Survey on mobility data. A supervised classification approach for area identification

Manuel Mendoza Hurtado[1] and Domingo Ortiz Boyer[2]

[1] Computational Intelligence and Bioinformatics Research Group
University of Cordoba
{i52mehum}@uco.es
[2] Computational Intelligence and Bioinformatics Research Group
University of Cordoba
{dortiz}@uco.es

This paper presents a comparative study between clustering analysis, which is typically used in mobility scenarios, and supervised classification for the identification of home and work zones of an area. We will use a mobility dataset from the city of Milan to achieve this. Using passive mobile positioning data offers a powerful tool to study the geography and the mobility of the population. With the available data, we will try to identify workplaces and residential areas using both supervised classification and clustering. In order to generate training data for the classification model, we manually label several sub-regions of the available grid, one with random cells and another with a 20-by-20 resolution. Experimental results show that the kNN algorithm provides an acceptable accuracy that could be able to predict if a cell represent a working or a residential area for the full grid, thanks to the semi-supervised approach used in learning from a manually-labeled region. However, the results provided with k-means and k-medoids clustering show that it is not able to accomplish the former idea, instead it focuses on identifying the mobile traffic distribution around the city.

1 Introduction

The generalization of mobile phone usage has turned it into a useful tool to gather mobility data and extract knowledge applicable to areas such as mobile networks optimizations [9], marketing [24], urban planning [8], service planning, public transport [5], tourism [1], among others.

In this paper we present a new focus on area identification with mobility data, by using supervised classification and making a comparative study between clustering analysis, which is typically used in mobility scenarios [7] [10] [12]. In order to compare them, we will use mobility data from the city of Milan [2]. Our main goal is to be able to identify working and residential zones from a given area with the available data. Using supervised classification for this goal is novel, to the best knowledge of the authors. In order to generate training data for the classification

model, we manually label several sub-regions of the available grid, which is detailed in the following sections.

We will study the clustering problem [26], which has been widely used in literature for places identification. Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to different clusters are different. Clustering represents an easy way to analyze and categorize data because it is unsupervised learning, and normally no difficult processing of the data is required. We have used the k-means algorithm [26] and k-medoids [23].

For k-means, the general objective is to obtain the fixed number of partitions that minimize the sum of squared Euclidean distances between objects and cluster centroids. The algorithm requires in advance the number of desired clusters, so in order to find an effective way to cluster the data, we will test Silhouette values [25] for different k and find the value that suits better to our data.

We also use k-medoids [23] because it could be more robust to noise and outliers as compared to k-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. In contrast to the k-means algorithm, k-medoids chooses actual data-points as centers.

Other works that attempt to identify meaningful places use clustering in order to associate several clusters to the most frequently visited places from the users and then those clusters are ranked and defined as home or work by the hour they have more events, as explained in [10]. In [12] data traces are also grouped with clustering and time-based clustering is used to categorise the places. These approaches are valid if you have fine-grained data, but with aggregated data from CDRs this is not a valid approach since you cannot get individual users' data and the clustering will be restricted to the mobile traffic distribution of the data available, making clustering very unlikely to identify home and work areas.

We will study the classification task as a novel way to identify home and work zones for aggregated mobility data. Classification is a category of supervised learning, in which we have a dataset with input labels and a desired output. The main objective is to build a model that is capable of predicting the desired output of a new object given the input features, in our case being the work-home label. In classification, labels are discrete, meaning that a clear distinction is given between categories, that are nominal and not ordinal. Also in supervised learning, there is always a clear distinction between a training set and the test set that must be inferred. For the classification task we have chosen the k-nearest neighbors (kNN) classifier [6]. This is a simple, yet effective classification algorithm that can be easily interpreted. We will attempt to identify the home and work areas training the model with manually labeled sub-grids and then predict the class for the complete grid. The main challenge is to choose the optimal value of neighbors k, since is highly data-dependent. For our data we have chosen 5 and 10 neighbors. The limitation of this model is the need of manually labeling a subset of the data in order to train the algorithm.

In the next sections we go into details about the processing of the CDR data and the settings for the classification and the clustering. Following that, we will evaluate the results obtained in the different studies and final conclusions are given.

2 Experimentation

2.1 Data

The generalized use of mobile phone and the exponential increase on Internet usage is generating enormous amounts of data that along with positioning data from GPS or RF-beacons are proving to be a very useful tool to identify population patterns such as workplaces, leisure, tourism, residences, etc. One important source of information comes in form of CDRs stored by the mobile network operators. Unfortunately, communication data is usually only available to research teams that sign non-disclosure agreements and research contracts with private companies. The lack of open datasets limits the number of potential studies in the area. In this context, research challenges that provide access to datasets to a large number of researchers are becoming a valuable framework. An example of that is Orange's 'Data for Development' (D4D) initiative in 2013 [3] and 2014-2015 [22], but unfortunately the data is not available anymore.

However, Telecom Italia in association with several universities and foundations has made available the data that will be used in this study. This dataset is unique because it is a rich, open multi-source aggregation of telecommunications, weather, news, social networks and electricity from the city of Milan and the province of Trentino. We will focus only on the telecommunication data that consists of CDR data aggregated in 10 minutes interval on a 100x100 grid that covers the city of Milan.

The CDR contains records for the following activities:

- **Received SMS** (smsin) a CDR is generated each time a user receives an SMS
- **Sent SMS** (smsout) a CDR is generated each time a user sends an SMS
- **Incoming Call** (callin) a CDR is generated each time a user receives a call
- **Outgoing Call** (callout) a CDR is generated each time a user issues a call
- **Internet** a CDR is generated each time a user starts an Internet connection or ends an Internet connection. During the same connection a CDR is generated if the connection lasts for more than 15 min or the user transferred more than 5 MB.

The information is already aggregated and anonymized so that it is not possible to identify individual mobility data.

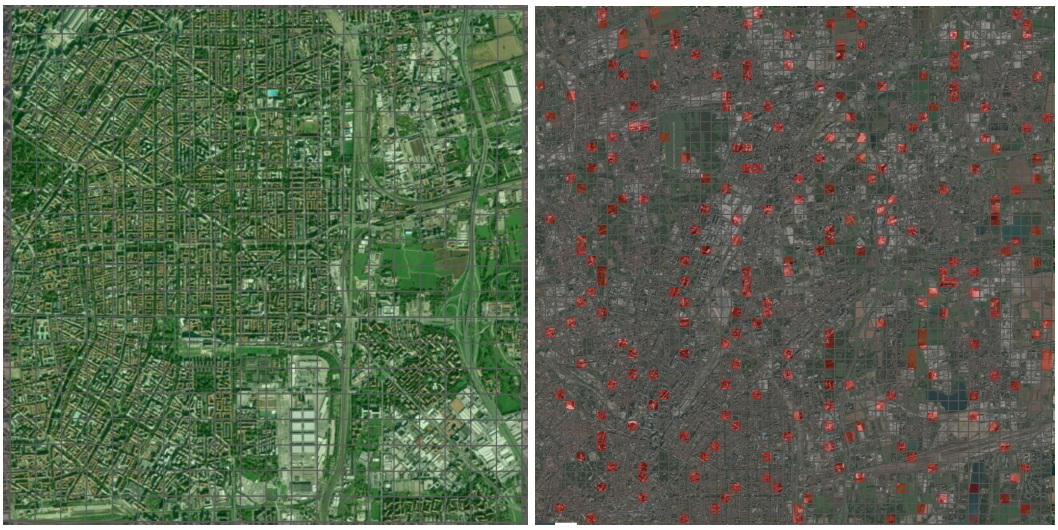
For our study, we select a typical week with no festivities over the 2-month period available to make the classification and clustering on the map. We will focus

on one week since what we are looking is to distinguish is the residential and work areas and one week is enough to study the mobile traffic patterns while having enough input data. We will also use a working day (monday) to compare the results from using a week of data. Having that into account, we also process the dataset by aggregating the measures into an hourly interval so that we have easier to manage data. In order to make the results normally distributed we normalize the data with a minmax scaler [11] so we can scale data into the [0,1] range.

As a mean to classify the complete grid, we create a subgrid of 20-by-20 resolution and another with 250 cells chosen randomly and manually label each cell with 1 if it represents a dominantly residential zone or 0 if it belongs to a work zone. This is possible to do thanks to the available file of the Milan grid which is a GeoJSON [4] that contains the coordinates of the 100x100 grid.

In order to have different samples to classify, we decided to create the aforementioned 20-by-20 grid that gathers an industrial zone clearly delimited and some high-density residential areas. It can be seen on figure 1a.

The other sample we have taken is choosing 250 cells randomly so we can check how the algorithms perform for this case. Figure 1b shows the random grid obtained.



(a) 20-by-20 grid visualization [14] **(b)** Random grid visualization [21]

Figure 1: Grids visualizations

In the next section we detail the process followed for the classification task.

2.2 Classification

We have decided to perform a k-nearest neighbors (kNN) classification because it could be effective due to the nature and shape of the data, and it is also the most commonly used technique. For the kNN classification we have chosen the following experiments:

- Classify the random grid with data for one working day.
- Classify the random grid with data for one week.
- Classify the 20x20 grid with data for one working day.
- Classify the 20x20 grid with data for one week.

We execute the classification algorithm with a 10-fold cross-validation and for each mobile traffic metric (incoming sms, outgoing sms, incoming calls, outgoing calls, internet traffic), with the neighbors parameter set to 5. Then we repeat all the classification but changing the neighbors parameter to 10 and present the new results in the next section. This will give us an idea on how the dataset is able to predict the home-work label for the defined sub-grids. After that, we will execute the kNN classification for the whole grid with the data for the 20x20 and random sub-grids as training for the algorithm. With this, we will check later if the algorithm is good for predicting the home and working zones for a bigger area.

2.3 Clustering

For the clustering task we have decided to use k-means as it converges fast and it is a commonly used algorithm. First, we execute k-means with 2 clusters as target to get a first idea on how the algorithm divides the grid in order to identify two classes, potentially being home and work.

The main question for k-means clustering is choosing the optimal cluster number (K) and in order to choose the best option for our data, we make a Silhouette analysis, calculating the S-Values for different K. From the range from k=2 to k=7 we have found that 5 clusters seems to be the best option for our data. Figure 2 shows the s-values distribution for the k-means with k=5.

Once we have the optimal k value for our data, we will proceed to execute the k-means clustering with 5 partitions. We will see how the partitions are distributed and if the follow some relation with the home and work zones present in the area of interest.

We are also going to study another clustering algorithm, k-medoids, since it could be more robust to noise and outliers compared to k-means. K-medoids works by minimizing a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances as in k-means. In contrast to the k-means algorithm, k-medoids chooses actual datapoints as centers.

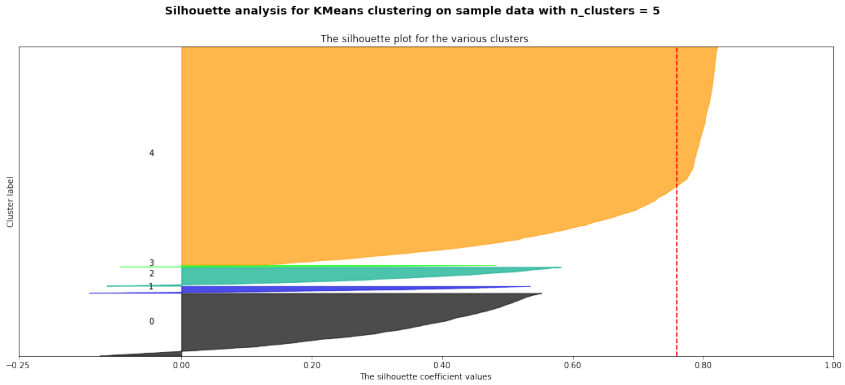


Figure 2: Silhouette analysis for k-means (k=5)

In the following section, we will detail the results obtained on the classification and clustering.

3 Results

Once we have explained the experiments that we have done in order to compare and find out which one gives us better results on detecting home and working areas, we present the results in this section. All the map visualization figures can be accessed with the reference links. First, we will start with the kNN classification with 5 and 10 neighbors (k=5 and k=10) for both of the sub-grids generated.

The results are shown in the following tables (accuracy and standard deviation values in percentage).

Table 1: kNN (neighbors=5) classification results

	smsin		smsout		callin		callout		internet	
	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std
Random - day	72.5	9.278	70.233	10.768	70.133	4.298	69.35	5.487	67.783	6.516
Random - week	71.72	6.374	66.517	5.158	70.55	6.396	67.33	4.58	69.767	7.82
20x20 - day	75.5	5.788	73.75	3.913	75.25	6.842	78	7.969	72.75	5.640
20x20 - week	77.75	5.75	75.5	6.305	77.5	5.7	77	6.689	69.25	6.986

Table 2: kNN (neighbors=10) classification results

	smsin		smsout		callin		callout		internet	
	Acc	Std	Acc	Std	Acc	Std	Acc	Std	Acc	Std
Random - day	70.533	7.578	71.383	10.462	71.05	7.555	69.3	6.03	68.217	7.882
Random - week	68.58	7.01	71.02	8.96	71.43	7.07	68.52	5.24	66.53	7.988
20x20 - day	77	7.228	74.75	4.535	74.75	6.842	78.5	7.263	72.75	6.169
20x20 - week	78.25	5.483	77.5	5.59	75.5	9.069	76	6.144	70.25	9.11

We can see that the 20x20 sub-grid performs better than the random sub-grid, and the accuracy achieved is good enough to consider the algorithm able to predict home and work areas. We will test this in a more visual way by showing the visualization of the home-work prediction for the 20x20 grid.

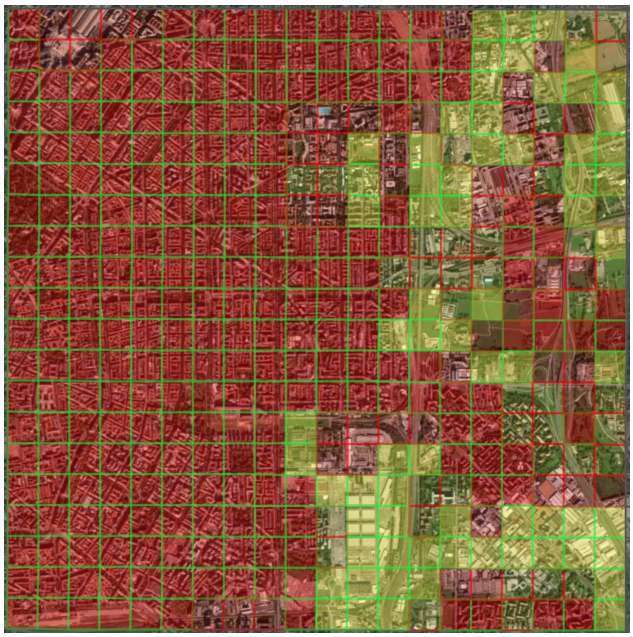


Figure 3: Prediction accuracy for the 20x20 grid [13]

In figure 3 it is shown the predicted grid for the 20x20 sample classified with kNN (k=10) with callout data. A grid in red represents a residential zone classified, and a yellow grid is a working zone. The grid border in green indicates a correct classification according to the training data (labels identified manually) and a red border indicates that the classification was incorrect. This figure gives us a visual idea of the performance of the classification performed.

Once we have tested the kNN algorithm with several settings (k=5 and k=10) and with several grids, we will proceed to predict the home-work label for the whole grid, using the 20x20 and random grid as training data for the algorithm.

The results are shown in figure 4, a red grid represents a zone identified as home and a yellow grid a working zone. We have found that, while it usually detects all industrial zones, it does not delimit the zones very well and several clearly defined residential zones are marked as working zones.

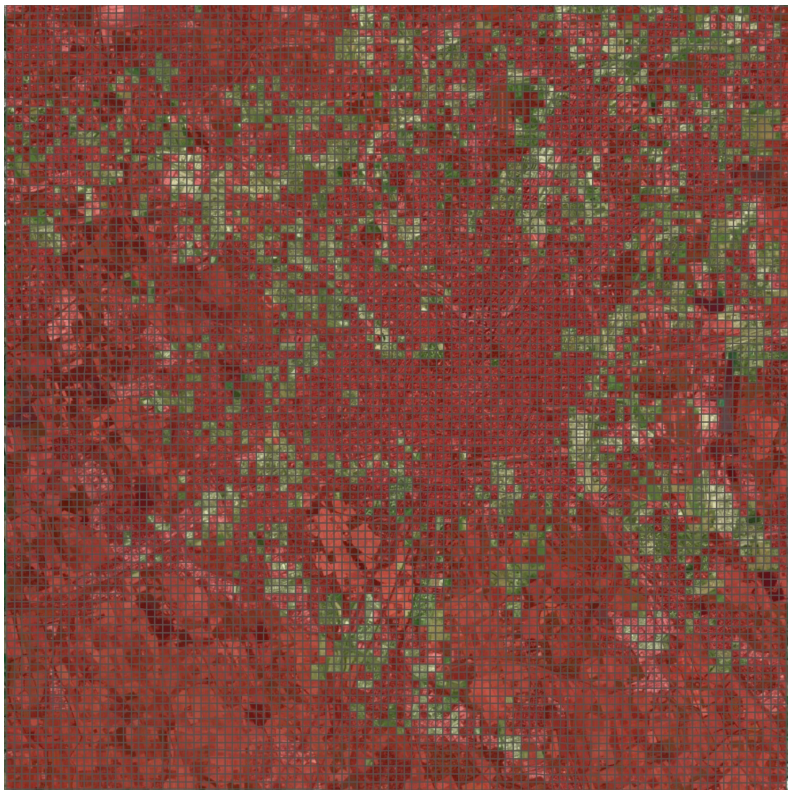


Figure 4: kNN home-work prediction for the full grid [15]

One problem that is made evident now is the base station (BS) coverage, because of the data being aggregated in the CDRs by grids, we have the following problem. There are many BS throughout the city and even more in the city center, and one grid will gather all the data from the different BS that it may cover. However, in less populated zones, the density of BS is lower and several grids can be covered by the same BS so they will have the same results, giving the same predictions in the algorithm while not being necessarily the same type of area.

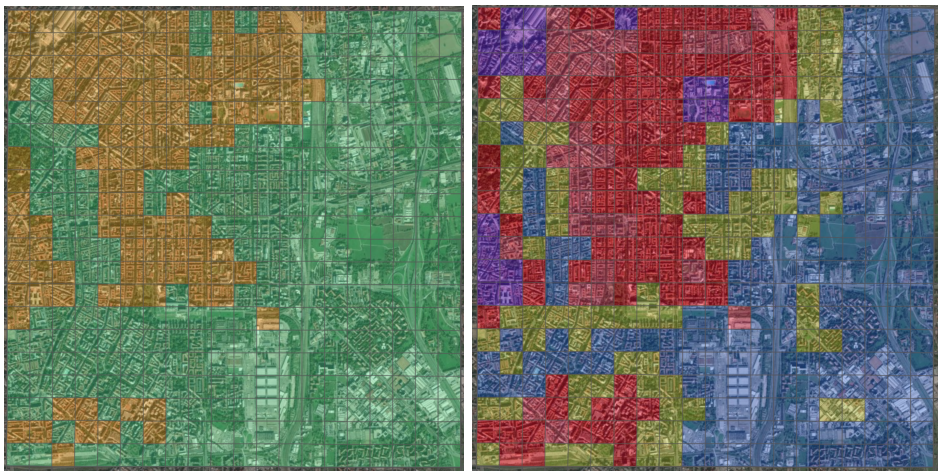
Taking this problem into account, the prediction with kNN gives us a good hint on the identification of the residential, working zones and industrial areas of the city of Milan.

Now, we will evaluate the predictions of the clustering algorithms and compare the results with the ones obtained with the kNN classification.

First we will study the k-means algorithm with 2 and 5 clusters on the 20x20 sub-grid in order to compare it with the predictions of the kNN classification.

In figure 5a we can see the 2 clusters resulting from the k-means algorithm and in figure 5b we can see the results with 5 clusters. In both cases, the clustering is not able to correctly differentiate into residential and work zones, as opposed to the kNN algorithm.

Next we will proceed to use clustering to the whole grid, and we start by distinguishing 2 partitions with k-means.



(a) k-means clustering result (k=2) [16] **(b)** k-means clustering result (k=5) [18]

Figure 5: Clustering on 20x20 sub-grid

As we can see in figure 6, the purple cluster represents the city center of Milan, which is an area with a dense mobile traffic and the rest of the map with less volume. But this does not satisfy our objective of detecting home and work areas, we will make another clustering with 5 clusters as found to be optimal in the Silhouette analysis.

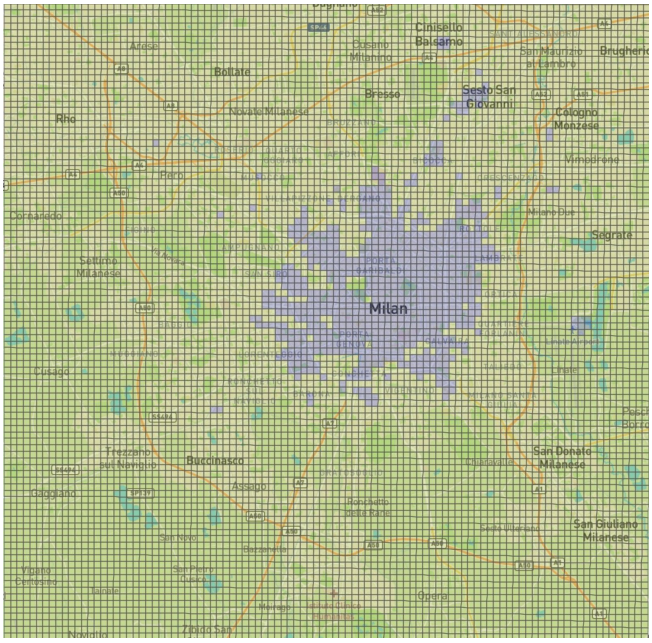


Figure 6: k-means clustering result (k=2) [17]

The results shown in figure 7, with 5 clusters we differentiate more widely the mobile traffic distribution along the city, but most of the identified clusters are grouping the city center of Milan. And we are still not able to clearly distinguish any home or work area.

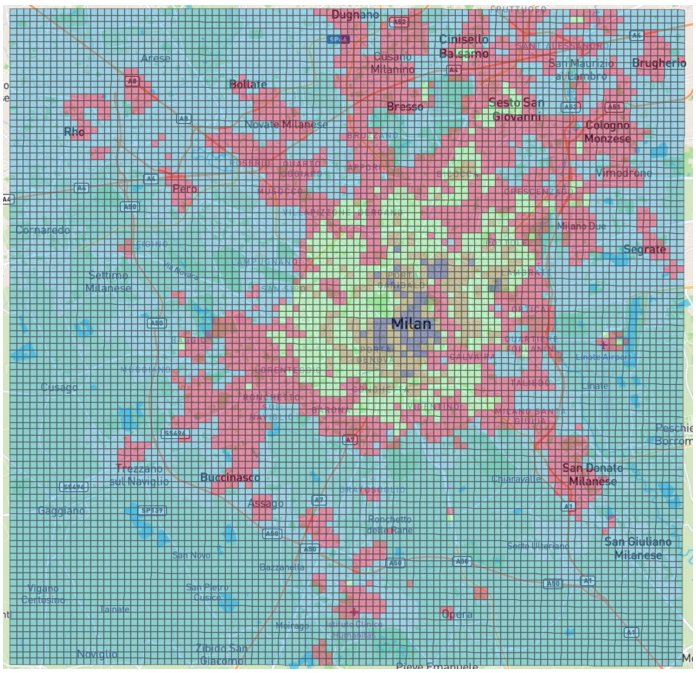
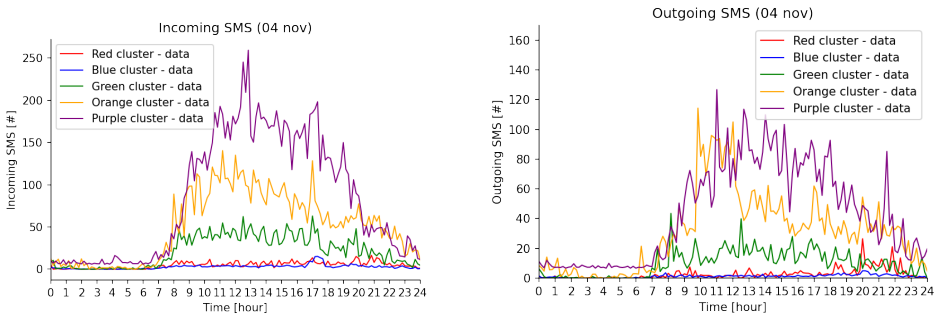


Figure 7: k-means clustering result (k=5) [19]

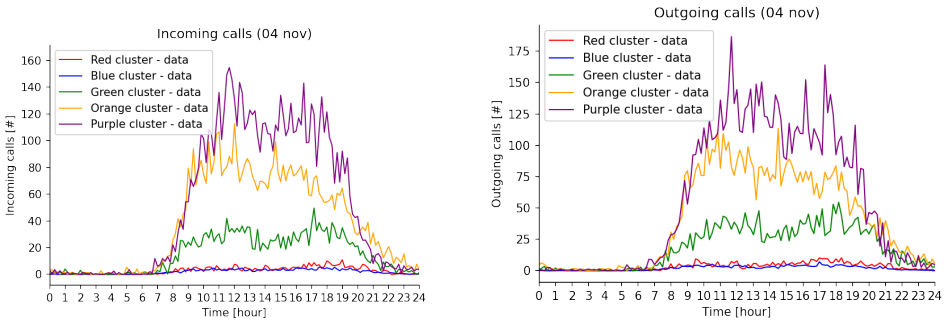
With these results we have made a mobile traffic analysis on each of the identified clusters, we will show the SMS (figures 8a, 8b), calls (figures 9a, 9b) and internet traffic (figure 10) on a working day among the 5 clusters.



(a) Incoming SMS analysis for k-means clusters

(b) Outgoing SMS analysis for k-means clusters

Figure 8: SMS traffic analysis



(a) Incoming calls analysis for k-means clusters

(b) Outgoing calls analysis for k-means clusters.

Figure 9: Call traffic analysis

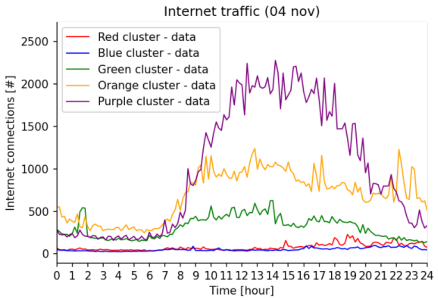


Figure 10: Internet traffic analysis for k-means clusters

We can conclude that each cluster represents the mobile traffic of each of the different zones of Milan, being the city center (purple, green and orange clusters) the zone that present more traffic on every kind of traffic analyzed. There are no meaningful differences on the k-means clustering using data for one week instead of a day.

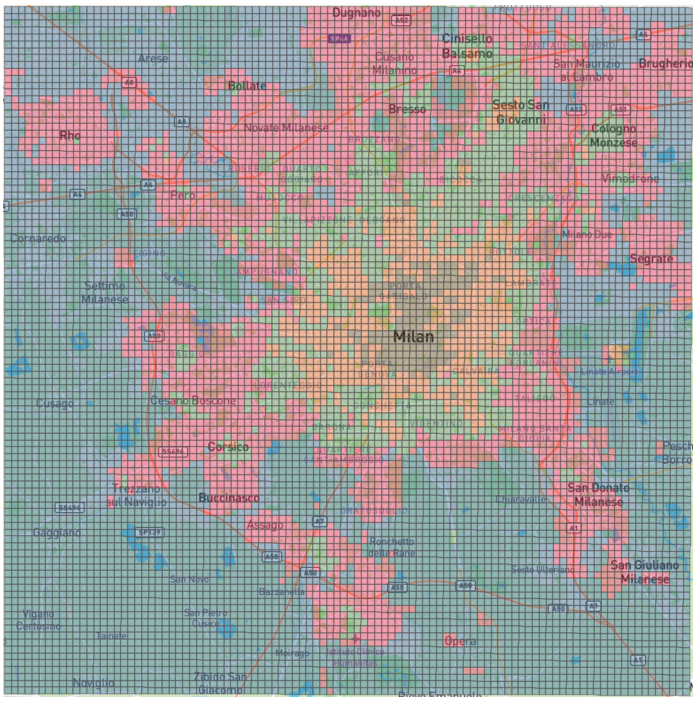


Figure 11: k-medoids clustering result (k=5) [20]

Lastly, we present the results from the k-medoids clustering. Figure 11 shows the predicted grid for k-medoids clustering with k=5 and data for one week with callout data.

We can see that this algorithm detects better the buildings, and that can be seen by comparing the outer clusters with the k-means, the external clusters in k-medoids model better the residential and industrial zones, but is still way behind the results obtained in the kNN classification since, as we have seen among all results, clustering focuses on detecting and grouping regions with similar mobile traffic distribution and that does not correlate with being a residential or a work zone.

4 Conclusions

In this paper we have performed a comparative study between the k-means and k-medoids clustering algorithms and the kNN supervised classification algorithm using an open mobility dataset containing CDR data from the city of Milan. We have seen that the kNN classification applied to several subsets of the whole grid have given good results on the 20x20 subgrid and thanks to that data, we were able to launch a prediction of home and work areas for the whole grid with the 20x20 and random grids as training data.

We have found that the prediction of the whole grid, although it is not perfect, it usually detects all the industrial zones, but it does not delimit the zones very

well and it marks some clearly-defined residential zones as working zones. This is probably due to the problem with the aggregated data which is related to the base station (BS) coverage, because the data used in the study was aggregated by grids. There are many BS throughout any city and even more in city centers, and one grid will gather all the data from the different BS that it may cover. However, in less populated zones, the density of BS is lower and several grids can be covered by the same BS so they will have the same results, giving the same predictions in the algorithm while not being necessarily the same type of area, so that is a very likely reason of the errors in the zone delimitation. However, the classification made with kNN has proven to be useful for identifying residential, industrial and working areas that could potentially help in urban planning and it is a novel method that has not been studied so far, to the best knowledge of the authors, for places identification with mobility data.

For the clustering analysis what we have distinguished are several zones of Milan by its mobile traffic, finding that the city center had the highest traffic, as expected, due to the touristic places and high-density houses. It was not possible to identify the desired categories of home and work areas due to the nature of unsupervised learning. This analysis did not differentiate between residential and working areas, but the information it provided about the mobile traffic in the city could result useful for telecommunication operators in order to optimize its infrastructure for better coverage and quality of service (QoS).

With this study we have shown the potential of using supervised classification in mobility problems, specially in zone identification with mobile telecommunication data, rather than using a clustering approach that has been commonly seen in the literature.

References

[1] R. Ahas, A. Aasa, Ü. Mark, T. Pae, and A. Kull. "Seasonal tourism spaces in Estonia: Case study with mobile positioning data". In: *Tourism Management* 28.3 (2007), pages 898–910. ISSN: 0261-5177. DOI: https://doi.org/10.1016/j.tourman.2006.05.010.

[2] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, B. Lepri, and et al. "A multi-source dataset of urban life in the city of Milan and the Province of Trentino". In: *Scientific Data* 2.1 (2015). DOI: 10.1038/sdata.2015.55.

[3] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. *Data for Development: the D4D Challenge on Mobile Phone Data*. 2013. arXiv: 1210.0137 [cs.CY].

[4] H. Butler, M. Daly, A. Doyle, S. Gillies, T. Schaub, and T. Schaub. *The GeoJSON Format*. RFC 7946. Aug. 2016. DOI: 10.17487/RFC7946.

[5] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. "Estimating Origin Destination Flows Using Mobile Phone Location Data". In: *IEEE Pervasive Computing* 10.4 (2011), pages 36–44. DOI: 10.1109/MPRV.2011.41.

[6] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (1967), pages 21–27. DOI: 10.1109/TIT.1967.1053964.

[7] Z. Duan, L. Liu, and S. Wang. "MobilePulse: Dynamic profiling of land use pattern and OD matrix estimation from 10 million individual cell phone records in Shanghai". In: *2011 19th International Conference on Geoinformatics*. 2011, pages 1–6. DOI: 10.1109/GeoInformatics.2011.5980928.

[8] L. Ferrari, M. Mamei, and M. Colonna. "Discovering events in the city via mobile network analysis". In: *Journal of Ambient Intelligence and Humanized Computing* 5.3 (2012), pages 265–277. DOI: 10.1007/s12652-012-0169-0.

[9] C. Fiandrino, C. Zhang, P. Patras, A. Banchs, and J. Widmer. "A Machine-Learning-Based Framework for Optimizing the Operation of Future Networks". In: *IEEE Communications Magazine* 58.6 (2020), pages 20–25. DOI: 10.1109/MCOM.001.1900601.

[10] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. "Identifying Important Places in People's Lives from Cellular Network Data". In: *Pervasive Computing*. Edited by K. Lyons, J. Hightower, and E. M. Huang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pages 133–151. DOI: 10.1007/978-3-642-21726-5_9.

[11] P. Juszczak, D. M. J. Tax, and R. P. W. Duin. "Feature scaling in support vector data description". In: (2002).

[12] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello. "Extracting Places from Traces of Locations". In: *SIGMOBILE Mob. Comput. Commun. Rev.* 9.3 (July 2005), pages 58–68. ISSN: 1559-1662. DOI: 10.1145/1094549.1094558.

[13] MapBox. *20x20 grid prediction*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/predictions_20co.geojson.

[14] MapBox. *20x20 grid visualization*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/test20by20.geojson.

[15] MapBox. *Full grid prediction*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/predict_fullgr2.geojson.

[16] MapBox. *k-means clustering (k=2) on 20x20 grid*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/clust20x20_km2.geojson.

[17] MapBox. *k-means clustering (k=2) on full grid*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/clusterfullgrk2callout.geojson.

[18] MapBox. *k-means clustering (k=5) on 20x20 grid*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/clust20x20_km5.geojson.

[19] MapBox. *k-means clustering (k=5) on full grid*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/clusterfullgrk5intnt.geojson.

[20] MapBox. *k-medoid clustering (k=5) on full grid*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/clustfullkmed5_co.geojson.

[21] MapBox. *Random grid visualization*. Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/rgridtest.geojson.

[22] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. *D4D-Senegal: The Second Mobile Phone Data for Development Challenge*. 2014. arXiv: 1407.4885 [cs.CY].

[23] H.-S. Park and C.-H. Jun. "A simple and fast algorithm for K-medoids clustering". In: *Expert Systems with Applications* 36.2, Part 2 (2009), pages 3336–3341. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.01.039.

[24] D. Quercia, G. Di Lorenzo, F. Calabrese, and C. Ratti. "Mobile Phones and Outdoor Advertising: Measurable Advertising". In: *IEEE Pervasive Computing* 10.2 (2011), pages 28–36. DOI: 10.1109/MPRV.2011.15.

[25] P. J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pages 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

[26] S. Singh and N. Singh Gill. "Analysis and Study of K-Means Clustering Algorithm". In: *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)* 2.7 (July 2013).