development and urban environment.

4 METHODS

In this section we describe how we measure spatial subdivisions in urban mobility. First, we define the metrics we evaluate. These metrics are local, in the sense that they cover a specific point or area and do not consider the context or their surroundings. Second, we perform a spatial analysis using established techniques, allowing us to take into account the spatial context and evaluate both local and global patterns to identify significant areas of over- and under-representation. And third, we define how to characterize the significant areas with respect to economic development and urban environment.

4.1 Cell Level Metrics



In our context, the city is partitioned by a grid which comprises neighboring cells. Cells may have edges and/or vertices in common but they do not overlap. Here we define cell-level metrics regarding the presence of women, elders, and tourists in them. The three metrics we define are: the women ratio G, the elder ratio E, and the tourist ratio T.

Women Ratio G. This metric captures the ratio of female visitors in an area i during the whole period under study. We first define the women ratio G' as:

$$G'_i = \frac{\text{\# women visiting area i}}{\text{\# total visitors in area i}}.$$

Note that it is likely that the mobile operator has a non-representative sample of the population. For instance, the sample ratio at the city level may not be 1, even though the population ratio may be close to it. To counter this effect we define a standardized version of the women ratio as

$$G_i = \frac{G'_i - \bar{G}'}{s(G')},$$

where s is the sample standard deviation function and $\bar{G}'$ corresponds to the mean of G'_i for all the cells. In this way, if $G_i = 0$, the area i has a women ratio equivalent to the average of the city. Positive and negative values of G indicate how many standard deviations the ratio deviates from the sample mean. Notice that, to focus on the floating population, we consider visitors between 8am and midnight.

Elder Ratio E. In a similar way to G', the elder ratio before standardization is defined as:

$$E'_i = \frac{\text{\# elder visitors in area i}}{\text{\# total visitors in area i}}.$$

We choose the threshold age to be considered elder as 65 years old or more, as defined by the *Ajuntament*.[3] Analogous to G, the metric E is the standardized version of E'.

[3] https://ajuntament.barcelona.cat/personesgrans/es/canal/la-gent-gran-de-barcelona

Tourist Ratio T. Our last metric is similar to the previous ones, as it represents the proportion of tourists in an area:

$$T'_i = \frac{\text{\# tourists (both foreign and national) in area i}}{\text{\# total visitors in area i}}.$$

Analogous to G and E, the metric T is the standardized version of T'.

4.2 Spatial Patterns

Our aim is to find places where each population group of interest is over- or under-represented according to its floating population patterns, expressed in the metrics G, E and T. To do so, we evaluate whether values of these metrics tend to concentrate in geographical terms, *i.e.*, if nearby areas have similar values. The Moran's I coefficient of spatial autocorrelation [23] measures this concentration. It is defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2},$$

where N is the number of spatial units (in our case, grid cells) under analysis, x_i is one of $\{G_i, E_i, T_i\}$, w_{ij} encodes the spatial weight of cell j into cell i, and W is the sum of all spatial weights. Note that w_{ij} is a matrix where $w_{ii} = 0$. The value of w_{ij} is a normalized version of the following schema:

$$w'_{ij} = \begin{cases} 1 & \text{if area i and j are contiguous.} \\ 0 & \text{otherwise.} \end{cases}$$

Here, contiguity between cells is defined as sharing an edge or sharing a vertex. This is coherent when using grids composed of square cells, as it is possible to move from one square to another through a corner. Then, w_{ij} is normalized in the following way:

$$w_{ij} = \frac{w'_{ij}}{\sum_j w'_{ij}}.$$

With these definitions, $I = -1$ when the variable under analysis is perfectly dispersed in space, $I = 1$ when it is completely clustered, and $I = 0$ when values are randomly arranged.

Next, for each metric, if indeed there is spatial autocorrelation, we proceed to estimate Local Moran's I, a coefficient that allows us to identify groups of areas that have high (or low) values that are surrounded by other areas with high (or low) values [3]. It is defined as follows:

$$I_i = \frac{x_i - \bar{x}}{s(x_i)^2} \sum_{j=1, j\neq i}^{n} w_{ij}(x_j - \bar{x}),$$

where $s(x_i)$ is the standard deviation of values of x of contiguous areas to area i. Note that, in global and local I, significance is estimated through permutation tests. Areas with significant high values of local I are known as *hot spots*, and areas with significant low values are known as *cold spots*. The other areas present neutral or average behavior.

4.3 Characterizing Hot and Cold Spots

After identifying areas of interest, our purpose is to characterize each type of spot. With this aim, we analyze the income of each spot and the availability of services and activities through Points of Interest (POIs).

Income. We estimate a mean income for all spatial units. Under the null hypothesis that income is independent of spatial subdivisions, one would expect that population income in hot spots has the same distribution as population income in cold spots. We test this hypothesis by comparing the income in all hot spots with the income for all cold spots using the two-sample Kolmogorov-Smirnov (KS) test. This non-parametric test evaluates whether two underlying one-dimensional probability distributions that generated those samples differ. If the result of a test is significant, it means that there is evidence to reject the null hypothesis of same income for both types of area for a given visitor metric.

Association and Diversity of Points of Interest. Next, we estimate how each category of POIs is associated with each area. This problem is analogous to document categorization in Information Retrieval where there are frequent words in many (if not all) documents that do not necessarily characterize them. In our context, areas are analogous to documents, and POIs are analogous to words. For instance, bus stops may be available in all areas of the city, but some areas may have more bus stops than others, while having less POIs of other categories. In that case, these latter areas have a stronger association with bus stops than other areas. Given that most areas have many kinds of POIs, we need to use a weighting schema that controls for frequency and variability. While a common technique to do so is Term Frequency–Inverse Document Frequency (TF-IDF), we resort to a technique that does not overweight elements with low frequencies. This method is known as Log-Odds ratio with Uninformative Dirichlet Prior [22]. It defines the weight of a word through the following point estimate:

$$\widehat{\delta}_{kw}^{(i)} = \log\left[\frac{(y_{kw}^{(i)}+\alpha_{kw}^{(i)})}{(n_k^{(i)}+\alpha_{k0}^{(i)}-y_{kw}^{(i)}-\alpha_{kw}^{(i)})}\right] - \log\left[\frac{(y_{kw}+\alpha_{kw})}{(n_k+\alpha_{k0}-y_{kw}-\alpha_{kw})}\right],$$

where kw is the frequency of the POI type at cell i, and α is the prior distribution. Positive values of this metric indicate positive association, while negative values indicate disassociation. Thus, we would expect larger amounts of specific kinds of categories in specific regions and some other categories that are rare in other regions. Values close to 0 indicate independence between the POI type and the cell.

Ref. [11] found relationships between accessibility gaps with the diversity of places, and between economic development and the diversity of social connections in places [10]. To explore this potential relationship, for each cell i we estimate its POI entropy, defined as the Shannon Information Entropy H:

$$H_i = -\sum_c p_c \log p_c,$$

where c is a POI category, a p_c is the fraction of POIs from category c within cell i.

By following this methodology, it is possible to identify spatial subdivisions in urban mobility according to who visits each area of the city, particularly women (G), elders (E), and tourists (T). The spatial subdivisions are defined as those areas identified as hot/cold spots of visits from these groups, which then can be characterized according to their economic development and urban environment.

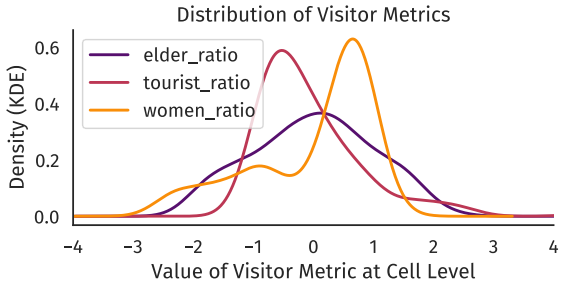


Figure 4: Probability density functions of cell-level metrics estimated with Kernel Density Estimation.

5 RESULTS

In this section we describe the results of applying our proposed methods to the data sets, with the aim of understanding spatial subdivisions in Barcelona, as seen from mobility data.

Cell-level Metrics. We estimated the women ratio G, elder ratio E, and tourist ratio T for all cells in the grid of Barcelona. Of all cells, 195 have the same size. However, a few of them have bigger sizes, because they were merged by the mobile phone operator to ensure privacy. We considered the number of regular cells as a scaling factor (i.e., the most common) that would fit in a merged cell. Thus, we divided the value of each cell according to its scaling factor.

The distributions of each observed metric (Figure 4) have different shapes. The elder ratio distribution is unimodal with fat tails on both sides. The tourist ratio distribution is positively skewed, having a negative mode. Conversely, women ratio distribution is negatively skewed with positive mode. It has a group on the negative values but the majority of the values are positive.

Global Spatial Autocorrelation. In spatial terms, the three metrics present spatial autocorrelation ($I_G = 0.25$, $I_E = 0.34$, $I_T = 0.39$, all significant with $p \leq 0.001$). Women and elders cover most of the densely populated areas of the city (see Figure 5 (a) and (b), respectively). However, the extent of elder concentration is smaller, thus having a greater autocorrelation than women. Tourists being the most concentrated group makes sense given the touristic attractiveness of the city, which tends to be concentrated in the historical districts, with few spots in other places such as the beaches, the highway that connects the airport to the city, and Barcelona's soccer team stadium (see Figure 5 (c)). Note, however, that all these concentrations are smaller than the concentration of income ($I_{RFD} = 0.83$, see Figure 1 (c)).

Local Spatial Autocorrelation. Next, we estimated the Local Moran's I coefficient of each area of the city for G, E, and T. Figure 6 shows the spatial location of all relevant areas, with hot spots of each metric in the top row, and cold spots of each metric in the bottom row. Color indicates the income level of each cell.

Both G and E hot spots are located mostly above the Diagonal Avenue (the avenue the goes from west to east), and they overlap in three different sectors. The first point corresponds to the Sants area (west). The second sector is north-west of the city, right above the Diagonal Avenue. It corresponds to a middle and high income

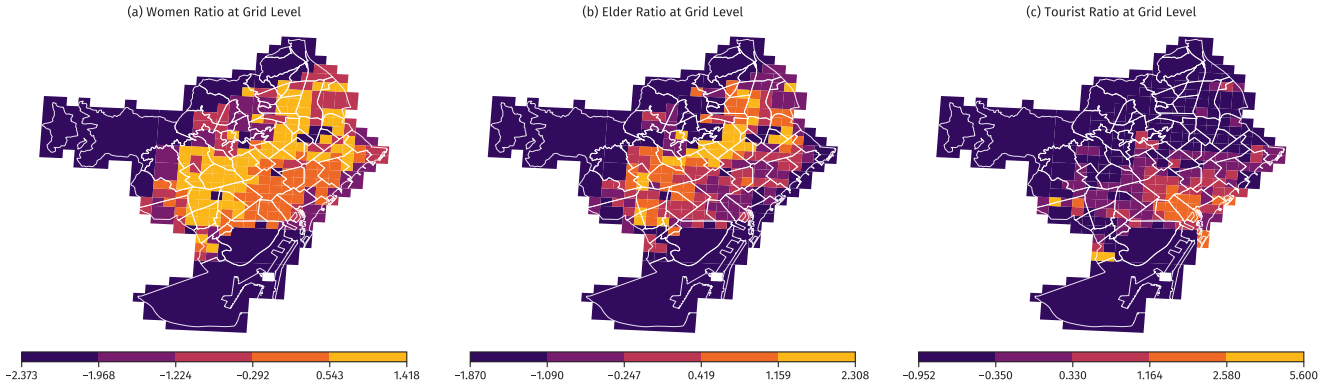


Figure 5: Maps of the spatial distribution for metrics G, E and T. Color of cell i corresponds to the value of G_i, E_i and T_i respectively. Notice that the scales are different for each one of the metrics.

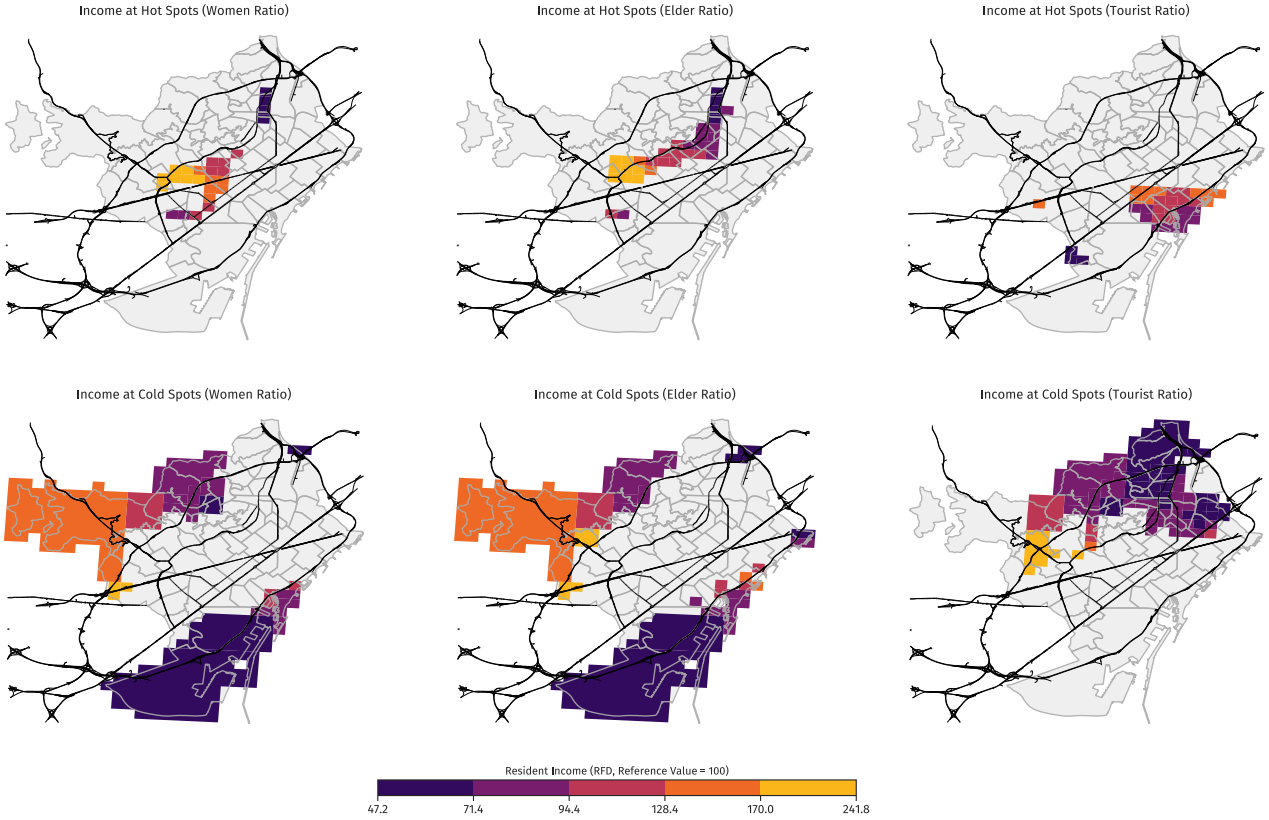


Figure 6: Hot/cold spots for each metric according to Local Moran's I metric. The first column is the gender ratio, the second column the elder ratio and the third column the tourist ratio. The top row contains hot spots, while the bottom row contains cold spots. The color represents the income of that area as in Figure 1.

area of the city, while the third sector, to the north of the city, corresponds to a low income area. It is hard to interpret this overlapping: we know the percentage of female population is larger with age [9], but also that women are the ones who dedicate more time to care-taking activities [9]. The E hot spot shows a unique cluster, mainly just below the middle beltway. Within this hot spot, we observe heterogeneity on the socio-economic status, having the south-western part a much larger income ratio. The city center does not show under or over-representation of women and elders, as both hot and cold spots are absent there. Conversely, the area in its

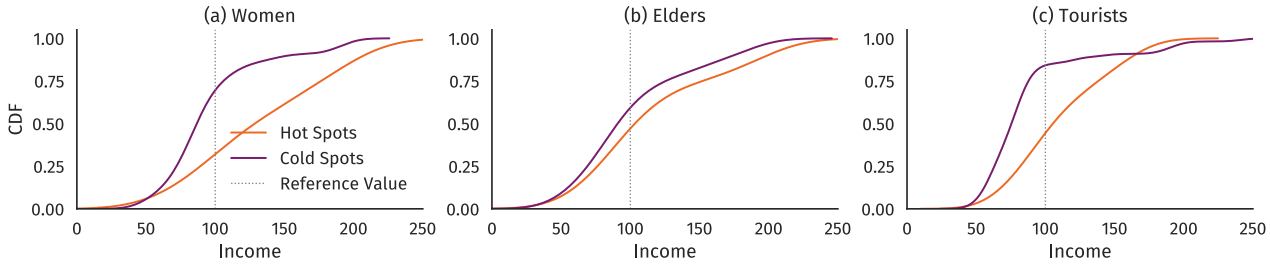


Figure 7: CDFs of the income on hot and cold spots for each of the metrics.

totality is signalled as a hot spot for tourists (T), encompassing the Old district and part of the adjacent Eixample, below the Diagonal Avenue. This observation seems reasonable given the density of historical sites and leisure spots in the area. Two more isolated areas show a hot spot of tourist activity. The one on the west is nearby the Barcelona Football Club Stadium. Particularly, it is not the cell that contains the stadium, although it contains one of the metro stations that is closer to it. At the same time, it contains the southern campus of Universitat Politecnica de Catalunya, which may also receive foreigners regularly. The other area is located on the south and contains the *Fira Gran Via Barcelona*, the biggest venue of the city for international congresses and expositions (including the Mobile World Congress). This is expected given that the mobile operator may use roaming connections to identify foreign tourists.

The cold spots are mostly spread around the periphery of the city and have different levels of income. There are three G cold spots. The smallest area is characterized by an infrastructural node that links motorways. The largest area covers mixed income but also low population density, as it is located in the periphery of the city, near the mountains. The third area, on the south, covers a leisure sector (the Montjuic hill), a working sector (the Port), and a densely populated neighborhood, La Barceloneta, which holds many touristic attractions, including restaurants and the beach. The E cold spots are similar to those of G. The western T cold spot corresponds to the richest areas, and to the toll access tunnels from the valley behind the Collserola chain mountains. In the north side, T cold spots comprehend the Sagrera high-speed train station and one of the main accesses to the city for road transport, with the infrastructural node between three motorways and two beltways. It is an area of low income and less leisure amenities than the rest of the city. Other areas of the city, such as the southern, are also characterized by similar income levels but they are not tourism cold spots.

Income Characterization. There is a variety of income levels in hot and cold spots. Women and tourist cold spots have different income distributions than their corresponding hot spots, according to a Kolmogorov-Smirnov (KS) test ($p_G = 0.004$, and $p_T < 0.001$, Bonferroni-corrected). Elders do not show that difference ($p = 0.798$, Bonferroni-corrected). To explore these differences visually, Figure 7 shows the cumulative distributions of income for hot/cold spots of each metric, estimated with Kernel Density Estimation (KDE). Hot spots tend to be shifted towards the larger income areas and cold spots appear to be on the low income areas.

Table 1: Kolmogorov-Smirnov tests (significance level $p < 0.05$, values show have been Bonferroni-corrected) for each of the metrics and significant POI categories on each studied subdivisions. Visitor metrics are women ratio (G), elder ratio (E), and tourist ratio (T).

Visitor Metric	POI Category	KS	# Cells (Hot)	# Cells (Cold)	p
G	Accommodation	0.703	26	22	< 0.001
G	Education	0.633	26	22	0.001
G	Food	0.587	26	22	0.006
G	Shops & Services	0.787	26	22	< 0.001
G	Nightlife	0.549	26	22	0.019
G	Leisure	0.580	26	22	0.007
G	Public Transport	0.657	26	22	0.001
E	Education	0.471	32	27	0.043
E	Shops & Services	0.678	32	27	< 0.001
E	Public Transport	0.524	32	27	0.007
T	Accommodation	0.500	27	55	0.003
T	Food	0.707	27	55	< 0.001
T	Leisure	0.502	27	55	0.003
T	Public Transport	0.593	27	55	< 0.001

POI Characterization. The distribution of POIs in the city exhibits different functional regions based on the activities and services available (see Figure 8). Categories such as accommodation, food, and nightlife are more concentrated than the others, while health, shops and services and education are more scattered, indicating that most of the city has access to a diversity of amenities and services.

To evaluate differences in POI (or *amenities*) association between hot and cold spots, we performed pairwise KS tests for each POI category and each metric (see Table 1). Then, we built swarm plots of each area type per metric, per POI category (see Figure 9). Every dot is a cell in a hot/cold spot of the associated variable, and its color represents its tendency (either hot or cold). Its y-position represents the corresponding POI association, while its x-position is only for legibility. Women (G) have the largest number of POI categories with significant differences between hot and cold spots association. Only the health category has the same distribution for hot and cold spots. Elders (E) present differences in the distribution of POI association for education, shops and services, and public transport. They present similar distributions to Tourists (T), where hot spots tend to be positively associated with amenities, except for the Public Transport category that presents some negative associations, similar to G and E. The hot spot association to amenities

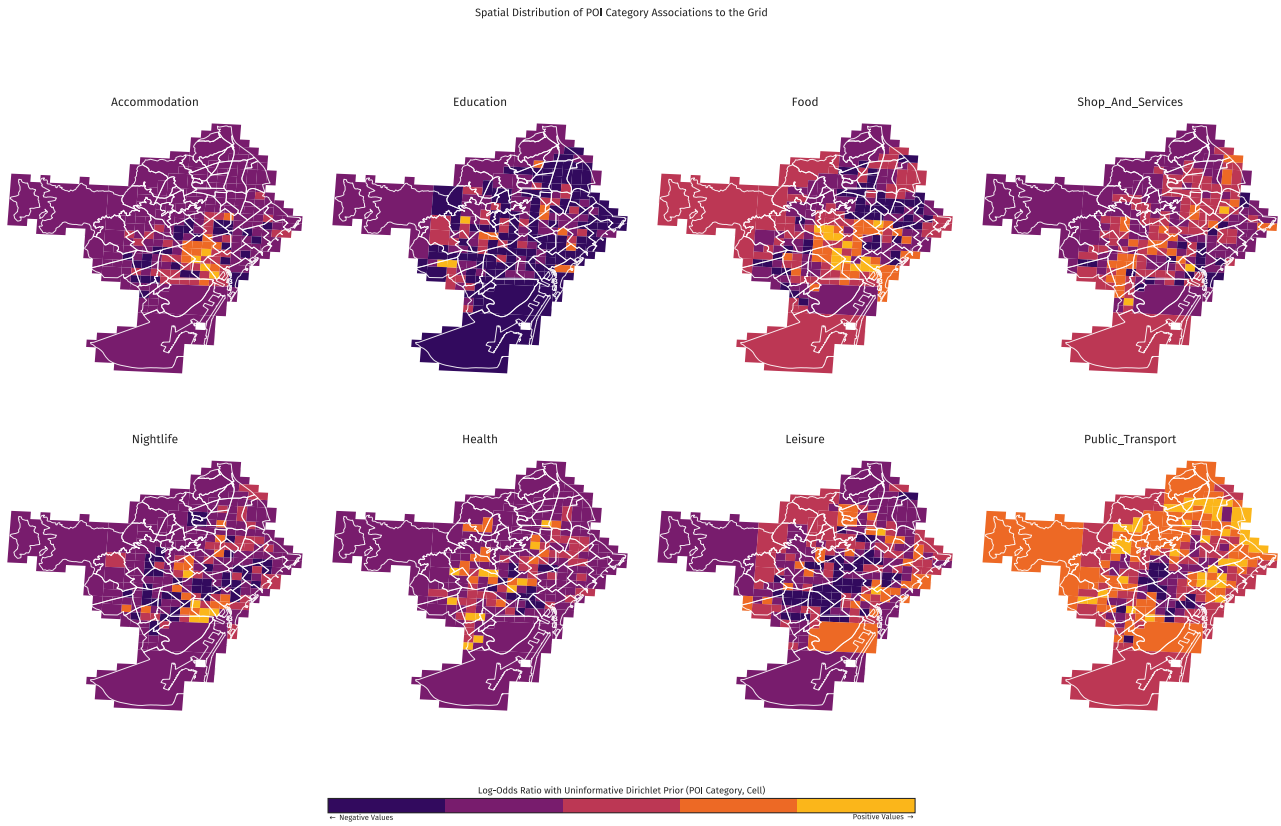


Figure 8: Log-odds Ratio with Uninformative Dirichlet Prior for each category of POI. Map colors according to the log-odds ratio value within each category.

may be related to gender or age based mobility behaviours, regarding trip chaining and trip purposes; however, we lack a clear understanding of the disassociation to public transport, which is arguably unexpected. Tourists (T) present differences between hot and cold spots association on accommodation, food, leisure, and public transport. The first three categories describe tourist attractors, as the hot spots are positively associated with these amenities. The public transport negative association to hot spots, similar to G and E, may be explained due to the historic district being comprised mostly by pedestrian streets. There are other associations that can be discussed, but we omit them due to space reasons.

Finally, regarding the diversity of POIs measured through entropy, cold spot areas are more associated with low diversity of POIs. The KS test was significant for the three pairwise comparisons ($p_G = 0.030$, $p_E = 0.005$, and $p_T < 0.001$, Bonferroni-corrected). The differences are illustrated through the KDE-based cumulative density functions on Figure 10.

In this section we explored how three population groups (women, elders, and tourists) were present in several areas of Barcelona in the year 2018. We observed that, indeed, there are areas of the city that tend to be visited by these groups (hot spots), as well as areas that tend to have an under-representation of them (cold spots), effectively creating subdivisions of over and under-representation in the city. Income and the availability and diversity of POIs play a role in characterizing these relevant areas. The most salient characterizations are two. On the one hand, cold spots of activity for women and tourist visitors are associated with less population income. On the other hand, hot spots of the three types of visitors are associated with less public transportation. Cold spots of all types are associated with a lesser diversity of POIs. We discuss further implications of these results in the next section.

6 DISCUSSION AND CONCLUSIONS

A city is experienced uniquely by each individual, although people with shared characteristics may experience it in similar ways. Urban disciplines have been studying these experiences for decades with the goal of improving quality of life in cities through urban planning and design. In this paper we have shown that aggregated mobile phone data allows us to identify relevant areas in terms of over- and under-representation of subpopulations such as women, elders, and tourists. Being a cost-effective source of data, our proposal brings knowledge of which places are relevant in terms of presence (or absence) of people from these groups as well as what characterizes these places in terms of the urban environment. Then, our methodology provides knowledge about under-represented groups in urban and policy design.

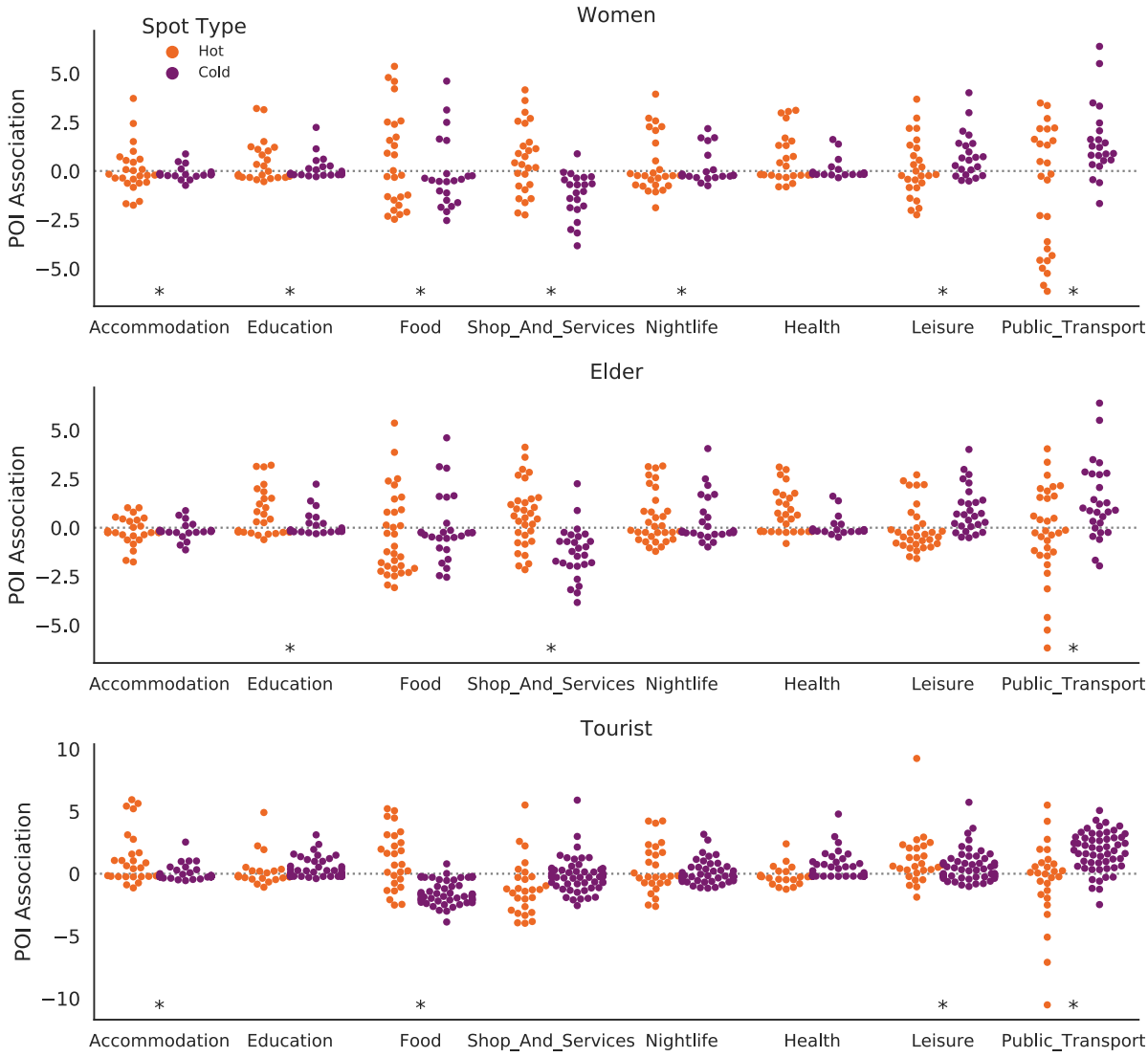


Figure 9: Swarm Plots of the POI association for each category of POI discerning hot and cold spots, for each kind of ratio. Plots marked with a star (*) indicate significant differences (according to K–S tests from Table 1) in POI association between Hot and Cold spots for the corresponding metric.

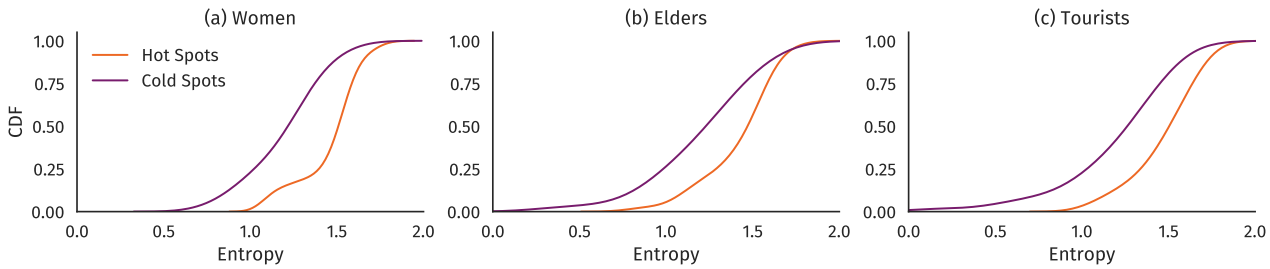


Figure 10: CDFs of the entropy for hot and cold spots on each of the ratios.

We have shown that the places visited by specific groups are related to income and the presence and diversity of amenities and services. By using mobile phone data, we were able to present these insights for the floating population of Barcelona, in contrast to and thus complementing traditional data sources that focus on the resident population only.

Our work has two main limitations. First, the analysis is bound to the market share of the mobile phone operator, which is likely to be biased toward specific socio-economic and demographic groups. Given that the data is aggregated and anonymized, we cannot control for this fact. This motivated the usage of standardized metrics to entice a clearer interpretation of our results. Second, there are intersections between the groups we analyzed, namely elderly female tourists. Hence, our analysis on income and POIs raises questions while providing preliminary answers which need further, deeper exploration, perhaps with more granular data.

In addition to improving the factors that limit the scope of this work, we devise three main lines of future work: the integration of additional area-level data sets, the definition of time-aware metrics, and multi-city analysis. Including data about crime or health would improve the characterization of hot/cold spots.

This aspect makes a time-aware analysis relevant, which would allow to measure the effect of urban interventions and seasonality according to our metrics. Finally, the issues studied here are not exclusive to one city only. In order to advance on the path to inclusive, safe, resilient and sustainable cities, quantitative methods are required to compare cities within and between them, as well as fine-grained data sets to which apply these methods to. This would allow us to distinguish between systematic subdivisions and those specific to a city.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857191 (IoTwins project). E. Graells-Garrido was partially funded by CONICYT Fondecyt de Iniciación project #11180913. We acknowledge the following libraries used in the analysis: *matplotlib* [14], *seaborn*, *PySAL* [27], *pandas* [18], and *geopandas*. Part of the map data used in this work is copyrighted by OSM contributors. Thanks to Leo Ferres for insightful discussion, and to Xavier Paradis for his help in proofreading. Finally, we thank the people from *Ajuntament de Barcelona* and Vodafone for providing access to the data and for useful discussions.

REFERENCES

[1] [n.d.]. About the Sustainable Development Goals. https://www.un.org/sustainabledevelopment/sustainable-development-goals/. Accessed: 2020-01-20.
[2] [n.d.]. Urban Mobility Plan of Barcelona. https://www.barcelona.cat/mobilitat/es/actualidad-y-recursos/nuevo-plan-de-movilidad-urbana-2019-2024. Accessed: 2020-01-20.
[3] Luc Anselin. 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis* 27, 2 (1995), 93–115.
[4] Mariano G Beiró, Loreto Bravo, Diego Caro, Ciro Cattuto, Leo Ferres, and Eduardo Graells-Garrido. 2018. Shopping mall attraction and social mixing at a city scale. *EPJ Data Science* 7, 1 (2018), 28.
[5] Mariano G Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. 2016. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science* 5, 1 (2016), 30.
[6] Orna Blumen. 1994. Gender differences in the journey to work. *Urban Geography* 15, 3 (1994), 223–245.
[7] Francesco Calabrese, Laura Ferrari, and Vincent D Blondel. 2014. Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Computing Surveys (CSUR)* 47, 2 (2014), 1–20.
[8] Sylvia Chant. 2013. Cities through a "gender lens": a golden "urban age" for women in the global South? *Environment and Urbanization* 25, 1 (2013), 9–29.
[9] Consorci Sanitari de Barcelona. 2019. La Salut a Barcelona 2018. (2019).
[10] Nathan Eagle, Michael Macy, and Rob Claxton. 2010. Network Diversity and Economic Development. *Science* 328, 5981 (2010), 1029–1031.
[11] Laetitia Gauvin, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. 2019. Gender gaps in urban mobility. *arXiv preprint arXiv:1906.09092* (2019).
[12] Eduardo Graells-Garrido, Leo Ferres, Diego Caro, and Loreto Bravo. 2017. The effect of Pokémon Go on the pulse of the city: a natural experiment. *EPJ Data Science* 6, 1 (2017), 23.
[13] Mordechai Haklay. 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design* 37, 4 (2010), 682–703.
[14] John D Hunter. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9, 3 (2007), 90.
[15] Regidoria de Feminismes i LGTBI de l'Ajuntament de Barcelona. 2016. Plan para la Justicia de Género 2016-2020. (2016). http://hdl.handle.net/11703/98743
[16] Tim Marshall. 2004. *Transforming Barcelona: The Renewal of a European Metropolis.* Routledge.
[17] Nancy McGuckin and Elaine Murakami. 1999. Examining Trip-Chaining Behavior: Comparison of Travel by Men and Women. *Transportation Research Record* 1693, 1 (1999), 79–85.
[18] Wes McKinney et al. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing* 14, 9 (2011).
[19] Graham McNeill, Jonathan Bright, and Scott A Hale. 2017. Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science* 6, 1 (2017), 24.
[20] David H Metz. 2000. Mobility of older people and their quality of life. *Transport Policy* 7, 2 (2000), 149–152.
[21] Claudio Milano, Marina Novelli, and Joseph M Cheer. 2019. Overtourism and degrowth: a social movements perspective. *Journal of Sustainable Tourism* 27, 12 (2019), 1857–1875.
[22] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis* 16, 4 (2008), 372–403.
[23] Patrick AP Moran. 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10, 2 (1948), 243–251.
[24] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. 2012. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE* 7, 5 (2012), e37027.
[25] Population Division of the United Nations. Department of Economic and Social Affairs (UN DESA). 2018. The 2018 Revision of the World Urbanization Prospects. (2018).
[26] Caroline Criado Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men.* Random House.
[27] Sergio J. Rey and Luc Anselin. 2010. PySAL: A Python Library of Spatial Analytical Methods. *Handbook of Applied Spatial Analysis* (2010), 175–193.
[28] Hannah Ritchie and Max Roser. 2018. Urbanization. *Our World in Data* (2018).
[29] Paula Vasquez-Henriquez, Eduardo Graells-Garrido, and Diego Caro. 2019. Characterizing Transport Perception using Social Media: Differences in Mode and Gender. In *Proceedings of the 10th ACM Conference on Web Science*. 295–299.
[30] Friederike Ziegler and Tim Schwanen. 2011. 'I like to go out to be energised by different people': an exploratory analysis of mobility and wellbeing in later life. *Ageing & Society* 31, 5 (2011), 758–781.

January 6, 2014

Measuring Socioeconomic Status (SES) in the NCVS: Background, Options, and Recommendations

Report

Prepared for

Bureau of Justice Statistics
U.S. Department of Justice
810 7th St NW
Washington, DC 20531

Prepared by

Marcus Berzofsky, DrPH
Hope Smiley-McDonald, PhD
Andrew Moore, MS
Chris Krebs, PhD
RTI International
3040 E. Cornwallis Road
Research Triangle Park, NC 27709

RTI Project Number 0213170.001.002.001

RTI INTERNATIONAL

Measuring Socioeconomic Status (SES) in the NCVS: Background, Options, and Recommendations

Prepared by
Marcus Berzofsky, DrPH
Darryl Creel, MS
Andrew Moore, MS
Hope Smiley-McDonald, PhD
Chris Krebs, PhD

Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the Bureau of Justice Statistics and the U.S. Department of Justice.

This document was prepared using Federal funds provided by the U.S. Department of Justice under Cooperative Agreement number 2011-NV-CX-K068. The BJS Project Managers were Lynn Langton, BJS Statistician, and Michael Planty, Victimization Unit Chief.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The National Crime Victimization Survey (NCVS) is the most important source of information on criminal victimization in the United States. Each year, data from a nationally representative sample of about 40,000 households comprising nearly 75,000 persons are obtained on the frequency, characteristics, and consequences of criminal victimization. The survey enables the Bureau of Justice Statistics (BJS) to estimate the likelihood of experiencing rape, sexual assault, robbery, assault, theft, household burglary, and motor vehicle theft victimization for the population as a whole as well as for segments of the population.

One of BJS's goals for the NCVS is to continually improve its utility so that victimization can be better understood as crime and its correlates change over time. Recently, BJS has been interested in assessing the measurement of variables that have long been associated with victimization, including factors such as socioeconomic status (SES). The goals of this paper are to (1) understand how other studies have measured SES and identify variables within the NCVS that could be used to measure or be proxies for SES, (2) explore options for creating an SES index that could enhance BJS's analysis of victimization and its correlates, and (3) assess how components of a potential SES index are currently measured in the NCVS and consider ways in which they can be improved.

Section 1 summarizes the literature on how SES has been measured in the scientific literature and how it relates to crime and victimization. ***Section 2*** summarizes the recommended approach for creating a measure of SES via an index that includes imputed income data. Finally, ***Section 3*** recommends changes to the NCVS that would address the current limitations and allow for better measurement of SES and its components.

SECTION 1. SOCIOECONOMIC STATUS—IMPORTANCE AND MEASUREMENT

1.1 Background

Socioeconomic factors are vital determinants of human behavior and functioning across the lifespan (American Psychological Association [APA], 2007). Most commonly referred to as socioeconomic status (SES) within the health, social, and criminological literature, the terminology for socioeconomic factors can vary widely (e.g., "socioeconomic position," "social disadvantage," and "socioeconomic deprivation"). Broadly, the APA describes SES as the social standing or class of an individual or group, often measured as a combination of education, income, and occupation (APA, 2007). SES also captures an individual's or a group's access to financial, social, cultural, and human capital resources (APA, 2007; National Center for Education Statistics, 2012; Shavers, 2007).

As Oakes and Rossi noted in 2003, there is a substantial gap between studies that evaluate how SES is measured and studies that include SES as a variable of interest, and this situation has not changed appreciably in the past decade. There is no standard method for measuring or deriving SES, and researchers use a variety of approaches depending on the conceptual modelor study design being employed and the data available (Braveman et al., 2005). The traditional indicators of SES are described below.

- **Income**: Gross household income is the most common measure of income used in calculations of SES. Rather than reporting salary as a continuous variable, most researchers define low, medium, and high income categories, often using the official Federal poverty line as a reference point (McLaughlin, Costello, Leblanc, Sampson, & Kessler, 2012) or dividing them into tertiles or quartiles, depending on the distribution of the sample.

- **Education**: Education is often considered a critical indicator of SES because it conveys information regarding earning potential across the lifespan, whereas income and occupation provide a snapshot of an individual's social and economic situation (Shavers, 2007).

- **Occupation**: Occupation, irrespective of salary, is a traditional indicator of SES because it is believed to convey information regarding an individual's power, income, and educational requirements associated with various positions in the occupational structure (APA, 2007). Most SES calculations using occupation specify categories of labor and rank those categories. For example, the Registrar General's Scale categorizes and ranks occupations as follows, from lowest to highest SES: Unemployed, Unskilled Manual Labor, Skilled Manual Labor, and Professional Labor (Szreter, 1984).

The traditional measures of SES—or the "big three" as they are sometimes called (National Center for Education Statistics, 2012)—can be measured at the individual, family, or household level. They have been referred to in the health literature as "compositional" approaches to measuring SES (Shavers, 2007). Other individual-level measures of SES can include indicators of accumulated wealth, which include savings, ownership of assets such as homes and vehicles, or both. Researchers who have used this additional indicator of SES argue that traditional SES indicators fail to account for contributions such as inheritances and savings, which can greatly improve an individual's social and economic situation (Kington & Smith, 1997).

When used in analyses, SES characteristics such as income and occupation are frequently used as potential confounders, correlates, or controls for examining certain phenomena. SES-related measures are sometimes used as single items or combined to create composite or index measures that can be applied to individuals, households, or both (Braveman et al., 2005; Cirino et al, 2002; Shavers, 2007). For example, some of the most commonly used SES scales (e.g., Hollingshead, Nakao and Treas, and Blishen scales; more information on these scales is provided later in this chapter) use averages or medians for determining the income in two-income households. In studies using single items as SES indicators, the main earner's occupation, educational attainment, income, or some combination of these have been used to represent the family or household (Lewis, Rice, Harold, Collishaw, & Thapar, 2011). Other studies that have examined SES in the context of families or households have taken both parents' educational attainment and occupation into account (Magklara et al., 2012).

Contextual SES measures have also been studied extensively (Shavers, 2007). Contextual measures of SES, which are designed to represent an individual's environment, can range from a neighborhood (identified via ZIP codes, census tracts, and census blocks) to areas as large as states and regions. The underlying assumption is that the physical environment has a bearing on individuals' health, behaviors, and functioning, as well as their access to goods and services (National Center for Education Statistics, 2012; Shavers, 2007). Common area-based attributes of SES include average home value, proportion of college-educated people, percentage of single-parent families, and unemployment, which have been used as single items or combined into scales. These types of methods have been used to produce values that are applied to individuals as well as households. In fact, contextual-level SES has been used to address a common problem in survey data, which is that in the absence of income data—because income typically has high nonresponse in comparison to education or occupation—contextual SES can be used as a proxy for individuals and for households.

Criminological research suggests that SES characteristics—such as employment, status, educational attainment, and income level—are associated with victimization both in the United States (Xie, Heimer, & Lauritsen, 2012) and abroad (Flatley, Kershaw, Smith, Chaplin, & Moon, 2010; Van Kesteren, Mayhew, & Nieuwbeerta, 2000). Studies like these and others (Baumer, Horney, Felson, & Lauritsen, 2003; Markowitz, 2003) show that violence is not randomly distributed across demographic or socioeconomic categories. In fact, there is evidence that victimization in the United States has become more concentrated in the poorest SES classes. For example, the crime drop in the United States in the mid-1970s was greater among upper-income households than among lower-income households (Thacher, 2004), and similar patterns have been found in Scandinavia (Aaltonen, Kivivuori, Martikainen, & Sirén, 2012; Nilsson & Estrada, 2006). SES also seems to affect the relationship between victimization and various health outcomes, such as general self-reported worsened health, pain, and anxiety (Winnersjö, Ponce de Leon, Soares, & Macassa, 2012).

The value of and interest in including SES measures in research is clear. In fact, an analysis of how SES has been used in health research yielded the conclusion that measures of economic resources, education, occupation, past SES, and neighborhood socioeconomic conditions have such a strong bearing on health outcomes that their absence in any health-related

analysis should be justified and the implications of these unmeasured factors discussed (Braveman et al., 2005). As noted above, there is no "gold standard" for deriving SES despite the growing body of multidisciplinary literature designed to address this gap that has emerged over the past decade (e.g., Braveman et al., 2005; Cirino et al., 2002; Deonandan, Campbell, Ostbye, Tummon, & Robertson, 2000; Demissie, Hanley, Menzies, Joseph, & Ernst, 2000; Shavers, 2007; Yabroff & Gordis, 2003). Indeed, the health literature alone demonstrates great variation in how SES has been measured and applied, which has, in turn, influenced the understanding about the relationships between SES characteristics and health outcomes (Braveman et al., 2005; Shavers, 2007). Unfortunately, there is a dearth of studies in the criminological literature that assess how accurately different approaches to measuring SES reflect the relationship between SES and victimization rates and crime outcomes.

The next sections summarize how SES has been measured and used in the crime and victimization literature specifically, how SES has been measured and used in other Federal surveys, how SES has been measured in the National Crime Victimization Survey (NCVS), and the ways in which alternative approaches could be used with NCVS data. SES has been conceptualized and measured using a wide variety of approaches, so rather than favoring or selecting one approach, the term "SES" is used throughout the remainder of ***Section 1*** in a somewhat general way to discuss the relationships between various socioeconomic characteristics and victimization within the criminological literature.

1.2 Use of SES in Crime and Victimization Literature

Many studies in the crime and victimization literature have established a relationship between measures of SES, crime, and victimization (Baumer et al., 2003; Faergemann Faergemann, Lauritsen, Brink, Skov, & Mortensen, 2009; Flatley et al., 2010; Markowitz, 2003; Khalifeh, Hargreaves, Howard, & Birdthistle, 2013; Van Kesteren et al., 2000; Xie et al., 2012). Although some studies show that the SES of victims varies by type of victimization, and within certain types of victimization, the findings are sometimes mixed. Generally, when the analysis focuses on serious violent victimization, the link between victimization and low SES is strong, whereas the association with SES is weaker for less serious violent acts (Aaltonen, Kivivuori, Martikainen, & Sirén, 2012; Magklara et al., 2012; Menard, Morris, Gerber, & Covey, 2011). Within the intimate partner violence (IPV) literature, Kiss and colleagues (2012) found that

women's risk of experiencing IPV was not influenced by socioeconomic factors, whereas an older review of the IPV literature concluded that SES, demographic characteristics, and alcohol use are important factors to consider in IPV analyses (Field & Caetano, 2004). Meanwhile, SES was found to be only minimally predictive of family violence and violence exposure in one study (Kassis, Artz, Scambor, Scambor, & Moldenhauer, 2012), but this relationship was robust in other research (Zinzow et al., 2009). Such variation in findings may be due in part to other research that shows that the strength of SES as a predictor for violence varies by sex. For example, a Finnish study showed that for all types of police-reported violence, violence against males was strongly associated with the SES of the offender, and male-to-female violence in private places was more associated with low SES than was violence in public places against males or females (Aaltonen, Kivivuori, Martikainen, & Sirén, 2012).

Across this body of work, individual- and household-level SES characteristics have been used to examine the SES–violence nexus. In a comprehensive review of risk factors for violence and victimization in the early 1990s, household or family income was highlighted as being inversely related to victimization, although the magnitude of the effect of family income was not as large as individual-level characteristics such as age, race, sex, and marital status (National Research Council, 1994). More recently, Khalifeh and colleagues (2013) found that lifetime physical IPV experienced by English women was associated with low SES (i.e., low household income and educational attainment, government-subsidized housing, low social class, and residence in a deprived area); physical IPV experienced by men was not associated with any of the socioeconomic characteristics studied.

Within the crime and victimization literature, it is not uncommon for researchers to use single measures to capture SES. In European studies, for example, occupation is most often used as a single-measure proxy for SES, and occupation has proven to be a more useful measure than other single measures like education (Cirino et al., 2002). Limitations associated with using occupation as a single measure include the heterogeneity associated with occupational classes, the lack of precise measurement, the difficulty in classifying certain groups (e.g., homemakers), and the racial/ethnic and gender differences in benefits that may arise from employment in the same occupation (Shavers, 2007).

Income has also been used as a single-measure indicator of SES within the crime and victimization literature, including in studies using NCVS data (e.g., Rennison & Planty, 2003). (Within the NCVS, household income is measured by a single categorical question that asks the respondent to choose from 14 different income response categories. The ranges of income choices are not uniform across the 14 categories, the upper income category is $75,000 or more, and the household respondent is asked about income every other interview wave.) In studies using NCVS income measures, income has been used as a categorical variable, as well as recoded into dummy and trichotomized variables. For example, the "20–20 ratio" has also been used to operationalize income in NCVS (Thacher, 2004) and with Swedish household victimization data (Nilsson & Estrada, 2006). The 20–20 ratio calculates the victimization rate for individuals in the poorest 20% of households and divides it by the rate for individuals in the wealthiest 20% of households. Although this method might be insensitive to changes in victimization in the middle of the income distribution, the 20–20 ratio has been found to show patterns of inequality similar to those of more sophisticated options, such as the Concentration or Gini index (which also uses a single measure of income; Thacher, 2004, p. 94).

There are limitations of using income as a single, individual-level measure for SES. Specifically, the relationship between income and violence has been shown to be nonlinear (Brownfield, 1986). Income is not an appropriate proxy for overall wealth (Braveman et al., 2005), which can provide an important buffer to problems such as unemployment, which has been found to be associated with victimization (Faergemann et al., 2009). Income is also age dependent and is typically less stable than measures of education or occupation (Shavers, 2007). Regardless of the single measure used, an added risk of using single measures is the possibility of spurious associations with outcome variables of interest.

Other research has used multiple measures to characterize SES. Khalifeh and colleagues (2013) measured individual or household deprivation across four domains, including housing tenure, household income, educational attainment, and social class. Each of the constructs was operationalized as a single ordinal categorical variable. In his analysis using NCVS data, Baumer (2002) used the 14-point ordinal income scale, the number of completed years of school, and home ownership (a dummy variable) for control variables in his analyses. In cases in which income information was missing within the NCVS, income was imputed on the basis of

education, employment status, job type, marital status, and age (Baumer, 2002). Similarly, Aaltonen, Kivivuori, Martikainen, and Salmi (2012) used three individual-level measures of SES—education, income, and unemployment history—to predict male-perpetrated violence against men and women. Unemployment was coded into three classes that took the number of days unemployed into account, including no employment, less than 1 year, and more than 1 year. A separate category was included for persons on disability retirement, retired, or otherwise outside of the workforce. Income, which was based on taxable income from both work and assets, was recoded into quintiles. Education was based on the highest degree completed.

Composite measures have also been developed to characterize SES, including the Duncan Socioeconomic Index (Duncan, 1961), Nakao and Treas Scale (Nakao and Treas, 1994), Blishen Index (Blishen, Carroll, & Moore, 1987; Pineo, Porter, & McRoberts, 1977), and Hollingshead Index (Hollingshead, 1975). Notably, the Nakao and Treas Scale, Blishen Index, and Hollingshead Index have all been applied to both individuals and to households. When applied to multiple-income households, the convention across each of these scales is to average the SES scores for multiple earners to obtain a single SES score to apply to the household and every household member.

Macro-level SES indicators have been studied in the context of victimization as well. Researchers using macro-level SES predictors suggest that understanding social structural forces is necessary to understanding root causes of phenomena of interest (Spriggs, Tucker Halpern, Herring, & Schoenbach, 2009). SES macro-level conditions may reflect an area's material resources and residents' access to municipal services, such as police protection (Cubbin, LeClere, & Smith, 2000), proximity to crime (Aaltonen, Kivivuori, Martikainen, & Sirén, 2012), and concentrations of offenders (National Research Council, 1994). The economic status of an area or neighborhood has generally been supported by social disorganization theory, in which poverty is considered a central tenet that lowers levels of social control. Other theories posit that economic deficiencies foster attitudes that support using violence and crime as a means of obtaining material goods and nonmaterial resources (Markowitz, 2003). Such studies have shown that low SES and violence are associated with neighborhood cohesion and collective efficacy (e.g., Markowitz, 2003). Although some studies have found no association between macro-level SES and victimization for IPV (Kiss et al., 2012) or violence in general (Baumer et

al., 2003), other research has found that area poverty is associated with victimization (Morenoff, Sampson, & Raudenbush, 2001; Spriggs et al., 2009) and homicide (Morenoff et al., 2001).

One review of the literature concluded that environmental or ecological SES can provide further insights into the study of risk factors for victimization, given the myriad studies showing that neighborhood levels of disruption and violence were related to victimization outcomes, even after controlling for individual-level factors such demographic characteristics and lifestyle measures (National Research Council, 1994). Baumer (2002) and Lauritsen (2001) used special releases of the NCVS data that contained census tract information to examine victimization and macro-level SES. Baumer (2002) found that neighborhood disadvantage does not significantly affect the likelihood of police notification among robbery and aggravated assault victims, whereas Lauritsen (2001) found that the persons most at risk for violence were in disadvantaged census tracts, but individual-level characteristics had a complex bearing on such risk. For example, Lauritsen (2001) showed that sex as a predictor for violent victimization was conditioned on whether the event occurred within a central-city location, occurred in an individual's neighborhood, or was limited to stranger events. Within central cities, men experienced higher rates of violent victimization than women, but outside these locations, women were as likely as men to experience violence within a mile of their homes (Lauritsen, 2001). Race was another individual-level characteristic that was predicated on location of the event (i.e., occurring in a central city, occurring in the individual's neighborhood, or being limited to stranger events). On the basis of this work, Lauritsen (2001) concluded that analyses that include multilevel SES can provide an enhanced understanding of how individual-level factors may interact with broader macro-level characteristics when violence and victimization are examined.

However, because census tracts are not available on the current, publicly available NCVS data files, creating contextual, community-level indicators of SES for the NCVS may not be feasible at this time but may be in the future. Therefore, for the purposes of this report, using macro-level SES is beyond the scope of the current effort. However, other ways of measuring SES with NCVS data are summarized in the next section.

1.3 Measuring SES in Other Federal Surveys

Poverty status, used as either a reference point for income categories or as a stand-alone measure, is also sometimes used in SES calculations. Federal agencies generally use indicators of poverty status that are based on both income and monetary resources. ***Table A-1*** in ***Appendix A*** shows how the Census Bureau and Bureau of Labor Statistics measure poverty status across the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), Panel Study of Income Dynamics (PSID), and the American Community Survey (ACS). In CPS and ACS reports in particular (e.g., Bishaw, 2012), poverty status reflects a set of income thresholds (money made before taxes, not including capital gains or noncash benefits) that vary by family size and composition to determine who is in poverty (more details of how poverty status is derived by the Census Bureau can be found in Table A-1 in ***Appendix A***). Thus, poverty status—because it does not account for occupation and educational attainment—does not represent an SES measure per se.

Notably, the Bureau of Labor Statistics and the Census Bureau have recently begun reporting on what is known as the "Supplemental Poverty Measure." Considered to be a work in progress, it adds geographic contextual data to analytic models to provide more indicators of macro-level SES (U.S. Census Bureau, 2010). Over the years, the Census Bureau has also defined several contextual variables that can be used to determine neighborhood or community characteristics, including

- ***social class***—the percentage of persons employed in 8 of the 13 Census-defined occupational groups;
- ***poverty area***—an area in which more than 20% of the persons are below the poverty level;
- ***working-class neighborhood***—a neighborhood in which more than two-thirds of employed persons work in working-class occupations; and
- ***wealth***—the percentage of households that own a home, that have one or more cars, and that have annual incomes of at least $50,000.

Others have provided insights into constructing other ways of measuring community-level SES. For example, the APA recommends deriving a community SES measure by including the percentage of individuals in the surrounding area who are unemployed, who are living at or below the Federally defined poverty level, and who lack a college degree (APA, 2007). Abroad, the United Kingdom uses deprivation indexes that assess SES in specific communities in England, Wales, and Scotland that can be applied to both individuals and households (Home Office, 2011; Page & Twist, 2011; Scottish Government, 2012).

Studies that incorporate macro-level or contextual analyses of SES have been criticized as showing contextual or group effects that may be due to the omission of individual-level variables related to the outcome or to the group characteristic under investigation (Diez-Roux, 1998). As Diez-Roux elaborates (1998, p. 219):

> [S]uppose that neighborhood violence level (measured by mean number of violent crimes in neighborhood each year) is associated with increased risk of hypertension after adjusting for age and gender. You could interpret it to mean that neighborhood violence, possibly through its effects on the stress levels experienced by individuals, is related to the development of hypertension. On the other hand, it is also possible that relevant individual-level variables have been excluded from the model and that the observed neighborhood effects are due to the low income of persons in the neighborhood who are at increased risk of hypertension because of diet, obesity, lack of exercise, and other factors and that the neighborhood effects disappear when individual-level income is included in the model.

SECTION 2. RECOMMENDATION FOR USING MEASURES OF SOCIOECONOMIC STATUS IN FUTURE REPORTS FROM THE NATIONAL CRIME VICTIMIZATION SURVEY

As noted in ***Section 1***, the relationship between SES and different types of victimization varies, which, in turn, underscores the importance of ensuring that a usable, appropriate, and meaningful SES measure is available for the NCVS. The goal of this section is to determine which variable or variables best capture the broader concept of SES. This measure could be a single measure, such as income or education level, or a derived measure that incorporates multiple components that represent SES concepts. Several considerations were evaluated before it was determined which measure best represents SES in the NCVS. The approach for determining the most appropriate measure, which is summarized below, includes a description of the process that was used and results from analyses of several potential SES measures.

2.1 Single Measure as a Proxy for SES

It is not hard to find instances in the victimization and crime literature in which single measures of SES are used. As a first step in assessing potential measures of SES, a review of all possible single measures in the NCVS questionnaire that could be used as SES proxies was conducted. For the comprehensive review of the possible single measures, all the measures that the literature clearly indicates are highly correlated with SES (e.g., income, education, occupation) were considered, as well as those that are not as well-documented in the literature but are available within the NCVS and potentially associated with SES (e.g., employment status, housing tenure). Additionally, macro-level factors, such as characteristics of the community in which the household resides, were considered.

2.1.1 Common Single Measures of SES

On the basis of the literature, the most obvious choices for single measures are the "big three" SES constructs: education, income, and occupation. The NCVS asks households about their incomes and individuals about their levels of education and current occupations.

Income. The 2010 distributions for the income measure are summarized in ***Table 2-1*** below (see ***Table A-2*** in ***Appendix A*** for victimizations rates by detailed crime categories and household income). As noted earlier, the current NCVS uses a single categorical question to

measure household income with 14 different income response choices. Household respondents are asked the income question every other interview wave. In the interview waves in which income is not asked, a carry-forward imputation method is used (i.e., the income response from the previous wave is used as the income level for the current wave). The carry-forward imputation assigns the reporting household income value to the current interview wave. For example, if a respondent reported a household income level of 3 during interview 5, an income level of 3 is assigned as the household income for interview 6.

Table 2-1. Victimization rates by type of crime and household income, 2010

Household income	Number of households	Percentage	All violent crimes (rate per 1,000 persons)	All property crimes (rate per 1,000 households)
Less than $15,000	17,185,600	14.0	28.4	159.3
$15,000–$34,999	30,206,400	24.6	22.9	132.2
$35,000–$49,999	19,406,900	15.8	18.2	121.6
$50,000–$74,999	20,965,200	17.1	17.6	109.8
$75,000 or more	35,121,200	28.6	14.9	114.4

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

As in other household surveys, such as the British Crime Survey (Home Office, 2011) and the ACS (U.S. Census Bureau, 2011), the NCVS income measure suffers from a high level of item nonresponse. In 2010, income was missing for 32.4% of households. In a separate Bureau of Justice Statistics (BJS) Working Paper, Berzofsky et al. (2014) recommend imputation methods for NCVS income data. For the purposes of the current paper, it is assumed that missing income data would be imputed before implementing any of the proposed strategies for creating an SES measure (as described in Section 3).

Table 2-1 uses the imputed household income values developed using the processes described in Berzofsky and colleagues (2014). The imputation process created five income categories that split the population into approximate quintiles. In 2010, the distribution for income across the five categories showed 14.0% of households with incomes of less than $15,000; 24.6% with incomes from $15,000 through $34,999; 15.8% with incomes $35,000–$49,999, 17.1% with incomes $50,000–$74,999, and 28.6% with incomes of $75,000 or more. Generally, as income increased, the rates for all violent and property crimes decreased.

Table 2-2 presents the unimputed distribution of income from 1998 to 2012, with every other year displayed. During this time, the distribution of income shifted from the lower income categories to the higher income categories due to inflation. Ideally, for analysis purposes, the income measure should be inflated to the current year's value using the Consumer Price Index (CPI). However, because the NCVS income variable is categorical, assumptions have to be made about the household's actual income before applying the inflation rates. These assumptions (e.g., household income has a uniform distribution within a category level) would introduce some additional error in the income estimate. Therefore, additional considerations need to be made before an inflation factor is applied. However, although the income distribution has shifted over time, the relationship between victimization and income has not. The victimization rates are higher in years before 2010, but the pattern (i.e., decreasing victimization rates as income increases) seen in Table 2-1 for both violent and property crime remains the same.

Table 2-2. Distribution of income among NCVS respondents, 1998–2012

	Income distribution by year							
Income category	**1998**	**2000**	**2002**	**2004**	**2006**	**2008**	**2010**	**2012**
Less than $15,000	21.8%	18.6%	16.9%	16.2%	14.8%	12.8%	13.8%	14.0%
$15,000–$34,999	32.2	29.8	28.3	27.2	25.0	23.6	24.7	24.6
$35,000–$49,999	17.0	17.0	17.1	16.1	15.9	16.5	16.1	15.7
$50,000–$74,999	15.4	16.7	17.0	17.3	18.4	17.8	17.5	17.0
$75,000 or more	13.5	17.8	20.8	23.1	25.8	29.4	27.9	28.7

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 1998–2012

Household income as a measure of SES can also be presented as a percentage of the Federal poverty level (FPL). The FPL for a household is a function of the household's total income and the number of people (adult and children) living in it. Using FPL rather than simply household income is an attractive measure for SES because it is used in other nationally representative surveys such as the ACS and the CPS.

Usually, when FPL is used as a measure of poverty on a survey such as the ACS, respondents provide their income as a number across several different types of income sources (e.g., on the ACS, the following income sources are included: wages, salary, commissions, bonuses, or tips from all jobs; self-employment income from nonfarm businesses or farm businesses, including proprietorships and partnerships; interest, dividends, net rental income,

royalty income, or income from estates and trusts). In contrast, the NCVS asks respondents to select one of 14 income categories, with some of the higher income levels having wide ranges (e.g., $50,000 to $74,999; see ***Table 2-3***). This makes constructing a sound poverty rate using the NCVS data challenging.

Table 2-3. NCVS income categories (question 12a)

Household income code	Income level
1	Less than $5,000
2	$5,000 to $7,499
3	$7,500 to $9,999
4	$10,000 to $12,499
5	$12,500 to $14,999
6	$15,000 to $17,499
7	$17,500 to 19,999
8	$20,000 to 24,999
9	$25,000 to $29,999
10	$30,000 to $34,999
11	$35,000 to $39,999
12	$40,000 to $49,999
13	$50,000 to $74,999
14	$75,000 or more

Source: 2010 NCVS-1 Basic Screen Questionnaire

As shown in ***Table 2-4***, the number of people who make up a household greatly affects the FPL for that household. Unfortunately, the income levels set forth in the Federal poverty guidelines do not correspond well with the NCVS income category cut points (e.g., some FPL cut points fall within an NCVS income category range). Therefore, for the distribution to be estimated accurately as a percentage of the FPL, a specific income value needs to be estimated for each household. To implement this, the distribution of income—controlling for age and race/ethnicity—from the ACS was used to generate the population parameters from a right-skewed log normal distribution (the distribution that income follows). The ACS provides income categories beyond $75,000+, allowing for better estimation of the distribution of income in households in the highest NCVS category. Specific income values were estimated on the basis of the respondent's reported or imputed income category. Once the household's actual income value was determined, it was assigned a percentage of FPL using the poverty levels for the

corresponding year of the survey (e.g., the 2010 NCVS survey year used the 2010 Federal poverty guidelines).

Table 2-4. 2012 Federal poverty level for the 48 contiguous States and the District of Columbia

Family size	Percent gross yearly income						
	50%	75%	100%	133%	175%	200%	250%
1	$5,585	$8,378	$11,170	$14,856	$19,548	$22,340	$27,925
2	7,565	11,348	15,130	20,123	26,478	30,260	37,825
3	9,545	14,318	19,090	25,390	33,408	38,180	47,725
4	11,525	17,288	23,050	30,657	40,338	46,100	57,625
5	13,505	20,258	27,010	35,923	47,268	54,020	67,525
6	15,485	23,228	30,970	41,190	54,198	61,940	77,425
7	17,465	26,198	34,930	46,457	61,128	69,860	87,325
8	19,445	29,168	38,890	51,724	68,058	77,780	97,225

Note: This table is modified from the table *2012 Federal Poverty Level* on the U.S. Department of Health & Human Services Office of the Assistant Secretary for Planning and Evaluation Web site (http://coverageforall.org/pdf/FHCE_FedPovertyLevel.pdf).

Next, in order to verify the process for assigning a percentage of FPL, the distribution for NCVS households was compared to the distribution reported by the CPS's Annual Social and Economic Supplement. ***Table 2-5*** presents this comparison for survey years 2012, 2011, 2010, 2009, and 2008. As Table 2-5 indicates, the distribution of the percentage of FPL is very similar between the two surveys for each survey year reviewed. This indicates that the process used for the NCVS is accurately assigning a household's income to its percentage of FPL category.

Table 2-5. Comparison of population distribution as a percentage of the Federal poverty level as estimated by the NCVS and the CPS, 2008, 2009, 2010, 2011, and 2012

Percentage of FPL	2008		2009		2010		2011		2012	
	NCVS	CPS	NCVS	CPS	NCVS	CPS	NCVS	CPS	NCVS	CPS
100% or less	12.3%	11.5%	12.8%	12.5%	13.8%	13.2%	14.0%	13.1%	14.6%	13.1%
101%–150%	9.2	8.5	9.8	8.6	10.0	8.7	10.1	9.1	10.2	8.9
151%–200%	9.4	8.8	9.8	8.9	9.6	8.9	9.7	9.2	10.0	9.2
201%–300%	17.2	17.4	17.9	17.4	17.2	17.1	17.5	16.8	17.5	16.5
301%–400%	13.1	14.2	13.2	13.7	12.5	13.5	12.1	14.0	12.3	13.8
401%–500%	8.9	10.9	8.7	10.8	8.6	10.9	8.5	10.2	8.1	10.7
Greater than 500%	29.9	28.7	27.7	28.1	28.3	27.7	28.1	27.6	27.3	27.8

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2008, 2009, 2010, 2011, and 2012; Bureau of Labor Statistics Current Population Survey (CPS) Annual Social and Economic Supplement, 2008, 2009, 2010, 2011, and 2012.

Table 2-6 presents the victimization rates for violent and property crime by percentage of FPL (see ***Table A-3*** in ***Appendix A*** for victimization rates by detailed crime categories and household income as a percentage of federal poverty level). In general, for both violent and property crime victimization, as the percentage of FPL increased, the rate of crime victimization decreased. For violent crime, the rate ranged from 29.5 crime victimizations per 1,000 persons with a percentage of FPL below 100% to 11.7 crime victimizations for persons with a percentage FPL 500% or greater. Similarly, for property crime, the rate was highest for households with a percentage of FPL below 100% (179.8 crime victimizations per 1,000 households) and lowest for households with a percentage of FPL of e401% to 500% (97.1 crime victimizations per 1,000 households).

Table 2-6. Victimization rates by type of crime victimization and percentage of Federal poverty level, 2010

Percentage of Federal poverty level	Number of households	Percentage	All violent crimes (rates per 1,000 persons)	All property crimes (rates per 1,000 households)
100% or less	16,979,800	13.8	29.5	179.8
101%–150%	12,226,600	10.0	23.6	166.6
151%–200%	11,842,100	9.6	22.5	129.9
201%–300%	21,132,500	17.2	19.6	117.2
301%–400%	15,335,100	12.5	19.0	103.6
401%–500%	10,564,000	8.6	17.8	97.1
Greater than 500%	34,805,100	28.3	11.7	106.0

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Education. Education has been used as a single measure of SES because it is often easier to measure in a survey than income or occupation (Shavers, 2007). As shown in ***Table 2-7*** (see ***Table A-4*** in ***Appendix A*** for victimization rates by detailed crime categories and education), only 2.2% of the 2010 data were missing for education. Nearly a quarter of respondents in 2010 indicated that they had less than a high school education (23.3%), and half (50.1%) reported having a high school degree, some college, or an associate's degree. Generally, as education increased, the rate of all reported violent victimizations decreased. However, this pattern was not entirely true for property crimes, because the rates were lowest among those with a bachelor's degree (89.5 per 1,000 households) rather than those with a master'ss, professional, or doctoral degree (103.5 per 1,000 households).

Table 2-7. Victimization rates by type of crime and education level, 2010

Education level	Number of persons	Percentage	All violent crimes (rate per 1,000 persons)	All property crimes (rate per 1,000 households)
Less than high school	59,533,500	23.3	23.8	211.7
High school or equivalent diploma, some college, or associate's degree	128,207,600	50.1	20.6	128.0
Bachelor's degree	43,868,200	17.1	14.5	89.5
Master's, professional, or doctoral degree	18,609,800	7.3	10.7	103.5
Unknown	5,742,800	2.2	7.8	53.2

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

As seen in ***Table 2-8***, the distribution of education has not shifted much during the period from 2004 through 2012. This table indicates that education level is comparable across years without any sort of adjustment factor. Furthermore, the relationship between victimization and education level has not changed over time. Specifically, although the rates themselves have fluctuated for this period, the pattern of victimization by education level (i.e., decreasing as education level goes up for violent crime and property crime) seen in Table 2-8 for 2010 is the same as other years.

Table 2-8. Distribution of level of education by year, 2004–2012

	Level of education distribution[a]				
Education level	2004	2006	2008	2010	2012
Less than high school	24.1%	23.4%	24.4%	23.8%	23.1%
High school or equivalent diploma, some college, or associate's degree	52.6	52.8	51.1	51.2	51.4
Bachelor's degree	15.6	15.7	16.2	17.5	18.0
Master's, professional, or doctoral degree	7.7	8.1	8.3	7.4	7.5

a Distribution excludes cases with an unknown value for level of education.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2004–2012.

Occupation. As a single measure for SES, occupation is often used because it reflects a person's level of education and income level. The NCVS Crime Screener Instrument (NCVS-1) asks all respondents aged 16 or older about their occupations.[1] The questionnaire allows a respondent to be placed in one of 27 different categories; however, seven of these categories are

[11] The NCVS Crime Incident Report (NCVS-2) provides a more detailed occupation measure. However, only victims are administered the NCVS-2, making it of limited value for an SES measure for the NCVS.

some form of an "other" response that does not allow for a specific occupation to be determined. In 2010, these "other" categories accounted for 51% of the responses. Furthermore, an additional 40% of respondents did not provide an answer. Therefore, the NCVS does not have useable occupation information on about 90% of its respondents. For this reason, using occupation as a measure of SES in the NCVS was not considered.

2.1.2 Alternative Single Measures for SES

Although less commonly found in the literature, other potential single measures of SES in the NCVS are worth considering. These measures include employment status and household tenure.

Employment. Unemployment, if prolonged, can be an indication of a lowered SES. The NCVS provides estimates of a person's past-week and past-6-months' employment status. Because a 1-week period of unemployment is not likely to negatively affect a person's overall SES, a person's 6-month employment status as a single measure of SES was the only employment measure considered. As shown in ***Table 2-9*** (see ***Table A-5*** in ***Appendix A*** for victimization rates by detailed crime categories and employment status), 7.4% of the 2010 data were missing for employment in the past 6 months. More than half of respondents (57.3%) were fully employed over the past 6 months, whereas more than a third (35.4%) were fully unemployed or employed only part of the time. The property crime victimization rate was higher among those who were employed during the previous 6 months than among those who were unemployed (265.8 vs. 58.3 per 1,000 population). The rates for all violent crimes were fairly similar.

Table 2-9. Victimization rates by type of crime and 6-month employment status, 2010

Employment status	Number of persons	Percentage	All violent crimes (rate per 1,000 persons)	All property crimes (rate per 1,000 households)
Employed	146,617,000	57.3	20.6	265.8
Unemployed	90,474,400	35.4	16.0	58.3
Unknown[a]	18,870,500	7.4	24.9	391.5

[a] Includes those under 18 years old

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Household tenure. Household tenure is a potential indicator of stability in a household. Families who own their homes may be less transient and have more assets and, therefore, may have a higher SES. ***Table 2-10*** shows the distribution for housing tenure (see ***Table A-6*** in ***Appendix A*** for victimization rates by detailed crime categories and housing tenure). The 2010 data show that 66.9% of respondents owned their houses and 33.1% rented their houses. The rate of violent crime among those who rented their homes was triple that of those who owned (36.1 vs. 12.0 per 1,000 households). Renters also had a higher rate of property crime than homeowners (169.4 vs. 103.6 per 1,000 households). Another housing-related measure in the NCVS that was considered as a proxy for SES was whether the household was designated as public housing (see ***Table A-7*** in ***Appendix A*** for victimization rates by detailed crime categories and public housing status). This measure was not used because in 2010 the NCVS estimated that 98% of households were not designated as public housing.

Table 2-10. Victimization rates by type of crime and housing tenure, 2010

Housing tenure	Number of households	Percentage	All violent crimes (rate per 1,000 persons)	All property crimes (rate per 1,000 households)
Own	82,203,700	66.9	12.0	103.6
Rent or no cash rent	40,681,400	33.1	36.1	169.4

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

2.1.3 Macro-level Single Measures of SES

Developing macro-level SES indicators that would provide a larger context for respondents' living situations was considered, but this approach was eventually concluded to be infeasible. Census tracts are commonly used in other studies to categorize neighborhoods and communities in SES terms. Although other researchers have been able to take advantage of special releases of NCVS data that included census variables indicating State, county, and the census tract in which the respondents reside (e.g., Lauritsen, 2001; Baumer, 2002), these data sets are outdated. Future special releases of census tract data or data that include ZIP codes are not anticipated; thus, to the extent that publicly available NCVS data are the only data that can be used, this avenue is not possible at this time.

Although urbanization level could be included, once the Census place size in which the sampled household resides is accounted for in the composite, the urbanization variable would not

add much because the information it provides is redundant with land use. Although it might be possible to identify more generic areas (e.g. large cities in the Northeast) by combining region and census land use population) to develop some average costs of living, this result might not be easily interpretable. Moreover, the health literature has shown that contextual variables do not often correlate well with individual measures (Shavers, 2007). Therefore, adding macro-level measures of SES was not considered.

2.2 Composite SES Measure for NCVS

Given the relationship between SES and victimization, it is clear that SES is not something that should be ignored when studying the relationship between characteristics of households and individuals and violent and property crime victimization. However, the numerous data limitations associated with income and occupation bring to bear several challenges with using single measures or constructing a poverty level that may be both appropriate and meaningful for any analyses that account for SES.

2.2.1 Factors Considered for Composite SES Indexes

The alternative to using a single-measure proxy for SES is to develop a household-level SES composite measure that incorporates the best SES elements that are available in the NCVS. All of the relevant NCVS variables that could be used to measure SES were considered, and the narrowed list of possibilities for the composite index included the following individual- and household-level characteristics:

- income as a percentage of FPL (reported and imputed)—household level
- education—individual level
- housing tenure (owned or being bought; rented for cash; no cash rent)—household level
- housing type (public housing vs. not)—household level
- employment in the last 6 months—individual level

As noted above, it was not possible to include occupation in the SES index. Employment status in the last week and employment in the last two consecutive weeks were considered as potential variables, but from a theoretical perspective, the 6-month perspective was more relevant and meaningful for the index, and the distribution of this variable for 2010 was sound enough to warrant inclusion and was supported by the victimization literature (e.g., Aaltonen, Kivivuori, Martikainen, and Salmi, 2012; Faergemann et al., 2009). Furthermore, including the measure related to the number of cars owned was considered, but it was generally concluded that it was not a good measure of assets (e.g., wealthy people in large cities like New York generally do not own cars, whereas those in poverty-stricken areas may own several aging cars). The literature indicates that assets are important for determining SES, and in the absence of other measures, using housing tenure has been supported by other studies using NCVS data (Baumer, 2002). Finally, adding household size to the SES index was also considered. Household composition is generally not included in SES index measures, but it is a factor in determining how the indexes are applied or used in analyses. To that end, the next section describes how household composition was used in calculations.

2.2.2 Considered SES Indexes

The goal of this section is to describe an SES composite measure that could be applied to all members of the household, as is done with other SES indexes (Cirino et al. 2002, Nakao & Treas, 1994; Blishen et al., 1987; Pineo et al., 1977; Hollingshead, 1975). In these scales, for families with multiple persons 18 years old or over, the individual SES scores are averaged to obtain a single SES score to apply to the household and everyone in it who is at least 18 years of age. Different potential household structures were taken into account by using averages for all persons 18 years and older in the household. For example, households with retirees, homemakers, or students over 18 would have their SES reduced because of lower income levels, but the household's SES would be increased if the education level of these individuals is relatively high (e.g., a retired person or homemaker with a college education or greater). Furthermore, even though 12- to 17-year-olds may provide some little income (through summer job, etc.) to the household, this group was excluded from the indexes because their income and education level would artificially dilute the average of their parents or guardians.

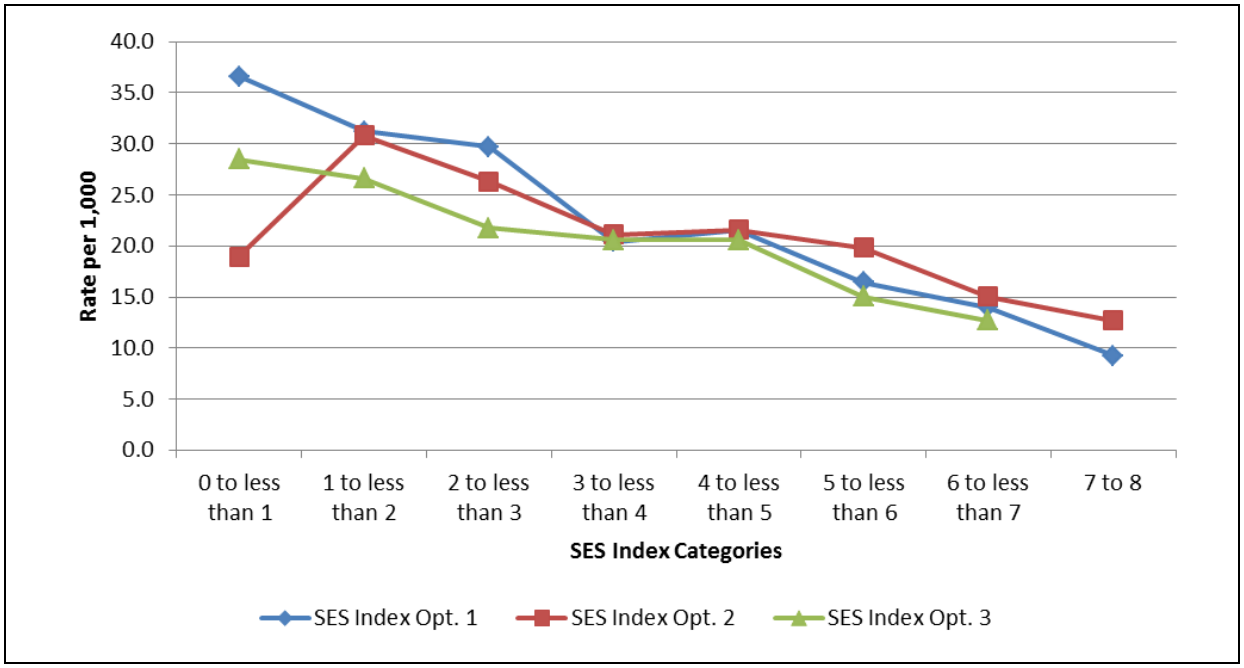
Three possible index options based on the variables listed in the bullets above were constructed. ***Table 2-11*** presents these three possibilities across the constructs examined. The SES indexes are weighted on the basis of the number of levels attributed to each characteristic. For example, in Index 1, income (as a percentage of FPL) and education have four levels, whereas employment and housing only have two. Therefore, income and education have equal weight and contribute two times more than employment and housing. Another approach is to assign a particular percentage of the index's weight to each characteristic (e.g., income counts as 50% of the score). This approach was not used because a suitable reference to what those weights should be was not identified.

Table 2-11. SES index options for NCVS

Measures	Index 1	Index 2	Index 3
Education	▪ 0: Less than high school ▪ 1: High school, some college, associate's degree ▪ 2: Bachelor's degree ▪ 3: Master's, professional, doctorate degree Possible range: 0–3	▪ 0: Less than high school ▪ 1: High school, some college, associate's degree ▪ 2: Bachelor's degree ▪ 3: Master's, professional, doctorate degree Possible range: 0–3	▪ 0: Less than high school ▪ 1: High school, some college, associate's degree ▪ 2: Bachelor's degree ▪ 3: Master's, professional, doctorate degree Possible range: 0–3
Income (percentage of Federal poverty level)	▪ 0: 100% or less ▪ 1: 101%–200% ▪ 2: 201%–400% ▪ 3: 401% or greater Possible range: 0–3	▪ 0: 100% or less ▪ 1: 101%–200% ▪ 2: 201%–400% ▪ 3: 401% or greater Possible range: 0–3	▪ 0: 100% or less ▪ 1: 101%–200% ▪ 2: 201%–400% ▪ 3: 401% or greater Possible range: 0–3
Employment	▪ 0: Unemployed past 6 months ▪ 1: Employed past 6 months Possible range: 0–1	▪ 0: Unemployed past 6 months ▪ 1: Employed past 6 months Possible range: 0–1	▪ 0: Unemployed past 6 months ▪ 1: Employed past 6 months Possible range: 0–1
Housing	▪ 0: Rent or no cash rent ▪ 1: Own Possible range: 0–1	▪ 0: Public housing ▪ 1: Non-public housing Possible range: 0–1	Not included
Possible range	0–8	0–8	0–7

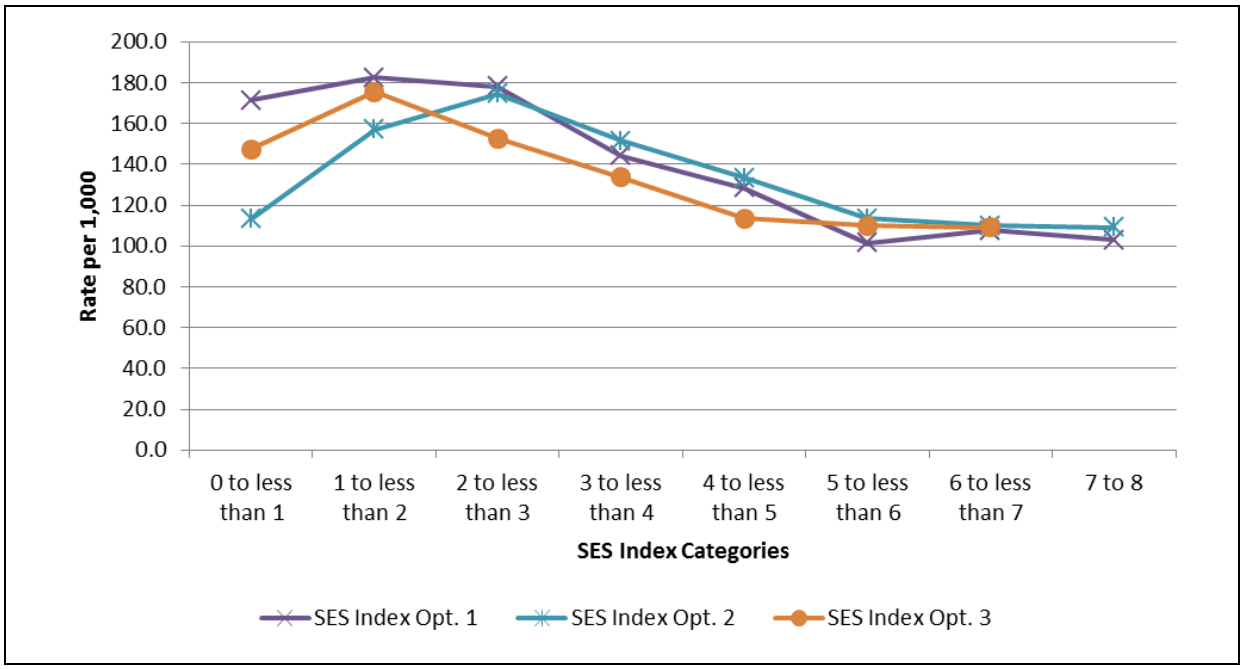
Figures 2-1 and ***2-2*** present victimization rates by the three SES options for violent and property crime, respectively. The figures show that the SES index options generally follow the same pattern in terms of their relationships with violent and property crime victimization. ***Table 2-12*** presents the weighted percent distribution and unweighted sample sizes of respondents by SES index level for each SES index option. In general, each level of the SES indexes has a large enough sample size so that suppression is not a concern. The smallest category occurs in SES index 2 for households with an index of 0 or 1 (301 respondents, 0.4% of weighted respondents).

Figure 2-1. Violent crime victimization rates by SES for three index options, 2010



Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Figure 2-2. Property crime victimization rates by SES for three index options, 2010



Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table 2-12. Unweighted sample size and weighted percent distribution of respondents by SES index options, 2010

	Index 1		Index 2		Index 3	
SES level	Unweighted sample size	Weighted percent	Unweighted sample size	Weighted percent	Unweighted sample size	Weighted percent
0–1	1,582	2.0	301	0.4	2,531	3.1
1–2	4,662	5.9	2,752	3.4	7,100	8.8
2–3	8,269	10.3	6,984	8.7	11,496	14.2
3–4	11,265	13.8	11,315	13.9	14,446	17.6
4–5	14,497	17.7	14,336	17.5	17,191	20.9
5–6	15,780	19.1	17,082	20.8	15,685	19.0
6–7	14,043	16.9	15,665	19.0	10,830	13.2
7–8	9,181	11.1	10,820	13.2	n/a	n/a
Missing	2,669	3.2	2,693	3.3	2,669	3.2

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table 2-13 presents the correlation matrix between each of the characteristics considered in one of the SES indexes. This table shows that none of these characteristics have a correlation with another characteristic greater than 0.35. Furthermore, all of the correlations are positive, except for the correlation between 6-month employment status and household tenure (which is near zero). The relatively small correlation between all of the characteristics suggests that there may be some benefit in using an index. Furthermore, the fact that no correlation between any two characteristics exceeds 0.35 indicates that none of the characteristics are redundant with each other in terms of explaining SES. In short, using all of these data in an index can capture an individual's SES better than any one of these characteristics individually.

Table 2-13. Correlation matrix between NCVS characteristics considered for an SES index

	SES characteristic				
SES characteristic	Household income (percentage of FPL)	Education level	Employment status	Household tenure	Public housing
Household income (percentage of FPL)	1.0000	0.3476	0.2021	0.3046	0.1366
Education level		1.0000	0.1920	0.1416	0.0821
Employment status			1.0000	-0.0122	0.0755
Household tenure				1.0000	0.1946
Public housing					1.0000

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table 2-14 presents the results of logistic models that regress crime victimization status on the SES index components. Separate models were run at the person level with violent crime and property crime as dependent variables. Models were run at the person level because highest level of education and 6-month employment status are person-level attributes. For the property crime model, if the reference person reported a property crime, all persons in the household were considered victims of a property crime. In both models, all four SES index components – income as a percent of the Federal poverty level, level of education, employment status, and household tenure – were significant predictors of crime victimization.

Table 2-14. Logistic regression of violent and property crime victimization by SES index characteristics

	Violent crime				Property crime[a]			
Index characteristic	**Odds ratio (OR)**	**OR lower bound**	**OR upper bound**	**Chi-square Wald p-value**	**OR**	**OR lower bound**	**OR upper bound**	**Chi-square Wald p-value**
Intercept	0.00	0.00	0.00		0.05	0.05	0.06	
Federal poverty level								
100% or less	1.58	1.28	1.95	0.0002	1.35	1.19	1.53	<0.0001
101%–200%	1.27	1.02	1.58		1.25	1.12	1.40	
201%–400%	1.09	0.90	1.33		0.97	0.88	1.07	
Greater than 400%[b]	1.00	1.00	1.00		1.00	1.00	1.00	
Education								
Less than high school	1.63	1.10	2.42	0.0004	1.31	1.16	1.49	<0.0001
High school or some college	1.69	1.19	2.39		1.19	1.07	1.32	
Bachelor's degree	1.28	0.84	1.95		1.05	0.93	1.18	
Master's, professional, or doctoral degree[b]	1.00	1.00	1.00		1.00	1.00	1.00	
Employment								
Unemployed in past 6 months	0.80	0.69	0.94	0.0057	0.75	0.70	0.79	<0.0001
Employed in past 6 months [b]	1.00	1.00	1.00		1.00	1.00	1.00	
Household tenure								
Rent or no cash rent	2.45	2.08	2.90	<0.0001	1.45	1.34	1.58	<0.0001
Own [b]	1.00	1.00	1.00		1.00	1.00	1.00	

[a] Model computed at the person level because education and employment are measured at the person level.

[b] Comparison group.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Tables A-8 through ***A-15*** in ***Appendix A*** present the crosstabs for each pair of index characteristics considered. In general, these crosstabs reflect the expected relationship between the index variables. For example, as income increases, the percentage of people that live in a household-owned home increases (Appendix Table A-9). However, some relationships are not as clear. For instance, the percentage of persons who are unemployed does not vary across household tenure (Appendix Table A-15). This result could be because this bivariate relationship does not take age into account. Therefore, it is possible for a younger person with a higher education level to earn less than a person who has a lower degree but is older and has worked longer. Furthermore, the income measure is a household measure, whereas the education measure is at the person level. Therefore, a higher wage earner with a higher education level in the household could mask the presence in the household of other adults with lower education levels. Situations like this are possible explanations for why the correlations between the variables are not higher than one might expect.

As Table 2-11 shows, education, income, and employment status were measured consistently across each of the three SES index options. Education was measured with four categories and a possible range of 0–3 (0 = less than high school; 1 = high school, some college, or associate's degree; 2 = bachelor's degree; 3 = master's, professional, or doctoral degree). Income was measured as a percentage of a household's FPL with four categories (0 = 0 – 100%, 1 = 101% - 200%, 2 = 301% - 400%, 3 = 400% or more). Employment status was measured based on whether a person was employed in the past 6-months (0 = not employed, 1 = employed). Beyond education, income, and employment status, the three SES index options differ in terms of what measures they include. Namely, SES Index 1 additional includes household tenure, SES Index 2 additionally includes public housing status, and SES Index 3 does not include any housing measure. As illustrated in Figure 2-1, although the differences in the index options did not alter the relationship between SES level and victimization rates, they could have substantive differences in how the levels of SES are interpreted. Therefore, comparing the SES index options on their substantive merits is worthwhile.

SES Index 1. Option 1 measures household income using the income as a percentage of FPL categories, with a range of 0–3, being collapsed as follows: 0 = 100% or less of FPL, 1 = 101% to 200%, 2 = 201% to 400%, and 3 = 401% or more. For the index, the number of income

as a percentage of FPL categories was collapsed from seven to four to reduce the weight that income has in the index when scores are summed across measures. Index 1 also incorporates measures of education, housing tenure, and 6-month employment status. The housing tenure measure seems preferable because it represents assets held by respondents, which has been identified as an important SES context in the literature (e.g., Braveman et al., 2005; Shavers, 2007).

Table 2-15 provides the victimization rates by type of crime and SES Index 1 categories (see ***Table A-16*** in ***Appendix A*** for more detailed crime categories for SES Index 1). Index 1 has an overall range of 0–8 for SES. As shown in Table 2-15, the rates for violent crimes decrease as SES increases. Property crimes show less of a pattern, ranging from 101.5 per 1,000 population (among those in the SES category 6) to 182.4 per 1,000 population (among those classified in SES category 2).

Table 2-15. Victimization rates by type of crime and SES Index 1, 2010

SES Index 1 categories[a]	Number of households	Percentage	All violent crimes (rate per 1,000 person)	All property crimes (rate per 1,000 household)
1	2,436,200	2.0	36.6	171.2
2	7,222,600	5.9	31.2	182.4
3	12,662,800	10.3	29.7	178.1
4	17,003,500	13.8	20.4	144.2
5	21,703,300	17.7	21.5	128.5
6	23,446,500	19.1	16.4	101.5
7	20,783,400	16.9	14.0	107.7
8	13,641,300	11.1	9.3	102.8
Unknown	3,985,600	3.2	6.2	40.4

[a] Because the SES index is averaged over all adults in the household, it does not result in whole numbers. The categories represent the results as follows: 1 = 0 to less than 1, 2 = 1 to less than 2, 3= 2 to less than 3, 4 = 3 to less than 4, 5 = 4 to less than 5, 6 = 5 to less than 6, 7 = 6 to less than 7, and 8 = 7 to 8.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

SES Index 2. Index 2 measures household income in the same way as Index 1 and also incorporates education and 6-month employment status. However, rather than including housing tenure, Index 2 includes whether the household is designated as public housing. The distribution of SES levels and the victimization rates by SES category are similar to Index 1. Therefore, detailed victimization rates are presented only in ***Appendix A, Table A-17***.

SES Index 3. Index 3 includes education, household income (as defined in Indexes 1 and 2), and 6-month employment status, but measures related to housing are excluded (see ***Appendix A, Table A-18*** for the detailed victimization rates).

2.2.3 Assessing the Quality of the SES Indexes

An index that measures some sort of latent variable or a variable that cannot be directly measured through a survey question must be assessed for quality to ensure that the results generated during its development are reproducible across similar samples and are not solely a function of the data used to create the index. For some indexes or scales that are based on reflective models, it is possible to look at the "internal consistency" of the index items. In these cases, all the items used in the index are consistently measuring the latent variable—that is, if the latent variable's value is changed, all the observed variables have similar changes in response. In such instances, it is possible to use a statistic, like Cronbach's alpha, to verify that all the observed items used to measure the latent variable are internally consistent.

Unfortunately, the indicators being used to measure SES and other latent variables are not reflections of the latent variable; they are components used to construct or form the latent variable. The models for such constructs are called formative models. Constructs typically measured with formative models are stress scales and SES indexes. In these models, the observed variables drive the latent variable by constructing it, rather than the other way around. The model does not assume that the observed variables consistently reflect any change in the latent variable, resulting in high correlations. In formative models, the observed indicators need not be correlated—in fact, they generally are not. For example, there are circumstances in which a person with a high level of education does not have a high level of income (e.g., the person has not been in the workforce very long). For this reason, an alternative approach to assessing the quality of the index needs to be employed.

Measuring the quality of a formative model is difficult. The observed indicators are not assumed to be correlated, so it is not possible to use a minimum level of correlation, as Cronbach's alpha does, to evaluate the measure. There is no error term in the model, so model fit cannot be used to determine whether the model is measuring what it is thought to measure. It is possible, however, to ask whether the model—that is, the loadings of the index on the observed

indicators—is consistently estimated across random subsets of the data. Put another way, this approach would see whether the relationships between indicators and index that form the loadings in the model are largely sample dependent (and therefore not consistent enough to be considered a useful model) or sample independent (consistently demonstrating a similar relationship between indicators and the index).

By implementing such an analysis, it is possible to see whether the correlations between the index measure and the items used in the index are consistent across samples. That is, if another random sample of households was provided (e.g., a year other than 2010), would the correlations between the index value and the item characteristics be the same?

In this type of analysis, the actual correlation is not as important as whether a similar correlation is produced across each of the samples. When a survey like the NCVS utilizes a panel design where households appear in multiple years, in order to ensure an independent set of comparison households, the correlation can be tested through the use of split samples. To implement this approach, the sample of households by interview (i.e., a household's two interviews during 2010 were not tied together for randomization purposes) was randomly split into two samples. Persons interviewed within a household were all assigned the same random sample. Correlations were weighted on the basis of the level of the characteristic (i.e., household income, tenure, and public housing used the household weight, whereas education level and employment status used the person-level weight).

Table 2-16 presents the results from this analysis. All indexes were tested for measuring SES in the NCVS; the correlations were consistent across both samples for all index items. For example, the correlation between household income as a percentage of FPL and SES Index 1 was 0.8546 in the first sample and 0.8571 in the second. Across all the indexes, the largest absolute difference between a pair of sample correlations was 0.0137 (employment status and SES Index 2). The small difference in the correlations leads to the conclusion that any of the three SES indexes would produce consistent results across NCVS data collection years.

Table 2-16. Correlations among SES index options and index characteristics by split sample

Characteristic	SES Index 1		SES Index 2		SES Index 3	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
Household percentage of Federal poverty level	0.8546	0.8571	0.8571	0.8586	0.8593	0.8610
Household tenure	0.4792	0.4887	--	--	--	--
Education	0.5499	0.5532	0.5750	0.5769	0.5779	0.5798
Employment	0.3495	0.3612	0.3899	0.4036	0.3903	0.4034
Public housing	--	--	0.2550	0.2608	--	--

— Not a component in the SES index.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

2.3 Conclusions and Recommendation

Both single measures and composite measures of SES in the NCVS provide advantages and disadvantages for interpreting the relationship between crime victimization and SES. Single measures offer an easily interpretable definition of SES, but they may not fully capture all aspects of SES. Composite indexes may offer a more complete representation of a household's SES, but their levels may be less interpretable. However, after considering the analyses presented in this section and the literature presented in ***Section 1***, the following conclusions and recommendations have been developed:

- Using single measures of SES increases the risk for spurious associations with outcome variables (i.e., crime victimization types) because single measures do not control for any other factors related to SES that may alter the relationship.

- If one had to identify one of the "big three" SES constructs to stand as the single measure proxy for SES, the literature suggests that occupation might be the best choice (e.g., Cirino et al., 2002; Fujishiro, Zu, & Gong, 2010). However, as noted earlier, current data restrictions limit the ability to use occupation for the NCVS.

- Using income as a single measure for SES is problematic because (1) it is age dependent; (2) it is relatively unstable compared with education or occupation; and (3) it does not necessarily account for the impact of wealth or other financial assets and resources (Shavers, 2007, p. 1014).

- It is possible to generate a household income as a percentage of FPL measure using the imputed income variable. FPL takes into account household size and, therefore, is a better measure of household wealth than an income measure that does not control for household size. It is recommended that percentage of FPL be used in any SES index for the income component.

- Myriad limitations surround education as a single measure for SES (Shavers, 2007, p. 1014). First, education has different social meanings and consequences in different cultures. Second, it is known that economic returns differ across racial/ethnic and gender groups because minorities and women realize lower returns than white men with the same educational backgrounds (Shavers, 2007; Braveman et al., 2005). Third, the relationship between SES and education is not consistently linear and may change over time (Shavers, 2007, p. 1014).

- Consistent with other studies (Braveman et al., 2005), income and education were not correlated strongly enough to justify using one as a proxy for the other (correlation of 0.3484 from Table 2-13).

- At best, the 6-month employment status and public housing status variables as they are currently identified in the NCVS survey provide a gross measure for employment. These variables are valuable to the extent that they can provide some SES context, but they are not strong or detailed enough to stand alone as SES single measures.

- None of the correlations between any two possible single measures of SES in Table 2-13 are above 0.35. This indicates that a composite measure or index that incorporates some of them may better represent SES and that no single measure is providing redundant information with another single measure.

- As seen in Table 2-14, two logistic regressions with any violent crime or any property crime occurring as the dependent variables, respectively, and the four proposed SES index characteristics as independent variables indicate that, given the other characteristics in the model, all four SES index characteristics are strongly associated

with victimization (i.e., p-values less than 0.001 for each SES index measure for both violent and property crime victimization).

- Index 2 includes public housing (public housing or not), in which only 2.1% of the 2010 sample indicated living (see Table A-7 in Appendix A for the public housing data by victimization types). The resulting analysis shows that this measure is not as meaningful as the household tenure measure and thus is not beneficial in an SES index.

- Between Index 1 and 3, Index 1 is preferable because it includes the housing tenure measure, which enables some measure of assets held by respondents, which has been identified as an important component of SES in the literature (e.g., Braveman et al., 2005; Shavers, 2007).

Therefore, the final recommendation for how a SES index should be constructed in the NCVS, given the limitations and quality of available data and assuming that the plan is to create an SES index, is SES Index 1.

As noted earlier, one key deficiency of an SES index is the interpretability of its levels. In order to alleviate this issue for the recommended index, it is possible to collapse the eight SES categories in Table 2-15 categories into low, middle, and upper SES when needed. ***Table 2-17*** presents a possible trichotomy with the 2010 data. The percentages, if graphed, would show a bell-curve, with about a fifth of all households falling into the low SES category (18.2%), 50.6% in the middle SES category, and 28.0% representing the highest SES category. Using the index in this manner will facilitate the creation of cross-tabulations and the interpretation of the relationships between various criminal victimization measures and SES.

Table 2-17. Option for reclassifying SES Index Option 1 into three categories

SES Index Option 1 categories	Percentage[a]	Collapsed income category	Combined percentage
1	2.0		
2	5.9	Low SES	18.2%
3	10.3		
4	13.8		
5	17.7	Middle SES	50.6%
6	19.1		
7	16.9	High SES	28.%
8	11.1		

Note: Because the SES index is averaged over all adults in the household, it does not result in whole numbers. The categories represent the results as follows: 1 = 0 to less than 1, 2 = 1 to less than 2, 3 = 2 to less than 3, 4 = 3 to less than 4, 5 = 4 to less than 5, 6 = 5 to less than 6, 7 = 6 to less than 7, and 8 = 7 to 8.

[a] Percentage distribution does not sum to 100 because cases with unknown SES are excluded

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2006–2010.

2.4 Operationalizing SES in Future Reports

One option is that the proposed SES Index 1 could be incorporated as a derived variable on the NCVS public use files (PUFs) available through the Interuniversity Consortium for Political and Social Research (ICPSR). As a component of the SES index, the imputed household income variables, along with a flag indicating which values were imputed, would be included on the PUF as well. This would need to be done in a two-step process: imputing household income (as described by Berzofsky et al., 2014) and creating the SES index as it was defined in ***Section 2.3***. Given that both of these steps were developed using PUFs, both of these steps could be conducted by BJS; the data collection agent, the U.S. Census Bureau; or a private contractor.

Both of these steps could be conducted during the development of the PUF. Based on the procedures developed to impute household income (Berzofsky et al., 2014) it is anticipated that implementing the imputation process for a given year's data would take about 2–3 weeks, including time for quality assurance. The derivation of the SES index could follow immediately after the imputations were complete and, therefore, would not add much time to the development schedule. Another option is that if an imputed income variable and a measure of FPL were added to the PUF, data users could independently create the SES Index.

2.5 How to Use the SES Index When Analyzing Crime Victimization Data

The derived SES index is a household level variable with the same value being applied to all persons residing in that household during a given interview wave (i.e., the SES index could change for a household across interview waves on an annual file). This variable could be merged onto the person file and incident file by household ID and year/quarter of interview, or it could be included during the creation of the PUFs.

If having eight levels of SES is too cumbersome for the analysis being conducted, the proposed index can be collapsed into three levels, as suggested in Table 2-17. Collapsing SES into these levels may make the interpretation of any crosstabs more easily understood by readers.

SECTION 3. IMPROVING THE MEASUREMENT OF SOCIOECONOMIC STATUS IN THE NATIONAL CRIME VICTIMIZATION SURVEY

The recommendations outlined in this section are put forth with the understanding of the current data restrictions. The section includes a summary of recommendations for modifying the current NCVS instrument, with an eye toward changes that would improve data elements related to SES and enhance the reporting of SES information in the future. The scope and cost of these recommendations vary widely, from simply moving a question (e.g., Recommendation 2) to conducting a field test (e.g., Recommendation 1).

Recommendation 1: Conduct a Field Test of Procedures That Would Yield a Revised and Improved Income Variable.

As noted earlier, the NCVS is not unique in its relatively high level of nonresponse to the question about household income. In 2006, the Centers for Disease Control and Prevention, conducted a field test of approaches for improving income response rates in the National Health Interview Survey (NHIS), a Federal household survey with income nonresponse rates exceeding 30%. For the test a sample of the NHIS respondents was selected to receive a new set of income questions, redesigned to reduce item nonresponse (Pleis & Cohen, 2007). The following paragraphs draw heavily from Pleis and Cohen's methodological report on this field experiment, which may be accessed at http://www.cdc.gov/nchs/data/nhis/income.pdf.

NHIS Field Test. As Pleis and Cohen (2007) describe, the original income question in the 1997–2006 NHIS instrument was, "*Now I am going to ask about the total combined income {for you/of your family} in {last calendar year}, including income from all sources we have just talked about such as wages, salaries, Social Security or retirement benefits, help from relatives, and so forth. Can you tell me the amount before taxes?*" In the 1997-2006 surveys, when respondents initially refused to answer the first exact income amount question a series of follow-up questions asking about income ranges. The closed-ended questions were designed so that each successive question homed in on a smaller range for the family's total income and identified whether families were reporting income below the poverty threshold. Specifically, if the respondent did not provide an answer to the exact amount question, the respondent was asked to provide the family' income in relation to $20,000 (greater or equal to $20,000 or less than $20,000). If an answer was given to this question, the respondent was provided with a list of

income intervals and asked to report the appropriate income interval. If the family's income was less than $20,000, the respondent was shown a list of intervals in $1,000 increments from $0 to $19,999. If the family's income was $20,000 or more, the respondent was shown a list of income intervals in $1,000 increments from $20,000 to $34,999 and in $5,000 increments starting at $35,000, up to a final category of $75,000 or more.

As part of the experiment, the follow-up questions used in the 1997-2006 NHIS were replaced with a different series of unfolding bracket questions with closed-ended income ranges when the respondent did not answer the exact income amount question. Rather than starting with $20,000, the experimental questions began with $50,000. Respondents who reported incomes of less than $50,000 were then asked if their income was less than $35,000. Among respondents reporting family income below $35,000, the poverty threshold for the family was prefilled by the computer-assisted instrument using the information on the family's size collected earlier in the interview. Respondents were then asked about the family's income in relation to the prefilled poverty threshold dollar amount. When the reported income was greater than $50,000, respondents were asked if it exceeded $100,000; if not, then respondents were asked if it exceeded $75,000 and the series of questions ended for the higher incomes.

The NHIS research team compared the percentages of unknown responses when calculating the poverty ratio, the percentage distributions of the poverty ratio for selected sociodemographic characteristics, and the percentage distributions of selected sociodemographic characteristics by poverty ration category, comparing results from the fourth quarter of the 2006 NHIS and the first quarter of the 2007 NHIS. The findings suggest that the weighted percentage of unknown responses for income for the "as usual" NHIS group was 29.6%, whereas the experimental group had a weighted percentage of 16.0%. The positive results from the experiment hastened the implementation of these revised questions starting with the 2007 NHIS.

The most current NHIS survey asked respondents who did not know or refused to state an income amount if their family's combined income in the previous calendar year was $50,000 or more, or less than $50,000. If they refused to answer or indicated that they did not know, they were not asked any additional income questions. If they answered the $50,000 question, then the survey followed the pattern described above as experimental (Sondik, Madans, & Gentleman,

2012). Thus, NHIS respondents were categorized into one of four income categories (Sondik et al., 2012): (1) those who supplied a dollar amount (83% of sample adults in 2011), (2) those who indicated a range for their income by answering all of the applicable follow-up questions (11% of sampled adults), (3) those who indicated a less precise range for their family's income by answering only some of the applicable follow-up questions (2% of sampled adults), and (4) those who provided no information about their income (4% of sampled adults). Based on these data, the Centers for Disease Control and Prevention is able to provide a poverty rate as defined by the Census Bureau.

Notably, the NHIS income questions are not analogous to what is currently in the NCVS. The current NCVS instrument has one income question, which asks the respondent to choose from 14 different income response categories. BJS may consider revising this lead-in question altogether, following up refusals with questions that that mirror the types of questions that the NHIS eventually adopted (e.g., *Was your total family income from all sources less than $50,000 or $50,000 or more?*), or both.

Implementing a similar experiment within the NCVS might be fruitful in light of the high item nonresponse rates. Given the importance of this variable and the high-profile nature of the NCVS, merely changing the income variable without testing the response is not recommended. It is possible to design and implement an effective experiment that reflects this recommendation.

Recommendation 2: Move the Income Question to Follow the Employment Section.

Currently, the income question is located near the beginning of the survey (question #12 on page 1). The current placement of income is far from ideal; the question is asked too early in the interview. A best practice in survey methodology is to ask sensitive questions like income toward the end of the survey, the rationale being that this provides interviewers ample time to develop rapport with the respondent and for the respondents to gain some comfort with the process and their responses (Sudman & Bradburn, 1982). For example, on the 2013 ACS and the 2013 National Survey on Drug Use and Health (NSDUH) surveys, the income questions are

last.[2] The rationale for moving the income question to follow the employment section is based on the following.

- The respondents will already be thinking of potential income sources as they think about their past employment, and this line of questions naturally follows this topic.

- The employment questions are located toward the back of the instrument, starting with question 47b; thus, any gains made in the rapport-building efforts by the interviewer and the related comfort levels of the respondents might have some positive effect on the item response rate.

Implementing this recommendation would presumably require some minor survey methodological work, training of the interviewers, and possibly overhauling some existing infrastructure at the Census Bureau. However, this may be one of the lowest-cost strategies to reduce the item nonresponse rate for income.

Recommendation 3: Expand the Upper-Bound Income Category.

Currently, the upper household income category in the NCVS is $75,000 or more. This category is the most common, and the distribution is highly skewed to the right, as demonstrated by Berzofsky and colleagues (2014). Unfortunately, the top tier income category does not provide much usable context given the wide range of SES that could be present at that income level. One-person households reporting an income of $75,000 or more may be very different from a five-person household reporting that same income. Thus, splitting this tier into additional categories to allow for finer grained analysis of victimization by income level, as well as a more accurate SES index measure, is strongly recommended.

Some Federal surveys (e.g., ACS) ask respondents to enter their total income rather than select a category, which means that, hypothetically, there is no upper-bound income level for

[2] To see these surveys, access the 2013American Community Survey at http://www.census.gov/acs/www/Downloads/questionnaires/2013/Quest13.pdf; see question 47at the end of the survey. The 2013 NSUDH survey may be accessed at http://www.samhsa.gov/data/2k12/NSDUH2013MRB/NSDUHmrbCAIquex2013.pdf; see pages 426–433 for the income variables.

those surveyed. As another example, the top two income categories on the 2013 NSDUH were (1) $75,000 to $99,999 and (2) $100,000 or more.

Should BJS decide to stay with the current income question as it is currently worded, adding at least one and possibly two or three additional income categories above the current top tier of $75,000 or more is recommended. The inclusion of more income categories at the top may also correct the skewed distribution so that it moves closer to the ideal bell-shaped curve. Because estimating the number of households is not a main goal of the NCVS, moving toward the categories currently being used by NSDUH is recommended, because it would be useful for comparability.

Recommendation 4: Ask the Census Bureau to Create Flags Indicating When Income Has Been Carried Over From Previous Survey Waves.

As noted in Berzofsky and others (2014), it is not currently possible to identify with certainty the household income values that came from a carried-forward value (i.e., the income value from the previous wave was inserted for a wave in which the household was not asked about income). Thus, it is recommended that these values be imputed in a manner similar to those who do not provide an income when asked. Furthermore, because an implicit imputation is being conducted, it is important that users of the data be made aware of which responses were populated through this process. For these reasons, it is recommended that a flag be included on the PUF allowing users and those imputing income data the ability to identify which income values were actually reported by a household during that interview and which income values were not actually reported in that particular wave.

Recommendation 5: Continue to Consider Household Size and Federal Poverty Level If and When Improvements to the Income Question Are Implemented.

As demonstrated in ***Section 2,*** as FPL increased, the rate of violent and property crime victimization decreased. Given the impact that household size can have on income, it is recommended that any SES measure that may be derived from an improved future income question should include household size and FPL. Therefore, it is further recommended that how the FPL changes based on household size be taken into account to the extent possible in any revised income question.

Recommendation 6: Add a Question That Asks for Number of Days of Unemployment During the Most Recent Reference Period.

The NCVS does not currently have a question that asks for the number of days of unemployment. This might be an important construct that could provide some interesting insights into unemployment patterns among victims, given research that shows an association between victimization and unemployment (Aaltonen, Kivivuori, Martikainen, and Salmi, 2012; Faergemann et al., 2009). More data on unemployment might also inform future imputation methods.

Recommendation 7. Revise the Occupation Questions.

As described above, the occupation data gleaned from the current NCVS question are not producing usable data; thus, it is not possible to use occupation as a factor in the proposed SES index. However, occupation represents one of the "big three" parts of SES, and the importance of being able to use occupation data for any analyses or imputations related to SES cannot be overstated. To that end, it is recommended that BJS make changes to the occupation question so that it will yield better data.

Recommendation 8: Ask for Additional Sources of Income Such as Food Stamps, Medicare, Medicaid, Disability, Pension, Other Measures of Wealth, etc.

The NCVS is one of the few national surveys that do not ask for additional sources of income (see Table A-1 in ***Appendix A*** for examples of surveys that include additional sources of income, including program participation). Adding these sources could provide a fuller picture of the economic status of respondents. Should BJS be interested in adding such questions to the NCVS, it would be preferable to add the questions that are on the Census Bureau and Labor Statistics surveys for comparability. However, given the volume of questions surrounding these sources of income, adding such a series of questions would be expensive and it is unclear whether doing so is feasible at this time.

A less expensive and less burdensome alternative would be to add one question that captures information on respondents' usage of or need for various social services or safety net sources. For example, Nilsson and Estrada (2006) describe a yes-or-no question that asked respondents whether they would need to borrow money (from friends or family, a bank, or another source) if an unexpected expense in the amount of €1,500 arose. Clearly, this does not

provide a comprehensive picture of the resources the respondents have, but it does provide a glimpse into the financial struggles faced by some respondents.

Recommendation 9: Ask the Census Bureau to Create a File That Could Be Used to Create a Macro-Level SES to Add to the Public Use Files.

Census tract information from the NCVS have been previously examined, with notable findings related to neighborhood disadvantage and police notification behaviors (Baumer, 2002) and to identifying those at greatest risk for experiencing violence (Lauritsen, 2001). Should the Census Bureau release a special file that contains Census tract information, macro-level neighborhood information could enhance BJS's ability to study the seemingly complex relationships between SES and criminal victimization. Moreover, micro- and macro-level SES indicators could be added to PUFs in a way that maintains the confidentiality of NCVS respondents.

Recommendation 10: Revise the Educational Attainment Question So That It Separates High School Graduates and Recipients of Equivalent Diplomas.

Currently, question 25a of the NCVS asks respondents to provide the highest grade they attained, and question 25b asks whether that year was completed. Thus, there is no distinction between those who completed high school and those who earned a high school equivalent diploma. Although many Federal agencies combine these two groups in surveys (e.g., see this presentation from the Department of Health and Human Services, slides 10–11: http://www.cdc.gov/nchs/ppt/nchs2012/SS-34_QUEEN.pdf), this practice may change in the coming years because recent studies show that these two groups' earnings and educational pursuits are very different. In short, GED recipients tend to earn less than high school graduates, and they are also less likely to pursue higher education. In light of reports at the Census Bureau (e.g., Crissey & Bauman, 2012; Ewert, 2012) and the Department of Labor (e.g., "Precis: Labor market returns of the GED," 2010) that acknowledge these differences, it may behoove BJS to recalibrate the NCVS educational attainment questions such that GED can be selected as an option and therefore a finer grained measure for education can be used in future analyses.

REFERENCES

Aaltonen, M., Kivivuori, J., Martikainen, P., & Salmi, V. (2012). Socio-economic status and criminality as predictors of male violence: Does victim's gender or place of occurrence matter? *British Journal of Criminology, 52,* 1192–1211.

Aaltonen, M., Kivivuori, J., Martikainen, P., & Sirén, R. (2012). Socioeconomic differences in violent victimization: Exploring the impact of data source and the inclusivity of the violence concept. *European Journal of Criminology, 9 (6),* 567–583.

American Psychological Association (APA), Task Force on Socioeconomic Status. (2007). *Report of the APA Task Force on Socioeconomic Status.* Washington, DC: Author.

Baumer, E. P. (2002). Neighborhood disadvantage and police notification by victims of violence. *Criminology, 40,* 579–616.

Baumer, E., Horney, J., Felson, R., & Lauritsen, J. L. (2003). Neighborhood disadvantage and the nature of violence. *Criminology, 41,* 39–72.

Berzofsky, M., Creel, D., Moore, A., Smiley-McDonald, H., & Krebs, C. (2014). Imputing NCVS income data. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC. Available at HYPERLINK

Bishaw, A. (2012). *American Community Survey research briefs: Poverty: 2010 and 2011*. Washington, DC: United States Census Bureau. Retrieved from http://www.census.gov/prod/2012pubs/acsbr11-01.pdf

Blishen, B. R., Carroll, W. K., & Moore, C. (1987). The 1981 socioeconomic index for occupations in Canada. *Canadian Review of Sociology, 24,* 465–488.

Braveman P. A., Cubbin, C., Egerter S., Chideva, S., Marchi, K. S., Metzler, M. & Posner, S. (2005). Socioeconomic status in health research: One size does not fit all. *JAMA, 294,* 2879–2888.

Brownfield, D. (1986). Social class and violent behavior. *Criminology, 24,* 421–437.

Cirino, P. T., Chin, C. E., Sevcik, R. A., Wolf, M., Lovett, M., & Morris, R. D. (2002). Measuring socioeconomic status: Reliability and preliminary validity for different approaches. *Assessment, 9,* 145–155.

Crissey, S. R., & Bauman, K. J. (2012, February). Measurement of High School Equivalency Credentials in Census Bureau Surveys (SEHSD Working Paper No. 2012-3). Washington, DC: U.S. Census Bureau, Social, Economic, and Housing Statistics Division. Retrieved from http://www.census.gov/hhes/socdemo/education/data/cps/GED_wp2012-3.pdf

Cubbin, C., LeClere, F. B., & Smith, G. S. (2000). Socioeconomic status and injury mortality: individual and neighborhood determinants. *Journal of Epidemiology and Community Health, 54,* 517–524.

Demissie, K., Hanley, J. A., Menzies, D., Joseph, L., & Ernst, P. (2000). Agreement in measuring socio-economic status: Area-based versus individual measures. *Chronic Diseases in Canada, 21(1).*

Deonandan, R., Campbell, K., Ostbye, T., Tummon, I., & Robertson, J. (2000). A comparison of methods for measuring socio-economic status by occupation or postal area. *Chronic Diseases in Canada, 21(3)*.

Diez-Roux, A. V. (1998). Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *American Journal of Public Health, 88,* 216–222.

Duncan, O. D. (1961). A socioeconomic index for all occupations. In J. Reiss, Jr. (Ed.), *Occupations and social status* (pp. 109–138). New York, NY: Free Press of Glencoe.

Ewert, S. (2012, February). What it's worth: Field of training and economic status in 2009. *Current Population Reports, Household Economic Studies* (No. P70-129). Washington, DC: U.S. Census Bureau. Retrieved from http://www.census.gov/prod/2012pubs/p70-129.pdf

Faergemann, C., Lauritsen, J. M., Brink, O., Skov, O., & Mortensen, P. B. (2009). Demographic and socioeconomic risk factors of adult violent victimization from an accident and emergency department and forensic medicine perspective: A register-based case-control study. *Journal of Forensic and Legal Medicine, 16,* 11–17.

Field, C.A., & Caetano, R. (2004). Ethnic differences in intimate partner violence in the U.S. general population. *Trauma, Violence, & Abuse, 5(4),* 303–317.

Flatley, J., Kershaw, C., Smith, K., Chaplin, R., & Moon, D. (Eds.). (2010). *Crime in England and Wales 2009/10. Findings from the British Crime Survey and police recorded crime* (3rd ed.). London, UK: Home Office.

Fujishiro, K., Xu, J., & Gong, F. (2010). What does "occupation" represent as an indicator of socioeconomic status: Exploring occupational prestige and health. *Social Science & Medicine, 71,* 2100–2107.

Hollingshead, A. B. (1975). *Four factor index of social status*. New Haven, CT: Department of Sociology, Yale University.

Home Office. (2011, October). *User guide to Home Office crime statistics*. Home Office Statistics, London, United Kingdom. Retrieved from http://www.homeoffice.gov.uk/publications/science-research-statistics/research-statistics/crime-research/user-guide-crime-statistics/user-guide-crime-statistics?view=Binary

Kassis, W., Artz, S., Scambor, C., Scambor, E., & Moldenhauer, S. (2012). Finding the way out: A non-dichotomous understanding of violence and depression resilience of adolescents who are exposed to family violence. *Child Abuse & Neglect*. Advance online publication. doi:10.1016/j.chiabu.2012.11.001

Khalifeh, H., Hargreaves, J., Howard, L. M., & Birdthistle, I. (2013). Intimate partner violence and socioeconomic deprivation in England: Findings from a national cross-sectional survey. *American Journal of Public Health, 103,* 462–472.

Kington, R. S., & Smith, J. P. (1997). Socioeconomic status and racial and ethnic differences in functional status associated with chronic diseases. *American Journal of Public Health, 87,* 805–810.

Kiss, L., Schraiber, L. B., Heise, L., Zimmerman, C., Gouveia, N., & Watts, C. (2012). Gender-based violence and socioeconomic inequalities: Does living in more deprived neighbourhoods increase women's risk of intimate partner violence? *Social Science & Medicine, 74,* 1172–1179.

Lauritsen, J. L. (2001). The social ecology of violent victimization: Individual and contextual effects in the NCVS. *Journal of Quantitative Criminology, 17,* 3–32.

Lewis, G., Rice, F., Harold, G.T., Collishaw, S., & Thapar, A. (2011). Investigating environmental links between parent depression and child depressive/anxiety symptoms using an assisted conception design. *Journal of the American Academy of Child & Adolescent Psychiatry, 50(5),* 451–459.

Magklara, K., Skapinakis, P., Gkatsa, T., Bellos, S., Araya, R., Styliandis, S., & Mavreas, V. (2012). Bullying behavior in schools, socioeconomic position, and psychiatric morbidity: A cross-sectional study in late adolescents in Greece. *Child and Adolescent Psychiatry and Mental Health, 6,* 8.

Markowitz, F. E. (2003). Socioeconomic disadvantage and violence: Recent research on culture and neighborhood control as explanatory mechanisms. *Aggression and Violent Behavior, 8,* 145–154.

McLaughlin, K. A., Costello, J. E., Leblanc, W., Sampson, N. A., & Kessler, R. C. (2012). Socioeconomic status and adolescent mental disorders. *American Journal of Public Health.* Advance online publication. doi:10.2105/AJPH.2011.300477

Menard, S., Morris, R. G., Gerber, J., & Covey, H. C. (2011). Distribution and correlates of self-reported crimes of trust. *Deviant Behavior, 32,* 877–917.

Morenoff, J. D., Sampson, R. J., & Raudenbush, S. W. (2001). Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. *Criminology, 39,* 517–560.

Nakao, K., & Treas, J. (1994). Updating occupational prestige and socioeconomic scores: How the new measures measure up. In P. Marsden (Ed.), *Sociological methodology 1994*. Washington, DC. American Sociological Association. Available from http://depts.washington.edu/socmeth2/

National Center for Health Statistics. (2012). *Multiple imputation of family income and personal earnings in the National Health Interview Survey: Methods and examples*. Atlanta, GA: Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/nchs/data/nhis/tecdoc11.pdf

National Research Council. (1994). *Understanding and Preventing Violence, Volume 3: Social Influences*. Washington, DC: The National Academies Press.

Nilsson, A., & Estrada, F. (2006). The inequality of victimization: Trends in exposure to crime among rich and poor. *European Journal of Criminology, 3,* 387–412.

Oakes, J. M., & Rossi, P.H. (2003). The measurement of SES in health research: Current practice and steps toward a new approach. *Social Science & Medicine, 56,* 769–784.

Page, L., & Twist, N. (2011). 2010/2011 Scottish Crime and Justice Survey: Technical report. Edinburgh, Scotland: Scottish Government Social Research. Retrieved from http://www.scotland.gov.uk/Resource/Doc/933/0122908.pdf

Pineo, P. C., Porter, J., & McRoberts, H. A. (1977). The 1971 census and the socioeconomic misclassification of occupations. *Canadian Review of Sociology, 14,* 91–102.

Pleis, J. R., & Cohen, R. A. (2007). *Impact of income bracketing on poverty measures used in the National Health Interview Survey's (NHIS) Early Release Program: Preliminary data from the 2007 NHIS*. Retrieved from http://www.cdc.gov/nchs/data/nhis/income.pdf

Precis: Labor market returns of the GED. (2010, October). *Monthly Labor Review,* 88–89. Retrieved from http://www.bls.gov/opub/mlr/2010/10/precis.pdf

Rennison, C., & Planty, M. (2003). Nonlethal intimate partner violence: Examining race, gender, and income patterns. *Violence and Victims, 18,* 433–443.

Scottish Government. (2012, December). *Scottish Index of multiple deprivation 2012: A National Statistics publication for Scotland*. Edinburgh, Scotland: Author. Retrieved from http://22fa0f74501b902c9f11-8b3fbddfa1e1fab453a8e75cb14f3396.r26.cf3.rackcdn.com/simd_448749_v7_20121217.pdf

Shavers, V. L. (2007). Measurement of socioeconomic status in health disparities research. *Journal of the National Medical Association, 99,* 1013–1023.

Sondik, E. J., Madans, J. H., & Gentleman, J. F. (2012). *Summary health statistics for U.S. adults: National Health Interview Survey, 2011.* Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/data/series/sr_10/sr10_256.pdf

Spriggs, A. L., Tucker Halpern, C., Herring, A. H., & Schoenbach, V. J. (2009). Family and school socioeconomic disadvantage: Interactive influences on adolescent dating violence victimization. *Social Science & Medicine, 68,* 1956–1965.

Sudman, S., & Bradburn, R. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco, CA: Jossey-Bass.

Szreter, S. R. S. (1984). The genesis of the Registrar General's social classification of occupations. *British Journal of Sociology, 35,* 523–546.

Thacher, D. (2004). The rich get richer and the poor get robbed: Inequality in U.S. criminal victimization, 1974–2000. *Journal of Quantitative Criminology, 20,* 89–116.

U.S. Census Bureau. (2010, March). *Observations from the Interagency Technical Working Group on developing a supplemental poverty measure.* Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/SPM_TWGObservations.pdf

U.S. Census Bureau. (2011). *American FactFinder: Table S1901: Income in the past 12 months (in 2011 inflation-adjusted dollars*). Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1901&prodType=table

U.S. Census Bureau. (2012a). *Description of income and poverty data sources*. Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/about/datasources/description.html#sippbckgrnd

U.S. Census Bureau. (2012b). *How the Census Bureau measures poverty*. Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/about/overview/measure.html

Van Kesteren, J., Mayhew, P., & Nieuwbeerta, P. (2000). *Criminal victimization in seventeen industrialised countries: Key findings from the 2000 International Crime Victims Survey*. The Hague, The Netherlands: Ministry of Justice, WODC [Research and Documentation Center].

Winnersjö, R., Ponce de Leon, A., Soares, J. F., & Macassa, G. (2012). Violence and self-reported health: Does individual socio-economic position matter? *Journal of Injury and Violence Research, 4,* 87–95.

Xie, M., Heimer, K., & Lauritsen, J. L. (2012). Violence against women in U.S. metropolitan areas: Changes in women's work status and work. *Criminology, 50,* 105–143.

Yabroff, K. R., & Gordis, L. (2003). Assessment of National Health Interview Survey-based method of measuring community socioeconomic status. *Annals of Epidemiology, 13,* 721–726.

Zinzow, H. M., Ruggiero, K. J., Resnick, H., Hanson, R., Smith, D., Saunders, B., & Kilpatrick, D. (2009). Prevalence and mental health correlates of witnessed parental and community violence in a national sample of adolescents. *Journal of Child Psychology and Psychiatry, 50,* 441–540.

APPENDIX A: REFERENCED TABLES

Table A-1. Description of how socioeconomic status is measured in selected studies

Survey agency	Panel design?	Description of the socioeconomic status data that are used
Current Population Survey (CPS) *Bureau of Labor Statistics; Census Bureau*	Y	**Overview:** The CPS focuses on "poverty status," which is derived by examining the data from a series of questions from more than 50 sources of income during the previous calendar year.[a] Under the Office of Management and Budget's (OMB's) Statistical Policy Directive 14 guideline, the CPS uses a set of money income thresholds that vary by family size and composition to detect who is poor. **Poverty Status Measures:** Income includes earnings, unemployment compensation, workers' compensation, Social Security, Supplemental Security Income (SSI), public assistance, veterans' payments, survivor benefits, pension or retirement income, interest, dividends, rents, royalties, income from estates, trusts, educational assistance, alimony, child support, assistance from outside the household, and other miscellaneous sources.[b] Noncash benefits (e.g., food stamps) do not count, and income is measured before taxes and excludes capital gains and losses. The questionnaire provides categorical responses for total income spanning 16 different income ranges, beginning with "Less than $5,000" and ending with "$150,000 or more." Poverty thresholds are the dollar amounts used to determine poverty status that do not vary geographically but are updated annually for inflation with the Consumer Price Index for All Urban Consumers (CPI-U). Each person or family is assigned one of 48 possible poverty thresholds.[b] **Computation:** If total family income is less than the threshold appropriate for that family, the family is in poverty, and all family members have the same poverty status. If total family income equals or exceeds the threshold, the family is not in poverty. For more information about the poverty measure, access http://www.census.gov/hhes/www/poverty/about/overview/measure.html. **Other Approaches:** Recently, the Census Bureau has begun reporting the "Supplemental Poverty Measure," which derives poverty thresholds from the Consumer Expenditure Survey on basic necessities (food, shelter, clothing, and utilities) and is adjusted for geographic differences in the cost of housing.[c] It is not intended to replace the official poverty measure but is supposed to provide an additional indicator of economic well-being and provide a better understanding of economic conditions and policy impacts.[c] For more information on the Supplemental Poverty Measure, access http://www.census.gov/hhes/www/poverty/SPM_TWGObservations.pdf.

(continued)

Table A-1. Description of how socioeconomic status is measured in selected studies (continued)

Survey agency	Panel design?	Description of the socioeconomic status data that are used
Survey of Income and Program Participation (SIPP) *Census Bureau*	Y	**Overview:** The SIPP collects monthly income for up to 81 sources of income and up to 73 individual income values.[d] SIPP estimates of annual income and annual poverty can be obtained by summing 12 months of family income and monthly poverty thresholds, both of which may vary month to month. Like the CPS, the SIPP uses a set of thresholds that vary by family size and composition to determine the poverty status of a household. If a family's total income is less than that family's threshold, then the family and every individual within it are considered to be in poverty.[e] **Poverty Status Measures:** At the individual level, income includes earned income, unearned or property income, and transfer program income.[f] Earned income comprises wage and salary income, self-employed earnings, and earnings from other work arrangements. Unearned or property income refers to all income generated from interest, dividends, lump-sum payments from insurance claims, and payments from annuities and retirement, as well as payments from trusts, estates, and royalties. Transfer program income refers to cash payments from social welfare programs, SSI, Temporary Assistance to Needy Families, and general assistance. Poverty thresholds are the dollar amounts used to determine poverty status. Each person or family is assigned one of 48 possible poverty thresholds. Thresholds vary according to the size of the family and the ages of the members. The same thresholds are used throughout the United States (i.e., they do not vary geographically) and are updated annually using the CPI-U. **Computation:** In the SIPP, because income is reported multiple times a year, annual poverty rates are calculated using the sum of family income over the year divided by the sum of poverty thresholds, which can change from month to month if one's family composition changes.[g] Notably, in one recent SIPP study,[h] the small portion of families reporting zero average monthly income for the year were excluded from the analysis because of concerns that "fixing it" would introduce bias because of the small proportion of households reporting zero income and the unlikelihood of a family with no annual income. **Other Approaches:** Beyond the computation that would be derived under OMB's Directive 14 (as described above), the Census Bureau (Anderson, 2011) has also examined poverty in other ways using the SIPP panel data, including the following: *Monthly Poverty Rate*—Percentage in poverty in a given month using monthly income and a monthly threshold. *Episodic Poverty Rate*—Percentage in poverty for 2 or more consecutive months. *Chronic Poverty Rate*—Percentage in poverty every month of the time frame being considered. *Length of Poverty Spell*—Number of months in poverty. The minimum spell length is 2 months, and spells are separated by 2 or more months of not being in poverty. Individuals can have more than one spell. Spells under way in the first interview month of the panel are excluded. *Poverty Entry*—On the basis of the annual poverty measures, people who were not in poverty in the first year of the panel but in poverty in one or more subsequent years. *Poverty Exit*—On the basis of the annual poverty measure, people who were in poverty in the first year of the panel but not in poverty in one or more subsequent years.

(continued)

Table A-1. Description of how socioeconomic status is measured in selected studies (continued)

Survey agency	Panel design?	Description of the socioeconomic status data that are used
Panel Study of Income Dynamics (PSID) *National Science Foundation, National Institute on Aging, and Eunice Kennedy Shriver National Institute of Child Health & Human Development*	Y	**Overview:** Like the CPS, the PSID uses threshold poverty values to assess socioeconomic well-being on the basis of combined family income, family size, the number of persons in the family under age 18, and the ages of the household members.[i] **Poverty Status Measures:** The "Total Family Income" variable is created for each data collection wave and represents an aggregation of all labor, asset, and government transfer income (cash welfare, Social Security, etc.) for the head, spouse, and all others living in the family unit at any point during the calendar year (Institute for Social Research, 2012). PSID family income reflects the income of all persons living in the family unit during calendar year *t*, regardless of whether that person was living in the family at the time of the interview in year *t+1*. Income for each family member includes only the amount accrued during the months that the person resided with the other family members. Notably, the PSID defines "family" more broadly and includes unrelated people who live together and share resources (like cohabiting partners). The broad income categories include head and wife taxable income; head and wife transfer income; other family member taxable income; other family member transfer income; head and wife Social Security income; and other family member Social Security income. In addition, respondents are asked to provide detailed information for each of the head's and wife's jobs. The employment data are combined with the income data to compute a wage rate for both the head and wife, and may also be used for imputing labor income when needed (Institute for Social Research, 2012). **Computation:** Variables representing the total value of wealth and its major subcomponents are used to derive an overall wealth indicator (Institute for Social Research, 2012).
American Community Survey (ACS) *Census Bureau*	N	**Overview:** Using a series of eight questions, the ACS asks about money income, plus one type of noncash benefit (food stamps), during the previous 12 months. ''Total income'' is the sum of the amounts reported separately for wage or salary income; net self-employment income; interest, dividends, net rental or royalty income, or income from estates and trusts; Social Security or railroad retirement income; SSI; public assistance or welfare payments; retirement, survivor, or disability pensions; and all other income. The estimates are adjusted for inflation using the Consumer Price Index.[j] **Poverty Status Measures.** Poverty statistics in ACS products adhere to the standards specified by OMB Directive 14. The Census Bureau uses a set of dollar value thresholds that vary by family size and composition to determine who is in poverty. Furthermore, poverty thresholds for people living alone or with nonrelatives (unrelated individuals) vary by age (under 65 years or 65 years and older). The poverty thresholds for two-person families also vary by the age of the householder. If a family's total income is less than the dollar value of the appropriate threshold, then that family and every individual in it are considered to be in poverty. Similarly, if an unrelated individual's total income is less than the appropriate threshold, then that individual is considered to be in poverty (Bishaw, 2012). **Computation.** In determining the poverty status of families and unrelated individuals, the Census Bureau uses thresholds (income cutoffs) arranged in a two-dimensional matrix. The matrix consists of family size (from one person to nine or more people) cross-classified by presence and number of family members under 18 years old (from no children present to eight or more children present). Unrelated individuals and two-person families are further differentiated by age of reference person (under 65 years old and 65 years old and over).

(continued)

Table A-1. Description of how socioeconomic status is measured in selected studies (continued)

Survey agency	Panel design?	Description of the socioeconomic status data that are used
		To determine a person's poverty status, the person's total family income in the last 12 months is compared with the poverty threshold appropriate for that person's family size and composition. If the total income of that person's family is less than the threshold appropriate for that family, then the person is considered "below the poverty level," together with every member of his or her family. If a person is not living with anyone related by birth, marriage, or adoption, then the person's own income is compared with his or her poverty threshold. The total number of people below the poverty level is the sum of people in families and the number of unrelated individuals with incomes in the last 12 months below the poverty threshold (Bishaw, 2012).

[a] U.S. Census Bureau. (2012a). *Description of income and poverty data sources*. Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/about/datasources/description.html#sippbckgrnd.

[b] U.S. Census Bureau. (2012b). *How the Census Bureau measures poverty*. Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/about/overview/measure.html

[c] U.S. Census Bureau. (2010, March). *Observations from the Interagency Technical Working Group on developing a supplemental poverty measure.* Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/SPM_TWGObservations.pdf

[d] U.S. Census Bureau. (2012a). *Description of income and poverty data sources*. Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/about/datasources/description.html#sippbckgrnd

[e] Anderson, R. J. (2011, March). Dynamics of economic well-being: Poverty 2004–2006 (Current Population Reports No. P70-123). Washington, DC: United States Census Bureau. Retrieved from http://www.census.gov/hhes/www/poverty/publications/dynamics04/P70-123.pdf

[f] Westat. (2001). *Survey of Income and Program Participation users' guide* (3rd ed.). Washington, DC: Author. Retrieved from http://www.census.gov/sipp/usrguide/sipp2001.pdf

[g] U.S. Census Bureau. (2012a). *Description of income and poverty data sources*. Washington, DC: Author. Retrieved from http://www.census.gov/hhes/www/poverty/about/datasources/description.html#sippbckgrnd

[h] Hisnanick, J. J. (2007). The dynamics of low income and persistent poverty among U.S. families. *Journal of Income Distribution, 16*(1), 116–132.

[i] Institute for Social Research, University of Michigan. (2012). PSID main interview user manual (Release 2012.1). Ann Arbor, MI: Author. Retrieved from http://psidonline.isr.umich.edu/data/Documentation/UserGuide2009.pdf

[j] Bishaw, A. (2012). *American Community Survey research briefs: Poverty: 2010 and 2011*. Washington, DC: United States Census Bureau. Retrieved from http://www.census.gov/prod/2012pubs/acsbr11-01.pdf

Table A-2. Victimization rates by type of crime and household income, 2010

Household income	Number of households	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
Less than $15,000	17,185,600	14.0%	1.6	4.0	5.6	17.2	28.4	45.6	4.4	159.3
$15,000–$34,999	30,206,400	24.6	0.8	3.0	4.2	14.9	22.9	29.0	5.5	132.2
$35,000–$49,999	19,406,900	15.8	0.8*	1.9	4.5	11.0	18.2	25.2	4.6	121.6
$50,000–$74,999	20,965,200	17.1	1.6	2.2	2.2	11.6	17.6	18.5	5.5	109.8
$75,000 or more	35,121,200	28.6	0.8*	1.2	2.0	10.9	14.9	18.2	4.5	114.4

Note: Rate per 1,000.

*Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-3. Victimization rates by type of crime and household income as percentage of Federal poverty level, 2010

Household income (percentage of Federal poverty level)	Number of households	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
100% or less	16,979,800	13.8%	1.8	5.0	5.8	16.8	29.5	45.4	6.3	179.8
101%–150%	12,226,600	10.0	0.7*	3.3	4.5	15.1	23.6	37.5	7.3	166.6
151%–200%	11,842,100	9.6	1.1*	1.9	3.2	16.3	22.5	29.7	4.0	129.9
201%–300%	21,132,500	17.2	0.7*	2.6	4.1	12.3	19.6	23.3	4.0	117.2
301%–400%	15,335,100	12.5	1.8*	1.0	2.2	14.0	19.0	18.9	4.4	103.6
401%–500%	10,564,000	8.6	--*	0.7*	2.3	14.6	17.8	14.0	6.4	97.1
Greater than 500%	34,805,100	28.3	0.9*	1.3	2.1	7.5	11.7	19.1	4.1	106.0

Note: Rate per 1,000. *Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

*Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

— Number less than 0.5.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-4. Victimization rates by type of crime and education level, 2010

Education level	Number of persons	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
Less than high school	59,533,500	23.3%	1.5	2.5	5.6	14.2	23.8	43.5	5.7	211.7
High school, some college, or associate's degree	128,207,600	50.1	1.2	2.6	3.0	13.8	20.6	27.7	5.3	128.0
Bachelor's degree	43,868,200	17.1	—*	1.6	2.1	10.6	14.5	16.9	4.5	89.5
Master's, professional, or doctoral degree	18,609,800	7.3	0.5*	1.0*	2.6*	6.6	10.7	16.8	2.8*	103.5
Unknown	5,742,800	2.2	—*	0.8*	—*	7.0*	7.8	12.1	3.2*	53.2

Note: Rate per 1,000.

*Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

— Number rounds to less than 0.5.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-5. Victimization rates by type of crime and 6-month employment status, 2010

Employment status	Number of persons	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household Burglary	Motor vehicle theft	All property crimes
Employed	146,617,000	57.3%	1.1	2.3	2.9	14.3	20.6	51.6	11.8	265.8
Unemployed	90,474,400	35.4	0.7	2.2	3.6	9.5	16.0	14.6	1.9	58.3
Unknown†	18,870,500	7.4	2.4*	1.6*	6.1	14.8	24.9	16.9*	2.1*	391.5

Note: Rate per 1,000.

*Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

†Includes those under 18 years old.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-6. Victimization rates by type of crime and housing tenure, 2010

Tenure	Number of households	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
Own	82,203,700	66.9%	0.6	1.0	2.0	8.4	12.0	21.1	3.7	103.6
Rent or no cash rent	40,681,400	33.1	2.0	5.1	6.4	22.6	36.1	35.5	7.5	169.4

Note: Rate per 1,000.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-7. Victimization rates by type of crime and public housing status, 2010

Public housing	Number of households	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
No	120,201,600	97.9%	1.0	2.2	3.3	12.5	19.0	25.4	4.9	125.0
Yes	2,630,200	2.1	1.7*	5.9*	3.3*	25.4	36.2	46.8	5.7*	143.6

Note: Rate per 1,000.

*Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 010.

Table A-8. Crosstab of persons by income level and employment status, 2010

	Employment status	
Income (percentage of Federal poverty level)	No (%)	Yes (%)
100% or less	53.2	46.8
101% to 200%	46.4	53.7
201% to 400%	35.4	64.6
401% or greater	26.3	73.7

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-9. Crosstab of persons by income level and household tenure, 2010

	Household tenure	
Income	Rent (%)	Own (%)
100% or less	58.5	41.5
101% to 200%	41.0	59.0
201% to 400%	26.2	73.8
401% or greater	17.0	83.0

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-10. Crosstab of persons by income level and public housing status, 2010

	Public housing	
Income	Yes (%)	No (%)
100% or less	6.0	94.0
101% to 200%	2.4	97.6
201% to 400%	0.7	99.3
401% or greater	0.3	99.7

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-11. Crosstab of persons by income level and education level, 2010

	Education			
Income	Less than high school (%)	High school, some college, or associate's degree (%)	Bachelor's degree (%)	Master's, professional, or doctoral degree (%)
100% or less	32.4	57.3	8.0	2.3
101% to 200%	24.7	63.0	9.8	2.6
201% to 400%	13.4	62.7	18.1	5.9
401% or greater	6.8	47.9	29.9	15.4

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-12. Crosstab of persons by education level and employment status, 2010

Education	Employment status	
	No (%)	Yes (%)
Less than high school	55.3	44.7
High school, some college, or associate's degree	37.1	62.9
Bachelor's degree	24.6	75.4
Master's, professional, or doctoral degree	23.6	76.4

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-13. Crosstab of persons by education level and household tenure, 2010

Education	Household tenure	
	Rent (%)	Own (%)
Less than high school	41.1	58.9
High school, some college, or associate's degree	31.1	68.9
Bachelor's degree	23.0	77.0
Master's, professional, or doctoral degree	17.3	82.7

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-14. Crosstab of persons by education level and public housing status, 2010

Education	Public housing	
	Yes (%)	No (%)
Less than high school	3.9	96.1
High school, some college, or associate's degree	1.6	98.4
Bachelor's degree	0.4	99.6
Master's, professional, or doctoral degree	0.2	99.8

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-15. Crosstab of persons by household tenure and employment status, 2010

Household tenure	Employment status	
	No (%)	Yes (%)
Rent	35.7	64.3
Own	37.0	63.0

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-16. Victimization rates by type of crime and SES Index 1, 2010

SES Index 1 categories†	Number of households	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
1	2,436,200	2.0	0.9*	1.0*	15.8	18.9	36.6	49.4	9.6*	171.2
2	7,222,600	5.9	1.8*	5.3	5.1	19.0	31.2	49.3	5.9	182.4
3	12,662,800	10.3	1.3*	5.2	5.0	18.2	29.7	39.7	6.5	178.1
4	17,003,500	13.8	1.3	2.4	3.9	12.9	20.4	33.0	5.1	144.2
5	21,703,300	17.7	1.6	2.1	3.5	14.2	21.5	27.7	4.9	128.5
6	23,446,500	19.1	--*	1.7	1.7	12.8	16.4	16.5	5.0	101.5
7	20,783,400	16.9	1.1*	0.8	3.2	8.9	14.0	19.9	4.3	107.7
8	13,641,300	11.1	--*	0.8*	1.3*	6.7	9.3	14.8	3.6	102.8
Unknown	3,985,600	3.2	--*	1.8*	--*	3.9*	6.2	8.7	2.4*	40.4

Note: Rate per 1,000.

†The Socioeconomic Status (SES) Index does not result in whole numbers. The categories represent the results as follows: 1 = 0 to less than 1, 2 = 1 to less than 2, 3 = 2 to less than 3, 4 = 3 to less than 4, 5 = 4 to less than 5, 6 = 5 to less than 6, 7 = 6 to less than 7, and 8 = 7 to 8.

*Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

— Number less than 0.5.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2006–2010.

Table A-17. Victimization rates by type of crime and SES Index 2, 2010

SES Index 2 categories†	Number of households	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
1	466,700	0.4	2.4*	2.5*	--*	14.0*	19.0*	28.0*	11.5*	113.1
2	4,148,200	3.4	1.3*	2.1*	9.8	17.7	30.8	49.0	7.4*	157.1
3	10,628,100	8.7	1.1*	4.3	5.9	15.1	26.3	48.5	5.2	174.5
4	17,103,400	13.9	1.2	3.6	3.1	13.2	21.1	32.2	5.3	151.4
5	21,475,200	17.5	1.0	2.2	3.8	14.5	21.6	27.4	5.6	133.3
6	25,504,200	20.8	1.3*	1.7	2.5	14.3	19.8	21.8	3.8	113.4
7	23,336,500	19.0	1.1*	1.5	2.1	10.3	15.0	18.9	5.2	110.0
8	16,200,600	13.2	0.5*	1.0	2.9	8.2	12.7	16.6	4.8	108.9
Unknown	4,022,200	3.3	--*	2.2*	0.8*	3.8*	6.8	8.7	2.3*	43.2

Note: Rate per 1,000.

†The Socioeconomic Status (SES) does not result in whole numbers. The categories represent the results as follows: 1 = 0 to less than 1, 2 = 1 to less than 2, 3 = 2 to less than 3, 4 = 3 to less than 4, 5 = 4 to less than 5, 6 = 5 to less than 6, 7 = 6 to less than 7, and 8 = 7 to 8.

*Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

— Number less than 0.5.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Table A-18. Victimization rates by type of crime and SES Index 3, 2010

SES Index 3 categories†	Number of households	Percentage	Rape and sexual assault	Robbery	Aggregated assault	Simple assault	All violent crimes	Household burglary	Motor vehicle theft	All property crimes
1	3,810,100	3.1	0.9*	1.2*	10.0	16.4	28.5	42.1	7.0*	147.2
2	10,795,900	8.8	1.3*	4.2	5.6	15.5	26.6	50.7	6.0	175.5
3	17,389,400	14.2	1.2	3.9	3.3	13.4	21.7	31.4	5.2	152.7
4	21,643,700	17.6	1.0	2.2	3.9	13.5	20.6	28.6	5.5	133.7
5	25,676,900	20.9	1.3*	1.6	2.5	15.1	20.6	21.7	3.7	113.4
6	23,368,200	19.0	1.1*	1.5	2.1	10.3	15.0	18.9	5.2	110.0
7	16,215,300	13.2	0.5*	1.0	2.9	8.2	12.7	16.6	4.8	108.9
Unknown	3,985,600	3.2	--*	1.8*	--*	3.9*	6.2	8.7	2.4*	40.4

Note: Rate per 1,000.

†The Socioeconomic Status (SES) does not result in whole numbers. The categories represent the results as follows: 1 = 0 to less than 1, 2 = 1 to less than 2, 3 = 2 to less than 3, 4 = 3 to less than 4, 5 = 4 to less than 5, 6 = 5 to less than 6, and 7 = 6 to 7.

* Interpret with caution; estimate based on 10 or fewer sample cases or coefficient of variation is greater than 50%.

— Number less than 0.5.

Source: Bureau of Justice Statistics National Crime Victimization Survey (NCVS), 2010.

Project: B: Identifying Individual Routines: URBAN BASICET ANALYSIS for TOURISTS

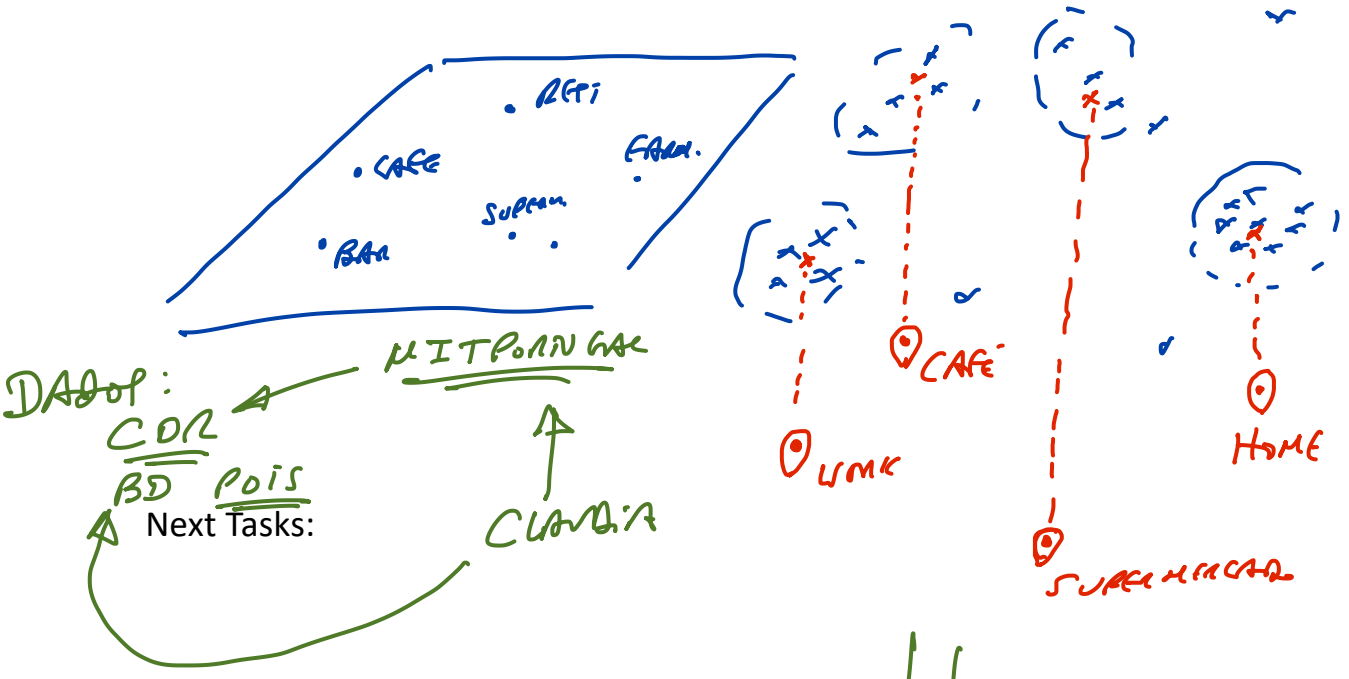
Week: 1/06

Team: DAVID FORTE
DAVID PALACIO ✓
RAFAEL GONÇALVES ✓

Progress (0..5): 4

This week:

- rotinas diárias
- What you did today ← CDR



DADOS: CDR, BD POIS → MITOMNGAE → CLUSTER

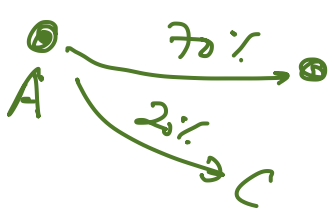
Next Tasks:

NEXT STEPS

- Analisar os CDRs...
- clustering CDRs TRAJECTO + POIS

CROWD

80%



Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

NextPlace: A Spatio-temporal Prediction Framework for Pervasive Systems

Salvatore Scellato[1], Mirco Musolesi[2], Cecilia Mascolo[1], Vito Latora[3], and Andrew T. Campbell[4]

[1] Computer Laboratory, University of Cambridge, UK
[2] School of Computer Science, University of St. Andrews, UK
[3] Dipartimento di Fisica, University of Catania, Italy
[4] Department of Computer Science, Dartmouth College, USA

Abstract. Accurate and fine-grained prediction of future user location and geographical profile has interesting and promising applications including targeted content service, advertisement dissemination for mobile users, and recreational social networking tools for smart-phones. Existing techniques based on linear and probabilistic models are not able to provide accurate prediction of the location patterns from a spatio-temporal perspective, especially for long-term estimation. More specifically, they are able to only forecast the next location of a user, but not his/her *arrival time* and *residence time*, i.e., the interval of time spent in that location. Moreover, these techniques are often based on prediction models that are not able to extend predictions further in the future.

In this paper we present NextPlace, a novel approach to location prediction based on nonlinear time series analysis of the arrival and residence times of users in relevant places. NextPlace focuses on the predictability of single users when they visit their most important places, rather than on the transitions between different locations. We report about our evaluation using four different datasets and we compare our forecasting results to those obtained by means of the prediction techniques proposed in the literature. We show how we achieve higher performance compared to other predictors and also more stability over time, with an overall prediction precision of up to 90% and a performance increment of at least 50% with respect to the state of the art.

1 Introduction

The ability to predict future locations of people allows for a rich set of novel pervasive applications and systems: accurate content dissemination of location related information such as advertisement, leisure events reports and notifications [20, 1] could be implemented in a more effective way, avoiding the delivery of information to uninterested users, and, therefore providing, a better user experience. For example, by exploiting the availability of future location information, Web search engines such as Google, Bing or Yahoo! and location-based social network services such as Facebook Places and Foursquare may provide "location-aware" sponsored advertisements together with search results that are relevant to the predicted user movement patterns.

K. Lyons, J. Hightower, and E.M. Huang (Eds.): Pervasive 2011, LNCS 6696, pp. 152–169, 2011.

The increasing popularity of smart-phones equipped with GPS sensors makes location-aware computing a reality. Even in the case of devices where this information is not currently available, location can be roughly estimated by means of triangulation and cell estimation techniques or by profiling places through the analysis of the MAC addresses advertised by nearby devices and 802.11 access points [17]. In addition, these devices are increasingly always connected to the Internet, at least in areas where GPRS/EDGE or WiFi connectivity is present. Therefore, information about the current positions of users can be transmitted to a back-end server, where analysis of the data can be performed at run-time in order to predict future location patterns.

In this paper we propose NextPlace, a new prediction framework based on *nonlinear* time series analysis [12] for forecasting user behavior in different locations from a *spatio-temporal* point of view. NextPlace focuses on the temporal predictability of users presence when they visit their most important places. We do not focus on the transitions between different locations: instead, we focus on the estimation of the duration of a visit to a certain location and of the intervals between two subsequent visits. The existing techniques are able to forecast the next location of a user, but *not* his/her *arrival* and *residence time*, i.e., the interval of time spent in that location. Moreover, these techniques are often based on prediction models that are not able to extend predictions further in the future, since they mainly focus on the next movement of a user [19, 26, 23, 14, 2, 16].

We focus instead on patterns of residence in the set of locations that are more frequently visited by users. We show that, at least in the datasets under analysis, human presence in important places is characterized by a behavior that, even if at first glance seems apparently random, can be effectively captured by nonlinear models. Predictions are based on the collection of movement data that can be of different types: sets of GPS coordinates, registration patterns to access points or also information about presence in locations by means of passive and active transponders (such as badges). In addition, check-ins performed in location-based social networking services can be exploited to acquire movement data.

The proposed prediction technique consists of two steps. Firstly, we need to identify significant locations among which users move more frequently. Secondly, we apply a model able to predict user presence within these locations and relative residence time by means of techniques drawn from nonlinear time series analysis [12]. More specifically, the contribution of this paper can be summarized as follows:

- We describe NextPlace, a novel approach to user location prediction based on nonlinear time series analysis of visits that users pay to their most significant locations. NextPlace estimates the time of the future visits and expected residence time in those locations.
- We analyze four datasets of human movements: two GPS-based (representing respectively the positions of the users involved in the deployment of the CenceMe application at Dartmouth College [21] and the locations of cabs in San Francisco [24]) and two containing registration patterns of WiFi access points (at Dartmouth College [15] and within the Ile Sans Fils wireless network in Montreal, Canada [18]). We identify regularity and, more specifically, some previously uncaptured degree of determinism in patterns of user visits to their significant places by means of nonlinear analysis.

– We evaluate NextPlace comparing it with a probabilistic technique based on spatio-temporal Markov predictors [26] and with a linear model [6]. We report an overall prediction precision over the four datasets of up to 90%, with precision of up to 65% even after a number of hours, and a performance increment of at least 50% over Markov-based predictors. We show how the adoption of a nonlinear prediction framework can improve forecasting precision with respect to other techniques even for long-term predictions.

The rest of this paper is organized as follows: Section 2 describes NextPlace and its novel approach to prediction based on nonlinear time series analysis as well as illustrates the techniques we use for the extraction of significant places. Section 3 presents the implementation issues and the validation of our approach using real-world measurements, also reporting the results of the evaluation of our method against other predictors. Section 4 discusses related work and Section 5 concludes the paper illustrating potential future work.

2 Predicting Spatio-temporal Properties of Mobile Users

Any prediction of future user behavior is based on the assumption of determinism. From a practical point of view, determinism simply means that future events are determined by past events, so that every time a particular configuration or situation is observed, the same (or a similar) outcome will follow. Since in human societies daily and weekly routines are well-established, human activities are characterized by a certain degree of regularity and predictability [8].

The intuition behind NextPlace is that the sequence of important locations that an individual visits every day is more or less fixed, with only minor variations that are also usually deterministically defined. As an example, if a woman periodically goes to the gym on Mondays and Thursdays, she may change her routine for those days, but the changed routine will be more or less the same over different weeks. Therefore, the sequence of events may still be predictable.

From a formal point of view, let us consider a certain number of mobile users, where user i freely moves among different locations. For the moment, we do not explicitly focus on how these locations can be identified, and only assume that the start time and the duration of each visit of a user to a given location can be determined. A visit of a user is simply defined by the tuple (u, l, t, d), where t and d are respectively the time of arrival and the residence time of user u in location l. It is worth noting that this approach does not model movements but, rather, residence time in some locations, hence, it can also be adopted in systems without any spatial or geographical information about locations, i.e., access points in 802.11 WLANs.

We now introduce the two steps of NextPlace and the basic theory behind them. We first describe how we isolate the user's significant places, exploiting the technique proposed by Kim et al. in [14]. Then, we describe our novel method for the estimation of future times of arrival and residence times in the different significant places and how we exploit this prediction to compute accurate estimation of where the user will be after a given time interval. Finally, we describe the mathematical details of the prediction techniques behind our approach.

(a) (b)

Fig. 1. Example of frequency map using GPS traces. Higher peaks in (a) reveal places where user spent most of their time and which represent its significant locations: in (b) we show some visits to these significant places reported on a geographical map.

2.1 Significant Places Extraction

In this section we present two methods we use to extract significant locations from both GPS information and WiFi association logs, the two most commonly available sources of data about user movements.

Extracting Places from GPS Data. Many solutions for the extraction of significant places from GPS measurements have been presented in the literature [11, 28, 2]. We choose one that is based on the residence time of a user to quantify the importance of a place for him/her: the intuition being that permanence at a place is directly proportional to the importance that is attributed to it by the user.

As proposed in [14], we apply a 2-D Gaussian distribution weighted by the residence time at each GPS point. This means that at each point the Gaussian distribution uniformly contributes also to nearby points, smoothing out values that are close together. The value of the variance for the Gaussian distributions that we choose is $\sigma = 10$ meters, which is related to the average GPS accuracy[1]. The resulting *frequency map* contains peaks which give information about the position of popular locations: we consider regions that are above a certain threshold T as significant places. The threshold T can be chosen as a fraction of the maximum value of the frequency map. We will show the application of this technique and how the value of the threshold T can be selected using two GPS-based datasets in Section 3.

In Figure 1(a) a close-up of a frequency map is shown: when a threshold is applied to the map, only higher peaks are selected and each peak generates an area defined

[1] `http://www.gps.gov/`

by a continuous boundary. All GPS points within that area result in visits to the same significant place. As an example, if we choose a threshold equal to 15% of the highest peak of the map, we obtain the visits to significant places shown over the area map in Figure 1(b).

Extracting Places from WiFi Logs. Alternatively, we can derive significant places from user registrations to 802.11 access points. Since these access points are fixed and easily identifiable from their globally unique MAC address, this information can be exploited to extract visit patterns to a set of locations in a straightforward manner. From this point of view, the most frequently seen access points are natural candidates to represent significant places. Hence, we can define as popular places for a user the access points he/she connects to more often, providing that a sufficient number of visits has been recorded to a given access point. More specifically, we define an access point as a significant place for a certain user if this user has a sequence of at least n visits to the access point, in order to filter out all the access points that are seldom visited and to have a sufficient number of observations from a statistical point of view. For the analysis presented in this paper, we select n equal to 20.

2.2 Predicting User Behavior

We now describe NextPlace's location prediction algorithm: in order to obtain an estimation of the future behavior, the history of visits of a user to each of its significant locations is considered. Then, for each location we try to predict when the next visits will take place and for how long they will last. After this estimation, the predictions obtained for different locations is analyzed, in order to produce a unique prediction of where the user will be at a given future instant of time. A theoretical foundation of this technique is described in Section 2.3.

For each user we keep track of all previous visits to a set of locations, that is, for each visit we consider the instant when it started and how long it lasted. The algorithm predicts the next visits to a given location by means of the previous history of visits $((t_1, d_1), (t_2, d_2), \ldots, (t_n, d_n))$:

1. two time series are created from the sequence of previous visits: the time series of the visit daily start times C and the time series of the visit durations D defined as follows:
$$C = (c_1, c_2, \ldots, c_n)$$
$$D = (d_1, d_2, \ldots, d_n)$$
where c_i is the time of the day in seconds corresponding to the time instant t_i (i.e., c_i is in the range $[0, 86400]$);
2. we search in the time series C sequences of m consecutive values $(c_{i-m+1}, \ldots, c_i)$ that are closely similar to the last m values $(c_{n-m+1}, \ldots, c_n)$[2];
3. the next value of time series C is estimated by averaging all the values c_{i+1} that follow each found sequence;

[2] We will discuss the choice of parameter m in the next section.

4. at the same time, in time series D the corresponding sequences $(d_{i-m+1}, \ldots, d_i)$ are selected; the sequences have to be located exactly at the same indexes as those in C;
5. the next value of time series D is then estimated by averaging all the values d_{i+1} that follow these sequences.

As an example, if the last three visits of a certain user to a location are Monday at 6:30pm, Monday at 10:00pm and Tuesday at 8:15am, we analyze the history of visits in order to find sequences that are numerically close to (6:30pm, 10:00pm, 8:15am), i.e. (6:10pm, 9:50pm, 8:35am) and (6:35pm, 10:10pm, 8:00am): then, assuming that the next visits that follow these subsequences start at 1:10pm and 12:40pm and last for 40 and 30 minutes respectively, we estimate the next visit at 12:55pm for 35 minutes, averaging both arrival times and duration times.

The main idea behind this algorithm is the assumption that human behavior is strongly determined by daily patterns: the sequence of visit start times is therefore mapped to a 24-hour time interval, focusing only on the start time of each visit. The choice of the value m has an impact on the accuracy of the prediction: in fact, this can be improved by taking into account more visits in order to identify particular patterns that may be present only in certain intervals of time such as specific days.

We can generalize this algorithm to predict not only the next visit to a location, but also successive visits in the future: in fact, we can choose to average together not only the next values of each subsequences but also values that are 2 or more steps ahead. However, the prediction of time series can become inaccurate when adopted to calculate further values in the future [12].

Since we can predict when the future visits to all significant locations will start and for how long they will last, we can design a simple method to predict the location where the user will be at a given time in the future. Let us suppose that at time T we want to predict in which significant location user i will be after ΔT seconds. Then, the following steps are performed:

1. for each location the sequence of the next k visits (starting with $k = 1$) are predicted and a global sequence of all predicted visits $(loc_1, t_1, d_1), \ldots, (loc_n, t_n, d_n)$ is created, with $t_1 \leq \cdots \leq t_n$;
2. if there is a prediction (loc_i, t_i, d_i) which satisfies $t_i \leq T + \Delta T \leq t_i + d_i$, then loc_i is returned as predicted location (in case several predictions exist which satisfy the predicate, we choose at random between them);
3. if no prediction satisfies the condition stated above, there are two cases: if the minimum start time t_1 of the current predicted visits is smaller than $T + \Delta T$, then prediction needs to be extended further in the future in order to find a suitable visit, thus the parameter k is doubled and the algorithm is repeated considering new predicted visits. Otherwise, extending the prediction provides visits which start after $T + \Delta T$ and which cannot be exploited for prediction: thus, the algorithm terminates returning that the user will not be in any significant location.

Note that it is realistic for a user to be predicted as being outside the set of significant places (e.g., maybe transitioning from one to another) and that our technique is also able to predict this state.

2.3 Nonlinear Prediction Framework: Key Concepts and Practical Implementation Issues

In this section we provide a brief overview of the key concepts at the basis of the forecasting framework and we discuss the practical issues in implementing it.

In this work we adopt a prediction technique inspired by *nonlinear time series analysis* [12]. A time series can be seen as a collection of scalar observations of a given system made sequentially in time and spaced at uniform time intervals, albeit this last assumption can be relaxed to allow any kind of temporal measurement pattern [6].

While the scalar sequence of values contained in a time series may appear completely unrelated to the underlying system, it is possible to uncover the characteristics of its dynamic evolution by analyzing sub-sequences of the time series itself. In order to investigate the structure of the original system, the time series values must be transformed in a sequence of vectors with a technique called *delay embedding*.

More formally, a time series $(s_0, s_1, \ldots, s_N)$ can be embedded in a m-dimensional space by defining an appropriate delay ν and then creating a *delay vector reconstruction* for the time series value s_n as follows:

$$\beta_n = [s_{n-(m-1)\nu}, s_{n-(m-2)\nu}, \ldots, s_{n-\nu}, s_n]$$

where all vectors β_n have m components and are defined in a so called *embedding space*. Note that m is the parameter used in the algorithm described in Section 2.2.

The values of the parameters m and ν greatly affect the accuracy of the representation. Nonetheless, a fundamental mathematical result (the so-called *embedding theorem* [12]) ensures that a suitable value for m does exist and is related to the complexity of the underlying system. At the same time, ν might be chosen to represent a suitable time scale of the phenomenon, since consecutive values in the time series should not be too strongly correlated to each other.

An effective predictive model can be generated directly from time series data through the delay embedding. Let us suppose that a prediction for the value $s_{N+\Delta n}$, a time Δn ahead of N, must be made for the time series $(s_0, s_1, \ldots, s_N)$. The steps of the prediction process are as follows:

1. The time series is embedded in a m-dimensional space by defining an appropriate time delay ν and then creating the related embedding space;
2. The embedding space is searched for all the vectors that are close, with respect to some given metric distance, to vector β_N: more formally, a neighborhood $U_\epsilon(\beta_N)$ of radius ϵ around the vector β_N is created;
3. Since determinism involves that future events are set causally by past events, and since all vectors $\beta_n \in U_\epsilon(\beta_N)$ describe past events similar to the past events of β_N, the prediction $p_{N+\Delta n}$ is taken as the average of all the values $s_{n+\Delta n}$

$$p_{N+\Delta n} = \frac{1}{|U_\epsilon(\beta_N)|} \sum_{\beta_n \in U_\epsilon(\beta_N)} s_{n+\Delta n}$$

where $|U_\epsilon(\beta_N)|$ denotes the number of elements of the neighborhood $U_\epsilon(\beta_N)$. The value of ϵ should be chosen in order to obtain a sufficient number of vectors for the prediction.

Intuitively, this algorithm searches the past history to find sequences of values that are very similar to the recent history: assuming that the evolution is ruled by deterministic patterns, a given state will always be followed by the same outcome.

In our implementation we have chosen $\nu = 1$, since we do not have to deal with particular time scales which require to skip some values of our time series. As suggested in [12], the radius ϵ of the vector neighborhood is chosen in order to be 10% of the standard deviation of each time series: this value allows us to obtain enough vectors to perform prediction and, at the same time, filters out vectors that are not close to β_N.

We note that for each prediction all vectors in the embedding space have to be considered and searched. For this reason, it is wise to use an efficient method to find nearest neighbors in the embedding space: the main computational burden is the calculation of the neighborhood $U_\epsilon(\beta_N)$ and the asymptotic complexity $O(N^2)$ can be reduced to $O(N \log N)$ with binary trees or even to $O(N)$ with a box-assisted search algorithm [25], which is the method we implement.

3 Validation of the Prediction Framework Using Real-World Measurements

In this section we introduce the datasets used in our analysis and we describe how we process them in order to extract significant places. Then, we investigate the predictability of the time series extracted from sequences of visits of each user to his/her significant locations, using standard metrics adopted in time series analysis. Finally, we compare NextPlace prediction performance against other prediction methods.

3.1 Datasets

For the evaluation of our approach we choose four different datasets of human movements:

1. **Cabspotting.** This dataset is composed of movement traces of taxi cabs in San Francisco, USA, with GPS coordinates of approximately 500 taxis collected over 30 days in the San Francisco Bay Area. Each vehicle is equipped with a GPS tracking device that is used by dispatchers to efficiently reach customers [24]. The average time interval between two consecutive GPS measurements is less than 60 seconds.
2. **CenceMe GPS.** This dataset was collected during the deployment of CenceMe [21], a system for recreational personal sensing, at Dartmouth College. The GPS data was collected by means of 20 Nokia N95 phones carried by postgraduate students and staff members from the Department of Computer Science and the Department of Biology.
3. **Dartmouth WiFi.** This dataset was extracted from the SNMP logs of the WiFi LAN of the Dartmouth College campus. The compact nature of the campus means that the signal range of interior APs extends to most of the campus outdoor areas. Between 2001 and 2004 data about traffic in the access points was collected through three techniques: syslog events, SNMP polls, and network sniffers [9, 15].

Table 1. Properties of the different datasets: total number of users N, total number of visits V, total number of significant places P, average number of significant places per user p, average number of visits per user v, average residence time in a place D (seconds), total trace length and average proportion of time spent by each user in significant places

Dataset	N	V	P	p	v	D [s]	Trace length	Significant time
Cabspotting	252	150612	6122	24.29	597	231	23 days	7.27%
CenceMe GPS	19	3832	225	11.84	201	696	12 days	14.74%
Dartmouth WiFi	2043	772217	539	17.87	377	2094	60 days	11.24%
Ile Sans Fils	804	142407	173	3.61	177	5296	370 days	0.18%

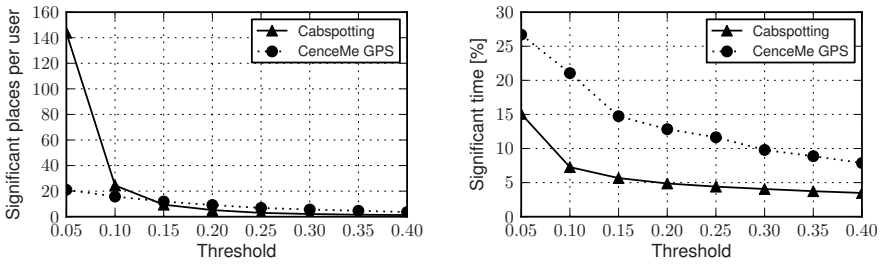
4. **Ile Sans Fils.** Ile Sans Fils [18] is a non-profit organization which operates a network of free WiFi hotspots in Montreal, Canada. It now counts over 45,000 users with 140 hotspots located in publicly accessible spaces. These hotspots are deployed mostly in cafes, restaurants and bars, libraries, but also outdoor to cover parks and sections of popular commercial streets.

We choose a subset of regularly active users for each original dataset, filtering out all the users that appear only a few times and for which prediction may be worthless. In Table 1 we report some important characteristics and metrics of the resulting datasets.

3.2 Practical Issues

In order to extract significant places for each user in the Cabspotting and CenceMe GPS datasets, which are composed of GPS measurements, we need to choose a suitable threshold T for the frequency map. Thus, we investigate how the average number of significant places per user changes as a function of the threshold itself. As reported in Figure 2(a), the average number of places decreases as the threshold increases: for the Cabspotting dataset a suitable choice is $T = 0.10$, where the curve changes its slope, which denotes the transition from a situation with many unimportant significant areas to a situation with less but probably more important places. However, in the case of the CenceMe GPS dataset such transition does not occur: hence, we investigate how the percentage of time spent in significant locations changes with T, as reported in Figure 2(b): this percentage quickly decreases with T but the steepness of the curve changes at $T = 0.15$. Hence, we choose the value of $T = 0.15$ for this dataset. These values of T result in an average number of about 24 and 12 places per user for the Cabspotting and CenceMe GPS datasets, respectively.

When dealing with GPS measurements, the duration of a visit can be computed as the difference between two consecutive GPS samples. However, the GPS measurement process usually involves a periodic sampling of the location. When the user is located for a long time interval inside the same region, this results in several successive short visits being recorded, whose length depends on the adopted sampling interval. The same problem may occur with WiFi association logs: since WiFi connectivity may be intermittently available and handoff mechanisms are in place in this type of network infrastructure, a long residence time may be split in several shorter sessions.



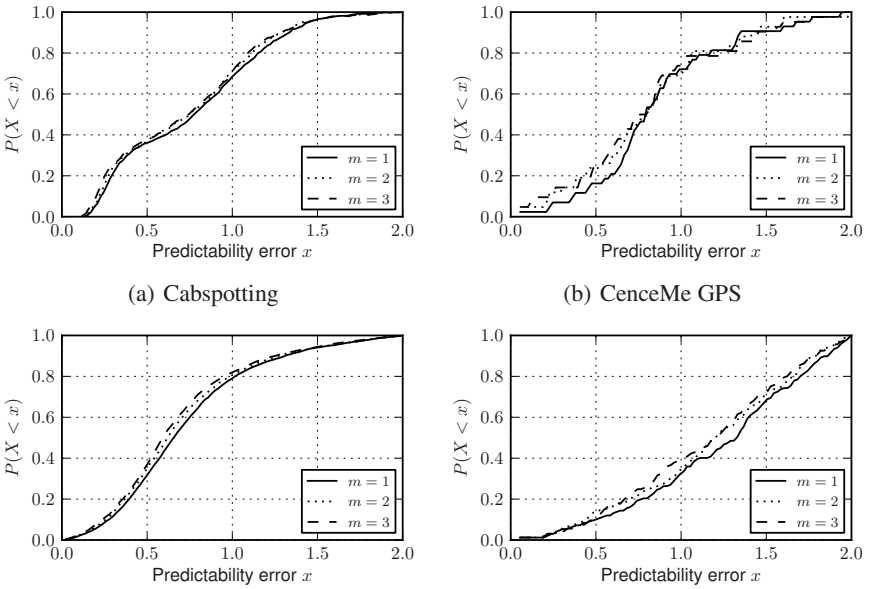
(a) (b)

Fig. 2. Average number of significant places per user (a) and percentage of time spent in significant locations (b) as a function of the threshold T of the frequency map for the Cabspotting dataset and the CenceMe GPS dataset

In order to infer a more accurate residence time of the user in a certain region, we apply a merging procedure to the dataset of the sequence of visits. Given a sequence of visits to the same location $(t_1, d_1), (t_2, d_2), \ldots (t_n, d_n)$, if the end time of a visit is close to the start time of the next one, that is if $t_{i+1} - (t_i + d_i) \leq \delta$, we merge them in a new visit starting at t_i and ending at $t_{i+1} + d_{i+1}$. In this way the visits obtained are more likely to mimic the real patterns of presence of users, thus improving prediction. We adopted the value of $\delta = 60$ seconds for the Cabspotting dataset and $\delta = 180$ seconds for the CenceMe GPS dataset, since these are the values of the scanning period for the GPS data acquisition. On the other hand, we apply the same merging procedure to WiFi association logs in the Dartmouth and Ile Sans Fils datasets with a value of $\delta = 300$ seconds, in order to filter out casual disconnections from the access point which may last for few minutes.

From a statistical point of view, these datasets show different characteristics, as reported in Table 1: while Cabspotting, Dartmouth WiFi and Ile Sans Fils contain measurements for hundreds or thousands of users, CenceMe GPS consists of data related to a smaller group of moving users. On average about 12 significant locations have been recorded for each user in the CenceMe GPS dataset. In the Dartmouth WiFi and Cabspotting datasets the number of significant places is 18 and 24, respectively. On the other hand, in the Ile Sans Fils dataset we have less than 4 significant locations per user. This is due to the fact that the Ile Sans Fils dataset contains association logs with access points located in public spaces, thus, a large portion of individuals are seen just in few locations. In fact, public access point are not likely to capture some important places for a given user, such as his/her home and working place. There are also differences in the residence time of users in their significant locations: while for Ile Sans Fils and Dartmouth WiFi the average residence time is about 90 and 30 minutes, in the Cabspotting and CenceMe GPS datasets it is about 5 and 10 minutes.

Finally, the amount of time spent in significant locations is crucial to the investigation of the performance of the location prediction technique. While in the CenceMe GPS and in the Dartmouth WiFi datasets each user spends on average 14.74% and 11.24% of their time in a significant location, this value drops to 7.27% in the Cabspotting dataset



(a) Cabspotting (b) CenceMe GPS

(c) Dartmouth WiFi (d) Ile Sans Fils

Fig. 3. Cumulative Distribution Function of the predictability error of the time series of the start instants extracted from the various datasets. We report the results for different values of the embedding dimension m adopted in the prediction method.

and to 0.18% in the Ile Sans Fils dataset, since it covers a longer period of time (more than one year) and many of its users are present less regularly than in the other datasets.

3.3 Time Series Predictability Test

In order to exploit time series techniques to predict user behavior, we first need to investigate if determinism is present in the extracted time series. In other words, we want to evaluate the *predictability* of these time series.

Let us consider a time series $(s_0, s_1, \ldots, s_N)$. If a real measurement for s_{N+1} is given, the prediction error is the difference between s_{N+1} and the predicted value p_{N+1}. Given a prediction technique, it is possible to obtain predicted values $(p_0, p_1, \ldots, p_N)$ for the whole time series. Then, the *mean quadratic prediction error* can be evaluated as $\varepsilon = \frac{1}{N} \sum_{n=1}^{N} (s_n - p_n)^2$. Large values of ε indicate that the prediction is not accurate and the time series is not predictable.

The evaluation of ε is based on the comparison to the variance σ^2 of the time series: thus, a convenient way of deciding whether ε is small or large is to take the ratio $\frac{\varepsilon}{\sigma^2}$, which is the *predictability error*: if this ratio is close to 1, then, the mean quadratic prediction error is large, while if it is close to 0, the mean quadratic prediction error is small. We refer to this ratio as the *predictability error* of a prediction algorithm. The absolute error value ε may be meaningless if not compared to the average amount of

fluctuations a time series exhibits: by dividing by the variance of the series we can normalize the error and compare the prediction accuracy for different time series.

We exploit this metric to evaluate whether the time series extracted from user visits in the different datasets are predictable. We divide each dataset in two halves: we use the first half to build the model and we compute predicted values of the second half and vice versa. A value equal to 1 means that no determinism is present in the time series, since in this case the predictor has the same accuracy of the simple average value, whereas a value closer to 0 indicates a high degree of determinism.

In Figure 3 we show the Cumulative Distribution Function of the predictability error for the time series of the visit start times for different values of the embedding dimension m. We have also investigated the predictability error for the time series of visit end times, obtaining similar results, which we do not show due to space limitations. On average, a large proportion of users exhibit predictability: in the Dartmouth WiFi dataset 80% of the time series show predictability error smaller than 1, whereas in the CenceMe GPS and Cabspotting datasets the same figure is 70% and it drops to 40% in the Ile Sans Fils dataset, which show less predictability than the others. This is due to the fact that visits may not occur every day with the same pattern for access points in public places, since different individuals are likely to show less regularity in public space than in more personal locations as living or working places, which are not present in this dataset. Moreover, in all datasets the predictability error is lower for higher values of the embedding dimension m: this confirms that nonlinear methods improve prediction quality, since they are able to capture and recognise specific patterns of visits and to estimate when the next visit will be. However, we have noticed that values of $m \geq 4$ show worse performance because we do not have sufficient statistics in order to make a correct prediction.

Interestingly, we expected to observe a lower degree of regularity in the Cabspotting traces, since the movements of a taxi are related to the destinations of the different customers and these destinations can be hardly predictable. Nonetheless, we were able to identify a set of places among which taxis move with more regular patterns. These places correspond to areas where taxi drivers periodically go and wait for new customers, such as touristic locations, shopping malls, cinemas, and they tend to exhibit regular and predictable patterns.

3.4 Evaluating Prediction Accuracy

We compare the performance of NextPlace with those of other two methods: a state-of-the-art Markov-based spatio-temporal predictor and a modified version of NextPlace, where time series of visits are predicted with linear methods rather than with nonlinear algorithms.

Methodology. Firstly, we compare NextPlace with a more sophisticated *spatio-temporal Markov predictor* derived by extending the techniques presented in [26]. To the best of our knowledge, this is the most accurate algorithm that has been presented in the literature for this class of prediction problems, because it combines spatial and temporal dimensions to estimate both next location and handover time for users in a cellular network.

Consider a user visit history among several locations $H = (t_1, d_1, l_1), \ldots, (t_n, d_n, l_n)$, where t_i is the time when the user arrived at location l_i and d_i is the residence time in that location. Then, from H we extract the location history $L = l_1, \ldots, l_n$ and the order-k location context $L_k = L(n-k+1, n) = l_{n-k+1}, \ldots, l_{n-1}, l_n$. The history L is searched for instances of the context L_k and, for each destination that follows an instance, we examine the duration of the previous residence time. More formally, we extract the following set of inter-arrival times A_x and set of durations D_x for each possible destination x:

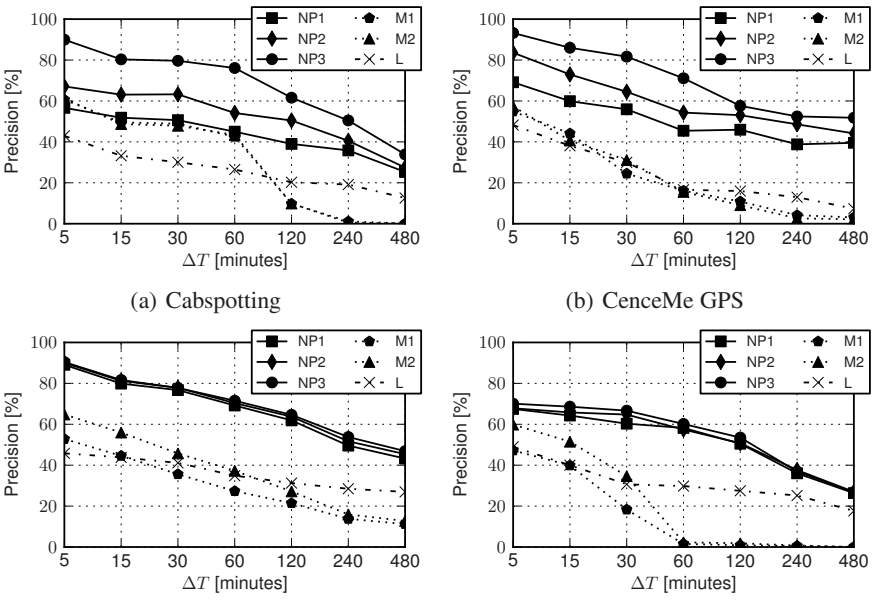
$$A_x = \{t_{i+1} - t_i \quad \text{if} \quad L(i-k+1, i+1) = (L_k, x)\}$$
$$D_x = \{d_{i+1} \quad \text{if} \quad L(i-k+1, i+1) = (L_k, x)\}$$

Then, we compute the estimated time when the user will move to location x and the estimated residence time in x by using a CDF predictor with probability $p = 0.8$ [26]. Moreover, a Markov predictor of order k is used to assign the probability of transition between the current location and the possible destinations. Finally, spatial and temporal information are combined to obtain the predicted location. In order to predict not only the next location but also the subsequent ones, we extend this approach taking the predicted location as the current one and computing again the next movement. We refer the interested reader to the original paper for further details [26].

To understand how largely NextPlace relies on the performance of the nonlinear time series predictor, we can design a linear version of our prediction technique. We use an *order-k running average predictor* instead of a nonlinear method to estimate the future values of a time series: given the sequence of previous visits of a user to a location, the last k visit duration times and k intervals between visits are averaged to obtain a prediction of future visits. Then, the future location is chosen among several predicted locations according to the same algorithm at the basis of the nonlinear predictor (presented in Section 2.2). However, this simplistic time series predictor ignores how user behavior changes over time, since high heterogeneity can be observed in visits occurring during different times of the day. Focusing only on recent data and not investigating these temporal aspects may not be sufficient to obtain accurate estimates.

Results. We now evaluate the performance of NextPlace with the nonlinear predictor presented in Section 2.2 compared to the other predictors previously described.

We use the following definition of correctness: if we predict, at time T, that the user i will be at location l at time $T_P = T + \Delta T$, the prediction is considered correct only if the user is at l at any time during the interval $[T_P - \theta, T_P + \theta]$, where θ is the error margin. It is important to note that each prediction algorithm can also estimate if the user will not be in any of her significant places: thus, a prediction may be correct whether the user is predicted to be in a particular location l and then he/she is in l or if the user is predicted not to be in any significant location and then, in fact, she is not. However, as reported in Table 1, the fraction of time that on average users spend in their significant locations ranges between 14.74% in the CenceMe GPS dataset and only 0.18% in the Ile Sans Fils dataset. Hence, it is not easy to understand if predictions are accurate because a method is performing well or because, on average, it is just easier to predict the user outside of all her significant locations.



(a) Cabspotting (b) CenceMe GPS

(c) Dartmouth WiFi (d) Ile Sans Fils

Fig. 4. Prediction precision as a function of time interval ΔT for the different datasets and for different predictors: NextPlace with nonlinear predictor for different values of embedding dimension $m = 1, 2, 3$ (NL1-NL2-NL3), first-order and second-order Markov-based (M1-M2) and NextPlace with linear predictor (L). Error margin is $\theta = 900$ seconds.

Therefore, we introduce an accuracy metric that takes into account this issue. We define the *prediction precision* as the ratio between the number of correct predictions and the number of all attempted predictions which forecast the user to be in a significant location. We do not consider for the evaluation any predictions which forecast the user outside her significant locations.

We report the performance of different predictors: we test NextPlace with different values of the embedding parameter $m = 1, 2, 3$, two order-1 and order-2 Markov-based predictors and the linear version of NextPlace with a running average predictor considering the last $m = 4$ values. For each dataset, we use the first half to build a prediction model and then we compute predictions during the second half and, for each user, we make 1000 predictions at uniformly distributed random instants. Finally, prediction precision is computed and we investigate how it changes with ΔT, using an error margin $\theta = 900$ seconds. All results are averaged over 20 runs with different random seeds.

We see in Figure 4 that for all datasets, NextPlace with its nonlinear predictor is always outperforming the other methods. We also note that using a higher value of m improves prediction quality, as it can be appreciated especially in the GPS-based Cabspotting and CenceMe GPS datasets. Similarly, Markov models are able to provide correct predictions when ΔT is smaller than 1 hour: however, except for the Ile Sans Fils dataset, the performance of the nonlinear NextPlace are at least about 50% better

of the Markov-based predictors, since they reach a maximum precision of 60% while NextPlace achieves a precision higher than 90%. Moreover, when ΔT increases, the precision of Markov predictors decreases rapidly and the performance gap with the nonlinear approach widens. This can be explained by the fact that Markov predictors are generally employed to predict the next movement and, thus, when predictions are extended in the future, movement after movement, a large error is accumulated.

If we substitute the nonlinear predictor in NextPlace with a linear one, we observe a similar trend but precision is considerably lower, since errors on time series prediction are larger and, hence, affect the location prediction. However, NextPlace with both nonlinear and linear predictors is less dependent on ΔT than Markov models, which show a lower precision when predictions are extended in the future. Again, this demonstrates how NextPlace, which focuses only on temporal information of visits in significant places, is more robust for long-term predictions.

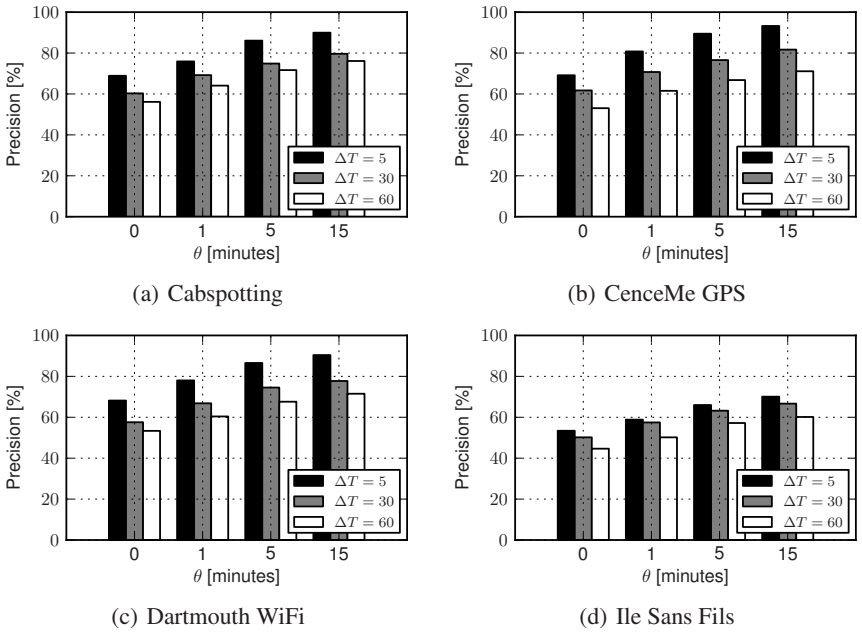
As discussed in Section 3, the Ile Sans Fils dataset exhibits less predictability. This is confirmed by the analysis of prediction precision, which shows the lowest figures among all the datasets. The other datasets score a precision equal to about 90% for $\Delta T = 5$ minutes and around 70% for $\Delta T = 60$ minutes. We also investigate the impact of the error margin θ on prediction results: prediction precision is lower for smaller error margins, but it shows the same trends for all predictors and for all the datasets. In Figure 5 we report how prediction precision of our nonlinear approach with $m = 3$ is affected by different error margins for some values of ΔT. Even with $\theta = 0$, which represents the worst case scenario, prediction precision is between 50% and 60% after $\Delta T = 60$ minutes for all datasets except Ile Sans Fils, where it is below 50%.

From a general point of view, our evaluation shows how NextPlace achieves high prediction accuracy, even for long-term predictions made some hours in advance. Furthermore, these results also show how focusing on spatial movements, as Markov models do, may be useful only for short-term predictions. Instead, focusing just on temporal information about recurrent patterns in significant places proves to be more robust both for short-term and long-term predictions, since NextPlace outperforms Markov models even for small values of ΔT.

4 Related Work

Pioneering work [3,4] has focused on the analysis of mobility traces in order to gain insight about human mobility patterns. Key papers in this area include studies on mobility and connectivity patterns, such as [5, 13]. The main findings are that contact duration and inter-contacts time between individuals can be represented by means of power-law distributions and that these patterns may be used to develop more efficient opportunistic protocols [10]. In addition, temporal rhythms of human behavior have been studied and modeled to discover daily activity patterns, to infer relationships and to determine significant locations [7]. This related body of work concentrates on the *statistical characterization* of temporal behavioral patterns of groups of users, whereas we concentrate on prediction of single users.

The evaluation of prediction techniques applied to the problem of forecasting the next location (but not the arrival time to that location and the corresponding residence



(a) Cabspotting (b) CenceMe GPS

(c) Dartmouth WiFi (d) Ile Sans Fils

Fig. 5. Prediction precision of NextPlace with nonlinear predictor with $m = 3$ as a function of error margin θ and for different values of ΔT

time) are presented in [27]. A prediction framework based on spatio-temporal patterns in collective mobility trajectories has been presented in [22]: this method attempts to predict the next location of a moving object by matching a new trajectory to a corpus of global frequent ones. While this prediction technique is more general, as it captures dependencies between visits at different places, our method includes time-of-day information and does not rely on global patterns, allowing prediction to be made also for users who deviate from collective behavior. In [2] the authors present a model of user location prediction from GPS data. A simple first-order Markov model to predict the transitions between significant places is used, albeit in this work temporal aspects are not taken into consideration. In [19] the significant places are extracted by means of a discriminative relational Markov network; then, a generative dynamic Bayesian network is used to learn transportation routines. Another system for the prediction of future network connectivity based on a second-order Markov model is BreadCrumbs [23]. Again, this system is able to predict only the next location of the user and not the time of the transitions and the interval of time during which users reside in that specific location. Similarly, Markov based techniques have also been applied to the prediction of the destinations (geographical locations) of vehicles using for example partial trajectories [16]. As we have shown in the evaluation section, this class of models is able to provide precise predictions only for instants of time close in the future, given the inherent memorylessness of Markov predictors.

5 Conclusions

In this paper we have presented NextPlace, a new approach to spatio-temporal user location prediction based on nonlinear analysis of the time series of start times and duration times of visits to significant locations. To the best to our knowledge, this is the first approach that not only allows to forecast the next location of a user, but also his/her *arrival* and *residence time*, i.e., the interval of time spent in that location. Moreover, existing models are not able to extend predictions further in the future, since they mainly focus on the next movement of a user.

We have evaluated NextPlace comparing it with a version based on a linear predictor and a probabilistic technique based on spatio-temporal Markov predictors over four different datasets. We have reported an overall prediction precision up to 90% and a performance increment of at least 50% over the state of the art. We have showed how the adoption of a nonlinear prediction framework can improve prediction precision with respect to other techniques even for long-term predictions.

As future work, there is a number of potential improvements that can be pursued. Regular collective human rhythms can be exploited to refine the prediction and a probabilistic framework can be used to choose between equally promising next locations, giving more flexibility to applications. Finally, we are interested in the investigation of prediction models which take into account human rhythms on a weekly basis, in order to better capture regular human behavior on a longer time scale.

References

1. Aalto, L., Göthlin, N., Korhonen, J., Ojala, T.: Bluetooth and WAP Push Based Location-aware Mobile Advertising System. In: Proceedings of MobiSys 2004, pp. 49–58 (2004)
2. Ashbrook, D., Starner, T.: Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. Journal of Personal and Ubiquitous Computing 7(5), 275–286 (2003)
3. Balachandran, A., Voelker, G.M., Bahl, P., Rangan, P.V.: Characterizing User Behavior and Network Performance in a Public Wireless LAN. In: Proceedings of SIGMETRICS 2002 (2002)
4. Balazinska, M., Castro, P.: Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network. In: Proceedings of MobiSys 2003, San Francisco, CA (May 2003)
5. Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R., Scott, J.: Impact of Human Mobility on Opportunistic Forwarding Algorithms. IEEE Transactions on Mobile Computing 6(6), 606–620 (2007)
6. Chatfield, C.: The Analysis of Time Series: An Introduction, 5th edn. Chapman & Hall/CRC, London (July 1995)
7. Eagle, N., Pentland, A.S.: Reality Mining: Sensing Complex Social Systems. Personal Ubiquitous Comput. 10(4), 255–268 (2006)
8. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding Individual Human Mobility Patterns. Nature 453(7196), 779–782 (2008)
9. Henderson, T., Kotz, D., Abyzov, I.: The Changing Usage of a Mature Campus-wide Wireless Network. In: Proceedings of MobiCom 2004, New York, NY, USA, pp. 187–201 (2004)
10. Jain, S., Fall, K., Patra, R.: Routing in a Delay Tolerant Network. In: Proceedings of SIGCOMM 2004 (2004)
11. Kang, J.H., Welbourne, W., Stewart, B., Borriello, G.: Extracting Places from Traces of Locations. SIGMOBILE Mobile Computing Communication Review 9(3), 58–68 (2005)

12. Kantz, H., Schreiber, T.: Nonlinear Time Series Analysis. Cambridge University Press, Cambridge (2004)
13. Karagiannis, T., Le Boudec, J.-Y., Vojnovic, M.: Power Law and Exponential Decay of Intercontact Times Between Mobile Devices. In: Proceedings of MobiCom 2007, pp. 183–194 (2007)
14. Kim, M., Kotz, D., Kim, S.: Extracting a Mobility Model from Real User Traces. In: Proceedings of INFOCOM 2006 (April 2006)
15. Kotz, D., Henderson, T., Abyzov, I.: CRAWDAD trace dartmouth/campus/movement/01_04 (v. 2005-03-08) (March 2005), http://crawdad.cs.dartmouth.edu/
16. Krumm, J., Horvitz, E.: Predestination: Inferring Destinations from Partial Trajectories. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 243–260. Springer, Heidelberg (2006)
17. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit, B.: Place Lab: Device Positioning Using Radio Beacons in the Wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 116–133. Springer, Heidelberg (2005)
18. Lenczner, M., Gregoire, B., Roulx, F.: CRAWDAD data set ilesansfil/wifidog (v. 2007-08-27) (August 2007), http://www.crawdad.cs.dartmouth.edu/ilesansfil/wifidog
19. Liao, L., Patterson, D.J., Fox, D., Kautz, H.: Building Personal Maps from GPS Data. In: Proceedings of IJCAI Workshop on Modeling Others from Observation (2005)
20. Marmasse, N., Schmandt, C.: Location-Aware Information Delivery with ComMotion. In: Thomas, P., Gellersen, H.-W. (eds.) HUC 2000. LNCS, vol. 1927, pp. 157–171. Springer, Heidelberg (2000)
21. Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S.B., Zheng, X., Campbell, A.T.: Sensing Meets Mobile Social Networks: the Design, Implementation and Evaluation of the CenceMe Application. In: Proceedings of SenSys 2008, pp. 337–350. ACM, New York (2008)
22. Monreale, A., Pinelli, F., Trasarti, R., Giannotti, F.: WhereNext: a location predictor on trajectory pattern mining. In: Proceedings of SIGKDD 2009, pp. 637–646. ACM, New York (2009)
23. Nicholson, A.J., Noble, B.D.: BreadCrumbs: Forecasting Mobile Connectivity. In: Proceedings of MobiCom 2008, pp. 46–57. ACM, New York (2008)
24. Piorkowski, M., Sarafijanovic-Djukic, N., Grossglauser, M.: CRAWDAD trace set epfl/mobility/cab (v. 2009-02-24) (February 2009), http://crawdad.cs.dartmouth.edu/epfl/mobility/cab
25. Schreiber, T.: Efficient Neighbor Searching in Nonlinear Time Series. International Journal on Bifurcations and Chaos 5, 349–358 (1995)
26. Song, L., Deshpande, U., Kozat, U.C., Kotz, D., Jain, R.: Predictability of WLAN Mobility and its Effects on Bandwidth Provisioning. In: Proceedings of INFOCOM 2006 (April 2006)
27. Song, L., Kotz, D.: Evaluating Location Predictors with Extensive Wi-Fi Mobility Data. In: Proceedings of INFOCOM 2004, pp. 1414–1424 (2004)
28. Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L.: Discovering Personally Meaningful Places: An Interactive Clustering Approach. ACM Trans. Inf. Syst. 25(3), 12 (2007)

International Journal of Geo-Information

MDPI

Article

Identification and Classification of Routine Locations Using Anonymized Mobile Communication Data

Gonçalo Ferreira [1,*], Ana Alves [1,2], Marco Veloso [1,3] and Carlos Bento [1]

[1] Centre of Informatics and Systems (CISUC), University of Coimbra, 3030-290 Coimbra, Portugal; ana@dei.uc.pt (A.A.); mveloso@dei.uc.pt (M.V.); bento@dei.uc.pt (C.B.)
[2] Instituto Superior de Engenharia de Coimbra (ISEC), Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal
[3] Escola Superior de Tecnologia e Gestão de Oliveira do Hospital (ESTGOH), Polytechnic Institute of Coimbra, 3030-199 Coimbra, Portugal
* Correspondence: gfferreira@student.dei.uc.pt

Abstract: Digital location traces are a relevant source of insights into how citizens experience their cities. Previous works using call detail records (CDRs) tend to focus on modeling the spatial and temporal patterns of human mobility, not paying much attention to the semantics of places, thus failing to model and enhance the understanding of the motivations behind people's mobility. In this paper, we applied a methodology for identifying individual users' routine locations and propose an approach for attaching semantic meaning to these locations. Specifically, we used circular sectors that correspond to cellular antennas' signal areas. In those areas, we found that all contained points of interest (POIs), extracted their most important attributes (opening hours, check-ins, category) and incorporated them into the classification. We conducted experiments with real-world data from Coimbra, Portugal, and the initial experimental results demonstrate the effectiveness of the proposed methodology to infer activities in the user's routine areas.

Keywords: call detail records; clustering algorithms; human mobility; meaningful places; mobile phone data; points of interest

Citation: Ferreira, G.; Alves, A.; Veloso, M.; Bento, C. Identification and Classification of Routine Locations Using Anonymized Mobile Communication Data. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 228. https://doi.org/10.3390/ijgi11040228

Academic Editors: Luca Pappalardo and Wolfgang Kainz

Received: 31 December 2021
Accepted: 25 March 2022
Published: 29 March 2022

1. Introduction

Human mobility has become a prominent research field in recent years. There is a growing need to understand how people move and use urban space in their daily routines. We know for a fact that human trajectories are characterized by a high degree of temporal and spatial regularity. For each individual, there is frequent time-independent travel distance and a significant probability of returning to a few highly frequented locations, with only minor deviations to visit new destinations [1]. These hidden patterns of motion have importance in applications such as urban planning, traffic forecasting and the spread of biological and mobile viruses. Ubiquitous computing has in fact unlocked the potential to determine the personal movements of the masses that previously were only modeled using household surveys and national or regional census.

In today's society, nothing could be more ubiquitous than mobile phones, each of which is a potential sensor providing a constant data stream. On this basis, specialized spatio-temporal datasets such as GPS records have shown enormous potential as a knowledge base about human mobility patterns. In spite of that, due to the overhead of collecting and analyzing such detailed and high frequency location logs at a large scale, many researchers have been urged to explore other data sources as potential proxies for human mobility.

With this in mind, we take special interest in call detail records. A call detail record (CDR) is a form of data that documents how a user interacts in detail with the cellular network. These records contain information fields such as origin/destination tower ID, user ID, time of start and duration on several types of communication. Collecting the

cellular tower ID that the user is connected to and corresponding Cartesian coordinates means that an accurate location is not possible, only an approximation. In the case of CDRs, records do not maintain regular time intervals, unlike GPS, only when a network or user event occurs (e.g., a call is made) the position of the tower used is known. This leads to data that is both spatially sparse and temporally irregular.

There are some compelling reasons for using call detail records (CDRs) versus more accurate location sources. Compared to other location log data types, the overhead for collection and analysis is inferior. Furthermore, no application has to be running, no additional battery life is consumed and data collection cannot be turned off by the user, it is generated by the regular usage of a mobile communication device and stored by the mobile network provider. Potentially every phone in one provider's network can be used as a data source, resulting in enormous amounts of information regarding a substantial percentage of the population that can be used for research purposes.

Despite the significant benefits of scale on this information type, the detail in the locations recorded is a challenge for research efforts focused on individual user analysis. Since we only have the position of a cellular tower, whose range of action spans between a few hundred meters to several kilometers, it is very difficult to accurately pinpoint a user's location. This uncertainty in location makes it very challenging to identify and classify each user's routine places. Although this question of routine places is far more stabilized with detailed information types, such as GPS [2,3], we feel that there is still room for improvement and innovation working with call detail records.

The main objective of this paper is to present an approach for attaching semantic meaning to a user's routine locations, inferring the motivations behind day-to-day mobility. The focus is on classifying activities (e.g., shopping, dining, outdoor recreation) outside of the normal home–work commute. This study uses call detail records from Portuguese citizens. Data were provided for this research by one of the largest telecommunication service providers in Portugal and records are from a four-month period between July and October 2020 by users who had a majority of mobile events in our study area, the district of Coimbra.

Using several types of complementary records, including those captured without human intervention (network-driven events), we hope to surpass some of the issues related to the sparse and irregular time intervals of events. Filling the gaps of mobility traces eases trajectory reconstruction and, as such, facilitates inferring people's routine places. Additionally, by using points of interest (POIs) datasets to classify routine places, we gather a wealth of information that can be used by other studies and applications. Analyzing each user's habits and tastes creates valuable parameters for recommendation systems, adapting marketing campaigns and improving the quality of service by taking into account clients profiling.

This paper is structured as follows: Section 2 comprises the state of the art, where we look over the current best practices and implemented methods. Section 3 provides an outline of the methodology including the research approach, data analysis techniques and description of the used algorithms. In Section 4, we report experiments conducted to test the behavior of the proposed methodology as well as describe and discuss the obtained results. Validation and evaluation are also presented since ground-truth data were made available from the mobile network provider. Finally, Section 5 addresses the conclusions and future work.

2. State of the Art

With the increasing popularity of personal mobile devices and location-based applications, large-scale trajectories of individuals are being recorded and accumulated at a faster rate than ever. Thus, it is possible to understand human mobility from a data-driven perspective. As a consequence of this continuously increasing availability of data, works based on spatial-temporal data have received a lot of interest, with a large spectrum of methods developed.

2.1. Call Detail Records

Despite the listed benefits, the use of this type of data raises questions regarding the validity of previous works and the obtained conclusions, seeing that CDRs provide limited accuracy along the spatial dimension. In fact, studies such as that in [4] were conducted with the objective of proving that identifying users' most significant locations, such as home and work, is possible with a high degree of success. Another research work, around the same theme, Ref. [5] compared CDR-based individual trajectories with reference information from GPS logs. They found these two types of information to match with a good enough accuracy for extracting the user's movements.

The analysis of CDRs has already revealed the spatial recurrence and temporal periodicity of the movement patterns of people, who show a strong tendency to return to previously visited locations [1]. This entails a high predictability potential for human mobility. Similarly, important places in our lives (e.g., home and workplace) can be inferred from CDRs [6,7]. Routine or hobby-related locations have also been detected with success [8–10]. The use of CDR data in travel and tourism is exemplified in [11] where tourism transportation demand in Shanghai is inferred by mobile phone data and a system to propose new routes is developed. Other examples of previous works making use of CDR analyses include the detection and modeling of aggregate mobility flows at large scales [12], the characterization of individual movement patterns [13], or the computation of origin–destination matrices in urban areas [14].

Preserving privacy and data protection is a concern working with what can be considered sensitive personal information. To that effect, the work presented in [15] demonstrated that, with certain protection techniques applied, the re-identification of an individual via frequently visited locations, co-location pairs or spatial temporal data points with a high probability is not possible. Proper storage and retrieval techniques were also discussed in [16]. Some new attempts are now proposing privacy preserving methods for trajectory data based on CDR information, ranging from building recommendation systems locally on a user device [3] or a novel stay-region based anonymization technique that caters to important locations of a user [17].

2.2. Points of Interest

A point of interest (POI) is an entity of interest with a well-defined location. Points of interest can range from famous landmarks (e.g., museums, churches, towers), natural attractions (e.g., bays, coasts, waterfalls) to commonplace spots (e.g., coffee shops, taverns) [18].

The spatial interactions and distributions of POIs reveal different urban functions which can be associated with activities. For different types of activities (e.g., sports, eating, shopping), people can usually go to specific areas. For the same reason, many scientific studies (e.g., [2,3,10,19–21]) have focused on extracting features for area classification. POIs can be collected from various online sources and are frequently available for free through application programming interfaces (APIs) and online map service providers. For example, Qihang et al. [2] proposed to match the visit of a user to a specific registered location using POIs extracted via Foursquare's API. In contrast, Proux et al. [3] mapped information from four different geographic databases (HERE, Foursquare, Grand-Lyon, IGN) and nine different social and cultural databases (PreditHQ, International Showtime, Evenbrite, Songkick, Allevents.in, Meetup, Sportradar, 10 times) to perform user profiling by detecting significant places and their semantic meaning with external sources of information.

2.3. Semantic Disambiguation of CDRs

The process of semantic enrichment and the disambiguation of places has mostly been seen with the use of GPS trajectories. Studies such as [2,3] center their effort on venue check-ins, where the goal is to match a user to a visit of a registered location (e.g., restaurant, hotel, etc.). These works perfectly showcase the main difficulty with trying to disambiguate stop locations without user input. Using CDRs further augments the issues, as the mobility traces are much less precise. Some authors classify geographic areas by categories and

match these areas with the user's routines, instead of trying to match sparse locations with a specific venue.

For example, in [19], a virtual grid with cells of 500 by 500 m was constructed. Fixed-size cells are a common approach for the classification of geographical areas [20–22]. Each grid cell was classified according to four main categories (eating, shopping, entertainment and recreational) and matched with the user's CDR locations. To obtain that classification, the number of POIs associated with each activity category was recorded for each cell, creating an activity distribution map. Each cell activity proportion was then normalized to a value between the 0 and 1 and the K-means clustering algorithm was applied to create four distinct groups. The final step was finding the most probable activity category for each of the k clusters with a probability function. No ground truth data and no validation were carried out, as the authors fixated in their conclusion on the difficulty and privacy concerns in obtaining such data.

In [10], the metropolitan area of Milan was divided into regions by using density clustering to aggregate the groups of proximate cell towers. In sequence, the POIs obtained from foursquare were used to classify these areas by the most frequent type of POIs present. Study area subdivisions were classified by the top level categories in the POI dataset (e.g., shop, food, nightlife). The main focus of the work was to categorize explorers, those who are inclined to break out of their daily mobility routine and explore new places, and find city areas associated with this behavior. Results and conclusions were based on exploratory analysis, finding, for example, the areas and categories most related to exploration and attempting to validate with existing knowledge of the study area.

Experiments were also conducted with Voronoi diagrams, road segmentation layers, transportation analysis zones (TAZ) and administrative layers [23]. This work primarily uses the statistics of CDRs to classify geographical areas and POIs as a complement. First, based on CDR data, they calculate the parameters required including weekday and weekend CDR density spikes, number of peak values, the intensity of peak values, and the distribution of peak values; this allows the travel behaviors and public cognition to be understood. Then, POI category density was used to complement the analysis of CDR events in each geographical division and based on that, the identification results are modified. Validation was made through existing knowledge of the study area and comparing the results obtained to that of the known reality.

Compared to the state of the art, the main novelties introduced by our work are the following: (i) the inference of mobility routines outside of tourist and exploration activities, using a dataset more tailored for this analysis (Facebook Places); (ii) the circular sector approach, to create signal areas relative to each antenna is, according to our research, an innovative way to subdivide space for classification; (iii) in contrast to previous work, our approach relies on multiple POI features (e.g., category percentage, popularity and opening hours) to match location data with geographical area classification.

3. Materials

This section presents, discusses and analyzes the data obtained or gathered in this work. Exploratory data analysis, using statistical graphics and other data visualization methods, was conducted to summarize the CDRs' main characteristics.

3.1. CDR Dataset

For the development work carried out in this paper, a dataset comprising 35,676 SIMs from the region of Coimbra, Portugal, was used. Data collection corresponded to a period of 4 months, from 1 July 2020 to 30 October 2020, totaling 41,371,218 unique events. This amount of data was large enough to experiment with and apply several state-of-the-art techniques and obtain representative results while still being acceptable and manageable in terms of size.

Data entries were a mix between event-driven and network-driven entries, which means that some events did not require user participation being generated periodically

without human intervention. Each entry in the data has: a user identification field; the timestamp of the event; a unique identifier for the cellular antenna; its corresponding location coordinates; the antenna's initial and final angle of action in degrees; as well as the estimated range value in meters.

Before being made available for research, the user's identifiers were pseudonymized. This means phone numbers were encrypted with a hash function. This function remained unknown to us, the researchers, preserving the anonymity of user identity.

Analyzing the obtained Call Detail Records with regard to the events made by each user, we can ascertain the values present in Table 1. For the 35,676 individuals, there is an average of approximately 1346 unique events recorded with a standard deviation of approximately 425. The number of users with more than 2000 events is small with the maximum recorded being 7288.

Table 1. Dataset analysis on number of events per user.

User Count	35,676
Mean Events	1346.32
Std	425.81
Min	1.000
25%	1109.00
50%	1351.00
75%	1561.00
Max	7288.00

Figure 1 represents a histogram of the events per user. As can be seen from Figure 1, the events do not follow a normal distribution. In fact, they have an appreciable positive skewness while peaking at around the 1300 events mark.

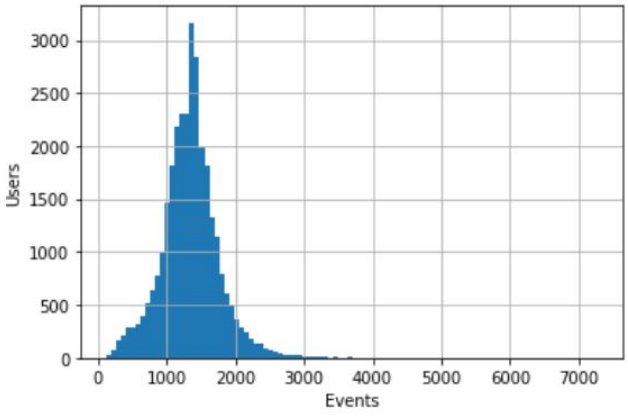


Figure 1. Histogram of the number of events per user.

The period of July/October 2020 represents, at the time of writing, the best possible chance for normal patterns of population movement from the data that can be made available to us since this represents a more relaxed period of COVID-19 confinement in Portugal [24]. Even though we know it will hardly give us an accurate indication of pre-pandemic mobility, it is, however, the closest we obtained since the data collection for this project began.

In search of a better indication of mobility in the data, we conducted an analysis of unique cell towers visited by each user. This would serve as a better indication if this data contains potential for mobility studies, or is compromised by the atypical situation lived throughout the year of 2020.

The information presented in Table 2 and Figure 2 shows that the average user has 25 different locations recorded and the standard deviation is relatively high, which should be expected as there is a big spread of values. This should certify that although time-frame for this type of study is not ideal, the dataset contains a good number of visits for each person.

Table 2. Dataset analysis on unique locations per user.

User Count	35,674
Mean (Unique Locations)	25.309
Std	18.406
Min	1.000
25%	12.000
50%	24.000
75%	34.000
Max	224.000

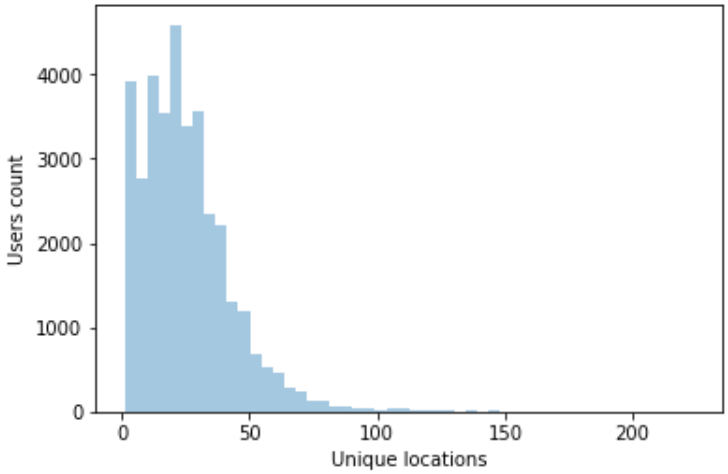


Figure 2. Histogram of unique locations per user.

3.2. POIs Dataset

POIs can be extracted via a single API or obtained and aggregated from several sources. In our particular case, it made more sense to use a single source since places are classified into various categories covering a variety of subcategories, and there are overlapping problems in different datasets, so it would be necessary to reconstruct and reclassify the POI data to join two or more sources.

Facebook is probably the most popular social networking site that makes it easy for people and/or businesses to connect and share with family, friends and clients online. Facebook Places is an associated geolocation service built into Facebook that is designed to help users share their favorite spots and discover new ones. Users can "check in" at various locations, from cities to small stores. Additionally, users are given the ability to create a new POI if the one they intend to 'check-in' or review does not already possess a Facebook Page. Business owners can claim and certify the pages created by a third party by following a verification process.

The main benefit of this data source when compared to Foursquare ([2,10,25]) is the wider reach of the Facebook platform and as such the amount of POIs is increased as expected. There is, in fact, a representation of categories that are not present in Foursquare's database, including organizations, societies, finance and healthcare. These categories, although not as important for tourism or leisure, are important to infer everyday mobility

motivations for the resident population. Furthermore, by observing both datasets, it was perceived that a bigger percentage of Facebook POIs contained information on opening and closing hours.

Furthermore, a dataset had already been constructed for the whole country in a previous work [26]. This allowed access to an extensive offline database using the code provided in the aforementioned work. In total, the dataset has 221,724 unique points spread over hundreds of categories of different hierarchies. An excerpt of the POIs data can be seen in Table 3.

Table 3. Sample of the Facebook Places POI table.

Name	Check-Ins	Hours	Latitude	Longitude
Restaurante Aviz	425	[[8, 0], [9, 0]]	39.82468	−7.4915
AZULMIR	15	[[9, 19], [9, 12]]	40.43211	−8.72678
B-Culture	0	[[9, 19], [9, 13]]	41.45011	−8.33808
...	...	...	...	...
Category	**City**		**Top Category**	
Portuguese Restaurant	Castelo Branco		Food and Beverage	
Wholesale and Supply Store	Mira		Shopping and Retail	
Medical and Health	Guimarães		Medical and Health	
...	...		...	

3.3. User Survey

To carry out validation on predicted user activities, a survey was made by the telecommunications service provider as the information needed was not present and could not be inferred by any data gathered to date. The survey was directed to the existing user pool of the original CDRs dataset in order to compare the knowledge obtained by our methods with reality. Since it was a voluntary questionnaire, it meant that not all users participated. From the total 35,676 users, only 574, or approximately 1.61%, voluntarily gave their answers. This questionnaire included information such as: the professional activity of the client, work schedule, if the client has a second home, where they spend the weekend, main interests/habits, and exercise frequency.

4. Proposed Approach

From this study and the analysis of the state-of-the-art research, we created an initial road map for experimentation and methods. The work can be divided into sections with the final goal being, with CDRs as input, to output a detailed table of routine areas and their classification.

4.1. CDR Pre-Processing

Pre-processing the dataset included retrieving and inferring additional data columns (e.g., the day of the week, workday/weekend) from the existing ones. This was intended to ease the detection of spatio-temporal patterns in the records. An integer for the day of the week (from 0 for Monday to 6 for Sunday) and a Boolean value for the workday or weekend (0 being a workday) were obtained from the timestamp columns. Furthermore, we adopted the time segment division found in [22]. For each entry, taking the timestamp, we verified the corresponding interval. One day is divided into eight time segments to capture the intraday variations in activity participation: early morning (3–6 a.m.); morning—peak hour (6–9 a.m.); morning—work (9 a.m.–12 p.m.); noon (12–2 p.m.); afternoon—work(2–5 p.m.); afternoon—peak hour (5–8 p.m.); night (8 p.m.–12 a.m.); and midnight (12–3 a.m.) [22].

As seen by data exploration in the CDR data description section, there were some cases where users had a lower number of events than average—even those users with less than one event per day. As expected, these will add little to no information for our research purpose, since we seek a higher number of events in order to infer spatial patterns. Thus, we created a simple function that, taking as input an event threshold, filters out all users with a number of events below that threshold. For example, removing users with less than one event per day resulted in a reduction of 0.12% of the dataset or 25,858 unique events.

Another step was the detection of cellular tower reselection in the middle of calls, or in very quick succession, creating impossible trajectories when taking in account the speed of movement. This is due to automatic network load balancing, a phenomenon often called load sharing [7]. With this in mind, distances between the network towers were computed and, consequently, the traveling speeds of users were estimated in consecutive records. For the detection of the load sharing effect, a speed-based method was implemented. A sequence is identified if the tower switching speed exceeds a given threshold. We set the value at 200 km/h inspired by the work of Iovan et al. [27].

After these initial steps and before we could search for routine activity patterns, we needed an accurate identification of each user's home and workplace locations. These are most likely the places where people spend the majority of their time and represent a large portion of their mobile records. Finding these locations first is important because it allows us to focus our attention on relevant records for our research of habits outside of these places. Thankfully this topic has been a subject of many prior studies and there are proven methods with good accuracy.

4.2. Home and Workplace Detection

Motivated by Vanhoof et al.'s work [6], a mixed approach of time filtering and density-based clustering is proposed. Firstly, we selected the temporal intervals of search when someone is not likely to be found in the places we want to identify. In this case, the home time interval was defined as the period from 7 p.m. to 9 a.m. as per [6]. However, because they did not try their method for workplace detection, we defined working hours as the period from 9 a.m. to 5 p.m., a common schedule of 8 h for day workers. Additionally, workplace CDRs were constrained to workdays. Given the state-of-the-art research, we opted for density-based spatial clustering, or DBSCAN, as per the works of [7,8]. DBSCAN is still to this day considered a competent algorithm for grouping CDRs and finding important areas. Its recurrent appearance throughout the literature supported our choice to use it our methodology.

After identifying and excluding homes and workplaces from the individual user's data, we are left with the remaining locations. From these, we then need to understand which are the most relevant to the daily routine, i.e., the most visited ones that account for a substantial time expenditure.

4.3. Other Routine Locations

The chosen method to detect the home and workplace using DBSCAN could also be used to find other routine locations. Without the time restrictions of home/workplace hours and by keeping all the clusters, rather than highlighting the one with the most events, it would be a good candidate solution. The issue found with using this density-based clustering is that we would lose additional precision in pinpointing the exact user position. Antenna locations already have great uncertainty when it comes to matching the user position, and clusters consisting of several antennas would substantially increase the challenge. For routine locations, we want to retain the maximum precision possible. The larger the area, the more difficult it will be to match a specific activity.

Inspired by the work of Quadri et al's [10], which divided users' locations in classes of importance with respect to the number of unique visit days, a similar approach was used. The three classes are: most visited places (MVPs), locations most frequently visited by the user; occasionally visited places (OVP), locations of interest for the user, but only visited

occasionally; exceptionally visited places (EVP): non-routine places. To classify places in these classes, a relevance metric was calculated for each place in the user's records. The initial relevance of a location l for a certain user u: $R(l, u)$ was calculated by the number of unique days that the user visited the location $d_{visit}(l, u)$ over their total number of active days $d_{total}(u)$. As our main goal is not only to detect routine locations but also to infer activities, the relevance metric was modified to accommodate the need for a time window and day type. We separated user places by coordinates, time interval and type of day (workday/weekend). The final metric for calculating the relevance of a location, $R(l_{tm,d}, u)$, is that presented in Equation 1. Instead of counting the unique days that the user visited location l, we counted the unique days that the user visited l in time interval tm and type of day d:

$$R(l_{tm,d}, u) = \frac{d_{visit}(l_{tm,d}, u)}{d_{total}(u)} \quad (1)$$

We used the calculated metric as input to a K-means clustering algorithm, this time with input value k = 3 to obtain the three distinct groups. Figure 3, a 3D scatter plot, shows coordinate points clustering by the relevance metric for one selected person in the data. Note that the Z axis represents the relevance metric while the X and Y are latitude and longitude, respectively. Purple color coded point, with the highest relevance score are MVPs, with orange points being OVPs and blue points EVPs.

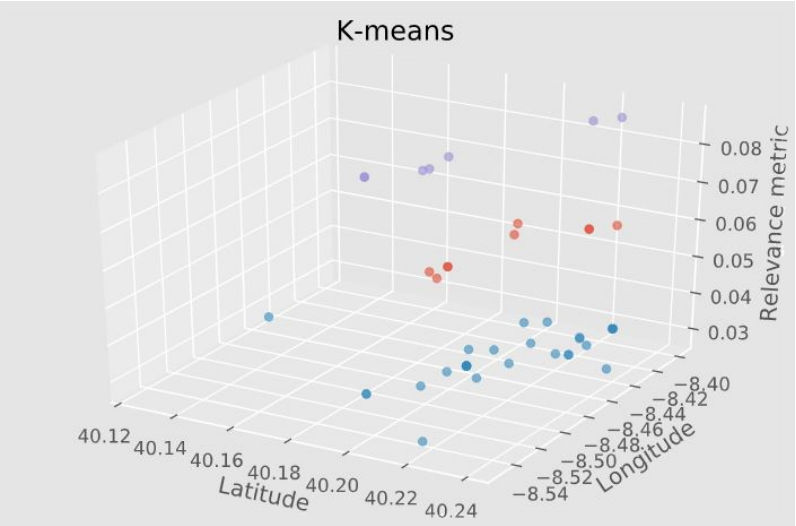


Figure 3. 3D scatter plot of the K-means clustering applied to user locations with K = 3.

Exploration or holiday-related activities (EVPs) do not entail a significant pattern in the data to be considered and are not analyzed further. The idea is that excluding home and work, we find other frequently visited places including MVPs and OVPs, that have significant importance for each user.

4.4. Geographic Regions Classification

To provide better insight into the motivations behind the mobility, at this point, we opted to subdivide the study area and classify the resulting geographic regions with the most likely activity. This is an important step in order to obtain the user's classified routine locations.

The selection of the regions is important as the size and shape can influence the final results. A slightly larger or differently shaped region can encompass more POIs, skewing the activity classification. We needed well-defined region boundaries that represented the search area in order for a function to return all contained POIs. Several approaches were considered, including fixed size ([19–22]) and dynamically sized [23]; however, we proposed a new type of region for activity classification using the antenna's signal attributes.

In our data, we accessed the values of the angle of coverage and the maximum expected range of each telecommunications antenna located in the study area. This resulted in the creation of circular sectors, corresponding to the antenna signal. In our understanding, these regions represent the user position with good estimation, as the user has to be within the boundaries of the signal range in order to connect to the corresponding antenna. Additionally, matching with the routine locations will be easier as these locations are also identified at the antenna level. The identified routine locations are associated with the antenna to which the user is connected, so we can infer that they must be within antenna's signal. An example of an antenna signal area and contained POIs can be seen in Figure 4.

It was necessary to use a specialized points of interest (POIs) dataset as the base data for the region POI mapping. We chose to use Facebook Places, a location-based social network (LSBN) that offers detailed information on 221,724 unique points spread over hundreds of categories of different hierarchies in Portugal [26]. The main attributes of each POI are: the name, Cartesian coordinates, city, opening and closing hours for both workdays and weekends as well as bottom-level category, top-level category and number of check-ins. POIs positioning is defined by the Cartesian coordinates present in the Facebook Places dataset. Although large POIs could theoretically encompass more than one antenna's signal, we only use a single pair of coordinates to infer its presence in those areas. An excerpt of the POI dataset can be seen in Table 3.

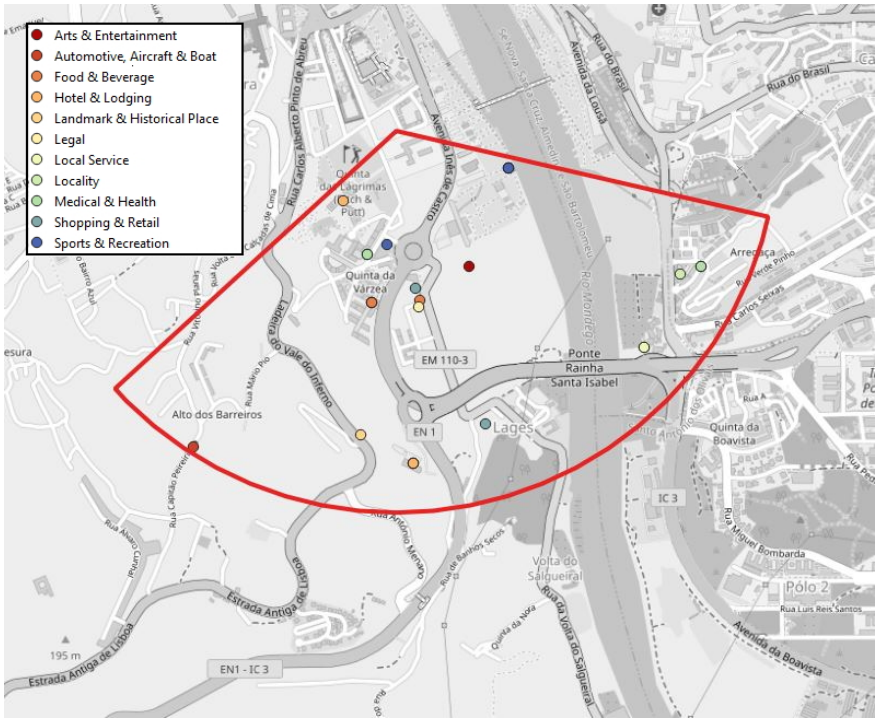


Figure 4. Example of an antenna's signal area.

Running a search function on the POI dataset, we managed to obtain a table with all contained points and respective attributes for each region. To obtain the final output of the classified regions, we need to filter and extract some information from the POIs.

As previously mentioned, in the pre-processing phase, a field was created to divide the CDRs into eight time intervals, as the time of day was taken into account and having a limited number of possible times instead of a continuous timeline facilitates classification. This means that we could classify all regions, at each specific time interval, for each day type (workday/weekend). To filter the POIs, we created a function to check for an intersection between each time interval and the opening hours of each POI. Let (a, b) symbolize the POI opening and closing hours and (x, y) the CDRs time interval; the base rule is to find the cases in which x is in-between (a, b) and the cases where a is in-between (x, y).

In our approach, we do not try to match the sparse user positioning with a single POI or a specific visit. We instead indicate the activity that the user is more exposed to in these locations, depending on the time of visitation and day of the week. Thus, we take into account the percentage of POIs of each category inside the area we are trying to classify. As such, the number of POIs is relevant but only when taking into account the relative percentage of a category when compared with the others. Our confidence measure for classification takes additional features such as opening hours and the popularity of each given POI. This also means that overlapping antennas should have the same activity for the same time periods and not affect the activity prediction for a user.

As points were already assigned to a category by the POI data source, we started by using those as our class labels. There is a problem, however. Lower level categories are sometimes over-specific and need to be grouped into broader classes to increase our chance of an accurate classification. For example, several types of restaurants (e.g., fast-food, Portuguese, Asian) can be grouped under one class label, food and beverage. This is the main reason why we decided to reclassify each POI according to the higher-level Facebook Places categories.

There was still one more value present in the POIs dataset that we could use in the hope of improving results, namely the number of check-ins. A check-in is a user registration of their presence in the location via a social network. As this value is related to the popularity of a given location, it could be inserted into the classification as a weight applied to that class.

From the region POI table, we counted the number of POIs from each label l. Those individual values were then divided by the total number of POIs in the region R, giving us a new area table with the label's percentage. In addition, for each class label l in region R, we sum the number of check-ins and then divided the individual results by the total number of check-ins in the region R. A new column was then added with this percentage to the existing table as can be seen in Table 4. To use both values in the calculation, we multiply the label percentage by the label check-ins to obtain our metric for classification. The resulting equation is written in Equation (2):

$$C(l,R) = \frac{Count_{POIS}(l,R)}{Count_{POIS}(R)} * \frac{Sum_{check-ins}(l,R)}{Sum_{check-ins}(R)} \quad (2)$$

Table 4. Example of a region classification calculation.

Facebook's Top Category	POIs Percentage	Check-Ins Percentage	Result
Food and Beverage	0.125	0.009709	0.001214
Beauty, Cosmetic and Personal Care	0.125	0.019417	0.002427
Religious Organization	0.125	0.064725	0.008091
Medical and Health	0.375	0.284790	0.106796
Shopping and Retail	0.250	0.621359	0.155340

Once we had the regions classified, including all time intervals and day possibilities, we reached the final step where we merged this information with the previous step. The matching key was the corresponding antenna identifier, the *cell id*. This identifier is present in both the region classification table, because regions are associated with an antenna, as well as the identified routine locations. From this merge, we can have a good idea of the user's routine patterns throughout the day, during workdays and weekends.

5. Results and Discussion

In this section, experimental results are summarized and discussed. Some results were validated with ground-truth data and qualitative evaluations were made in cases in which data were not obtained.

5.1. Experimental Results on Regions

Starting with the results in region classification, because our approach for geographic area subdivision does not match any existing administrative or area segmentation diagrams, obtaining ground-truth data for comparison and subsequent validation was not possible. We instead used a discussion approach found in other works with similar objectives [20,23] using knowledge from the study area to make a qualitative analysis of the results.

We take the antenna visualized in Figure 4 with the POI distribution present in Figure 5 as an example. The achieved classification for the antenna is present in Table 5. The predominant activity classification throughout workdays is *Medical and Health* and *Sports and Recreation*. This is due to a high number of POIs pertaining to these categories open in the outlined map region, including care centers, health clinics, a physical fitness center and some outdoor Padel fields. This region also intersects with the locations of some popular restaurants, but is only classified as *Food and Beverage* in the time interval past midnight until 3 a.m., when it is already past closing hours for most other POIs. This is consistent with our knowledge of the region, since POIs related to *Food and Beverage*, which include bars, cafes and restaurants, tend to close at a later hour than health- or sports-related businesses. Furthermore, consistent with known reality is the fact that gyms and sport activities open earlier than other types of POIs, for people that prefer to take part in these activities early in the morning, giving strength to the 6 a.m.–9 a.m. classification. The last point to focus in this area is the change in activity on the weekends between 2 p.m. and 8 p.m. The explanation could be related to the presence of the Hotel Quinta das Lágrimas inside the outlined region, whose gardens are a popular tourist location. As it might be a more frequent choice for a visit during the weekends, the check-ins count could increase its relevance at this particular schedule.

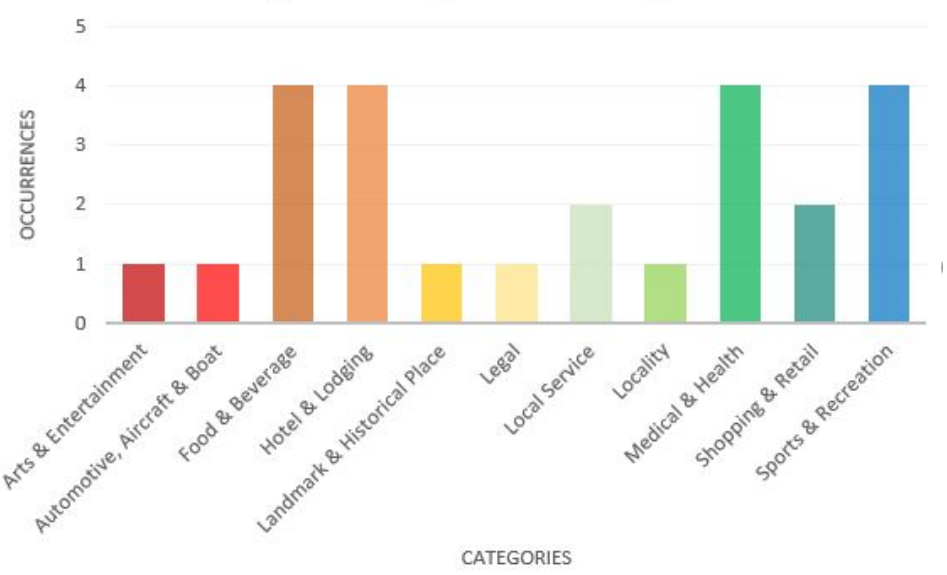


Figure 5. Histogram of high-level categories present in the example antenna of Figure 4.

As this process was repeated with other antennas, subjecting them to the same evaluation, we find that the classifications obtained generally matched our expectations for that particular region. Although the Facebook Places dataset does not contain every existing POI, it contains enough relevant ones to capture the main activities or motivations that might lead an individual to a visit.

Table 5. Region classification by type of day and time interval.

Temporal Interval	Classification	Temporal Interval	Classification
(0, 3)	Food & Beverage	(0, 3)	Food & Beverage
(3, 6)	None	(3, 6)	None
(6, 9)	Sports and Recreation	(6, 9)	Sports and Recreation
(9, 12)	Medical and Health	(9, 12)	Medical and Health
(12, 14)	Medical and Health	(12, 14)	Medical and Health
(14, 17)	Medical and Health	(14, 17)	Hotel and Lodging
(17, 20)	Medical and Health	(17, 20)	Hotel and Lodging
(20, 24)	Sports and Recreation	(20, 24)	Sports and Recreation

5.2. User Routines Validation

To date, excluding home and work, we find other frequently visited places, most visited places (MVPs) and occasionally visited places (OVPs), that have significant importance for each user. In these locations, the time of visitation and the greater offer/popularity of services available in that area may indicate that the user is more exposed to one activity. Even if the user did not visit any of the POIs in the area, the frequency of visitation (not being your home) shows that that place is important to the user, and depending on the time of day, this same place will have different services available.

It is important to reiterate that region classification is done before matching with the detected routine locations. This is the final step in the process, giving us the predicted user activities. This merging of both data tables allows us, for each user, to obtain a visualization similar to the one present in Figure 6, where routine locations are separated and labeled according to the prevalent activity. Some locations might repeat for different time intervals having a different classification. These routines are inferred from 4 months of activity, between July and October of 2020.

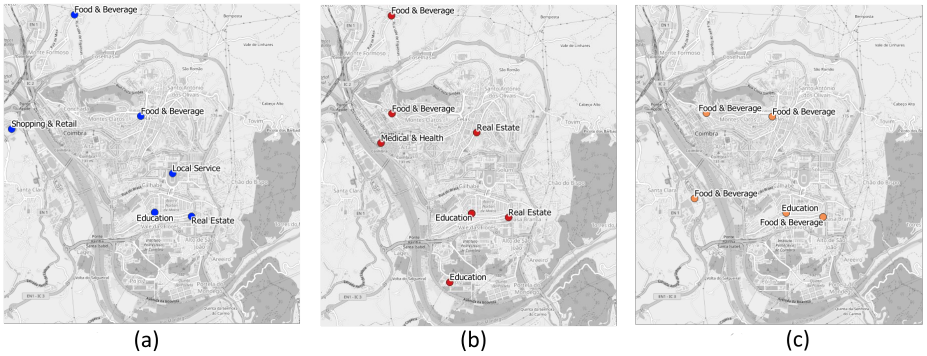


Figure 6. Routine locations classified for one user: (**a**) time interval from 9 a.m.to 12 a.m.; (**b**) time interval from 12 a.m.to 2 p.m.; and (**c**) time interval from 2 p.m. to 5 p.m.

Another way to visualize the activity patterns of users is by grouping the results by the type of day and time interval, as demonstrated in Table 6. This table shows, in each time interval, for both workdays and weekends, what were the activities identified according to the routine places. The user used for this example was the same as that in Figure 6.

Table 6. Example of a user's predicted activities.

Workdays		Weekends	
Temporal Interval	**Activities**	**Temporal Interval**	**Activities**
(9, 12)	[Local Service, Real Estate, Education, Food and Beverage]	(9, 12)	[Real Estate, Shopping and Retail, Food and Beverage]
(12, 14)	[Education, Real Estate, Medical and Health, Food and Beverage]	(12, 14)	[Food and Beverage]
(14, 17)	[Beauty, Cosmetic and Personal Care, Real Estate, Education, Food and Beverage]	(14, 17)	[Food and Beverage]
(17, 20)	[Real Estate, Education, Food and Beverage, Medical and Health, Local Service]	(17, 20)	[Food and Beverage]
(20, 24)	[Food and Beverage, Education, Medical and Health, Shopping and Retail]	(20, 24)	[Food and Beverage]

The main question of interest in the survey had users choose from a list of habits which ones were part of their weekly routine. Results could then be compared between our predictions of the user activities (as present in Table 6) and their given answers. However, the habits present in the question did not exactly match with our classification labels, so to that end, a match had to be made between the two. For example, the Facebook Places category 'arts and entertainment' contains the POIs associated with movie theaters and theatrical plays so we connected it to the survey interest of 'theatre and cinema'. The process was repeated for all six main survey interests: sports; automotive; bars and restaurants; beauty and fashion; theatre and cinema; and travel.

To validate our methods, we transformed the survey interest columns from the several possible answers (yes; no; do not know; blank) into true or false Boolean values. Accordingly, using the matching table, we can verify whether the real interests in an assigned true value are present in our predictions. If we find a match between the user answers and our predictions, we consider it a correct prediction, meaning that accuracy values can be calculated. Figure 7 shows the accuracy results for all six categories. Some categories' performances are worse than others, such as 'bars and restaurants'. This can mostly be explained by the fact that there are many POIs pertaining to this category which lead to an overestimation in the prevalence of this activity.

As concluded in the work of [25], one of the main factors that negatively impact user activity prediction is spatial uncertainty. An uncertain user position means a larger area of search for POIs, giving less importance to those actually visited, which could result in incorrect predictions. Consequently, we conducted experiments in order to reduce the spatial uncertainty by removing regions above a certain threshold of radius. Our x axis in Figure 7 is the average region radius for several radius thresholds. As revealed by the observed values, accuracy tends to increase for certain categories when the average radius is lower. In other cases, accuracy maintains or fluctuates its value very slightly. The correlation between accuracy and spatial granularity resemble that found in [25]. We can infer that if we were to obtain even more precise and relevant regions, results could improve further.

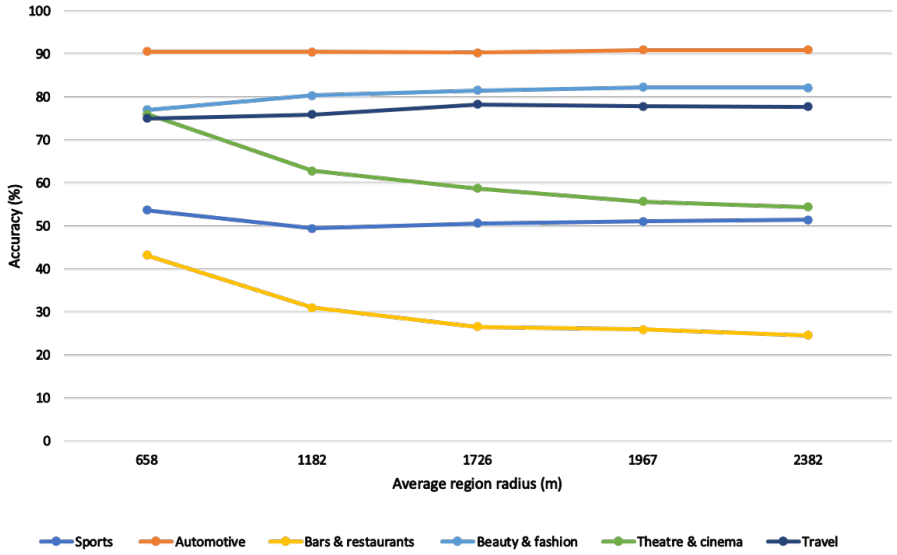


Figure 7. Accuracy of predictions in relation to average antenna's radius.

6. Conclusions

In this work, we addressed the objective of identifying and classifying routine locations using data from anonymized mobile communications events. We presented a group of interacting methods and developed a system to analyze individual mobility behavior. Such an analysis relies on the locations present in call detail records (CDRs) and intends to improve the understanding of user's regular places and their associated routine activities. Throughout the work, we focused on several main steps, data pre-processing, the detection of home and workplace, the detection of other routine places and finally, classification of geographic regions.

The review of the state of the art was essential to understanding the challenges faced by others and potentially explore parts that could be improved in those works. The circular sector approach, to create signal areas relative to each antenna is, according to our research, an innovative way to subdivide space for classification. This allows, in our understanding, for an improved match of the area where a user is when a mobile telecommunication event is made.

Comparing our results with those present in the state of the art, we conclude that for similar categories and taking into account an approximate spatial granularity, our work is on par with state-of-the-art methods being slightly above or below depending on the activity we are trying to identify.

However, we also faced some challenges. Due to the fact that data collection by the telecommunication service provider started in July 2020, the data provided originated from a time period during a global pandemic. Even though the months between July and October of 2020 had a higher user mobility than the stricter confinement period that followed, we know that they do not possess an accurate representation of pre-pandemic mobility.

It is important to reiterate some limitations of this work. For one, we assume that, during the period of study, we are not dealing with phones shared by more than one user and that no particular user changes or has more than one SIM card. Furthermore, there is the possibility that users, for any reason, did not make or receive calls in their workplace or home, precluding the correct detection of these places. The pandemic situation creates further possibilities that users mostly worked from home during the period of data collection.

Future Work

Some possible improvements were found by conducting the analysis of area's activity classification. Manually giving a weight to POIs of certain types to increase their importance depending on the time of day, e.g., for restaurants at regular meal hours, would possibly change the activities to better mirror population tendencies. The same effect could also be achieved with a popularity/check-in value that was hour dependent, but as far as we know, no POI dataset contains this information. There is still the question of points missing from the used dataset, as they might not be registered in the used data provider. One possible solution would be the combination of several POI datasets with the added difficulty of merging completely different category hierarchies into one.

The telecommunications service provider data currently contain cells from 2G to 4G; however, we do not differentiate between these cells. We understand that the signal areas of antennas of different technologies overlap. However, in our approach, these cells will have the same classification, and should not affect the predicted user routines we would like to explore a way to merge overlapping antennas and their records.

In the future, in addition to considering the frequency of visitation, its temporal distribution (every day, weekly, biweekly) could be a factor to take into account regarding the type of activity.

The areas created for classification, although closer to the reality of where the user might be, still remain too large to have a good percentage of certainty in terms of user activity. Newer information sources that have been discussed with the telecommunication service provider for future work have the potential to improve the user's location even further. Doing so allows for a smaller search area and generally more accurate methods. The arrival of 5G networks, with more precise smaller radius antennas, could be the next evolution step in mobility analysis using call detail records. All methods and created and implemented algorithms have the foresight of easy adaptation for future technologies allowing continuation work to be carried out.

Author Contributions: Conceptualization, Ana Alves, Marco Veloso and Carlos Bento; formal analysis, Gonçalo Ferreira; investigation, Gonçalo Ferreira; methodology, Gonçalo Ferreira and Ana Alves; project administration, Carlos Bento; resources, Carlos Bento; software, Gonçalo Ferreira; supervision, Ana Alves, Marco Veloso and Carlos Bento; validation, Gonçalo Ferreira; visualization, Gonçalo Ferreira; writing—original draft, Gonçalo Ferreira; writing—review and editing, Ana Alves and Marco Veloso. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of the data. Data were obtained from a third party and are available from the authors with the permission of said third party.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
CDRs	Call Detail Records
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EVPs	Exceptionally Visited Places
GPS	Global Positioning System
MVPs	Most Visited Places
OVPs	Occasionally Visited Places
POIs	Points of Interest

References

1. Gonzalez, M.C.; Hidalgo, C.; Barabasi, A.L. Understanding Individual Human Mobility Patterns. *Nature* **2008**, *453*, 779–782. [CrossRef] [PubMed]
2. Gu, Q.; Sacharidis, D.; Mathioudakis, M.; Wang, G. Inferring Venue Visits from GPS Trajectories. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017. [CrossRef]
3. Proux, D.; Roulland, F. Mobile Recommendation Challenges within a Strong Privacy Oriented Paradigm. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising, Chicago, IL, USA, 2019; Association for Computing Machinery: New York, NY, USA, 5–8 November 2019. [CrossRef]
4. Zhang, D.; Huang, J.; Li, Y.; Zhang, F.; Xu, C.; He, T. Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. In Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, Maui, HI, USA, 7–11 September 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 201–212. [CrossRef]
5. Ranjan, G.; Zang, H.; Zhang, Z.L.; Bolot, J. Are Call Detail Records Biased for Sampling Human Mobility? *Mob. Comput. Commun. Rev.* **2012**, *16*, 33–44. [CrossRef]
6. Vanhoof, M.; Reis, F.; Smoreda, Z.; Ploetz, T. Detecting home locations from CDR data: Introducing spatial uncertainty to the state-of-the-art. *arXiv* **2018**, arXiv:1808.06398.
7. Ayesha, B.; Jeewanthi, B.; Chitraranjan, C.; Perera, A.S.; Kumarage, A.S. User Localization Based on Call Detail Record. In *Intelligent Data Engineering and Automated Learning–IDEAL 2019*; Yin, H., Camacho, D., Tino, P., Tallón-Ballesteros, A.J., Menezes, R., Allmendinger, R., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 411–423.
8. Yang, P.; Zhu, T.; Wan, X.; Wang, X. Identifying Significant Places Using Multi-Day Call Detail Records. In Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, Washington, DC, USA, 10–12 November 2014; pp. 360–366. [CrossRef]
9. Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Martonosi, M.; Rowland, J.; Varshavsky, A. Identifying Important Places in People's Lives from Cellular Network Data. In *Pervasive Computing*; Lyons, K., Hightower, J., Huang, E.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 133–151.
10. Quadri, C.; Zignani, M.; Gaito, S.; Rossi, G.P. On Non-Routine Places in Urban Human Mobility. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018; pp. 584–593. [CrossRef]
11. Qian, C.; Li, W.; Duan, Z.; Yang, D.; Ran, B. Using mobile phone data to determine spatial correlations between tourism facilities. *J. Transp. Geogr.* **2021**, *92*, 103018. [CrossRef]
12. Csáji, B.C.; Browet, A.; Traag, V.; Delvenne, J.C.; Huens, E.; Van Dooren, P.; Smoreda, Z.; Blondel, V.D. Exploring the mobility of mobile phone users. *Phys. A Stat. Mech. Appl.* **2013**, *392*, 1459–1473. [CrossRef]
13. Hess, A.; Marsh, I.; Gillblad, D. Exploring communication and mobility behavior of 3G network users and its temporal consistency. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 5916–5921. [CrossRef]
14. Lenormand, M.; Picornell, M.; Cantú-Ros, O.G.; Tugores, A.; Louail, T.; Herranz, R.; Barthelemy, M.; Frías-Martínez, E.; Ramasco, J.J. Cross-Checking Different Sources of Mobility Information. *PLoS ONE* **2014**, *9*, e105184. [CrossRef] [PubMed]
15. Ding, J.; Ni, C.C.; Gao, J. Fighting Statistical Re-Identification in Human Trajectory Publication. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017. [CrossRef]
16. Vancea, F.; Vancea, C.; Popescu, D.; Zmaranda, D.; Gabor, G. Secure Data Retention of Call Detail Records. *Int. J. Comput.* **2010**, *5*, 961–967. [CrossRef]
17. Yang, J.; Dash, M.; Teo, S. PPTPF: Privacy-Preserving Trajectory Publication Framework for CDR Mobile Trajectories. *ISPRS Int. J. Geo Inf.* **2021**, *10*, 224. [CrossRef]
18. Khosrow-Pour, D.B.A. *Encyclopedia of Information Science and Technology*, 3rd ed.; IGI Global: Hershey, PA, USA, 2015. [CrossRef]
19. Phithakkitnukoon, S.; Horanont, T.; Di Lorenzo, G.; Shibasaki, R.; Ratti, C. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In *Human Behavior Understanding*; Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 14–25.
20. Zhou, X.; Liu, J.; Yeh, A.G.O.; Yue, Y.; Li, W. The Uncertain Geographic Context Problem in Identifying Activity Centers Using Mobile Phone Positioning Data and Point of Interest Data. In *Advances in Spatial Data Handling and Analysis: Select Papers from the 16th IGU Spatial Data Handling Symposium*; Springer International Publishing: Cham, Switzerland, 2015; pp. 107–119. [CrossRef]
21. Wang, F.; Chen, C. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 58–74. [CrossRef] [PubMed]
22. Diao, M.; Zhu, Y.; Ferreira, J.; Ratti, C. Inferring individual daily activities from mobile phone traces: A Boston example. *Environ. Plan. Plan. Des.* **2015**, *43*, 920–940. [CrossRef]
23. Yuan, G.; Chen, Y.; Sun, L.; Lai, J.; Li, T.; Zhuo, L. Recognition of Functional Areas Based on Call Detail Records and Point of Interest Data. *J. Adv. Transp.* **2020**, *2020*, 1–16. [CrossRef]
24. Available online: https://www.portugal.gov.pt/pt/gc22/governo/comunicados-do-conselho-de-ministros?p=13 (accessed on 13 February 2021).

25. He, R.; Cao, J.; Zhang, L.; Lee, D. Statistical Enrichment Models for Activity Inference from Imprecise Location Data. In Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications, Paris, France, 29 April–2 May 2019; pp. 946–954. [CrossRef]
26. Andrade, R.; Alves, A.; Bento, C. POI Mining for Land Use Classification: A Case Study. *Int. J. Geo Inf.* **2020**, *9*, 493. [CrossRef]
27. Iovan, C.; Olteanu-Raimond, A.M.; Couronné, T.; Smoreda, Z. Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies. In *Geographic Information Science at the Heart of Europe*; Springer International Publishing: Cham, Switzerland, 2013; pp. 247–265. [CrossRef]

Efficient neighbor searching in nonlinear time series analysis

Thomas Schreiber
Department of Theoretical Physics, University of Wuppertal,
D–42097 Wuppertal

July 18, 1996

We want to encourage the use of fast algorithms to find nearest neighbors in k–dimensional space. We review methods which are particularly useful for the study of time series data from chaotic systems. As an example, a simple box–assisted method and possible refinements are described in some detail. The efficiency of the method is compared to the naive approach and to a multidimensional tree for some exemplary data sets.

1 Introduction

Finding nearest neighbors in k–dimensional space is a task encountered in many data processing problems. In the context of time–series analysis *e.g.* it occurs if one is interested in local properties in a reconstructed phase space. Examples are predictions, noise reduction or Lyapunov exponent estimates based on local fits to the dynamics, or the calculation of dimension estimates. Other applications in physics include simulations of molecular dynamics with finite range interactions, where a box–oriented approach is used called "link–cell algorithm."[Fincham & Heyes, 1985, Form *et al.*, 1992]

As long as only small sets (say $n < 1000$ points) are evaluated, neighbors can be found in a straightforward way by computing the $n^2/2$ distances between all pairs of points. However, numerical simulations and to an increasing degree experiments are able to provide much larger amounts of data. With increasing data sets efficient handling becomes more important.

Neighbor searching and related problems of computational geometry have been extensively studied in computing science, with a rich literature covering both theoretical and practical issues. General references include [Sedgewick, 1990, Preparata & Shamos, 1985, Gonnet & Baeza–Yates, 1991, Mehlhorn, 1984]. In particular, the tree–like data structures are studied in [Omohumdro,

1987, Bentley, 1980, Bentley, 1990], and the bucket (or box) based methods in [Noga & Allison, 1985, Devroye, 1986, Asano *et al.*, 1985].

Although considerable expertise is required to find and implement an optimal algorithm, we want to demonstrate in this paper that with relatively little effort a substantial factor in efficiency can be gained. The use of any intelligent algorithm can result in reducing CPU time (or increasing the maximal amount of data which can be handled with reasonable resources) by orders of magnitude, compared to which the differences among these methods and the gain through refinements of an existing algorithm are rather marginal. Thus we give a simple and general algorithm which is worth the effort even for sets of only moderate size.

The box–assisted algorithm given here has been heuristically developed in the context of time series analysis [Grassberger, 1990, Grassberger *et al.*, 1991]. Similar procedures are proposed for this purpose in [Theiler, 1987, Kostelich & Yorke, 1988], while the k–d–tree (Sec. 2) seems to be the most popular approach [Bingham & Kot, 1989, Farmer & Sidorowich, 1988]. In Sec. 3 we describe a very simple version of a box–assisted algorithm for finding all points closer than a given distance ϵ. In Sec. 4 we describe how it can be improved by using *linked lists*. To illustrate the usefulness of this data structure, a very fast sorting method is presented. Furthermore we describe how to modify the basic algorithm in order to find a given number of neighbors rather than a neighborhood of fixed diameter. In Sec. 5 we will discuss some examples of the performance of the algorithms described. For comparison we give results obtained using the k–d–tree algorithm described in [Bingham & Kot, 1989], which represents a similar level of sophistication.

Both the box–assisted and the tree implementation used were chosen rather for simplicity than for optimality. The reader who wants to go beyond this will find some suggestions in Sec. 6.

2 The classical approach: multidimensional trees

How to find nearest neighbors in k–dimensional space? The textbook answer is to use multidimensional trees. From the point of view of computing theory they have the advantage that they can be proven to have an asymptotic number of operations proportional to $n \log n$ for a set of n points, which is the best possible performance for sets of arbitrary distribution.

We will see later that there exist algorithms which are faster (the number of operations is asymptotically proportional to n) provided the set is not too singularly distributed. Nevertheless, since trees are very widely applicable and accordingly very popular, we want to recall the basic idea. The algorithm is described in more detail in [Bingham & Kot, 1989].

Neighbor searching consists of two steps, first a data base is built and then this base is scanned in an efficient way to extract the required points. In this

section we want to use a tree–like structure as a data base. At each branching point, the remaining data set is split into two branches according to the result of a comparison. Say we have n vectors of Cartesian coordinates in k dimensions. Since there is no ordering relation in $k > 1$ dimensions we use one of the k coordinates at each level of the tree to determine into which branch a point falls. We cyclically use all the coordinates. The method is illustrated in Fig. 1. Since we will have n branching points and on average $\log_2 n$ levels, the construction of the tree requires $O(n \log n)$ operations.

Once the tree has been built up, we can now use it to find all points closer than ϵ to a given reference point $\mathbf{x}_0$. Instead of computing the distances to all the other points in the set we can now exclude branches which fall completely outside an ϵ–cube around $\mathbf{x}_0$.

3 Box methods: The basic algorithm

Consider a set of n vectors $\mathbf{x}_i$ in k dimensions, for simplicity rescaled to fall into the unit cube. For each x_i we want to determine all neighbors closer than ϵ,[1] or, strictly speaking, determine the set of indices

$$\mathcal{U}_i(\epsilon) = \{j : \ ||\mathbf{x}_j - \mathbf{x}_i|| < \epsilon\} . \tag{1}$$

This is preferable for technical reasons since we do not want to move the whole vectors around, in particular not if they are obtained as delay coordinates from a scalar time–series.

The idea of box–assisted methods is the following. Divide the phase space into a grid of boxes of side length ϵ. Then each point falls into one of these boxes. All its neighbors closer than ϵ have to lie in either the same box or one of the adjacent boxes. Thus in two dimensions only 9 boxes have to be searched for possible neighbors, *i.e.* for $\epsilon < 1/3$ we have to compute less distances than in the simple approach.

The technical problem is how to put the points into the boxes without wasting memory. How to provide the right amount of memory for each box without knowing beforehand how many points it will contain? The conceptually simplest solution is to obtain this information by first filling a histogram which tells how many points fall into which box. With this information one can divide an array of n elements so that each box is assigned a contiguous section of memory exactly as large as to hold all the points in the box (see Fig. 2). For each box one has to store the position where this section starts. The end of the section is given by the starting position for the next box. Once the indices of all the points are stored in the section corresponding to the appropriate box, scanning a box simply means scanning its section of the array.

We suggest the following implementation.

[1] Throughout this paper we use the sup norm. Neighbors in the Euclidean or the L^1 norms are also neighbors in sup norm.

1. To save memory we use two–dimensional boxes regardless of the dimension k, unless the set itself is of very high dimension. Neighbors in two dimensions are the candidates for neighbors in k dimensions. The two least correlated coordinates promise most information in two dimensions.

2. Provide an array `BOX` of size $n \times n$. If $n\epsilon > 1$ we waste memory. If $n\epsilon < 1$ we have to "wrap around" the set since the boxes do not cover the whole area. Thus we take the coordinates $\{x_j^{(k)} \bmod (n\epsilon)\}$ instead of $\{x_j^{(k)}\}$ to determine into which box a point falls.

3. Provide a linear array `POINTS` with n elements to contain the indices of all points.

4. Use `BOX` to make a histogram of how many points fall into each box.

5. Replace the histogram in `BOX` by an accumulated histogram, going through `BOX` row by row, say from left to right, and sum the number of points to be stored so far. Now each element of `BOX` tells where the section of `POINTS` corresponding to this box *ends*.

6. For every point determine into which box it falls, say `(I,J)`. Store the index of this point at the location on `POINTS` given by the current value of `BOX(I,J)`. Count down `BOX(I,J)` one step.

7. When all points are inserted, the elements of `BOX` tell where the corresponding sections of `POINTS` *start*.

8. In order to scan a box, just scan the corresponding section of `POINTS`. It ends where the section of the next box begins. Each candidate is tested whether it is a neighbor in the desired norm in k dimensions.

The resulting FORTRAN program is given as algorithm 1. It consists of two subroutines. The first one, `BOXIT`, scatters the points into a square array `BOX` and the appropriate sections of array `POINTS`. The second one, `FIND`, finds $\mathcal{U}_i(\epsilon)$ for a given index i. The indices of these points are stored in array `FOUND`.

The method works most efficiently if the expectation value of the number of points per box is $O(1)$. For small box size ϵ this can be achieved by providing roughly as many boxes as there are points. Thus we need $2n$ storage locations in addition to the data. For 2^{16} points or less, the arrays `BOX` and `POINTS` can be 2–byte integers.

4 Linked lists and sorting

The above algorithm can be slightly accelerated by the use of *linked lists* [Knuth, 1973] instead of linear arrays. We can save the time needed for forming and accumulating the histogram, while the storage needed is exactly the same. With

linked lists points can be added to the data base at any time, a useful feature for real–time applications. Thus it is also possible to compress filling of boxes and retrieval of neighbors into one sweep through the data, as was done in [Grassberger, 1990].

A one–directional linked list is a data structure in which each element contains a data item (or a pointer to it) and a pointer to the next element in the list. A pointer to a *null* location indicates the end of the list. To access the list we have to know where its first element is stored. A new element is inserted into the list by redirecting the pointer of the preceding element and letting the new element point to the following one (see Fig. 3). Note that we do not want to move the data items (the phase space vectors $\mathbf{x}_i$) around. We can save the storage for the pointers to these items by exploiting the one–to–one correspondence between list elements and data items: list elements and data items have the same index.

To illustrate the usefulness of this data structure let us for a moment think about the problem of ranking (or sorting) n numbers $x_i \in [0,1]$. To begin with, one could implement the method of *straight insertion*. First the list has only one element, x_1. Now the other numbers are added to the list consecutively. Each new number is either inserted between a smaller and a larger element, added at the end if it is larger than the last element, or added at the beginning if it is smaller than the first element. The number of comparisons required by this procedure is $\propto n^2$.

A box–assisted technique can be used to accelerate the algorithm to an operation count $\propto n$. The idea is to guess from the size of x_i where it has to be inserted. We divide the interval into n boxes and store for each box where the section of the list corresponding to this value starts. Inserting x_i we have to determine into which box it falls and scan only through the corresponding section of the list. For non–singularly distributed points the sections have a mean length of n/m. Sorting n points with the help of $m \propto n$ boxes yields an algorithm requiring $\propto n$ operations. For non–uniformly distributed points, the operation count is $\propto n^2/m^{d_2}$, where d_2 is the correlation dimension of the distribution.

Algorithm 2 shows a FORTRAN subroutine for ranking `NMAX` real numbers stored in the array `X`, and storing their ranks in the array `LIST`. It does this by means of an auxiliary array `BOX` of size `NBOX`+1. After calling the subroutine, `LIST(I)` contains the rank of `X(I)`, which is defined as {1 + the number of `X(J)` less than `X(I)`}. For equal values the time order is taken. It is easy to modify this algorithm so that it gives ordered linked lists as output.

Ranking the 50,000 points in the data set (iii) with `NBOX` = `NMAX`/2 takes 0.48 sec. of CPU time on a SPARCstation 1+ (average value of several runs). The fastest known comparison–based algorithm, quicksort (as implemented in [Press *et al.*, 1988]), takes 1.07 sec, slower by a factor of more than 2. For more evenly distributed data, this factor can increase up to 4.

In the neighbor search problem in k dimensions we use the same idea, only

now the lists starting at each box do not have to be sorted. The lists corresponding to each box all end with a *null* pointer to mark where to stop scanning.

Efficient neighbor searching with a fixed grid of boxes requires the box size to be equal to the radius of the neighborhoods we want to find. However, if one wants to find a given number K of nearest neighbors (neighborhoods of *fixed mass*), the procedure has to be modified. It is quite inefficient to choose a small box size ϵ and then scan more and more boxes until at least K neighbors are found. We rather suggest a slight modification of a scheme applied in [Schreiber & Grassberger, 1991] in the context of noise reduction. It consists of the following steps:

1. Start with a very fine partition of box length ϵ.
2. For all points find the neighbors within a distance less than ϵ.
3. Some of the points will have less than K neighbors with this resolution ϵ. These points are marked for the next sweep.
4. Coarsen the partition by a linear factor $\alpha > 1$, $\epsilon' = \alpha\epsilon$. With $\alpha = 2^{1/d}$, where d is the dimension of the set, one roughly doubles the mass of the neighborhoods. A new data base is created containing all points in boxes of size ϵ'.
5. Now only for the marked points find all the neighbors within a distance less than ϵ'.
6. Coarsen the partition until enough neighbors have been found for all points.

Note that we first satisfy the requirement of K neighbors for points in dense regions. Points in sparse regions are served last. If one needs to find neighborhoods for central points according to a given order, the method fails. In this case the algorithm of the last section can be modified by mapping the linear array `POINTS` to the square array `BOX` via a space filling Peano curve [Simmons, 1963]. Then boxes of different sizes $2^l\epsilon, l = 0, 1, \ldots$ map to contiguous sections of the array. Although this can be implemented efficiently, we prefer the use of trees for this problem.

5 Results and comparison

In this section we will give some typical results for the CPU time required by the methods described above. How long it takes to find nearest neighbors depends on the number of points n, the dimension of the set, the embedding dimension and either the radius ϵ of the neighborhoods or the minimal number of neighbors required. Rather than scanning through a four–dimensional parameter space we

will concentrate on three data sets, (i) a data set consisting of 1,000 iterations of the Ikeda map, (ii) 10,000 iterates of the same map, and (iii) 50,000 sampless taken at one site of a coupled map lattice which is in the state of space–time chaos, representing a high dimensional example.

We embed the first two samples in 4 dimensions and the third in 3, 5 and 10 dimensions. All CPU times were obtained on a SPARCstation 1+ but even their relative magnitudes will be slightly machine dependent.

We implemented a k–d tree as described in [Bingham & Kot, 1989], also using FORTRAN for comparison. If a language supporting pointers and recursive function calls (*e.g.* C) is used, it is about as easy to implement as the algorithm we present. (It will usually not be faster than the FORTRAN implementation though). For possible improvements of this algorithm see Sec. 6.

Even for the short set (i) the box method pays compared to the naive approach below $\epsilon = 1/3$ where both take about 7s. Only for very small $\epsilon < 1/8$ the tree is faster than the naive method, while it is consistently a factor of two slower than the box method. The time to find 50 neighbors is 11s for the naive, 6s for the tree and 4s for the box methods.

The results for the longer ikeda sample (n=10,000) are given in Table 1. Three different values of the fixed diameter ϵ were chosen: a quite small value (1/32 of the total attractor size), a typical value below which scaling could be expected (1/16), and a larger value (1/8). Also with this data set the box method was somewhat faster than the tree.

For the high–dimensional data in set (iii), the naive approach would take several hours of CPU time due to the large number of points. On this set the box method is fastest for fixed size neighborhoods below 1/16 of the linear extension of the phase space, where it takes about 60s to find all neighbors. For larger ϵ the tree wins slightly.

For fixed mass neighborhoods the situation depends on the embedding dimension. While for the three dimensional embedding the two algorithms were about equally fast (*e.g.*, finding 50 neighbors takes 760s), for the high embedding dimensions the tree is clearly faster (by a factor of 1.5–2), even if a grid of boxes in up to 5 dimensions is applied. In our method each box requires a storage location even if it is empty. Thus the dimension of the grid can only be increased under coarsening of the resolution, unless one has a lot of memory to spend. Circumventing this problem is addressed in the following section.

6 Possible improvements

The programs we applied in the above sections are very basic implementations of more general concepts. In this section we give some pointers to the literature which develops this further.

The problem of finding nearest neighbors is related to a whole range of problems of computational geometry [Gonnet & Baeza–Yates, 1991, Mehlhorn,

1984, Bentley, 1990, Asano *et al.*, 1985], problems in which a data base has to be built from a set of points in some geometric space. After that, certain queries have to be supported by this data base. One wants to devise algorithms optimal in terms of the size of storage and number of operations needed for a given task. The asymptotic scaling of the operation count with the number of points n is called the *complexity* (more specific, the *computational complexity*) of an algorithm. In the class of problems under consideration the total operation count comprises the effort to built the data base from the unstructured data and the effort to perform a query.

For the naive method of finding neighbors the effort to build the data base is actually zero since queries are answered using the raw data. The latter takes $O(n)$ operations for each query. With the simple k–d tree we used, the data base (*i.e.* the tree) can be built in $O(n \log n)$ operations on average and a query costs $O(\log n)$ operations. To fill the boxes in our box–algorithm always takes $O(n)$ operations but the query time depends on how many boxes we provided. We can trade storage for time and obtain optimal behavior for $O(n)$ boxes, where we expect to find $O(1)$ points to scan through.

Which algorithm is optimal depends on how many queries will be performed. For *in sample* applications like noise reduction, dimensions etc. the number of queries is equal to the number of data points. This can be different in forecasting. If only a few forecasts are asked for, there is no point in applying a fast algorithm at all. If many forecasts will be asked from a limited data base it will pay to optimize accessability of the data base even if the setup of the data base might be slower.

6.1 Refinements for tree methods

For tree methods several improvements have been proposed [Bentley, 1990, Sproull, 1991]. For all but the simplest implementation of trees it is highly recommended to use a programming language supporting structured data types, pointers and dynamical storage allocation.[2] The simplest k–d tree recursively splits the available space at each node into two halves, the splitting line at level J being parallel to the (J mod k)–th coordinate axis.

Balancing: To get optimal query times, the tree should be *balanced*, *i.e.* branches of the same level should contain about equally many points. However, in more than one dimensions it is hard to construct a set which would lead to a grossly unbalanced tree. This is different from the situation in one dimension, where a simple sorting algorithm based on a binary tree could perform very badly on presorted data.

[2] On some systems the last feature is formally supported but very slow. One can avoid its use as long as the number of nodes etc. is predictable from the number of data points, which is the case for all algorithms mentioned here.

Bucketing: A more relevant improvement is obtained by limiting the spatial resolution of the tree. Thus a *leaf* of the tree no more corresponds to a single data point but to a whole *bucket* of points close to the last node. This in itself reduces the query time to $O(\log m)$ where m is the number of buckets. A further improvement is gained by providing a pointer to the parent node for each node within the tree structure. The thereby possible *bottom up* queries take only $O(1)$ operations.

Principal cuts: While cuts parallel to the coordinate axes are natural if neighborhoods are defined in the sup norm, queries in the euclidean norm can be accelerated by cuts along principal axes of the distribution of points. Of course, this requires more work in constructing the data base. See [Sproull, 1991] for details.

6.2 Refinements for box methods

Clever high–dimensional meshes: As we have seen working on data set D, the algorithm given in this paper works best on not too high–dimensional problems. The reason is that in order to cover a high–dimensional space with $O(n)$ boxes of size ϵ we have to wrap around the set several times. Eventually only a small fraction of the points we find in a box will be real neighbors, most will be wrapped images of far away points. The box–assisted method proposed by Theiler [Theiler, 1987] circumvents this problem. An arbitrarily fine grid of boxes can be provided by storing only a table of the non–empty boxes. Thus to scan a neighboring box one has to look up the location of this box in the table. If the table has been organized by a *lexicographical sort*, queries can be done with $O(\log n)$ operations. This and the higher ($O(n \log n)$) cost for constructing the data base pays mainly in higher dimensions.

Adaptive buckets: The sorting algorithm we gave performs best on fairly uniform sets. For singularly distributed sets the use of multiple keys or *adaptive buckets* will be necessary (see [Noga & Allison, 1985]).

6.3 Alternatives for special problems

The algorithms described so far can be used as general purpose routines. However, one often has some additional knowledge about the data or one has to deal with a simpler task. For some cases the algorithms given above can be accelerated, one might even think of implementing a special purpose algorithm for the case at hand.

For some problems one does not really need the set of neighbors itself but only the number of neighbors within a given maximal distance. With the algorithms described so far this can accelerate queries if we can access the number

of points in each box or, using trees, store at each node the size of the subtree starting at this node.

One frequently deals with data of low significant resolution, either because the data is given in coarsely discretized form or only part of the given information is relevant due to noise. In this case it can be possible to form a histogram at the resolution of the data and use this histogram to count neighbors. If the number of distinguishable states is too high to provide bins for all of them, one can again store only the nonempty bins.

An efficient algorithm for finite resolution data can be obtained as a special case of the box–algorithm by Theiler [Theiler, 1987] mentioned above. If one box is provided for each distinguishable state, both the setup of the data base and the retrieval of neighbors can be done very fast.

On a wide range of problems even the simple tree– and box–assisted methods are both found to be quite efficient, the differences being rather marginal. In some cases (finding fixed mass neighborhoods in sets of high dimension) tree methods are faster, in other cases (sets of lower dimensions, fixed radius neighborhoods) box methods win. The differences are in no case strong enough to favour one of the methods in general.

However, we feel that the barrier of implementation of any fast neighbor search method is quite low with the simple algorithm given here, which we found much easier to implement (and to modify) than even the simple multidimensional tree [Bingham & Kot, 1989] we used for comparison.

Acknowledgements

I am grateful to Holger Kantz and Peter Grassberger who closely accompanied the work presented in this paper. Neil Gershenfeld drew my attention to alternative approaches for finite resolution data. The author receives a grant within the framework of the SCIENCE programme of the Commission of the European Communities under contract no B/SC1*-900557.

References

[Asano *et al.*, 1985] T. Asano, M. Edahiro, H. Imai, M. Iri and K. Murota, *Practical Use of Bucketing Techniques in Computational Geometry*, Computational Geometry, G. T. Toussaint, ed., Elsevier (1985).

[Bentley, 1980] J. L. Bentley, *Multidimensional Divide–and–Conquer*, Communications of the ACM, **23**, 214 (1980).

[Bentley, 1990] J. L. Bentley, *K-d Trees for Semidynamic Point Sets*, Sixth Annual ACM Symposium on Computational Geometry, 91, San Francisco, (1990).

[Bingham & Kot, 1989] S. Bingham and M. Kot, *Multidimensional trees, range searching, and a correlation dimension algorithm of reduced complexity* Phys. Lett. **A 140**, 327 (1989).

[Devroye, 1986] L. Devroye, *Lecture Notes on Bucket Algorithms*, Progress in Computer Science no. 6, Birkhäuser, Boston, (1986).

[Farmer & Sidorowich, 1988] J.D. Farmer and J. Sidorowich, *Exploiting Chaos to Predict the Future and Reduce Noise*, in "Evolution, Learning and Cognition", Y.C. Lee, ed., World Scientific, (1988).

[Fincham & Heyes, 1985] D. Fincham and D. M. Heyes, in "Dynamical Process in Condensed Matter – Advances in Chemical Physics," vol. LXIII, M. W. Evans, ed., Wiley, New York, (1985).

[Form *et al.*, 1992] W. Form, N. Ito, and G. A. Kohring, *Vectorized and parallelized algorithms for multi–million particle MD–simulation*, to appear in Int. Jour. Mod. Phys. C, (1992).

[Gonnet & Baeza–Yates, 1991] G. H. Gonnet and R. Baeza–Yates, *Handbook of Algorithms and Data Structures, in Pascal and C*, Addison–Wesley, Wokingham (1991).

[Grassberger, 1990] P. Grassberger, *An optimized box–assisted algorithm for fractal dimensions*, Phys. Lett. **A 148**, 63 (1990).

[Grassberger *et al.*, 1991] P. Grassberger, T. Schreiber and C. Schaffrath, *Nonlinear time sequence analysis*, Int. J.Bifurcation and Chaos **1**, 521 (1991).

[Knuth, 1973] D. E. Knuth, *The art of computer programming*, Vol 1. *Fundamental algorithms* and Vol 3., *Sorting and Searching*, Addison–Wesley, Reading, (1973).

[Kostelich & Yorke, 1988] E. J. Kostelich and J. A. Yorke, *Noise reduction in dynamical systems*, Phys. Rev. **A 38**, 1649 (1988).

[Mehlhorn, 1984] K. Mehlhorn, *Data Structures and Algorithms 3: Multi–dimensional Searching and Computational Geometry*, Springer (1984).

[Noga & Allison, 1985] M. T. Noga and D. C. S. Allison, *Sorting in Linear Expected Time*, Bit 25 (1985) 451–465.

[Omohumdro, 1987] S.M. Omohumdro, *Efficient Algorithms with Neural Network Behavior*, Complex Syst. **1**, 273 (1987).

[Preparata & Shamos, 1985] F. P. Preparata and M. I. Shamos, *Computational Geometry, an Introduction*, Springer, New York (1985).

[Press *et al.*, 1988] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes*, Cambridge Univ. Press, (1988).

[Sedgewick, 1990] R. Sedgewick, *Algorithms in C*, Addison–Wesley, Reading, (1990).

[Schreiber & Grassberger, 1991] T. Schreiber and P. Grassberger, *A simple noise–reduction method for real data*, Phys. Lett. **A 160**, 411 (1991).

[Simmons, 1963] G. F. Simmons, *Introduction to Topology and Modern Analysis*, McGraw–Hill, Tokyo, (1963).

[Sproull, 1991] R. F. Sproull, *Refinements to nearest–neighbor searching in k-d trees*, Algorithmica **6** (1991) 579.

[Theiler, 1987] J. Theiler, *Efficient algorithm for estimating the correlation dimension from a set of discrete points*, Phys. Rev. **A 36**, 4456 (1987).

Figure captions

1. How the plane is divided into branches of a tree.
2. How points within each box are stored in sections of an array.
3. How a new element is inserted into an existing list.

Table caption

1. CPU times (seconds) for data set (ii), n=10,000

Fig. 1

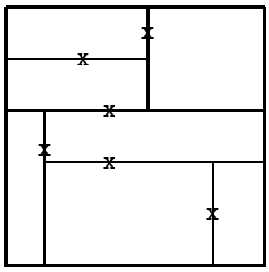


Fig. 2

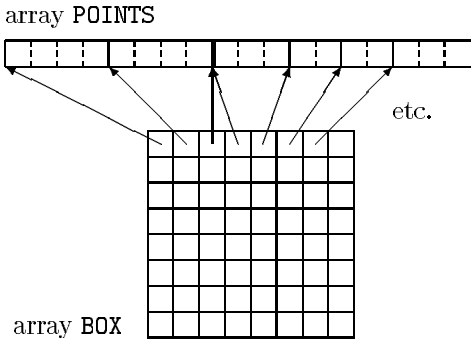
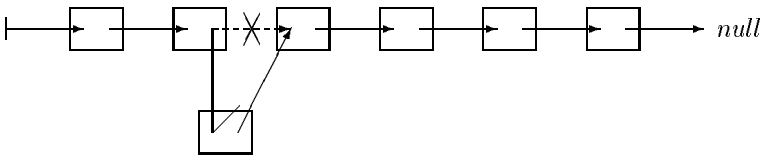


Fig. 3



task		box	tree	naive
fixed radius				489
	$\epsilon = 1/32$	5.6	9.8	
	$\epsilon = 1/16$	19.9	21.0	
	$\epsilon = 1/8$	72.2	75.9	
fixed mass	$m = 50$	45	64	675

Algorithm 1

```
      SUBROUTINE BOXIT(NMAX,X,K,EPS,BOX,POINTS)
```

This routine sets up the box data base for neighbor search in K+1 dimensional delay coordinates. Give a REAL array X containing NMAX points and the box size EPS. The data base is stored in arrays BOX and POINTS. The bitwise .AND.IM is used to to take an integer modulo IM+1

```
      PARAMETER(IM=127)   ! IM must be 2**l-1 for integer l
      REAL X(NMAX)
      INTEGER BOX(0:IM,0:IM), POINTS(NMAX)

      EPSINV=1./EPS

      DO 1000 I1=0,IM
         DO 1000 I1=0,IM
1000        BOX(I1,I2)=0

      DO 2000 N=1,NMAX-K   ! make histogram
         I1=INT(X(N)*EPSINV).AND.IM
         I2=INT(X(N+K)*EPSINV).AND.IM
2000     BOX(I1,I2)=BOX(I1,I2)+1

      DO 3000 I1=0,IM       ! accumulate it
         DO 3100 I2=1,IM
3100        BOX(I1,I2)=BOX(I1,I2)+BOX(I1,I2-1)
3000     IF(I1.LT.IM) BOX(I1+1,0)=BOX(I1+1,0)+BOX(I1,IM)

      DO 4000 N=1,NMAX-K   ! fill boxes
         I1=INT(X(N)*EPSINV).AND.IM
         I2=INT(X(N+K)*EPSINV).AND.IM
         POINTS(BOX(I1,I2))=N
4000     BOX(I1,I2)=BOX(I1,I2)-1
      END
```

```
      SUBROUTINE FIND(NMAX,X,N,K,EPS,BOX,POINTS,FOUND,NFOUND)
```

Given the data base set up by BOXIT, this routine finds all neighbors closer than EPS for the point with index N. The indices of the NFOUND neighbors found are stored in INTEGER array FOUND

```
      PARAMETER(IM=127)    ! same value as in BOXIT
      REAL X(NMAX)
      INTEGER BOX(0:IM,0:IM), POINTS(NMAX), FOUND(NMAX)

      EPSINV=1./EPS
      NFOUND=0
      I1=INT(X(N)*EPSINV).AND.IM
      I2=INT(X(N+K)*EPSINV).AND.IM

      DO 1000 J2=I2-1,I2+1    ! search neighboring boxes
         DO 1000 J1=I1-1,I1+1
            L1=J1.AND.IM
            L2=J2.AND.IM
            IF(L2.LT.IM) THEN     ! determine end of section
               IEND=BOX(L1,L2+1)
            ELSE IF(L1.LT.IM) THEN
               IEND=BOX(L1+1,0)
            ELSE
               IEND=NMAX-K
            ENDIF

            DO 1100 ISCAN=BOX(L1,L2)+1,IEND  ! scan box
               IPOINT=POINTS(ISCAN)
               DO 1110 M=0,K                 ! sup norm in k+1 D
1110              IF(ABS(X(N+M)-X(IPOINT+M)).GE.EPS) GOTO 1100
               NFOUND=NFOUND+1               ! it's a neighbor
               FOUND(NFOUND)=IPOINT
1100           CONTINUE
1000        CONTINUE
      END
```

Algorithm 2

```
      SUBROUTINE RANK(XMAX,NMAX,X,LIST)
```

Rank NMAX points given in X in the inteval [0,XMAX]. Ranks are stored in array LIST. It is assumed that in a composed logical expression containing .OR. or .AND. the first part is evaluated first. Depending on the result, the second part should not be evaluated. For other compilers the IF statements containing .OR. or .AND. have to be split into two statements.

```
      PARAMETER(NBOX=100000)
      REAL X(NMAX)
      INTEGER LIST(NMAX),BOX(0:NBOX)

      NL=MIN(NBOX,NMAX/2)
      SC=(NL-1)/XMAX
      DO 1000 I=0,NL
1000     BOX(I)=0

      DO 2000 N=1,NMAX
         XN=X(N)
         I=INT(XN*SC)
         IP=BOX(I)
         IF ((IP.EQ.0).OR.(XN.LE.X(IP))) THEN
            BOX(I)=N
         ELSE
  2         IPP=IP
            IP=LIST(IP)
            IF ((IP.GT.0).AND.(XN.GT.X(IP))) GOTO 2
            LIST(IPP)=N
         ENDIF
2000     LIST(N)=IP

      N=0
      DO 3000 I=0,NL
         IP=BOX(I)
  3      IF (IP.GT.0) THEN
            N=N+1
            IPP=IP
            IP=LIST(IP)
            LIST(IPP)=N
            GOTO 3
         ENDIF
3000  CONTINUE
      END
```

International Journal of Geo-Information

MDPI

Article

Location Extraction and Prediction Method Based on Floating Car Spatial-Temporal Trajectory

Shaoming Pan, Ziying Li and Yanwen Chong *

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; pansm@whu.edu.cn (S.P.); zy.lili@whu.edu.cn (Z.L.)
* Correspondence: ywchong@whu.edu.cn

Received: 30 March 2020; Accepted: 4 May 2020; Published: 7 May 2020

Abstract: Predicting the next important location by mining the user's historical spatial-temporal trajectory can be done for behavioral analysis and path planning. Since extracting the important location precisely is the premise of next location prediction, an enhanced location extraction algorithm is proposed to meet the requirements of dynamic trajectory via dynamic parameter estimation. To realize the estimation of parameters dynamically, the differences of floating car velocity in terms of spatial distribution and behavior in time distribution are considered in the location extraction algorithm. Then, an improved recurrent neural network (RNN) model is designed to mine the variation law of floating car trajectories to improve the accuracy of important location prediction under different conditions. Different from the traditional prediction model considering only the constraint of distance, the attention mechanism and semantic information are considered simultaneously by the proposed prediction model. Finally, the floating car trajectory of Beijing is selected for our experiments, and the results show that the proposed location extraction algorithm can meet the requirements of a dynamic environment and our model achieves high prediction accuracy.

Keywords: location extraction and prediction; spatial-temporal distribution; dynamic location extraction; RNN; self-attention

1. Introduction

With the popularity of smartphones and other equipment that can be used to record locations via the Global Navigation Satellite System (GNSS) [1], capturing and recording trajectory data based on such equipment becomes more convenient. Thus, much research on location-based service (LBS) has been proposed to improve service capabilities by mining the laws from recorded trajectory data, such as for targeted advertising, road assistance and navigation, personnel tracking, point-of-interest (POI) recommendations [2], and so on. Obviously, a complete LBS system can not only obtain a user's current location, but also predict the next possible location based on their historical trajectory information, so as to plan the navigation path in advance or analyze the user's behavior. Since recording and obtaining trajectory data have become easy and convenient, and predicting the next location based on historic behavior to find future destinations is a key factor for complete LBS [2], so it is important to accurately predict the next location.

In fact, the research indicates that people's mobility is highly dependent on their historical behaviors [3]. Location prediction can be divided into two types: personalized and popularized [4]. Personalized location prediction mainly looks at the historical trajectory of a single user, while popularized location prediction mines the travel habits of many users. Since personalized location prediction can be used to provide guidance for navigation and personality recommendations, it plays an important role in the research of location prediction, so this paper also pays attention to that research. It is clear that the trajectory of different users is basically different even if their visited locations might

be similar. Since the personalized location prediction is based on user's own behaviors, building a separate prediction model for each user might be preferable [5]. In addition, the different characteristics of user behavior have different influences on prediction. For example, predicting dynamic behavior is more difficult than predicting fixed behavior, and predicting many possible destinations is also more difficult than predicting only several possible destinations. Much different from pedestrian, bus, subway, train, and airplane travel, where the behavior features are mostly fixed or the destinations are mostly limited, not only are the behaviors of floating cars dynamic, but also their destinations are hugely variable. Thus, research based on trajectory data recorded in floating cars is more meaningful and more challenging.

Moreover, a location prediction algorithm mainly predicts the user's next important location, such as a mall, school, important road intersection, residential area, or attraction. In order to make meaningful predictions, discarding as much of the unimportant data as possible and keeping the most important data is a priority due to the massive and redundant location data recorded by floating cars. Thus, it is necessary to propose an important location extraction algorithm to filter and cluster the original trajectory data in order to reduce the number of trajectory points and obtain important locations as the input of the prediction model.

Based on the above analysis, an important location extraction algorithm and a precise location prediction model are the two aspects of location prediction research for LBS, the former for reducing the size of the dataset and data redundancy, and the latter for precise prediction of behaviors.

In order to obtain important locations for the prediction model, many traditional location extraction methods can be used, such as manual tagging algorithms [2,6] and clustering methods [7–9]. Among them, attention-based spatiotemporal gated recurrent unit (ATST-GRU) introduced a method to realize POI recommendations by users sharing locations and check-in information with others in location-based social networks (LBSNs) [2]. Tree-based hierarchical graph (TBHG) also introduced a method that marks important locations by matching the trajectory with the map [6]. It is clear that both of these methods require geographical information, which will lead to heavy workloads. Furthermore, it is difficult to obtain the check-in information of users because of personal privacy issues.

Hybrid methods and clustering algorithms are two typical models that can be used to extract locations and avoid personal privacy issues. A typical hybrid method simply annotates POIs with meaningful information from LBSNs [10]. Since an important location represents a geographic area where a user stayed for a certain interval of time, extracting the important location based on two scale parameters, time, and distance, instead of collecting check-in information can also avoid personal privacy issues. Therefore, many clustering algorithms have been proposed to extract locations by selecting appropriate distance and time thresholds [11], such as density-based spatial clustering of applications with noise (DBSCAN) [7–9], K-means [12], and ordering points to identify the clustering structure (OPTICS) [13] algorithms.

Although the above methods can be used to obtain satisfactory results in location extraction, artificially configuring the threshold and parameters, which have a great influence on the clustering results, is their obvious disadvantage. Due to the dynamic behaviors of different floating cars, obtaining common parameters for different users is impossible. Thus, an adaptive algorithm that can meet the dynamic requirements of different users should to be considered. In this paper, a dynamic DBSCAN (D-DBSCAN) algorithm based on the traditional DBSCAN algorithm is proposed to dynamically adjust the value of *Eps* (the neighborhood radius) in order to extract important locations via different parameters based on different situations. The D-DBSCAN algorithm considers the dynamic information of trajectories and can effectively filter strip clusters. This has been proved by location extraction experiments.

As the traditional domain for trajectory analysis, location prediction based on floating car trajectory data was a kind of trip-matching process in early approaches [14]. Sub-trajectory synthesis (SubSyn) [14] divided the trajectories into sub-trajectories and then fused the sub-trajectories of different users to match current trips. Different from the above trip-matching algorithm, various Markov chain

models offered a kind of destination-matching method where possible next locations were collected in advance and the prediction result was the probability of different important locations. Among them, Ashbrook et al. proposed a prediction method using the Markov model, where each node was an important location clustered from Global Positioning System (GPS) data, and a transition between two nodes represented the probability of the user traveling between those two important locations [7]. On this basis, Simmons et al. [15] proposed a hidden Markov model (HMM) to predict destinations. Furthermore, Ashbrook et al. [16] built models by adding the concept of support points to the hidden Markov model to improve prediction performance. Although the experimental results showed that the strategy of support points can be used to improve the performance of the algorithm, the model still cannot achieve high accuracy because a different choice of strategy will invariably lead to different performance, and no appropriate method has been proposed to match the various requirements in different situations. It is possible that a passenger may go to a totally different place where the floating car has never been. Thus, not only trip-matching algorithms but also destination-matching methods cannot meet the requirements of new next-location prediction. Furthermore, the above approaches will be ineffective when there is an absence of prior knowledge.

With the development of machine learning, deep learning technology has been widely used for next-location prediction. First, a kind of Bayesian network model was designed, and descriptive information of the trajectory was added to the model as a feature to perform destination prediction [14]. In addition, a location-prediction algorithm mainly makes predictions based on contextual information of the trajectory and the current location. Since the historical trajectory information is typical time series data, the location-prediction algorithm is also a typical time series prediction process. Recently, recurrent neural networks (RNNs) [17] have been adopted in machine translation [18], target recognition [19], video behavior recognition [20], sentiment classification, and image caption generation [21] and show promising performance in processing time series prediction compared with traditional methods. Therefore, RNNs can be used to predict next important locations [22,23]. Liu and co-workers [22] and Al-Molegi and colleagues [23] focused on using a set of features to obtain a good prediction performance. In these models all historical trajectory points have the same importance whether they are located in an intersection or in the middle of a straight road. In fact, trajectory points located in intersections, which greatly affect the turning direction, may play a more important role in next-location prediction. Some attention mechanisms need to be further considered for prediction algorithms based on the traditional RNN model in order to track the dramatic changes of floating car trajectories.

To remedy the two problems, this paper proposes a D-DBSCAN location extraction algorithm to cluster important locations and an attention-RNN location prediction algorithm to predict the next location. The main contributions are summarized as follows:

(1). A novel dynamic important location extraction algorithm based on DBSCAN is proposed to extract important locations via different parameters based on different situations to meet the requirements of dynamic situations. The algorithm can dynamically adjust the parameters by tracking different user behaviors and can effectively filter out strip shape clusters in order to avoid using such invalid data as input for the prediction algorithm.
(2). We propose attention-RNN location prediction, which can assign a different level of attention to historical track points to grasp the spatial characteristics of trajectories and closely track the user trajectory.
(3). A time-step window mechanism is added to the attention-RNN model to reduce the time consumption and computational complexity.

The rest of this paper is organized as follows. We briefly introduce some related work on important location extraction and prediction models in Section 2. The dynamic important location extraction algorithm (D-DBSCAN) and attention-RNN location prediction model are proposed in Section 3. Section 4 reports the experimental design, performance metrics, and extensive experimental results and discusses the merits and drawbacks of our results and the baseline methods (RNN [17], space

time features based RNN (STF-RNN) [23], and spatial-temporal RNN (ST-RNN) [22]). Finally, the conclusions are drawn in the last section.

2. Related Works

The DBSCAN algorithm does not need to set a fixed number of clusters in advance and has advantages over other algorithms, and the clustering results are not constrained by the cluster shape. However, the traditional algorithm still has some disadvantages, including that the parameters of the neighborhood radius *Eps* and the minimum number of points contained in the neighborhood (*Minpts*) must be set artificially, and the selection of parameters has a crucial effect on clustering results [8]. Unfortunately, it is difficult to select an appropriate value of *Eps* for the traditional DBSCAN algorithm, especially in massive nonuniform distribution conditions. For example, some important locations may be merged, or a single important clustering location may be split into many different clusters. In order to solve the parameter setting problem of the traditional DBSCAN model, Zhou et al. proposed a DBSCAN algorithm based on data partitioning [24]. The algorithm partitions the data according to density and establishes an *R** tree in each region to obtain a *K-dist* map. According to the *K-dist* map, the appropriate values are selected in different partitions, and finally the clustering results of each region are combined to obtain the final results.

The partition-based DBSCAN algorithm can be used to obtain clustering locations that are not constrained by shapes, and small areas can also be preserved so as to avoid discarding some important location information. However, due to the massive dataset, the partition-based DBSCAN algorithm can obtain many strip clusters that are formed by points on the road. Most of these strip clusters are meaningless locations and should not be input into the prediction model. On the other hand, some clusters that are formed by congestion points and intersection points play an important role in the location prediction application and these clusters need to be retained.

Based on the above analysis, a dynamic DBSCAN (D-DBSCAN) algorithm is necessary to dynamically adjust the value of *Eps* and extract important locations via different parameters in different situations, such as different velocity and time information, rather than artificially setting the parameters. The details of the D-DBSCAN algorithm are presented in Section 3.

In order to deal with the problem of trajectory location prediction, Brébisson et al. [5] inputted trajectory points one by one into an RNN network and used the memory function of the hidden layer to achieve the prediction purpose. Liu et al. extended the RNN and proposed a novel method called spatial temporal RNN (ST-RNN) [22]. ST-RNN can model local temporal and spatial contexts in each layer with time-specific transition matrices for different time intervals and distance-specific transition matrices for different geographical distances [22]. Al-Molegi et al. [23] proposed a method to leverage RNN to model people's movement behaviors in order to predict their next location. Space and time are included in the network as features, where their internal representations are learned by the network itself rather than relying on a manmade representation.

In the above algorithms, all historical trajectory points have the same importance whether they are located in an intersection or in the middle of a straight road. In fact, trajectory points located in intersections, which greatly affect the turning direction, may play a more important role in next-location prediction. In order to fully consider the different influence weights of historical trajectory data on predicted locations, an attention mechanism module is added in the traditional RNN network in this paper. At the same time, in order to solve the problem that the performance of the model will deteriorate rapidly as the length of the input sequence increases, a time-step window should be considered to reduce the time consumption and computational complexity and improve training efficiency. In addition, an attention mechanism and the trajectory semantic information must be fully considered to grasp the spatial characteristics of trajectories.

3. Methodology

3.1. Dynamic Location Extraction Method

The D-DBSCAN algorithm integrates the track point velocity information dynamically, which can adjust the key parameters of the algorithm dynamically. The main principles of the D-DBSCAN algorithm are as follows: (1) it calculates the instantaneous velocity of each trajectory point according to time sequences, (2) divides the points into different regions based on their velocity so as to reduce the velocity difference in the same region to better set *Eps*, and (3) it takes the velocity of the track point as an assist, dynamically adjusting the value of *Eps*.

Generally, denote $P = \{p_1(x_1, y_1),\ p_2(x_2, y_2), \ldots, p_n(x_n, y_n)\}$ and $T = \{t_1,\ t_2, \ldots,\ t_n\}$ as the trajectory data and the corresponding timestamp data of a certain user, then the velocity of each trajectory point $V = \{v_1,\ v_2, \ldots,\ v_n\}$ can be computed based on P and T. By sorting V in descending order to obtain the sorted velocity $V' = \left\{v'_1,\ v'_2, \ldots v'_n\right\}$ and then subtracting the sorted velocity one by one to get the velocity differences $\Delta V = \{\Delta v_1,\ \Delta v_2, \ldots,\ \Delta v_{n-1}\}$, the maximum velocity $v_{max} = v'_1$ and the minimum velocity $v_{min} = v'_n$ can be easily found from V'. After that the maximum $N-1$ velocity differences $\Delta V' = \left\{\Delta v'_1, \Delta v'_2, \ldots, \Delta v'_{N-1}\right\}$ is calculated ΔV, where all elements of $\Delta V'$ are more than a minimum velocity difference threshold ω. It is clear that the different values of ω may lead to obtain different numbers of the maximum velocity differences.

In order to divide the points into different regions, velocity partition thresholds can be computed firstly based on the velocity differences $\Delta V'$. $\Delta v'_i$ is obtained by subtracting v'_m and v'_{m+1}, and without loss of generality, denotes the smaller (v'_{m+1}) of two corresponding velocities (v'_m and v'_{m+1}) of each maximum velocity differences $\Delta v'_i$ as a velocity threshold, then the corresponding $N-1$ velocity thresholds of the maximum $N-1$ velocity differences can be noted as $\{vt_1,\ vt_2, \ldots,\ vt_{N-1}\}$, and thus, by sorting VT in ascending order to obtain the sorted $VT = \{vt_0, vt_1,\ vt_2, \ldots,\ vt_{N-1},\ vt_N\}$. VT can be defined as velocity partition thresholds, where $vt_0 = v_{min}$ and $vt_N = v_{max}$. Then, all trajectory points can be easily divided into N regions $R = \{R_1, R_2, \ldots, R_N\}$ based on VT. Thus, the mean velocity of each region $\overline{V} = \{\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_N\}$ can be computed.

In order to clearly demonstrate the above process to obtain velocity partition thresholds, a simple example is shown in Table 1. It is also clear that the different number of maximum velocity differences may lead to obtaining a different number of regions.

Table 1. Example of computing velocity partition thresholds.

Parameters	Value
V	{20, 80, 96, 100, 98, 60, 10, 8, 3|m/s}
V'	{100, 98, 96, 80, 60, 20, 10, 8, 3|m/s} and $v_{max} = 100$ m/s, $v_{min} = 3$ m/s
ΔV	{2, 2, 16, 20, 40, 10, 2, 5|m/s}
$\Delta V'$	{16, 20, 40, 10|m/s} when assuming $\omega = 6$(m/s) and all elements greater than 6 are selected from ΔV as the elements of $\Delta V'$.
VT	{v_{min}, 10, 20, 60, 80, v_{max}|m/s} = {3, 10, 20, 60, 80, 100, |m/s}. Since $\Delta v'_4 = 10$ is obtained via $v'_6 - v'_7 = 20 - 10$, the smaller of two corresponding velocities $v'_7 = 10$ is then selected as the one element of VT.

Obviously, the velocities of the different trajectory points V which belong to the same region are similar and thus the same parameter of *Eps* can be used. Therefore, the parameter *Eps* of each region can be dynamically calculated by Equation (1), and then clustering in each region. Finally, the clustering results of each region are combined to obtain the final results. The complete procedure is described as Algorithm 1:

$$Eps_i = \varepsilon \frac{1}{\overline{v}_i},\ i = 1,\ 2, \ldots,\ N \tag{1}$$

where ε is the influence factor which can be simply estimated via calculating the *K-dist* map [24] (the appropriate value of ε is selected as 0.06 in our experiment). Moreover, although the velocity based

partition method can reduce the velocity difference of the trajectory points in the same region so as to more appropriately set *Eps* parameter for each region, partitioning too many regions based on a too small velocity difference threshold ω may lead to splitting a single important clustering location into several fragmented location parts and which will reduce the accuracy of important location clustering as well as increase computational complexity. By the results of an ablation experiment, $\omega = 3(\mathrm{m/s})$ is defined as the minimum velocity difference threshold in our experiment so as to obtain the optimal region number N via the velocity differences ΔV.

Algorithm 1: D-DBSCAN algorithm for location extraction

Input: Given the trajectory data P, and the timestamp data T.
(1) Calculate V based on P and T, and sort V in descending order to get V'.
(2) Obtain v_{max} and v_{min} from V' and subtract V' one by one to get ΔV.
(3) Get the velocity partition thresholds VT based on ΔV, ω, v_{max}, and v_{min}.
(4) Partition all trajectory points into N regions R based on VT:
For $v_i \in V$
  If $vt_0 \le v_i \le vt_1$
    $v_i \in R_1$
  else if $vt_1 < v_i \le vt_2$
    $v_i \in R_2$
  ...
  else if $vt_i < v_i \le vt_{i+1}$
    $v_i \in R_{i+1}$
  ...
  else if $vt_{N-1} < v_i \le vt_N$
    $v_i \in R_N$
  End if
End for
(5) Calculate the mean velocity of each region $\overline{V}$.
(6) Update *Eps* via Equation (1) for each region, and cluster by region.
(7) Combine the clustering results of each region and calculate the latitude and longitude of the center point of each cluster.
Output: Location extraction results (sequence of locations).

As shown in Algorithm 1, a larger instantaneous velocity of the trajectory point indicates a smaller *Eps* based on Equation (1). Thus, the D-DBSCAN algorithm can realize dynamic clustering under different velocity conditions by adjusting *Eps* dynamically, and then key locations, such as intersection points, can be retained and strip clusters can be filtered out, and finally a better location extraction result can be obtained to provide a better guarantee for location prediction. Moreover, the algorithm takes full account of the dynamic characteristics of trajectories, thus better location extraction results can be obtained.

3.2. Attention-RNN Location Prediction

3.2.1. Traditional RNN Prediction Network

The architecture of an RNN is a recurrent structure, and traditional RNN includes the Elman network and the Jordon network. The Elman network feeds the hidden layer (h_t) back into the recurrent structure, while the Jordon network feeds the output of the network (o_t) back into the recurrent structure. Many network variants are derived from the Elman network, so in general, when it comes to RNN, it refers to the Elman network. The Elman network is adopted in this paper, and the hidden layer (h_t) is fed back into the recurrent structure. At each time t, we can predict the hidden layer (h_t) by

the previous moment hidden layer (h_{t-1}), and then feed the new hidden layer (h_t) back into the next hidden status. The formulation of the hidden layer in RNN is:

$$h_t = g\left(W^{(x)}x_t + W^{(h)}h_{t-1}\right) \tag{2}$$

where the activation function $g(\bullet)$ is a *tanh* function, $W^{(x)}$ is the input layer weight, and $W^{(h)}$ is the hidden layer weight.

The final result o_t of the network can be obtained by using the appropriate activation function $\sigma(\bullet)$ and the output layer weight $W^{(o)}$ on the hidden layer state h_t generated from the hidden layer.

$$y = o_t = \sigma\left(W^{(o)}h_t\right). \tag{3}$$

3.2.2. Attention-RNN Location Prediction

Due to the unequal time interval from the predicted location and different spatial adjacencies, different historical trajectory data have unequal influence weights on the predicted location. For example, the next location is greatly affected by the trajectory point at that turning point when the direction of the track changes. The attention mechanism module is added to the traditional RNN network to fully reflect the different influence weights of the historical trajectory data on the predicted location. The architecture of the proposed location prediction network in this paper is represented in Figure 1.

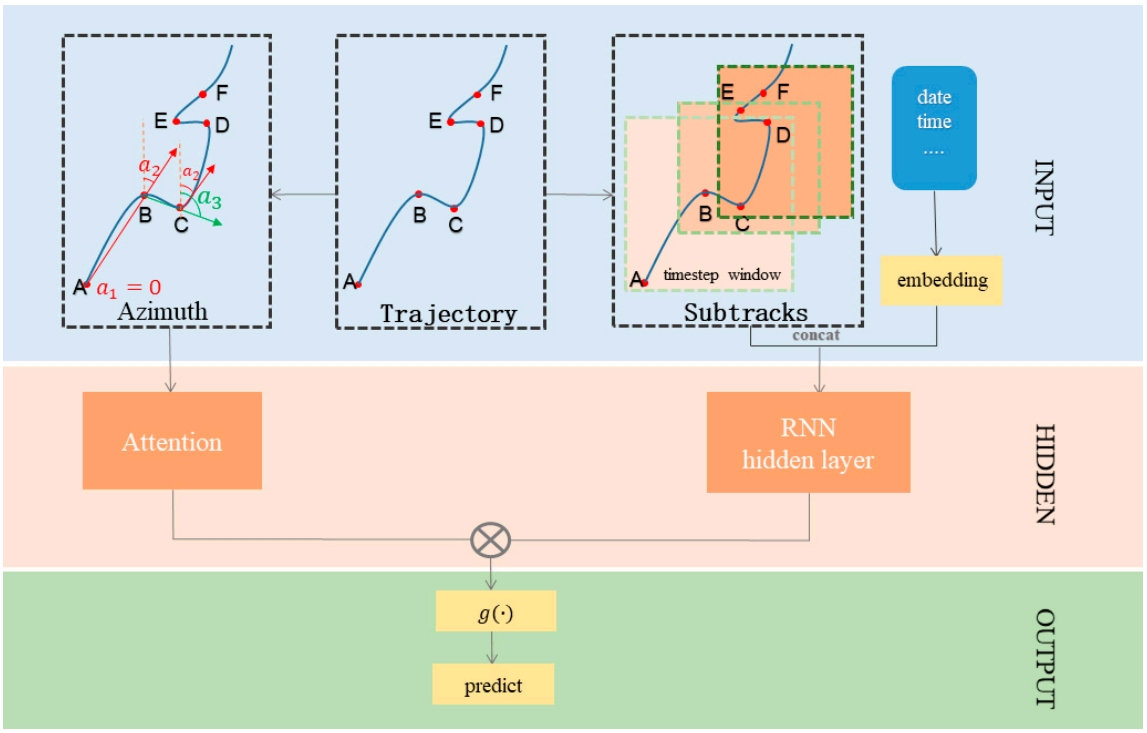


Figure 1. Attention recurrent neural network (RNN) structure.

As shown in Figure 1, a time-step window model is designed to divide the original trajectory sequence into several sub-tracks to reduce the amount of data and improve the training efficiency of the network, and the trajectory semantic information (SI) is extracted by the embedding layer. The attention coefficient of the proposed self-attention mechanism (SAM) can be calculated via azimuth information, which can also be obtained from the original trajectory sequence.

After that, a feature vector of the network input layer can be obtained by concatenating the sub-tracks with the trajectory description information, which goes through the embedding layer, and

the attention coefficients are added into the RNN network; also, the attention mechanism is adopted to make the network better grasp the spatial characteristics of trajectories. Finally, the prediction location (latitude and longitude) is obtained by the activation function. The different parts of the attention-RNN location prediction method are described in detail as follows:

(a) Time-step window: In location prediction experiments, padding based on the length of the longest track is usually used to solve the problem of unequal track sequence lengths, then track points are inputted into the RNN one by one. However, because some of the tracks are collected by the user for 24 hours or longer, padding all tracks will result in a sharp increase in data volume, and it turns out that the influence of historical trajectory points on predicted location gradually weakens over time. If the training is not controlled, it will not only be a waste of meaningless resources, but also will likely be counterproductive to the prediction results. Inspired by the concept of mask convolution in pixel convolutional neural network (CNN) [25], a time-step window is added to divide the original trajectory into several sub-tracks so that all tracks are guaranteed to have the same number of timestamps as the window size and the scope of the historical information is controlled in a certain range. The experimental results show that adopting an appropriate time-step window size can improve not only the efficiency of network training but also the prediction accuracy.

(b) Trajectory semantic information: Using only the latitude and longitude of the GPS point for location prediction, the network can learn poorly because the feature dimension is too small, and the accuracy of the predicted result is not sufficient. In the location prediction experiment, not only the latitude and longitude coordinates of the trajectory data but also some description information should be utilized. Considering that there are certain differences in people's travel destinations between weekends and workdays, an embedding layer is added to the network to mine deeper semantic information (SI) from the data, and to determine whether the trajectory occurs on a weekend or weekday. After the embedding layer, the time information is concatenated with the trajectory sequence, and then input into the RNN.

(c) Self-attention mechanism: In order to fully reflect the different influence weights of the historical trajectory data on the predicted location, this paper introduces a self-attention mechanism (SAM) to learn a set of weight information based on the azimuth changes produced by historical trajectories. In the process of training, each historical track point is assigned a different level of attention.

In the process of obtaining the attention coefficient, the concept of coordinate azimuth is introduced [26]. Usually the coordinate azimuths of the trajectory data $P = (p_1, p_2, \ldots, p_n)$ are defined as $A = (a_1, a_2, \ldots a_n)$. The azimuth of trajectory point p_i is a_i, and azimuth a_i is the angle formed by clockwise rotation from the north to the line connected by track point p_i and its previous point p_{i-1} based on the definition of azimuth. Thus, the azimuths of trajectory P are as follows:

$$A = (a_1, a_2, \ldots a_n | 0 \le a_i \le 2\pi) \tag{4}$$

where $a_1 = 0$, and a_i is the azimuth of p_i. Then the angle of the entire track changes can be denoted as:

$$\beta = (\beta_1, \beta_2, \ldots \beta_n) = (a_1, |a_2 - a_1|, |a_3 - a_2|, \ldots, |a_n - a_{n-1}|) \tag{5}$$

When β_i is taken as π, the track direction changes the most, and the degree of change is symmetrically distributed from π to 0 and from π to 2π. Therefore, the correspondence between attention coefficients s and β is:

$$s_i = \delta sin\left(\frac{\beta_i}{2}\right) \tag{6}$$

where δ is a scaling factor, and the network's powerful self-learning ability is used to adjust the attention coefficient and obtain $S = (s_1, s_2, \ldots s_n)$. Using the attention factor obtained and the attention layer weight $W^{(s)}$, Equation (3) is modified to update the hidden layer:

$$h_t = g\left(W^{(s)} S + W^{(x)} x_t + W^{(h)} h_{t-1}\right). \tag{7}$$

4. Experimental Results and Analyses

4.1. Dataset

The GPS trajectory dataset used in this paper was collected by the Microsoft Asia Research Institute Geolife project [27,28] from April 2007 to August 2012, including trajectory data of 182 users. There are 73 labels of the trajectories according to transportation mode such as driving, taking a bus, riding a bike, or walking. The total distance and duration of different transportation modes are listed in Table 2.

Table 2. Total distance and duration of transportation modes.

Transportation Mode	Distance (km)	Duration (h)
Walk	10,123	5460
Bike	6495	2410
Bus	20,281	1507
Car and taxi	32,866	2384
Train	36,253	745
Airplane	24,789	40
Other	9493	404
Total	140,304	12,953

Data collection duration of different users varies, and the specific distribution is shown in Figure 2.

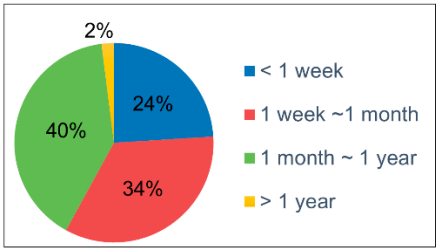


Figure 2. Duration of user data collection.

As shown in Figure 2, according to the collection duration, the experimental data can be classified into three categories: users with a long collection duration (more than 1 month), users with a medium collection duration (1 week to 1 month), and users with a short collection duration (less than 1 week). In order to verify the adaptability of the model under different trajectory lengths, 30% of users in each category were randomly selected for the experimental data. Eventually, experimental results of each category are listed.

4.2. Data Preprocessing

The quality of GPS data is affected by sensors, transmission paths, etc., and data leakage, data loss, and other noise pollution are prone to occur, which will adversely affect subsequent data analysis. Therefore, noise point filtering is required before trajectory prediction. Although the user trajectories in the Geolife dataset are widely distributed, most of them were collected in Beijing, China. Before the conventional noise processing, the latitude and longitude range (115.70° N–117.37° N, 39.40° E–41.03° E) was used to select the data inside the study area. Only floating car trajectory data according to transportation mode were selected according to the analysis in Section 1.

Several typical preprocessing methods for GPS data noise were selected: outlier filtering, smoothing, and data interpolation. The data sampling interval used in the experiment was small and the gross error in the data had the largest influence on the trajectory prediction. Zhang et al. [29] compared several filtering methods under several different types of noise. The results showed that

median filtering works better when dealing with gross errors, so median filtering was used for data filtering.

4.3. Evaluation Metrics

In order to evaluate the validity of the proposed location extraction method, the silhouette coefficient (SC) was used [30]. The SC is a metric that does not need to know the true labeling of the dataset. The value of SC ranges from −1 to +1, and a high SC value indicates a good location extraction result.

The haversine distance between two points was used to measure the error between predicted result $\hat{y}$ (longitude and latitude of a predicted location) and true value y (longitude and latitude of a true destination location). The haversine distance between point $x(lon_x, lat_x)$ and point $y\left(lon_y, lat_y\right)$ is:

$$d_{haversine}(x, y) = 2R{\cdot}arctan\left(\frac{f(x, y)}{f(x, y) - 1}\right) \tag{8}$$

where R is the radius of the Earth, and $f(x, y)$ is defined as:

$$f(x, y) = sin^2\left(\frac{lat_y - lat_x}{2}\right) + \cos(lat_x)\cos\left(lat_y\right)sin^2\left(\frac{lon_y - lon_x}{2}\right) \tag{9}$$

In this paper, mean absolute error (*MAE*) and root mean square error (*RMSE*) are used to measure the experimental results of the prediction algorithm. *MAE* is the average haversine distance, which can be calculated as:

$$MAE = \frac{1}{n}\sqrt{\sum_{i=1}^{n} d_{haversine}(y, \hat{y})} \tag{10}$$

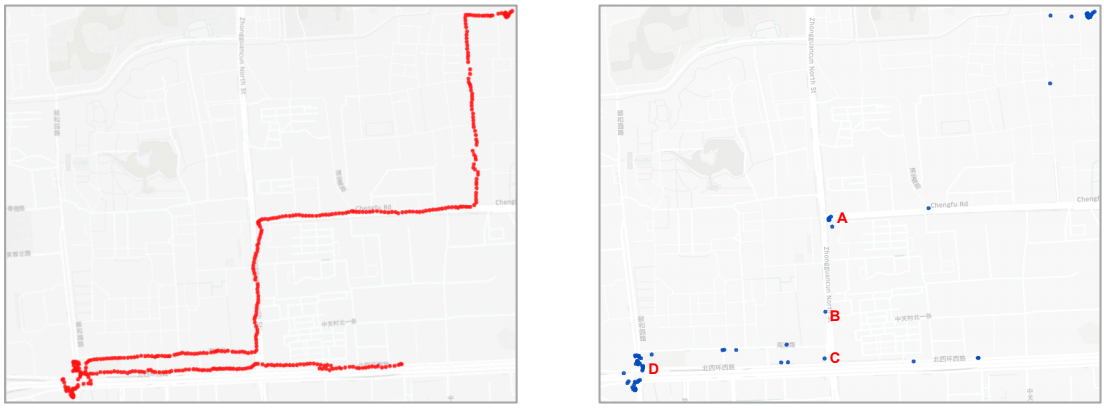
RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_{haversine(y,\hat{y})}{}^2} \tag{11}$$

For the sake of simplicity, all experimental results in this paper are expressed as integers.

4.4. Location Extraction Experimental Results

In order to evaluate the precision of the proposed D-DBSCAN algorithm compared with the traditional DBSCAN, the experiment of important location extraction is demonstrated with a randomly selected user's trajectory, and the results of different clustering algorithms can be visualized in Figure 3.



(**a**) (**b**)

Figure 3. Location extraction results of: (**a**) density-based spatial clustering of applications with noise (DBSCAN) and (**b**) dynamic DBSCAN (D-DBSCAN) algorithms.

As shown in Figure 3, the proposed D-DBSCAN method effectively filters out strip shape clusters and preserves some small clusters. It is found that the proposed method can be used to dynamically adjust the clustering parameters, so it can take a relatively small *Eps* to filter out strip shape clusters when the user is driving on the road at high velocity. As shown in Figure 3a,b, the D-DBSCAN algorithm can obtain better performance in filtering strip shape clusters than DBSCAN, which can reduce disturbance to the prediction model and help to improve the precision. Furthermore, the parameters can also be dynamically adjusted to take a relatively large value when the velocity is low to preserve these important location points (e.g., points A, C and B, D in Figure 3b, where A and C are intersections and B and D are congestion points in the trajectory).

Accuracy comparisons can be found in Table 3, which provides the numbers of track points during clustering and obtained locations based on both algorithms.

Table 3. Location extraction results of different clustering algorithms. SC, silhouette coefficient.

Method	Total Track Points	Number of Clusters	Number of Locations	SC
DBSCAN	592,912	189,180	383	−0.429
D-DBSCAN	592,912	104,312	1093	0.113

As the results shown in Table 3, from the evaluation metric perspective, D-DBSCAN has a higher SC (last row) than DBSCAN. From the perspective of the needs of the prediction model, similar to the previous analysis, D-DBSCAN can effectively filter strip shape clusters, so a lot of unimportant data are discarded and only the most important clusters are preserved, thus, as shown in Table 3 (second row), fewer track points remain with the D-DBSCAN clustering algorithm. At the same time, erroneously merging some different important locations due to data nonuniform distribution are effectively avoided based on the dynamic parameters' strategy, and more important locations are more accurately obtained by the D-DBSCAN algorithm than DBSCAN, which can also be seen in Table 3 (third row).

It is obvious that discarding as many unimportant track points as possible can reduce calculation cost; moreover, preserving the important locations and dividing locations into as many different clusters as possible are two key aspects of providing sufficient input data for the prediction model and avoiding disturbance to the model. The results show that the proposed D-DBSCAN can achieve the best performance in both aspects.

Furthermore, a series of velocity differences threshold ω were used in the experiment so as to check the influence of different minimum velocity difference threshold ω to the result of important location clustering, and the comparison experimental results are shown in Figure 4.

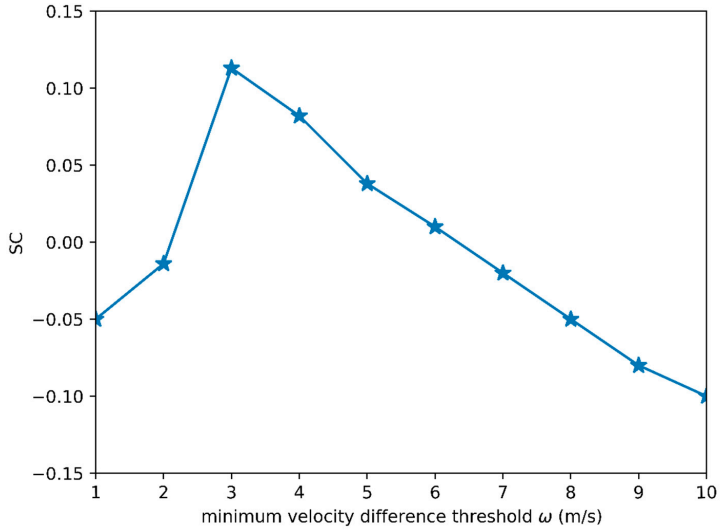


Figure 4. Comparison experimental results with different minimum velocity difference thresholds.

According to the analysis above, partitioning trajectory points into different regions with an optimal minimum velocity difference threshold ω is the key to both obtaining a velocity similarity in the same region and avoiding the incorrect extracting important positions by appropriately setting *Eps* for each region. It can be seen from Figure 4 that the SC value gradually increases as the value of the minimum velocity difference threshold increases from $1(\mathrm{m/s})$ to $3(\mathrm{m/s})$. However, the SC value contrarily decreases as the value of the minimum velocity difference threshold continuously increases from $3(\mathrm{m/s})$ to $10(\mathrm{m/s})$. Since the value SC indicates the accuracy of important location clustering, $\omega = 3(\mathrm{m/s})$ is the optimal minimum velocity difference threshold in our experiments.

4.5. Location Prediction Experiment Results

Users have different behaviors and interests in different time periods and regions. According to the collection duration, the experimental data can be classified into three categories. In order to verify the prediction ability and adaptability of the proposed algorithm, 30% of users in each category were randomly selected for the location prediction experiment. The average accuracy of each category is listed. The size of time-step window is 10 in our experiment.

We compared our model with some baseline approaches: RNN [17], STF-RNN [23], and ST-RNN [22]. The results are shown in Table 4.

Table 4. Comparison of experimental results between users. MAE, mean absolute error; RMSE, root mean square error; STF-RNN, space time features based RNN; ST-RNN, spatial temporal RNN.

Methods	Long Collection Duration Users		Medium Collection Duration Users		Short Collection Duration Users		Mean	
	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)
RNN	1110	2502	1591	3772	1685	1889	1462	3621
STF-RNN	471	788	562	1509	264	623	284	973
ST-RNN	164	453	265	1164	152	363	194	660
Ours	61	312	172	789	150	287	128	463

It can be seen from the results in Table 4 that the proposed method is superior to the others in trajectory prediction based on users with historical trajectories of various lengths. In this paper, the attention mechanism and trajectory semantic information are added into the RNN model. Therefore, the prediction accuracy has an average increase of 87% compared to RNN. The more effective time-specific transition matrices and distance-specific ones are adopted in ST-RNN to extract temporal and spatial information, which makes it have higher prediction accuracy than STF-RNN. Since more semantic information is used in the proposed method and the attention mechanism fully reflects the different influence weights of the historical trajectory data on the predicted location, the prediction accuracy has an average increase of 55% compared to STF-RNN and has an average increase of 34% compared to ST-RNN.

4.6. Ablation Experiment Results

The size of the time-step window determines the size of the history track reference range. The larger the step window, the more historical track data is needed; the prediction accuracy will be affected accordingly, and the model training time will also become longer. In order to prove that applying an appropriate time-step window can reduce prediction error, different time-step windows were used in the experiment. The comparison experimental results with different time-step windows are shown in Figure 5.

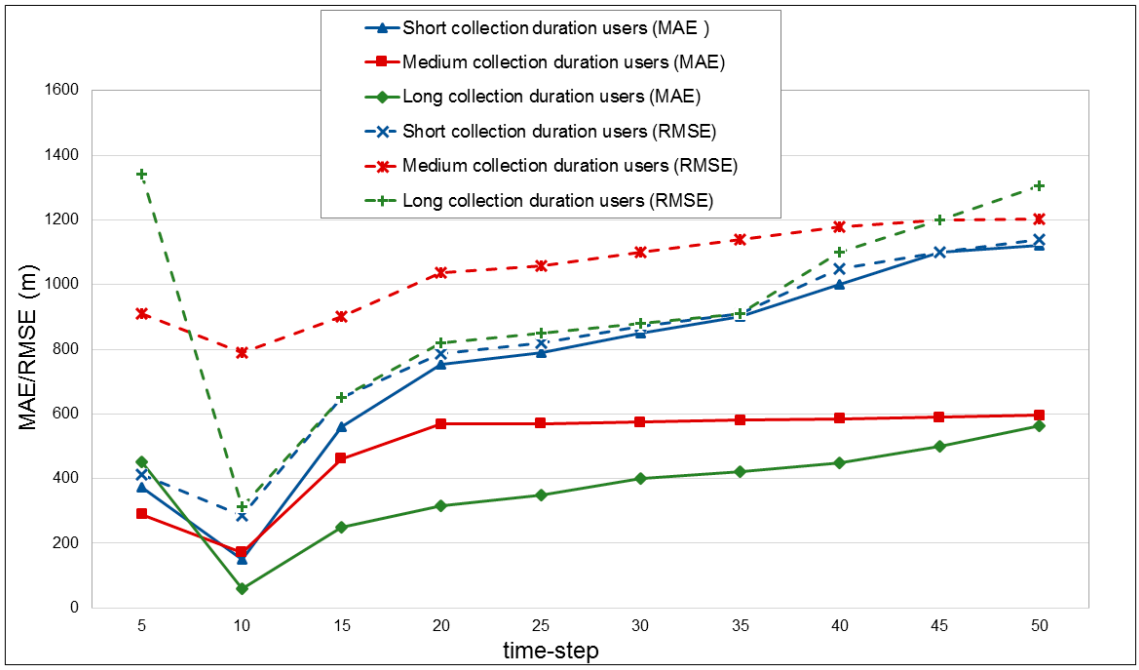


Figure 5. Comparison experimental results with different time-step windows.

It can be seen from Figure 5 that the error of the prediction network is decreased sharply, and the prediction accuracy is increased when the time step is changed from 50 to 10. In addition, reducing the size of the time-step window can contribute to a reduction in the amount of data during network training and then reduce training consumption. When the time step is changed from 10 to 5, although the time consumption is reduced, the prediction error also increases, and the accuracy of the prediction model drops sharply. The reason for this result is that when the time step is 5, the network has a small history track reference range and there is too little historical information to learn the user's travel habits. It can be concluded that applying a suitable time-step window size can not only improve accuracy but also reduce time consumption.

Furthermore, in order to observe the contribution of each part of the model, an ablation experiment was conducted, and the experimental result is shown in Table 5, where RNN is the base model and SI and SAM represent the semantic information strategy and self-attention mechanism model, respectively.

Table 5. Results of ablation experiment. SI, semantic information; SAM, self-attention mechanism.

Method	Long Collection Duration Users		Medium Collection Duration Users		Short Collection Duration Users		Mean	
	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)	MAE (m)	RMSE (m)
RNN	1110	2502	1591	3772	1685	1889	1462	2721
RNN + SI	269	566	325	862	517	605	370	678
RNN + SI + SAM	61	312	172	789	150	287	128	463

It can be seen from the results in Table 5 that each part of the proposed model plays its due role in improving prediction accuracy based on users with historical trajectories of various lengths. Since most users travel between their fixed residences and workplaces on workdays but are free to go anywhere they want on weekends, their behaviors are quite different between weekends and workdays. Thus, deeper semantic information from the data can better match the interest points of users to improve prediction performance. As the results in Table 5 show, the prediction accuracy of all

different collection duration users is significantly improved after adding semantic information, and the average prediction accuracy is improved by 75%. Moreover, the directions of some key locations, such as the road intersections of the track, may play an important role in deciding whether users go to work or back home. Thus, paying close attention to the changes of track direction is another key factor to follow users' behaviors, so the additional SAM strategy can be used to further improve the accuracy of prediction. It can be seen from Table 5 that the average prediction accuracy is improved by 64% after applying self-attention mechanism.

5. Conclusions

In this paper, a novel dynamic important location extraction algorithm based on DBSCAN is proposed to meet the requirements of dynamic situations. Visualizing the clustering results, we can see that a better location extraction result is obtained. In addition, in the prediction model, a time-step window is added based on the RNN to shorten the training time and reduce the equipment requirements, and a self-attention mechanism and trajectory semantic information are added to the RNN model. It is proven by experiments that using a suitable time-step window size can not only improve accuracy but also reduce time consumption and adding a self-attention mechanism and trajectory semantic information can improve prediction accuracy. In the future, extracting more rich semantic information from the trajectory data to improve the adaptive ability of the model and realize the prediction of unknown important points is a possible research direction. Furthermore, location information belongs to the user's private information, so starting simultaneously collaborative research on location prediction and privacy protection is necessary and meaningful.

Author Contributions: Conceptualization, Shaoming Pan, Ziying Li, and Yanwen Chong; methodology Shaoming Pan, Ziying Li, and Yanwen Chong; software, Shaoming Pan, Ziying Li, and Yanwen Chong; validation, Shaoming Pan, Ziying Li, and Yanwen Chong; formal analysis, Shaoming Pan, Ziying Li, and Yanwen Chong; investigation, Shaoming Pan, Ziying Li, and Yanwen Chong; resources, Shaoming Pan, Yanwen Chong; data curation, Shaoming Pan, Ziying Li, and Yanwen Chong; writing—original draft preparation, Shaoming Pan, Ziying Li, and Yanwen Chong; writing—review and editing, Shaoming Pan, Ziying Li, and Yanwen Chong; visualization, Shaoming Pan, Ziying Li, and Yanwen Chong; supervision, Shaoming Pan, Ziying Li, and Yanwen Chong; project administration, Shaoming Pan, Ziying Li, and Yanwen Chong; funding acquisition, Shaoming Pan, Yanwen Chong. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant nos. 41671382, 61572372, and 41271398) and partly supported by the National Key Research and Development Program of China under grant no. 2017YFB0504202.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Camacho-Lara, S. Current and Future GNSS and Their Augmentation Systems. In *Handbook of Satellite Applications*; Springer Science and Business Media LLC: Berlin, Germany, 2013; pp. 617–654.
2. Liu, C.; Liu, J.; Wang, J.; Xu, S.; Han, H.; Chen, Y. An Attention-Based Spatiotemporal Gated Recurrent Unit Network for Point-of-Interest Recommendation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 355. [CrossRef]
3. Lu, X.; Wetter, E.; Bharti, N.; Tatem, A.J.; Bengtsson, L. Approaching the Limit of Predictability in Human Mobility. *Sci. Rep.* **2013**, *3*, 2923. [CrossRef] [PubMed]
4. Naserian, E.; Wang, X.; Dahal, K.; Wang, Z.; Wang, Z. Personalized location prediction for group travellers from spatial-temporal trajectories. *Future Gener. Comput. Syst.* **2018**, *83*, 278–292; [CrossRef]
5. De Brébisson, A.; Simon, É.; Auvolat, A.; Vincent, P.; Bengio, Y. Artificial Neural Networks Applied to Taxi Destination Prediction. *arXiv* **2015**, arXiv:1508.00021.
6. Zheng, Y.; Zhang, L.; Xie, X.; Ma, W.-Y. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th International Conference on Computer Systems and Technologies—CompSysTech '17, Ruse, Bulgaria, 23–24 June 2009; p. 791.
7. Ashbrook, D.; Starner, T. Using GPS to learn significant locations and predict movement across multiple users. *Pers. Ubiquitous Comput.* **2003**, *7*, 275–286. [CrossRef]

8. Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Knowledge Discovery and Data Mining*; Institute for Computer Science, University of Munich: Munich, Germany, 1996; pp. 226–231.
9. Zhou, C.; Frankowski, D.; Ludford, P.; Shekhar, S.; Terveen, L. Discovering personally meaningful places. *ACM Trans. Inf. Syst.* **2007**, *25*, 12. [CrossRef]
10. Zou, H.; Zhu, Z.Z.; He, X.; Zhu, A.-X. An Automatic Annotation Method for Discovering Semantic Information of Geographical Locations from Location-Based Social Networks. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 487. [CrossRef]
11. Xie, K.; Deng, K.; Zhou, X. From trajectories to activities. In Proceedings of the 2009 International Workshop on Location Based Social Networks, Seattle, WA, USA, 3 November 2009; p. 25. [CrossRef]
12. Zhou, A.W.; Yu, Y.F. The Research about Clustering Algoritlm of K-Means. *Coll. Comput. Sci. Technol.* **2011**, *21*, 62–65.
13. Ankerst, M.; Breunig, M.M.; Kriegel, H.P.; Sander, J. *OPTICS: Ordering Points to Identify the Clustering Structure*; ACM: Philadelphia, PA, USA, 1999; pp. 49–60.
14. Xue, A.Y.; Zhang, R.; Zheng, Y.; Xie, X.; Huang, J.; Xu, Z. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In Proceedings of the 2013 IEEE 29th International Conference on Data Engineering (ICDE), Brisbane, QLD, Australia, 8–12 April 2013; pp. 254–265.
15. Simmons, R.; Browning, B.; Zhang, Y.; Sadekar, V. Learning to Predict Driver Route and Destination Intent. In Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference, Toronto, ON, Canada, 17–20 September 2006; pp. 127–132.
16. Ashbrook, D.; Starner, T. Magic: A motion gesture design tool. In *Sigchi Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2010; pp. 2159–2168.
17. Goodfellow, I.; Bengio, Y.; Courville, A. *InDeep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 1, pp. 367–415.
18. Cho, K.; Van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078, 1724–1734.
19. Hakan, B.; Gelenbe, E. Feature-based RNN target recognition. In *Algorithms for Synthetic Aperture Radar Imagery V*; International Society for Optics and Photonics: Orlando, FL, USA, 1998.
20. Kajal, K. CARF-Net: CNN attention and RNN fusion network for video-based person reidentification. *J. Electron. Imaging* **2019**, *28*, 023036.
21. Rehab, A.; Chung, H.P.; James, H. Sequence-to-sequence image caption generator. In Proceedings of the Eleventh International Conference on Machine Vision (ICMV 2018), Munich, Germany, 1–3 November 2019. [CrossRef]
22. Liu, Q.; Wu, S.; Wang, L.; Tan, T. *Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts*; AAAI Press: Palo Alto, CA, USA, 2016.
23. Al-Molegi, A.; Jabreel, M.; Ghaleb, B. STF-RNN: Space-Time Features-based Recurrent Neural Network for Predicting People's Next Location. In Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 6 December 2016.
24. Zhou, S.; Zhou, A.; Cao, J. A Data-Partitioning-Based DBSCAN Algorithm. *J. Comput. Res. Dev.* **2000**, *37*, 1154–1159.
25. Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. *arXiv* **2016**, arXiv:1601.06759, 1747–1756.
26. Rutstrum, C. *The Wilderness Route Finder*; University of Minnesota Press: Minneapolis, MN, USA, 2000; ISBN 0-8166-3661-3.
27. Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; Ma, W.-Y. Understanding mobility based on GPS data. In Proceedings of the 10th International Conference on Information Integration and Web-Based Applications & Services—iiWAS '08, Linz, Austria, 24–26 November 2008.
28. Zheng, Y.; Xie, X.; Ma, W. GeoLife: A Collaborative Social Networking Service among User, location and trajectory. *IEEE Data Eng. Bull.* **2010**, *33*, 32–40.

29. Zhang, Z.; Zhu, J.; Kuang, C.; Ke, Y. Comparative Study and Improvement on Several De-noising Methods For Different Noise. *J. Geodesy Geodyn.* **2014**, *1*, 127–130.
30. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]

Building Personal Maps from GPS Data

Lin Liao and **Donald J. Patterson** and **Dieter Fox** and **Henry Kautz**
Department of Computer Science & Engineering
University of Washington
Seattle, WA 98195

Abstract

In this paper we discuss a system that can learn personal maps customized for each user and infer his daily activities and movements from raw GPS data. The system uses discriminative and generative models for different parts of this task. A discriminative relational Markov network is used to extract significant places and label them; a generative dynamic Bayesian network is used to learn transportation routines, and infer goals and potential user errors at real time. In this paper we focus on the basic structures of the models and briefly discuss the inference and learning techniques. Experiments show that our system is able to accurately extract and label places, predict the goals of a person, and recognize situations in which the user makes mistakes (*e.g.*, taking a wrong bus).

1 Introduction

A typical map consists of significant places and road networks within a geographic region. In this paper, we present the concept of a *personal map*, which is customized based on an individual's behavior. A personal map includes personally significant places, such as home, a workplace, shopping centers, and meeting places and personally significant routes (*i.e.*, the paths and transportation modes such as foot, car, or bus, that the person usually uses to travel from place to place). In contrast with general maps, a personal map is customized and primarily useful for a given person. Because of the customization, it is well-suited for recognizing an individual's behavior and offering detailed personalized help. For example, in this paper we use a personal map to

- Discriminate a user's activities (is she dining at a restaurant or visiting a friend?);
- Predict a user's future movements and transportation modes, both in the short term (will she turn left at the next street corner? will she get off the bus at the next bus stop?) and in terms of distant goals (is she going to her workplace?);
- Infer when a user has *broken his ordinary routine* in a way that may indicate that he has made an error, such as failing to get off his bus at his usual stop on the way home.

We describe a system that builds personal maps automatically from raw location data collected by wearable GPS units. Many potential applications can be built upon the system. A motivating application for this work is the development of personal guidance systems that helps cognitively-impaired individuals move safely and independently throughout their community [Patterson *et al.*, 2004]. Other potential applications include *customized* "just in time" information services (for example, providing the user with current bus schedule information when she is likely to need it or real time traffic conditions on her future trajectories), intelligent user interface (instructing a cell phone not to ring when in a restaurant or at a meeting), and so on.

This paper is focused on the fundamental techniques of learning and inference. We develop probabilistic models that bridge low level sensor measurements (*i.e.*, GPS data) with high level information in the personal maps. Given raw GPS data from a user, our system first finds a user's set of significant places, then a Relational Markov Network (RMN) is constructed to recognize the activities in those places (*e.g.*, working, visiting, and dining out); as discriminative models, RMNs often outperform their corresponding generative models (*e.g.*, HMMs) for classification tasks [Taskar *et al.*, 2002]. The system then uses a dynamic Bayesian network (DBN) model [Murphy, 2002] for learning and inferring transportation routines between the significant places; such a generative model is well-suited for online tracking and real-time user error detection.

The paper is organized as follows. In the next section, we discuss related work. In Section 3, we present the discriminative model for place extraction and labeling, followed by the generative DBN model in Section 4. We show experimental evaluations before concluding.

2 Related Work

Over the last years, estimating a person's activities has gained increased interest in the AI, robotics, and ubiquitous computing communities. [Ashbrook and Starner, 2003; Hariharan and Toyama, 2004] learn significant locations from logs of GPS measurements by determining the time a person spends at a certain location. For these locations, they use

frequency counting to estimate the transition parameters of Markov models. Their approach then predicts the next goal based on the current and the previous goals. Our system goes beyond their work in many aspects. First, our system not only extracts places, but also recognizes activities associated with those places. Second, their models are not able to refine the goal estimates using GPS information observed when moving from one significant location to another. Furthermore, such a coarse representation does not allow the detection of potential user errors. In contrast, our hierarchical generative model is able to learn more specific motion patterns of transpiration routines, which also enables us to detect user errors.

In the machine learning community, a variety of *relational probabilistic models* were introduced to overcome limitations of propositional probabilistic models. Relational models combine first-order logical languages with probabilistic graphical models. Intuitively, a relational probabilistic model is a *template* for propositional models such as Bayesian networks or MRFs (similar to how first-order logic formulas can be instantiated to propositional logic). Templates are defined over object classes through logical languages such as Horn clauses, frame systems, SQL, and full first-order logic. Given data, these templates are then *instantiated* to generate propositional models (typically Bayesian networks or MRFs), on which inference and learning is performed. Relational probabilistic models use high level languages for describing systems involving complex relations and uncertainties. Since the structures and parameters are defined at the level of classes, they are shared by the instantiated networks. Parameter sharing is particularly essential for learning from sparse training data and for knowledge transfer between different people. As a popular relational probabilistic model, Relational Markov Networks (RMN) define the templates using SQL, a widely used query language for database systems, and the templates are instantiated into (conditional) Markov networks, which are *undirected* models that do not suffer the cyclicity problem and are thereby more flexible and convenient. Since their introduction, RMNs have been used successfully in a number of domains, including web page classification [Taskar *et al*., 2002], link prediction [Taskar *et al*., 2003], and information extraction [Bunescu and Mooney, 2004].

In the context of probabilistic plan recognition, [Bui *et al*., 2002] introduced the abstract hidden Markov model, which uses hierarchical representations to efficiently infer a person's goal in an indoor environment from camera information. [Bui, 2003] extended this model to include memory nodes, which enables the transfer of context information over multiple time steps. Bui and colleagues introduced efficient inference algorithms for their models using Rao-Blackwellised particle filters. Since our model has a similar structure to theirs, we apply the inference mechanisms developed in [Bui, 2003]. Our work goes beyond the work of Bui *et al*. in that we show how to learn the parameters of the hierarchical activity model, and their domains, from data. Furthermore, our low level estimation problem is more challenging than their indoor tracking problem.

The task of detecting abnormal events in time series data (called *novelty detection*) has been studied extensively in the data-mining community [Guralnik and Srivastava, 1999], but remains an open and challenging research problem. We present the results on abnormality and error detection in location and transportation prediction using a simple and effective approach based on comparing the likelihood of a learned hierarchical model against that of a prior model.

3 Extracting and Labeling Places

In this section, we briefly discuss place extraction and activity labeling. For full technical details of the activity labeling refer to [Liao *et al*., 2005].

3.1 Place Extraction

Similar to [Ashbrook and Starner, 2003; Hariharan and Toyama, 2004], our current system considers significant places to be those locations where a person typically spends extended periods of time. From the GPS data, it first looks for locations where the person stays for a given amount of time (*e.g*., 10 minutes), and then these locations are clustered to merge spatially similar points. An extension of the approach that takes into account more complex features is discussed in future work (Section 6).

3.2 Activity Labeling

We build our activity model based on the Relational Markov Network (RMN) framework [Taskar *et al*., 2002]. RMNs describe specific relations between objects using clique templates specified by SQL queries: each query C selects the relevant objects and their attributes, and specifies a *potential* function, or clique potential, ϕ_C, on the possible values of these attributes. Intuitively, the clique potentials measure the "compatibility" between values of the attributes. Clique potentials are usually defined as a log-linear combinations of *feature* functions, *i.e*., $\phi_C(\mathbf{v}_C) = \exp\{\mathbf{w}_C^T \cdot \mathbf{f}_C(\mathbf{v}_C)\}$, where $\mathbf{v}_C$ are the attributes selected in the query, $\mathbf{f}_C()$ is a feature vector for C, and $\mathbf{w}_C^T$ is the transpose of the corresponding weight vector. For instance, a feature could be the number of different homes defined using aggregations.

To perform inference, an RMN is *unrolled* into a Markov network, in which the nodes correspond to the attributes of objects. The connections among the nodes are built by applying the SQL templates to the data; each template C can result in several cliques, which share the same feature weights. Standard inference algorithms, such as belief propagation and MCMC, can be used to estimate the conditional distribution of hidden variables given all the observations.

Relational activity model

Because behavior patterns can be highly variable, a reliable discrimination between activities must take several sources of evidence into account. More specifically, our model defines the following templates:

1. *Temporal* patterns: Different activities often have different temporal patterns, such as their duration or the time of day. Such local patterns are modeled by clique templates that connect each attribute with the activity label.
2. *Geographic* evidence: Information about the types of businesses close to a location can be extremely useful to determine a user's activity. Such information can be extracted from geographic databases like

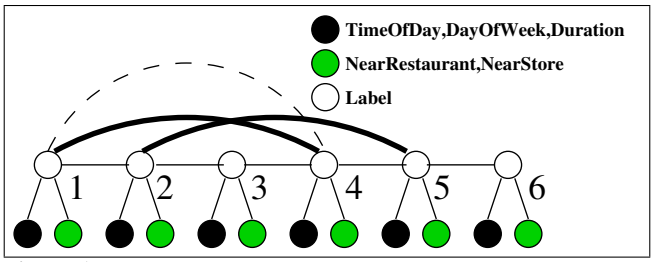


Figure 1: An example of unrolled Markov network with six activities. Solid straight lines indicate the cliques generated by the templates of temporal, geographic, and transition features; bold solid curves represent spatial constraints (activity 1 and 4 are associated with the same place and so are 2 and 5); dashed curves stand for global features, which are label-specific cliques (activity 1 and 4 are both labeled as 'AtHome' or 'AtWork' at this moment).

`http://www.microsoft.com/mappoint`. Since location information in such databases is not accurate enough, we consider such information by checking whether, for example, a restaurant is within a certain range from the location.

3. *Transition* relations: First-order transitions between activities can also be informative. For example, going from home to work is very common while dining out twice in a row is rare.
4. *Spatial* constraints: Activities at the same place are often similar. In other words, the number of different types of activities in a place is often limited.
5. *Global* features: These are soft constraints on the activities of a person. The number of different home locations is an example of the global constraints. Such a constraint is modeled by a clique template that selects all places labeled as home and returns how many of them are different.

Inference

In our application, the task of inference is to estimate the labels of activities in the *unrolled* Markov networks (see Fig. 1 for an example). Inference in our relational activity models is complicated by the fact that the structure of the unrolled Markov network can change during inference. This is due to the fact that, in the templates of global features, the label of an object determines to which cliques it belongs. We call such cliques *label-specific cliques*. Because the label values are hidden during inference, such cliques potentially involve all the labels, which makes exact inference intractable.

We perform approximate inference using Markov Chain Monte Carlo (MCMC) [Gilks *et al*., 1996]. We first implemented MCMC using basic Gibbs sampling. Unfortunately, this technique performs poorly in our model because of the strong dependencies among labels. To make MCMC mix faster, we developed a mixture of two transition kernels: the first is a block Gibbs sampler and the second is a Metropolis sampler (see [Liao *et al*., 2005] for details). The numbers of different homes and workplaces are stored in the chains as global variables. This allows us to compute the *global* features *locally* in both kernels. In order to determine which kernel to use at each step, we sample a random number u between 0 and 1 uniformly, and compare u with the given threshold γ ($\gamma = 0.5$ in our experiments).

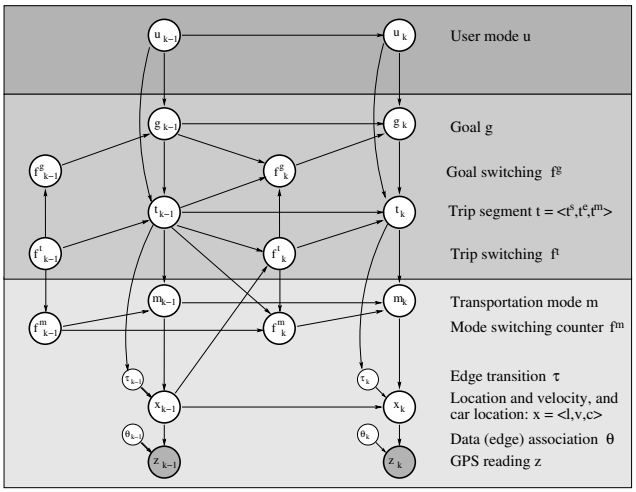


Figure 2: Hierarchical activity model representing a person's outdoor movements during everyday activities. The upper level estimates the user mode, the middle layer represents goals and trip segments, and the lowest layer is the flat model, estimating the person's location, velocity, and transportation mode.

Learning

The parameters to be learned are the weights $\mathbf{w}$ of the features that define the clique potentials. To avoid overfitting, we perform maximum a posterior (MAP) parameter estimation and impose an independent Gaussian prior with constant variance for each component of $\mathbf{w}$. Since the objective function for MAP estimation is convex, the global optimum can be found using standard numerical optimization algorithms [Taskar *et al*., 2002]. We apply the quasi-Newton methods to find the optimal weights because they have been found to be very efficient for CRFs [Sha and Pereira, 2003]. Each iteration of this technique requires the value and gradient of the objective function computed at the weights returned in the previous iteration. In [Liao *et al*., 2005], we presented an algorithm that *simultaneously* estimates at each iteration the value and its gradient using MCMC.

Although parameter learning in RMNs requires manually labeled training data, parameter sharing makes it easy to transfer knowledge. For example, in our system, we can learn a *generic* model from people who have manually labeled data, and then apply the model to people who have no labeled data. Generic models in our system can perform reasonably well, as we will show in the experiments.

4 Learning and Inferring Transportation Routines

We estimate a person's activities using the three level dynamic Bayesian network model shown in Fig. 2. The individual nodes in such a temporal graphical model represent different parts of the state space and the arcs indicate dependencies between the nodes [Murphy, 2002]. Temporal dependencies are represented by arcs connecting the two time slices $k-1$ and k. The highest level of the model indicates the user

mode, which could be typical behavior, user error, or deliberate novel behavior. The middle level represents the person's goal (*i.e.*, next significant place) and trip segment (defined below). The lowest level is the *flat* model, which estimates the person's transportation mode, location and motion velocity from the GPS sensor measurements. In this section, we explain the model from bottom up; refer to [Liao *et al.*, 2004] for more details.

4.1 Locations and Transportation Modes

We denote by $x_k = \langle l_k, v_k, c_k \rangle$ the location and motion velocity of the person, and the location of the person's car [1] (subscripts k indicate discrete time). In our DBN model, locations are estimated on a graph structure representing a street map. GPS sensor measurements, z_k, are generated by the person carrying a GPS sensor. Since measurements are given in continuous xy-coordinates, they have to be "snapped" to an edge in the graph structure. The edge to which a specific measurement is "snapped" is estimated by the association variable θ_k. The location of the person at time k depends on his previous location, l_{k-1}, the motion velocity, v_k, and the vertex transition, τ_k. Vertex transitions τ model the decision a person makes when moving over a vertex in the graph, for example, to turn right when crossing a street intersection.

The mode of transportation can take on four different values $m_k \in \{\text{BUS}, \text{FOOT}, \text{CAR}, \text{BUILDING}\}$. Similar to [Patterson *et al.*, 2003], these modes influence the motion velocity, which is picked from a Gaussian mixture model. For example, the walking mode draws velocities only from the Gaussian representing slow motion. *BUILDING* is a special mode that occurs only when the GPS signal is lost for significantly long time. Finally, the location of the car only changes when the person is in the *CAR* mode, in which the car location is set to the person's location.

An efficient algorithm based on Rao-Blackwellised particle filters (RBPFs) [Doucet *et al.*, 2000] has been developed to perform online inference for the flat model. In a nutshell, the RBPF samples transportation mode ${m_k}^{(i)}$, transportation mode switch $f_k^{m(i)}$, data association ${\theta_k}^{(i)}$, edge transition ${\tau_k}^{(i)}$, and velocity ${v_k}^{(i)}$, then it updates the Gaussian distribution of location ${l_k}^{(i)}$ using a one-dimensional Kalman filter. After all components of each particle are generated, the importance weights of the particles are updated. This is done by computing the likelihood of the GPS measurement z_k, which is provided by the update innovations of the Kalman filters [Doucet *et al.*, 2000].

We apply expectation maximization (EM) to learn the model parameters. Before learning, the model has no preference for when a person switches mode of transportation, or which edge a person transits to when crossing a vertex on the graph. However, information about bus routes, and the fact that the car is either parked or moves with the person, already provide important constraints on mode transitions. At each iteration of EM, the location, velocity, and mode of transportation are estimated using the Rao-Blackwellised particle filter of the flat model. In the E-step, transition counts of a forward and a backward filtering pass through the data log are combined, based on which we update the model parameters in the M-step. In our Rao-Blackwellised model, edge transitions are counted whenever the *mean* of a Kalman filter transits the edge. The learned flat model encodes information about typical motion patterns and significant locations by edge and mode transition probabilities.

[1] We include the car location because it strongly affects whether the person can switch to the car mode.

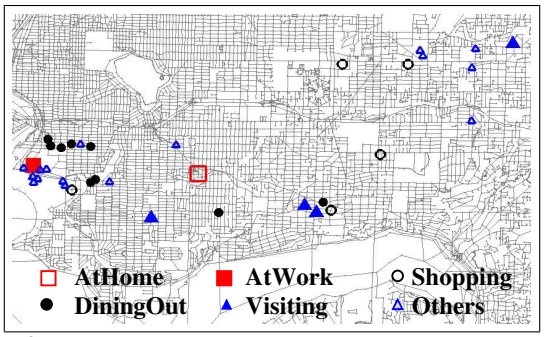


Figure 3: Part of the locations contained in the data set of a single person, collected over a period of four months (x-axis is 8 miles long).

After we estimate the mode transition probabilities for each edge, we find *mode transfer locations*, *i.e.*, usual bus stops and parking lots, by looking for those locations at which the mode switching exceeds a certain threshold.

4.2 Goals and Trip Segments

A trip segment is defined by its start location, t_k^s, end location, t_k^e, and the mode of transportation, t_k^m, the person uses during the segment. For example, a trip segment models information such as "she gets on the bus at location t_k^s and takes the bus up to location t_k^e, where she gets off the bus". In addition to transportation mode, a trip segment predicts the route on which the person gets from t_k^s to t_k^e. This route is not specified through a deterministic sequence of edges on the graph but rather through transition probabilities on the graph. These probabilities determine the prediction of the person's motion direction when crossing a vertex in the graph, as indicated by the arc from t_k to τ_k.

A goal represents the current target location of the person. Goals include the significant locations extracted using our discriminative model. The transfer between trip segments and goals is handled by the boolean switching nodes f_k^t and f_k^g, respectively.

To estimate a person's goal and trip segment, we apply the inference algorithm used for the abstract hidden Markov memory models [Bui, 2003]. More specifically, we use a Rao-Blackwellised particle filter both at the low level and at the higher levels. Each sample of the resulting particle filter contains the discrete and continuous states described in the previous section, and a joint distribution over the goals and trip segments. These additional distributions are updated using exact inference.

Because we have learned the set of goals using the discriminative model and the set of trip segments using the flat model, we only need to estimate the transition matrices at all levels: between the goals, between the trip segments given the goal, and between the adjacent streets given the trip segment.

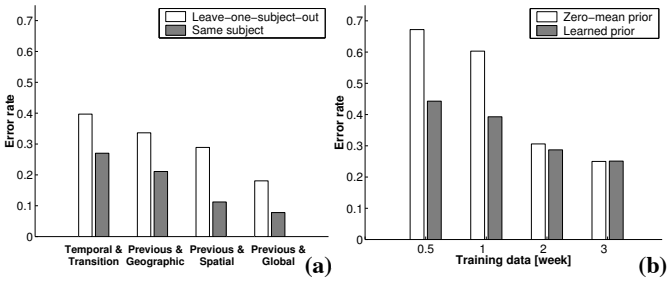


Figure 4: (a) Error rates of cross-validation of the generic models and customized models using different feature sets; (b) Zero-mean prior vs. learned model as prior mean (shown the error rates over the new places only).

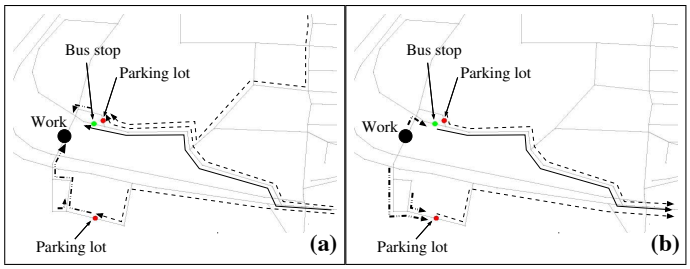


Figure 5: Learned model zoomed into the area around the work place and the very likely transitions (probability above 0.75). Dashed lines indicate car mode, solid lines bus, and dashed-dotted lines foot. (a) Given that the goal is the work place. (b) Given that the goal is home.

Again, we use EM in the hierarchical model, which is similar to that in the flat model. During the E-steps, smoothing is performed by tracking the states both forward and backward in time. The M-steps update the model parameters using the frequency counts generated in the E-step. All transition parameters are smoothed using Dirichlet priors.

4.3 User Modes

To detect user errors or novel behavior, we add the variable u_k to the highest level, which indicates the user's behavior mode $\in \{Normal, Novel, Erroneous\}$. Different values of u_k instantiate different parameters for the lower part of the model. When user mode is typical behavior, the model is instantiated using the parameters learned from training data. When a user's behavior is $Erroneous$, the goal remains the same, but the trip segment is set to a distinguished value "unknown" and as a consequence the parameters of the flat model (*i.e.*, transportation mode transitions and edge transitions) are switched to their a priori values: An "unknown" trip segment cannot provide any information for the low level parameters. When a user's behavior is $Novel$, the goal is set to "unknown," the trip segment is set to "unknown," and the parameters of the flat model are again set to their a priori values.

To infer the distribution of u_k, we run two trackers simultaneously and at each time their relative likelihood is used to update the distribution. The first tracker uses the hierarchical model with learned parameters and second tracker uses the flat model with a priori parameters. When a user is following her ordinary routine, the first tracker has higher likelihoods, but when the user makes error or does something novel, the second tracker becomes more likely. Unless the true goal is observed, the system cannot distinguish errors from novel behavior, so the precise ratio between the two is determined by hand selected prior probabilities. In some situations, however, the system knows where the user is going, *e.g.*, if the user asks for directions to a destination, or if a caregiver indicates the "correct" destination, and thus the goal is fixed, treated as an observed, and therefore *clamped*. After we have clamped the goal, the probability of novel behavior becomes zero and the second tracker just determines the probabilities of an error.

5 Experiments

To evaluate our system, we collected two sets of location data using wearable GPS units. The first data set contains location traces from a single person over a time period of four months (see Fig. 3). It includes about 400 activities at 50 different places. The second data set consists of one-week of data from five different people. Each person has 25 to 30 activities and 10 to 15 different significant places. We extracted from the logs each instance of a subject spending more than 10 minutes at one place. Each instance corresponds to an activity. We then clustered the nearby activity locations into places.

5.1 Evaluating the RMN Model

For training and evaluation, the subjects manually labeled the data with their activities from the following set: {**AtHome, AtWork, Shopping, DiningOut, Visiting, Other**}. Then, we constructed the unrolled Markov networks using the templates described above, trained the models, and tested their accuracy. Accuracy was determined by the activities for which the most likely labeling was correct.

In practice, it is of great value to learn a *generic* activity model that can be immediately applied to new users without additional training. In the first experiment, we used the data set of multiple users and performed leave-one-subject-out cross-validation: we trained using data from four subjects, and tested on the remaining one. The average error rates are indicated by the white bars in Fig. 4(a). By using all the features, the generic model achieved an error rate of 20%. Note that the global features and the spatial constraints are very useful. To gauge the impact of different habits on the results, we also performed the same evaluation using the data set of single subject. In this case, we used one-month data for training and the other three-month data for test, and we repeated the validation process for each month. The results are shown by the gray bars in Fig. 4(a). In this case, the model achieved an error rate of only 7%. This experiment shows that it is possible to learn good activity models from groups of people. It also shows that if the model is learned from more "similar" people, then higher accuracy can be achieved. This indicates that models can be improved by grouping people based on their activity patterns.

When estimating the weights of RMNs, a prior is imposed in order to avoid overfitting. Without additional information, a zero mean Gaussian is typically used as the prior [Taskar *et al.*, 2002]. Here we show that performance can also be improved by estimating the *hyper-parameters* for the means of the weights using data collected from other people. Similar to the first experiment, we want to learn a customized model

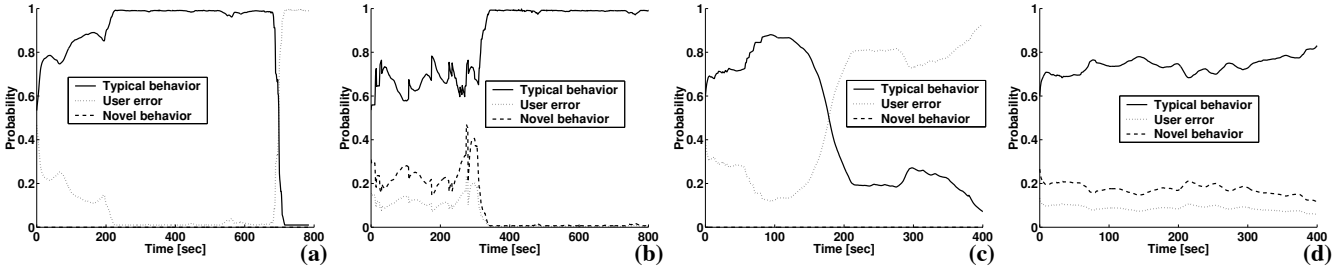


Figure 6: The probabilities of user mode in two experiments (when goal is unclamped, the prior ratio of typical behavior, user error and novel behavior is 3:1:2; when goal is clamped, the probabilities of novel behavior are always zero): (a) Bus experiment with goal clamped; (b) Bus experiment with goal unclamped; (c) Foot experiment with goal clamped; (d) Foot experiment with goal unclamped.

for a person **A**, but this time we also have labeled data from others. We could simply ignore the others' data and use the labeled data from **A** with a zero-mean prior. Or we can first learn the weights from the other people and use that as the mean of the Gaussian prior for **A**. We evaluate the performance of the two approaches for different amounts of training data from person **A**. The results are shown in Fig. 4(b). We can see that using data from others to generate a prior boosts the accuracy significantly, especially when only small amounts of training data are available.

Using the Bayesian prior smoothly shifts from generic to customized models: on one end, when no data from the given subject is available, the approach returns the generic (prior) model; on the other end, as more labeled data become available, the model adjusts more and more to the specific patterns of the user and we get a customized model.

5.2 Evaluating the DBN Model

The learning of the generative model was done completely unsupervised without any manual labeling. Fig. 5 show the learned trip segments and street transitions zoomed into the workplace. The model successfully discovered the most frequent trajectories for traveling from home to the workplace and vice-versa, as well as other common trips, such as to the homes of friends.

As we described, an important feature of our model is the capability to capture user errors and novel behavior using a parallel tracking approach. To demonstrate the performance of this technique, we did the following two experiments:

In the first experiment, a user took the wrong bus home. For the first 700 seconds, the wrong bus route coincided with the correct one and the system believed that the user was in $\{u_k = Normal\}$ mode. But when the bus took a turn that the user had never taken to get home, the probability of errors in the clamped model dramatically jumped (see Fig. 6(a)). In contrast, the unclamped model cannot determine a user error because the user, while on the wrong bus route, was on a bus route consistent with other previous goals (see Fig. 6(b)).

The second experiment was a walking experiment in which the user left his office and proceeded to walk away from his normal parking spot. When the destination was not specified, the tracker had a steady level of confidence in the user's path (see Fig. 6(d), there are lots of previously observed paths from his office), but when the goal was specified, the system initially saw behavior consistent with walking toward the parking spot, and then as the user turned away at time 125, the tracker's confidence in the user's success dropped (see Fig. 6(c)).

6 Conclusions and Future Work

In this paper we have described a system that can build personal maps automatically from GPS sensors. More specifically, the system is able to: recognize significant locations of a user and activities associated with those places, infer transportation modes and goals, and detect user errors or novel behavior. The system uses a Relational Markov Network for place classification and a hierarchical Dynamic Bayesian Network for online tracking and error detection. This technique has been used as the basis for both experimentation and for real context-aware applications including an automated transportation routing system that ensures the efficiency, safety, and independence of individuals with mild cognitive disabilities (see [Patterson *et al*., 2004]).

In our future work we plan to improve the place extraction. The current approach only relies on measuring the time periods a person stays at each place and uses a fixed threshold to distinguish significant places from insignificant ones. However, it is hard to find a fixed threshold that works for all significant places. If we set the threshold too big (say 10 minutes, as in our experiments), some places could be missed (*e.g*., places a user stops by to get coffee or pick up his kids); if we set the value too small (*e.g*., 1 minute), some trivial places (such as traffic lights) may be considered significant. Therefore, to extract more places accurately, we will take into account more features besides stay duration. For example, transportation mode is a very useful indicator: if a user switches to *foot* at some place during a *car* trip, that place is likely to be significant. Since transportation mode itself has to be inferred, we must design a model that considers all these uncertainties comprehensively. In order to do that, we plan to extend the existing relational probabilistic languages so that we can model complex relations and still perform efficient inference and learning.

Acknowledgments

This research is supported in part by NSF under grant number IIS-0433637 and SRI International subcontract 03-000225 under DARPA's CALO project. We also thank anonymous reviewers for their helpful comments.

References

[Ashbrook and Starner, 2003] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. In *Personal and Ubiquitous Computing*, 2003.

[Bui *et al*., 2002] Hung H. Bui, Svetha Venkatesh, and Geoff West. Policy recognition in the abstract hidden markov models. *Journal of Artificial Intelligence Research*, 2002.

[Bui, 2003] H. H. Bui. A general model for online probabilistic plan recognition. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.

[Bunescu and Mooney, 2004] R. Bunescu and R. J. Mooney. Collective information extraction with relational markov networks. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[Doucet *et al*., 2000] Arnaud Doucet, Nando de Freitas, Kevin Murphy, and Stuart Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2000.

[Gilks *et al*., 1996] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1996.

[Guralnik and Srivastava, 1999] Valery Guralnik and Jaideep Srivastava. Event detection from time series data. In *5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[Hariharan and Toyama, 2004] R. Hariharan and K. Toyama. Project Lachesis: parsing and modeling location histories. In *Geographic Information Science*, 2004.

[Liao *et al*., 2004] L. Liao, D. Fox, and H. Kautz. Learning and inferring transportation routines. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2004.

[Liao *et al*., 2005] Lin Liao, Dieter Fox, and Henry Kautz. Location-based activity recognition using relational Markov networks. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[Murphy, 2002] Kevin Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, 2002.

[Patterson *et al*., 2003] Donald J. Patterson, Lin Liao, Dieter Fox, and Henry Kautz. Inferring high-level behavior from low-level sensors. In *International Conference on Ubiquitous Computing (UbiComp)*, 2003.

[Patterson *et al*., 2004] Donald J. Patterson, Lin Liao, Krzysztof Gajos, Michael Collier, Nik Livic, Katherine Olson, Shiaokai Wang, Dieter Fox, and Henry Kautz. Opportunity Knocks: a System to Provide Cognitive Assistance with Transportation Services. In Itiro Siio Nigel Davies, Elizabeth Mynatt, editor, *Proceedings of UBICOMP 2004: The Sixth International Conference on Ubiquitous Computing*, volume LNCS 3205, pages 433–450. Springer-Verlag, October 2004.

[Sha and Pereira, 2003] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of Human Language Technology-NAACL*, 2003.

[Taskar *et al*., 2002] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.

[Taskar *et al*., 2003] B. Taskar, M. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

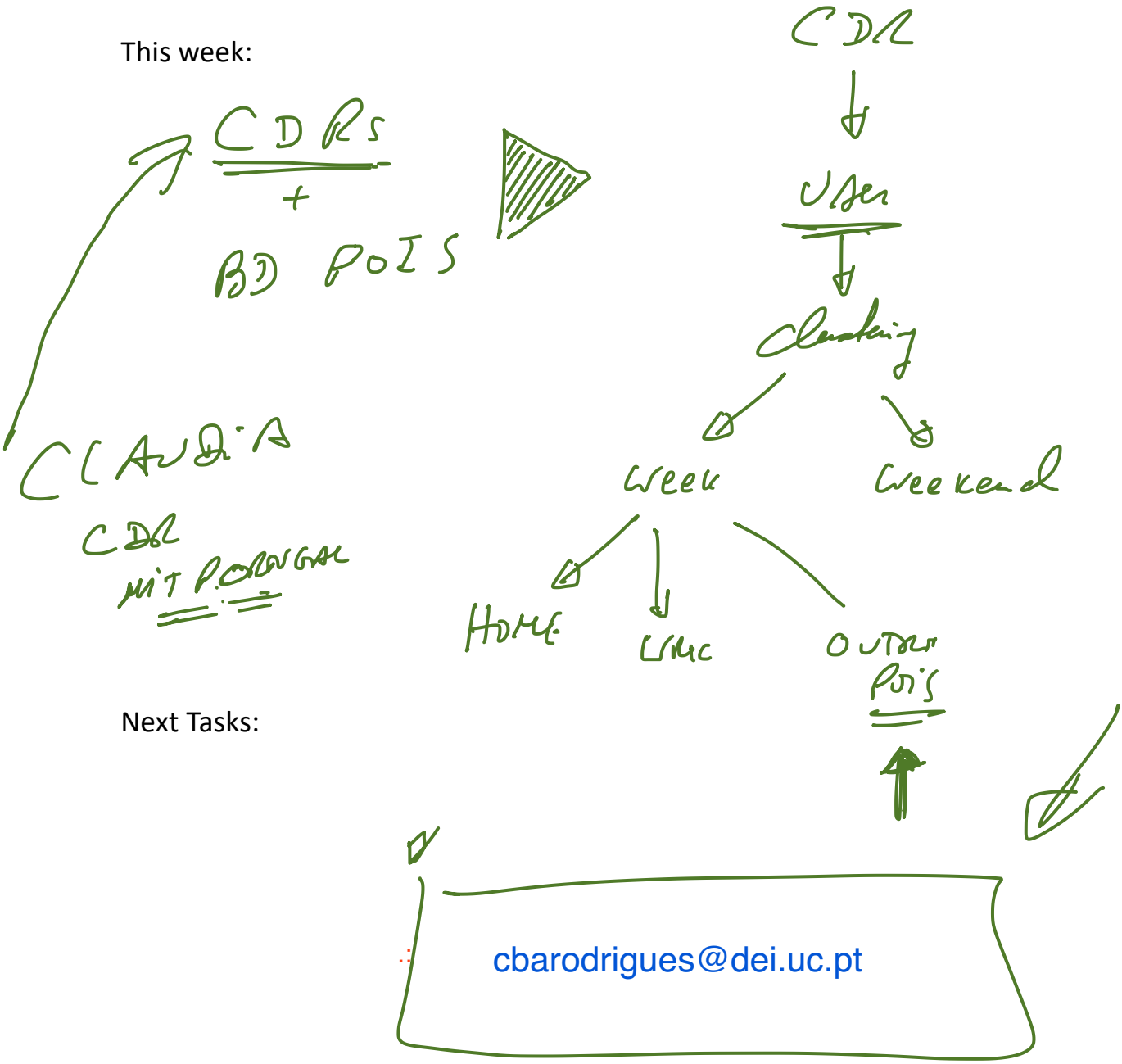
Project: B: Identifying Individual Routines: URBAN BASICET ANALYSIS for TOURISTS

Week: WO 6

Team: ALEXANDRE LEOPOLDO
SANCHO SIMÕES
TIAGO VENTURA

Progress (0..5): 4

This week:



Next Tasks:

cbarodrigues@dei.uc.pt

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: D: Understanding citizens transport modality choice

Week: W06

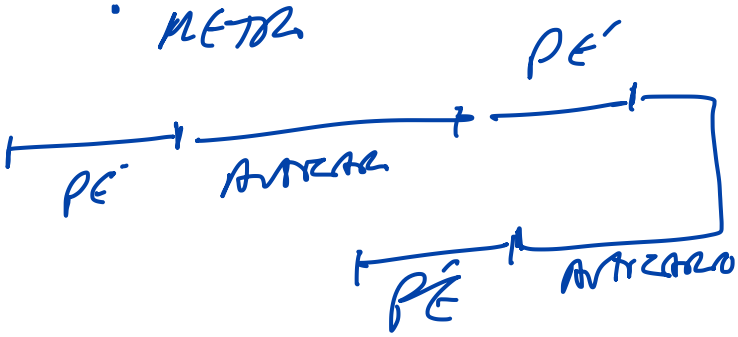
Team: ADRIANA BERNARDO
PEDRO HENRIQUES

Progress (0..5): 3

This week:

- GPS
- CAR
- MAP

⇒ em cada trajeto
- A PÉ
- AUTOCARRO
- CARRO
- METRO



Next Tasks:

CLAUDIA
GPS dataset IST ← PEDRO CLAUDIA

PAPERS 5/
Mode de transporte ← CLB

⇒ algoritmo + escolher
TRAÇO IDEIAS INFORMAÇÃO RELEV. SEG.

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

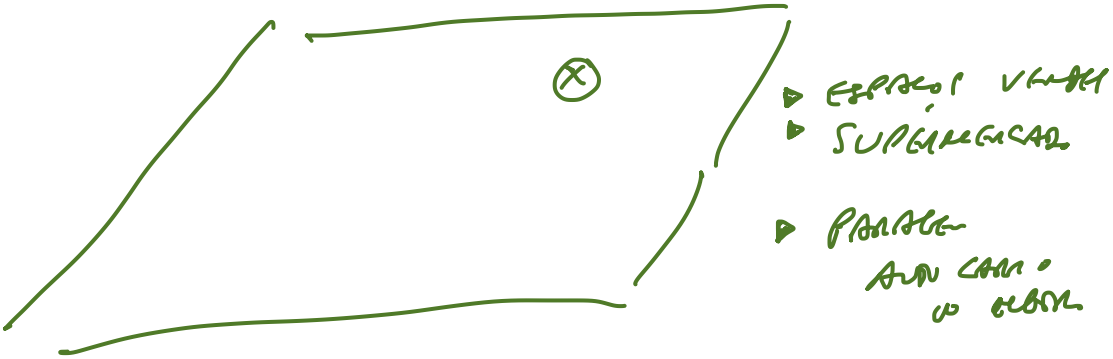
Project: E: Understanding the Accessibilities of a Urban Area (urban and transport planning ... real estate)

Week: W06

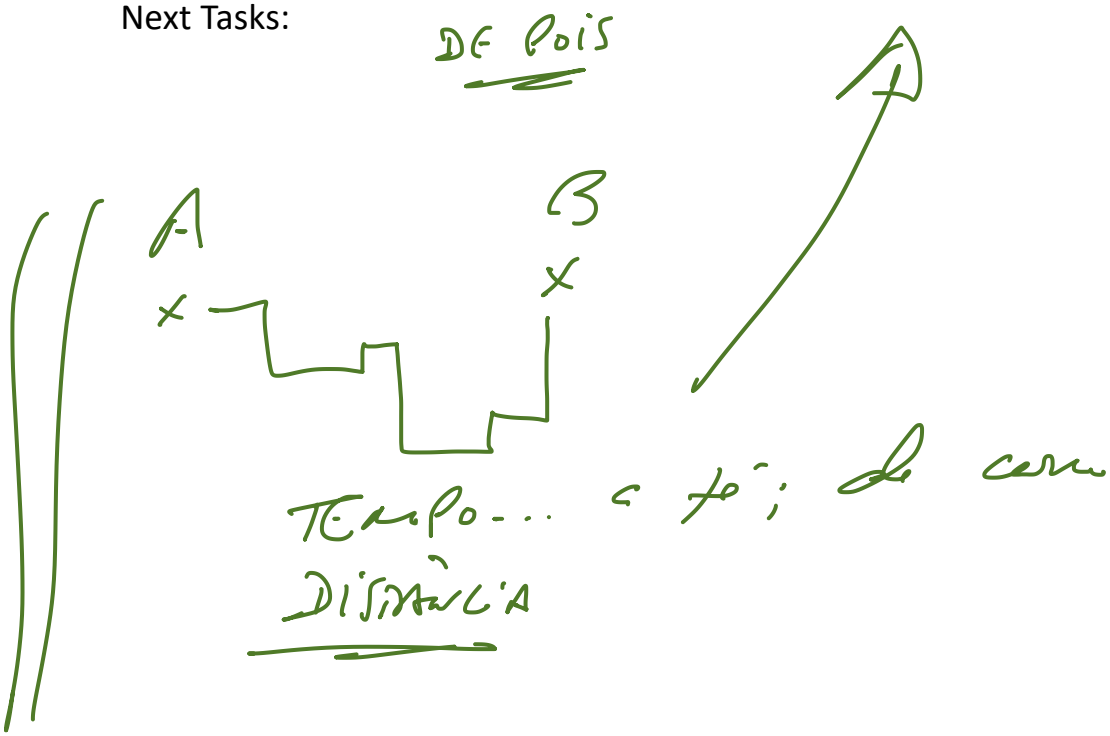
Team: JOÃO FERNANDES ✓
SAMUEL PIRES ✓

Progress (0..5): 4

This week:



Next Tasks:



Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks:

Project: Week:

Team:

Progress (0..5):

This week:

Next Tasks: