International Journal of Geo-Information

MDPI

Article

SAMPLID: A New Supervised Approach for Meaningful Place Identification Using Call Detail Records as an Alternative to Classical Unsupervised Clustering Techniques

Manuel Mendoza-Hurtado *, Juan A. Romero-del-Castillo and Domingo Ortiz-Boyer

Computer and Numerical Analysis, Campus de Rabanales, Albert Einstein Building, University of Cordoba, 14071 Cordoba, Spain; jaromero@uco.es (J.A.R.-d.-C.); dortiz@uco.es (D.O.-B.)
* Correspondence: mmendoza@uco.es

Abstract: Data supplied by mobile phones have become the basis for identifying meaningful places frequently visited by individuals. In this study, we introduce SAMPLID, a new Supervised Approach for Meaningful Place Identification, based on providing a knowledge base focused on the specific problem we aim to solve (e.g., home/work identification). This approach allows to tackle place identification from a supervised perspective, offering an alternative to unsupervised clustering techniques. These clustering techniques rely on data characteristics that may not always be directly related to classification objectives. Our results, using mobility data provided by call detail records (CDRs) from Milan, demonstrate superior performance compared to applying clustering techniques. For all types of CDRs, the best results are obtained with the 20 × 20 subgrid, indicating that the model performs better when supplied with information from neighboring cells with a close spatial relationship, establishing neighborhood relationships that allow the model to clearly learn to identify transitions between cells of different types. Considering that it is common for a place or cell to be labeled in multiple categories at once, this supervised approach opens the door to addressing the identification of meaningful places from a multi-label perspective, which is difficult to achieve using classical unsupervised methods.

Keywords: machine learning; land use; mobile network data; call detail records; urban mobility; human mobility; clustering; classification

check for updates

Citation: Mendoza-Hurtado, M.; Romero-del-Castillo, J.A.; Ortiz-Boyer, D. SAMPLID: A New Supervised Approach for Meaningful Place Identification Using Call Detail Records as an Alternative to Classical Unsupervised Clustering Techniques. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 289. https://doi.org/10.3390/ijgi13080289

Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 28 May 2024
Revised: 2 August 2024
Accepted: 15 August 2024
Published: 17 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The generalization of mobile phone usage has turned mobile phones into useful tools for gathering mobility data and extracting knowledge applicable to areas such as mobile network optimization [1], the identification of human mobility patterns [2,3], marketing [4], urban planning [5,6], service planning, public transport [7], and tourism [8], among others.

These data can be obtained from the mobile's global positioning system [9,10], the proximity to a radio-frequency beacon such as a Wi-Fi antenna or beacon [11], or call detail records (CDRs) stored by mobile phone network operators [12–15].

CDRs contain information that identifies the type of activity (internet, incoming/outgoing call, SMS message, etc.), its time of occurrence, and the radio base station (RBS) that provides the service. This information allows the user to be geolocated within the area covered by the RBS.

Although the positioning accuracy obtained by using CDRs depends on the extent of the area covered by the RBS, the large amount of information provided by these records makes them a useful tool for identifying population patterns such as those related to work, leisure, home, travel, and tourism [3,7,12–15]. Several other works with the use of mobile network data and CDRs have been conducted, such as in [16], where Gergő et al. demonstrate how mobile network data can provide an efficient, automated alternative to traditional census methods for analyzing commuting patterns in the Budapest Metropolitan

Area. And in [17], Ferreira et al. present a methodology to identify individual users' routine locations and attach semantic meanings to these locations using anonymized CDRs from mobile network data.

In any population analysis, the identification of meaningful places or locations [18,19] regularly visited by individuals or anchor points is of special interest. Home and work are the most common anchor points [12].

Clustering techniques [20] have been widely used in the literature for place identification [19,21,22]. Cluster analysis, also known as clustering, is the unsupervised process of partitioning a set of unlabeled data points into distinct and mutually exclusive groups (clusters) based on their inherent similarities. This process aims to maximize the similarity of data points within each cluster while minimizing the similarity between data points belonging to different clusters. Csáji et al. [21] use *k*-means, a clustering algorithm to identify locations exhibiting similar weekly calling patterns and to identify which places correspond to work or home based on the calling patterns. In the work of Isaacman et al. [15], clustering is used to associate several clusters with the places most frequently visited by users, and then those clusters are ranked and defined as home or work based on the hour during which they have the most events. For each cluster, they calculate five variables using the CDR data and a score using logistic regression, with coefficients calculated using the reported locations of the 18 volunteers. The resulting model for identifying home and work concludes, as expected, that the determining variables are only two: the number of events that occur during work hours (between 1 p.m. and 5 p.m.) and home hours (weekends or weekdays between 7 p.m. and 7 a.m.). Although the model uses the knowledge provided by the 18 volunteers, this only serves to demonstrate that the variables calculated from the CDR data fulfill the purpose for which they were defined. In [9], data traces are also grouped with clustering, and time-based clustering is used to categorize the identified places.

Recently, more advanced and refined clustering techniques have been introduced, such as the method described in [23], which combines Particle Swarm Optimization (PSO) with Fuzzy c-means (FCM) clustering. This approach aims to overcome the limitations of traditional FCM, including its sensitivity to noise and reliance on initial centroid selection.

In a *k*-means algorithm, the general objective is to obtain *k* partitions that minimize the sum of squared Euclidean distances between objects and cluster centroids. Although this algorithm is efficient, it is sensitive to noise and outliers. In contrast to *k*-means, the *k*-medoids algorithm [24] chooses actual data points as centers and minimizes the sum of general pairwise dissimilarities, making it more robust to noise and outliers.

Clustering represents an easy way to analyze and categorize data because it is an unsupervised learning technique and, typically, no difficult data processing is required. However, the groupings are based on data characteristics, which are not always related to the classification objectives [25]. That is, clusters made based on a measure of similarity or proximity of the data do not always have to correspond well with the classes into which we want to classify the data.

As an alternative to clustering and unsupervised learning, we can use supervised classification to identify population patterns that are learned by the classifier from correctly labeled data provided to it during the training phase. In supervised learning, 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 supervised learning, there is always a clear distinction between a training set and the test set that must be inferred.

The *k*-nearest neighbors (kNN) classifier [26] is among the highest-performing methods for classification tasks. This simple yet effective classification algorithm can be easily interpreted. The main challenge is to choose the optimal value of the *k* neighbors, which is highly data dependent.

In this paper, we present SAMPLID, a new Supervised Approach for Meaningful PLace IDentification using mobility data provided by call detail records of Milan. The idea

underlying our approach is based on addressing place identification problems from a supervised perspective, providing the model with representative information that allows it to learn from the places labeled by a expert, rather than through analysis and categorization of unlabeled data. In this way, we can train a classifier based on the knowledge provided and focused on the problem we aim to solve.

Providing a solid knowledge base to the classifier is essential. To generate training data for the classification model, we manually label several subregions using different sampling strategies such as random selection or selecting a large area containing all types of places to identify as detailed in the following sections. We attempt to identify the home and work places by training the model with manually labeled subgrids, based on the predominant type of buildings for each cell, and then predict the class for the complete grid.

To compare these approaches, we use mobility CDR data from the city of Milan [27] to identify home and work places and make a comparative study against clustering analysis, which is typically used in mobility scenarios [9,14,15]. We use *k*-means and *k*-medoids algorithms as unsupervised classifiers and SAMPLID-kNN using kNN classifier as a supervised classifier.

Overall, the contributions of our work are as follows.

- We propose to address the problem of meaningful place identification from a supervised perspective to improve the results obtained with unsupervised methods, such as clustering techniques, which are traditionally used when dealing with large amounts of unlabeled data like those provided by CDRs.
- To achieve this, we propose selecting a representative portion of the available data and correctly labeling it to use as a knowledge base for training a supervised model or classifier. This will enable us to accurately classify the entirety of the data for which we do not know the labels.
  We use mobility CDR data from the city of Milan to identify home and work places. To achieve this, we correctly label the data corresponding to a 20×20 subgrid and a series of random cells into which we divided the city of Milan.
- We compare the results obtained using *k*-means and *k*-medoids algorithms as unsupervised classifiers and SAMPLID-kNN using the kNN classifier as a supervised classifier.
- Finally, we not only assess the effectiveness of the proposed method compared to the alternatives considered, but we also draw additional conclusions, such as what information contained in the CDRs is most useful for identifying home and work places; whether better results are obtained using the knowledge provided by the 20×20 subgrid, the selected random cells, or their combination; the advantages and limitations of the proposed method; and the direction for future research.

In the next sections, we describe our proposed method (Section 2), the data used to perform the experiments and the experimental setup that we used to test the goodness of our proposed method (Section 3). We present our results in Section 4, and in Section 5, we state the conclusions of our work and future research lines.

2. SAMPLID: Supervised Approach for Meaningful Place Identification

In this section, we explain how to approach the problem of identifying areas of interest through supervised learning as an alternative to the traditionally used unsupervised methods.

The first problem we encounter in unsupervised classification is determining the number of classes c to identify. We can set this number based on our objectives or use a statistical procedure, letting the classifier analyze the data and classify them based on their characteristics. However, these data characteristics do not always have meaning or utility for the classification objectives, and the classifier may identify curious but impractical spurious correlations.

SAMPLID does not seek similarities among the data for classification; instead, it tries to find or learn a function that, when applied to the input variables associated with a pattern or sample, indicates the class to which the data belong.

Let T be a dataset consisting of n cells x_i and its associated classes y_i, $T = \{(x_i, y_i)\}$, $1 \leq i \leq n, (x_i \in \mathcal{X}, y_i \in \mathcal{Y})$. Here, $\mathcal{Y} = \{y_1, \ldots, y_c\}$ and $x_i = (x_{i1}, \ldots, x_{ij}, \ldots x_{im})$, where c is the number of classes and m is the number of variables available to classify each cell. h is a classifier, and $h(x_i) = y_i$ is the predicted class for the example x_i [28].

Here, we need a set of representative patterns or cells correctly classified $T = \{(x_i, y_i)\}$ that allows us to infer a function $h(x_i) = y_i$ and evaluate its efficiency before using it to classify unlabeled patterns x'_i. If such a set of patterns is not available, it will be necessary to generate a set using a specific strategy to select and classify such patterns.

We can choose to create a random selection of cells from the entire surface or part of the surface under study, select one or several representative subgrids that integrate all types of cells, or a combination of both. A random selection will, a priori, contain more information about the characteristics of the cells based on their distribution on the map, whereas a selection of subgrids will contain more information about the boundaries between areas of different types.

The classification of the learning set will depend on the characteristics of the problem to be solved. In the case of identifying home and work places, cells can be classified with on-site observation, based on maps that provide information about building types, signage in the areas, businesses, etc. The classification accuracy of the learning set is crucial for inferring the classifier. The importance and complexity of this task determine the size of the learning set.

The next step is to use a supervised classification method that allows us to infer the function $h(x_i) = y_i$. Taking into account the characteristics of the problem at hand, the kNN classifier may be a good choice. This classifier is well known and widely used due to its good generalization and easy implementation [29,30]. Although simple, the kNN classifier can usually match, and even surpass, the performance of more sophisticated and complex methods. The kNN algorithm utilizes a reference set of data points, known as prototypes, to encode the domain knowledge relevant to the problem. To classify a cell x_i using SAMPLID-kNN, the k cells that are more similar to x_i are obtained, denoted as the k nearest neighbors, and x_i is classified into the class that is most frequent in this set of k neighbors.

The time complexity of the algorithm in the prediction phase is $O(n \cdot m + n \log n)$.

If the classifier obtained and evaluated with correctly classified data is sufficiently accurate, we will apply the classifier to the study area. The method for evaluating the goodness of the results obtained in unlabeled cells will depend on the characteristics of the problem and actual evidence based on observation or available data.

Algorithm 1 presents SAMPLID in a formal manner and describes the steps for addressing the problem of meaningful place identification using supervised classifiers. A flowchart of the algorithm is represented in Figure 1.

Algorithm 1: SAMPLID: Supervised Approach for Meaningful Place Identification.

```
Select a representative set of cells x_i;
Label the selected set correctly T = {x_i, y_i}, taking into account the places we
  want to identify;
Train a supervised classification method to infer the function h(x_i) = y_i;
Evaluate the accuracy of the obtained results y'_i compared with the expected
  values y_i;
if accuracy > threshold then
    Apply h to unlabeled cells x'_i, h(x'_i) = y'_i;
    Evaluate the obtained results y'_i using real evidence;
```

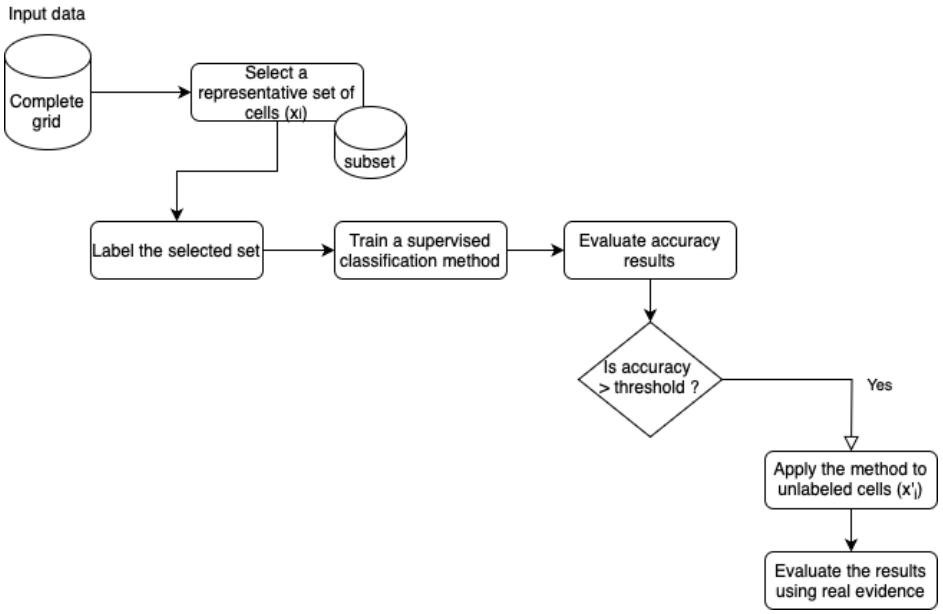


Figure 1. Graphical representation of SAMPLID.

3. Experimental Setup

Next, we describe the characteristics of the data, the supervised classification method used to evaluate the goodness of SAMPLID, and the clustering algorithms used for comparison with our proposed method.

3.1. Data Description

Access to communication data is often restricted, typically requiring research teams to enter into non-disclosure agreements and research contracts with private companies. This limited availability of open datasets presents a significant challenge for researchers seeking to conduct studies in this field.

In this context, research challenges providing access to datasets to a large number of researchers are becoming a valuable framework. An example is Orange's "Data for Development" (D4D) initiative in 2013 [31] and 2014–2015 [32], but unfortunately, these data are no longer available.

However, Telecom Italia, in association with several universities and foundations, has made available the data that will be used in this study. This dataset stands out due to its open- and multi-source aggregation of data from diverse sources, such as telecommunications, weather, news, social networks, and electricity information for the city of Milan [27]. We will focus on the telecommunication data, which consist of CDR data aggregated in 10 min intervals on a 100-by-100 grid (squared cells with an approximate size of 235 $\times$ 235 m) that covers the city of Milan (Figure 2a).

Each time a user interacts with the telecommunications network, the operator assigns an RBS and creates a CDR containing the RBS that handled the interaction and the time at which the interaction occurred. Each RBS covers a specific area of the territory v and allows its users to be geolocated using RBS coverage maps $RBSc_map$ (Figure 2b).

A square of the grid can be composed of several portions of the coverage areas of the RBS and vice versa. To perform spatial aggregation, the number of interactions or records $S_i(t)$ in a grid square i at time t is calculated as follows [27]:

$$S_i(t) = \sum_{v \in RBSc_map} R_v(t) \frac{A_{v \cap i}}{A_v}, \tag{1}$$

where $R_v(t)$ is the number of records in v at time t, A_v is the area of v, and $A_{v \cap i}$ is the area of the intersection between v and the square i.

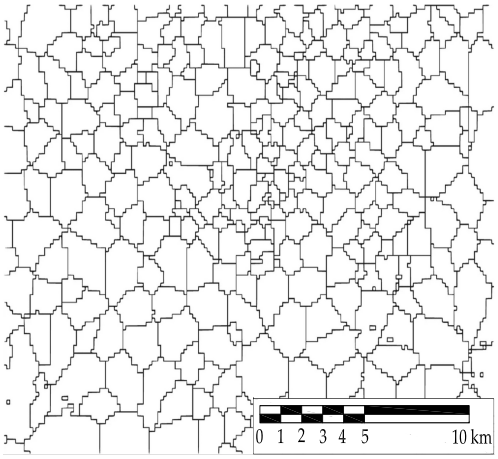
The number of records in the datasets $S'_i(t)$ is calculated as follows:

$$S'_i(t) = S'_i(t)\kappa, \tag{2}$$

where κ is a constant that hides the true number of records.

9901	9902	···	9999	10,000
9801	···	···	9899	9900
···	···	···	···	···
101	102	···	···	200
1	2	3	···	100

(**a**)



(**b**)

Figure 2. (**a**) Grid system of Milan. (**b**) An example of a RBS coverage map of Milan.

The CDR contains records for the following activities:

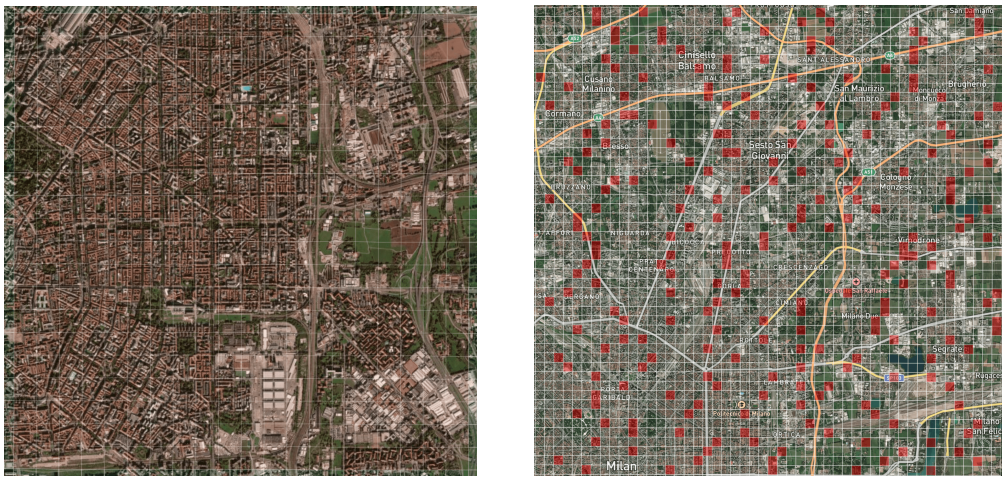
- **Incoming SMS** (smsin). A dedicated CDR is generated for each received SMS message.
- **Sent SMS** (smsout). Similarly, a single CDR is generated for each sent SMS message.
- **Incoming call** (callin). Every received call triggers the creation of a corresponding CDR.
- **Outgoing call** (callout). Likewise, a CDR is generated for each initiated call.
- **Internet**. Mobile internet activity is represented in the CDR. A CDR is generated at the start and end of each internet connection. Additionally, during an ongoing connection, a new CDR is created if the duration exceeds 15 min or if data transfer surpasses 5 MB.

The information is already aggregated and anonymized so that it is not possible to identify individual mobility data. The data are provided in text files, in which each line contains the following fields: cell identifier, date, time, number of records ($S'_i(t)$) of incomng SMS messages (smsin), sent SMS messages (smsout), incoming calls (callin), outgoing calls (callout) and internet.

For our study, we select a typical week with no festivities over the two-month period available to perform classification and clustering on the map. We focus on one week because we are aiming to distinguish home and work places; one week is sufficient for studying mobile traffic patterns while providing enough input data. We also use a working day (Monday) for comparison with the results obtained from using a week of data.

We process the dataset by aggregating the measures into hourly intervals and scaling their values into a range of [0, 1] using a min-max scaling [33], for ease of data management.

As a means of classifying the complete grid, we manually label two sets of cells as home and work places. This labeling is possible because of the available file of the Milan grid, which is a GeoJSON file [34] that contains the coordinates of the 100 × 100 grid. The first set is formed by cells belonging to a 20 × 20 subgrid that includes an industrial zone that is clearly delimited and some high-density residential areas (Figure 3a (Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/test20by20.geojson accessed on 14 August 2024). The second set consists of 250 cells randomly chosen from the upper left 50 × 50 subgrid (Figure 3b (Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/rgridtest.geojson accessed on 14 August 2024)).



(**a**) (**b**)

Figure 3. Satellite view of (**a**) a 20 × 20 subgrid and (**b**) randomly chosen cells, in red, of the upper right 50 × 50 subgrid.

3.2. Supervised Classification Method

To validate SAMPLID, we use the kNN classifier [26] because it is one of the most commonly used, simple, effective, and easily interpretable classifiers, with a high performance that reaches the level of more complex classifiers. To evaluate the performance of SAMPLID-kNN, we use $k = \{5, 10\}$ as the number of neighbors, and six different datasets:

- The 20 × 20 subgrid with data for one week;
- The 20 × 20 subgrid with data for one working day;
- The subgrid formed by 250 random cells with data for one week;
- The subgrid formed by 250 random cells with data for one working day;
- The 20 × 20 subgrid and 250 random cells with data for one week;
- The 20 × 20 subgrid and 250 random cells with data for one working day.

To obtain results, we perform a 10-fold cross-validation for each mobile traffic metric (incoming SMS, outgoing SMS, incoming calls, outgoing calls, and internet traffic) and each value of $k = \{5, 10\}$.

The results obtained will indicate the most appropriate values for k and the mobile traffic metric, the most appropriate training set, and whether it is better to use data for one week or one working day. Using this information, we apply SAMPLID-kNN to the whole grid and assess whether the method is effective in predicting home and work places for the whole grid.

3.3. Clustering Method

To compare the results obtained with our model, we use a k-means clustering algorithm whose objective is to minimize the sum of the squared Euclidean distances between elements of the clusters and their centroids. This algorithm converges rapidly and is widely used in the literature for detecting meaningful places [21].

We also apply the k-medoids clustering algorithm [24], which minimizes a sum of general pairwise dissimilarities and chooses actual data points as centers. This algorithm may be more robust to noise and outliers than the k-means algorithm.

First, we apply the k-means and k-medoids algorithms with two clusters to identify two classes, potentially being home and work. Next, we conduct a silhouette analysis [35] using the silhouette coefficient [36] to calculate the optimal number of clusters k; we then apply the clustering algorithms again with the optimal number of clusters, k. Here, we utilize values between $k = 2$ and $k = 7$.

4. Results

After describing the experiments that we have performed in order to identify which approach achieves superior results in detecting home and work places, we present the results in this section. All of the map visualization figures can be accessed at the reference links.

Tables 1 and 2 show the accuracy and standard deviation values obtained by SAMPLID-kNN with 5 or 10 neighbors, respectively, for each dataset considered. The values marked in bold represent the best result for each category of incoming or outgoing SMS messages, calls, and internet. A *t*-test [37] with an alpha of 0.05 is conducted to identify significant differences among the results. We also calculate Cohen's *d* [38] to determine whether the significant differences are very small ($d \leq 0.01$), small ($d \leq 0.2$), medium ($d \leq 0.5$), large ($d \leq 0.8$), very large ($d \leq 1.20$), or huge ($d \leq 2.0$).

Table 1. SAMPLID-kNN (k = 5) classification results.

	smsin		**smsout**		**callin**		**callout**		**internet**	
	Acc	**Std**	**Acc**	**Std**	**Acc**	**Std**	**Acc**	**Std**	**Acc**	**Std**
20 × 20—week	**0.777**	**0.057**	0.755	0.063	**0.775**	**0.057**	0.770	0.668	0.692	0.069
20 × 20—day	0.755	0.057	0.737	0.039	0.752	0.068	**0.780**	**0.079**	**0.727**	**0.056**
Random—week	0.717	0.063	0.665	0.051	0.705	0.063	0.673	0.045	0.697	0.078
Random—day	0.725	0.092	0.702	0.107	0.701	0.042	0.693	0.054	0.677	0.065
Random & 20 × 20—week	0.716	0.058	0.688	0.050	0.708	0.054	0.728	0.057	0.716	0.057
Random & 20 × 20—day	0.715	0.052	0.721	0.054	0.719	0.059	0.715	0.07	0.716	0.066

Table 2. SAMPLID-kNN (k = 10) classification results.

	smsin		**smsout**		**callin**		**callout**		**internet**	
	Acc	**Std**	**Acc**	**Std**	**Acc**	**Std**	**Acc**	**Std**	**Acc**	**Std**
20 × 20—week	**0.782**	**0.054**	**0.775**	**0.055**	0.755	0.090	0.760	0.061	0.702	0.091
20 × 20—day	0.770	0.072	0.747	0.045	0.747	0.068	**0.785**	**0.072**	**0.727**	**0.061**
Random—week	0.685	0.070	0.710	0.089	0.714	0.070	0.685	0.052	0.665	0.079
Random—day	0.705	0.075	0.713	0.104	0.710	0.075	0.693	0.603	0.682	0.078
Random & 20 × 20—week	0.739	0.071	0.704	0.043	0.718	0.069	0.738	0.082	0.707	0.057
Random & 20 × 20—day	0.724	0.042	0.724	0.048	0.738	0.055	0.727	0.056	0.715	0.080

The best results, in absolute terms, are obtained for the 20 × 20 subgrid with data for outgoing calls (*callout*) for one working day with the number of neighbors $k = 10$; in general, the results obtained are sufficient to consider this algorithm capable of predicting home and work places.

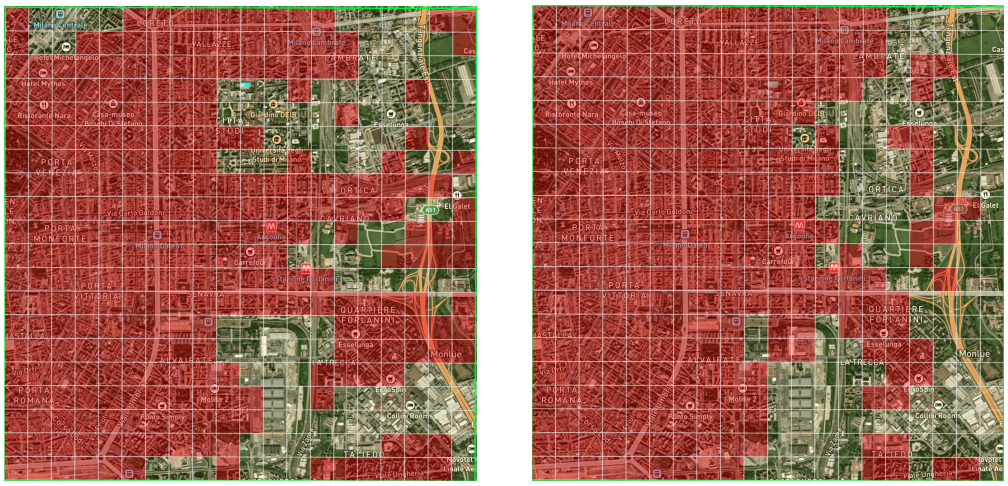
For incoming messages (*smsin*), outgoing messages (*smsout*), and incoming calls (*callin*), the best results are obtained with the 20 × 20 week dataset. For outgoing calls (*callout*) and *internet*, the best results are obtained with the 20 × 20 day dataset.

We can see that the 20 × 20 subgrid performs better than the random subgrid. This behavior may result from the greater facility in classifying cells concentrated in a zone with clear transitions between areas, compared with randomly chosen cells with diverse spatial relationships due to the extent of the territory covered. However, the use of both datasets does not improve the results, indicating that the information provided to the classifier by groups of close cells is essential for achieving good results.

The differences obtained for $k = 5$ and $k = 10$ are not significant for *smsin*, *callout*, or *internet*, but these differences are significant for *smsout* ($k = 10$) and *callin* ($k = 5$). The *Cohen's d* values for *smsout*, $d = 0.318$, and *callin*, $d = 0.339$ indicate that these significant differences are moderate. Therefore, we can conclude that the results obtained for $k = 5$ and $k = 10$ are similar.

In a more visual way, Figure 4a on the left, (Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/true20x20.geojson accessed on 14 August 2024) shows the manually labeled subgrid,

and on the right, Figure 4b (Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/predictions_2020co.geojson accessed on 14 August 2024) shows the same predicted subgrid for the 20 × 20 sample classified with SAMPLID-kNN ($k = 10$) using callout data from one working day. The red cells represent a residential zone, and transparent cells represent a working zone. Both figures visually indicate the good performance of the classification. The main discrepancies occur, as expected, in cells that are difficult to classify because they share both types of places or they are border cells, small islands, or cells that have a shared use as in a university campus.



(**a**) (**b**)

Figure 4. (**a**) True and (**b**) predicted classification for the 20 × 20 subgrid classified with SAMPLID-kNN ($k = 10$) using callout data from one working day.

If a group of cells is covered by a single RBS, the cells will have the same data; if they belong to different classes, the classifier will assign the same class to all cells. If cells of different classes are covered by different RBSs, their data will be different, and the cells will be more easily classified.

Once the goodness of the results obtained in the labeled cells using SAMPLID-kNN has been evaluated, we proceed to predict home/work labels for the whole grid. To do this, we use the configuration for which we obtained the best results in the previous phase, i.e., the 20 × 20 subgrid with *callout* data for one working day, as a training set, with $k = 10$.

The results are shown in Figure 5 (Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/predictions_fullgrid_co.geojson accessed on 14 August 2024). A red grid represents a zone identified as home, and a transparent grid indicates a working zone. The city center is primarily classified as home. Only some large areas covered by RBSs that primarily cover working zones are classified as such. When we move away from the city center, the opposite trend occurs, and the residential zones are difficult to identify when their extension or density is not sufficient. As we move away from densely populated areas, the extent of the area covered by the RBSs increases, and it becomes more complex to detect the islands of cells belonging to residential zones. Given the limitations of the data used, the SAMPLID-kNN prediction provides valuable insight into the identification of home and work places in the city of Milan.

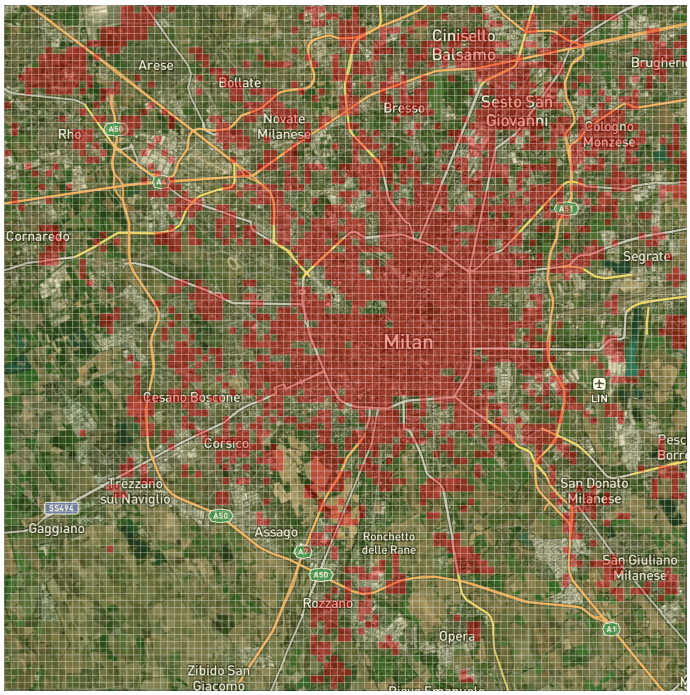


Figure 5. SAMPLID-kNN home/work prediction for the full grid.

Now, we evaluate the predictions of the clustering algorithms and compare the results with those obtained via SAMPLID-kNN.

Tables 3 and 4 show the results obtained by the *k*-means and *k*-medoids clustering algorithms using two clusters. The best results, shown in bold, are obtained with the *k*-medoids algorithm and the 20 × 20 subgrid with *callout* data for one working day; however, these results are never better than those obtained by the kNN classification.

Table 3. Cluster accuracy results obtained by *k*-means using two clusters.

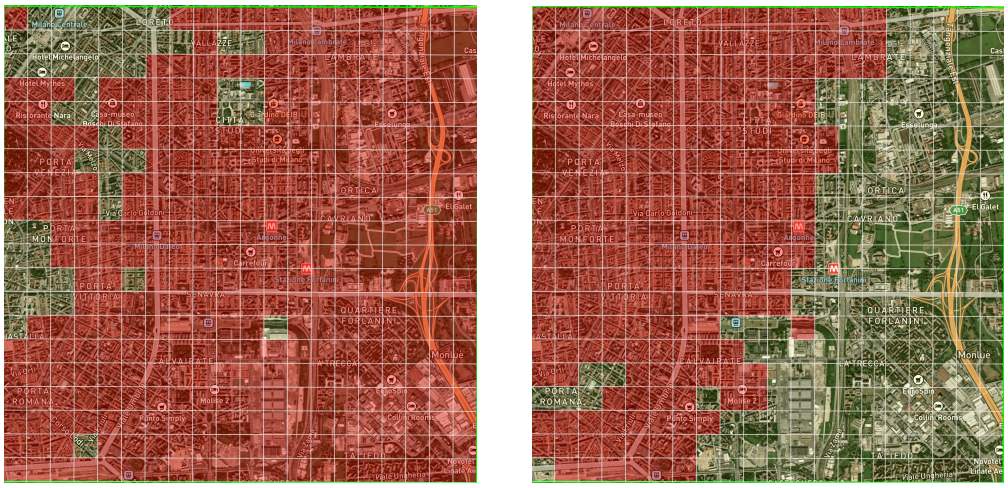
	smsin	**smsout**	**callin**	**callout**	**internet**
20 × 20—week	0.540	0.528	0.505	0.570	0.515
20 × 20—day	0.638	**0.635**	0.503	0.503	0.525
Random—week	0.629	0.617	**0.637**	**0.605**	0.657
Random—day	**0.641**	0.621	**0.637**	0.597	**0.661**
Random & 20 × 20—week	0.631	0.630	0.563	**0.605**	0.543
Random & 20 × 20—day	0.634	0.635	0.546	0.539	0.543

Table 4. Cluster accuracy results obtained by *k*-medoids using two clusters.

	smsin	**smsout**	**callin**	**callout**	**internet**
20 × 20—week	**0.650**	0.630	0.643	0.728	0.548
20 × 20—day	0.640	**0.635**	**0.653**	**0.740**	**0.588**
Random—week	0.573	0.585	0.577	0.552	0.573
Random—day	0.552	0.585	0.581	0.569	0.565
Random & 20 × 20—week	0.557	0.610	0.542	0.619	0.515
Random & 20 × 20—day	0.577	0.554	0.563	0.605	0.540

Figure 6a (available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/clust20x20_2.geojson accessed on 14 August 2024) shows the best results obtained for the 20 × 20 subgrid and *k*-means using smsin data from one working day. Similarly, Figure 6b (available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendoz

amanu/milan-mobility/main/geojsons/clust20x20_km2.geojson accessed on 14 August 2024) shows the best classification obtained for the same area and *k*-medoids using callout data from one working day. Graphically, we can see that the *k*-means classifies almost all cells as home, even those that are clearly work. In the case of *k*-medoids, a clear border is established between the left part of the image, home, and the right part, work. In both cases, the clustering is not able to improve the classification performed by the proposed model SAMPLID using kNN (Figure 4b).



(**a**) (**b**)

Figure 6. Classification for the 20 × 20 subgrid with (**a**) *k*-means using smsin data from one working day and (**b**) *k*-medoids using callout data from one working day.

Next, we apply clustering to the whole grid using the configuration for which we obtain the best results: *callout* data for one working day with the *k*-medoids algorithm.

The results obtained are shown in Figure 7 (Available at http://geojson.io/#data=data:text/x-url,https://raw.githubusercontent.com/mendozamanu/milan-mobility/main/geojsons/clust_fg_km2_callout.geojson accessed on 14 August 2024). We can see that the clustering algorithms have more difficulty in identifying residential areas far from the city center and large areas of the city center without homes than the method proposed in this work (Figure 5).

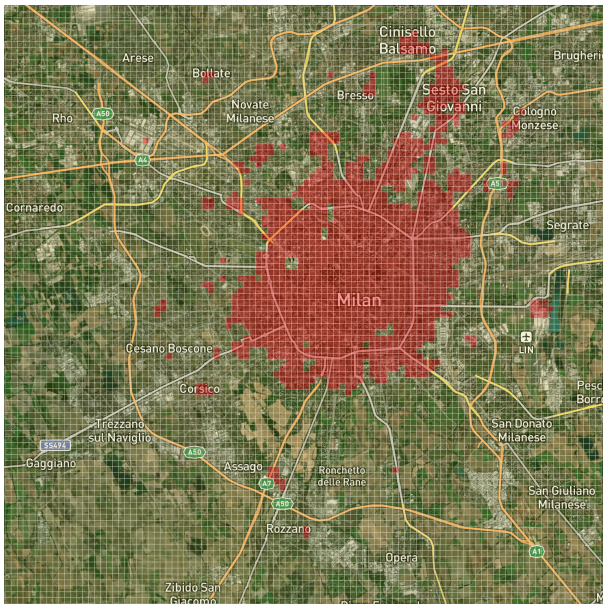


Figure 7. *k*-means clustering result (*k* = 2). Cluster 1 in red and Cluster 2 in transparent/yellow.

To determine whether it is possible to improve the clustering results obtained for the whole grid algorithm by applying better data partitioning, we determine the silhouette coefficient for k values ranging between $k = 2$ and $k = 7$, as shown in Table 5 for each dataset. The values marked in bold represent the best results, which are those closest to unity. In general, the best results are always obtained for the $k = 2$ cluster, which indicates that the best way to group the data is in two clusters as we did before. Thus, it is not necessary to conduct additional runs of the clustering algorithms.

Table 5. Values of the silhouette coefficient for k clusters.

Dataset	k Clusters	smsin	smsout	callin	callout	internet
20 × 20 week	2	**0.726**	**0.668**	**0.742**	**0.764**	0.769
	3	0.715	0.593	0.737	0.703	**0.771**
	4	0.721	0.605	0.670	0.708	0.747
	5	0.616	0.600	0.683	0.652	0.624
	6	0.626	0.474	0.669	0.619	0.635
	7	0.630	0.483	0.606	0.584	0.622
20 × 20 day	2	**0.834**	**0.801**	0.740	**0.727**	0.749
	3	0.747	0.645	**0.747**	**0.727**	**0.772**
	4	0.743	0.663	0.702	0.691	0.752
	5	0.662	0.666	0.714	0.704	0.655
	6	0.662	0.558	0.711	0.677	0.621
	7	0.643	0.567	0.641	0.615	0.649
Random week	2	**0.863**	**0.841**	**0.882**	**0.858**	**0.887**
	3	0.855	0.777	0.855	0.809	0.835
	4	0.724	0.755	0.789	0.793	0.799
	5	0.692	0.503	0.699	0.688	0.673
	6	0.665	0.498	0.696	0.683	0.671
	7	0.658	0.512	0.687	0.687	0.680
Random day	2	**0.874**	**0.849**	**0.885**	**0.856**	**0.886**
	3	0.849	0.805	0.858	0.834	0.828
	4	0.822	0.776	0.816	0.821	0.748
	5	0.709	0.777	0.739	0.724	0.674
	6	0.711	0.584	0.725	0.707	0.675
	7	0.702	0.587	0.725	0.706	0.671
Random & 20 × 20 week	2	**0.852**	**0.817**	**0.791**	**0.828**	**0.808**
	3	0.746	0.705	0.748	0.721	0.799
	4	0.718	0.572	0.756	0.733	0.723
	5	0.634	0.548	0.691	0.648	0.649
	6	0.622	0.546	0.682	0.613	0.660
	7	0.631	0.551	0.611	0.588	0.644
Random & 20 × 20 day	2	**0.859**	**0.843**	**0.784**	**0.777**	**0.794**
	3	0.709	0.719	0.778	0.759	**0.794**
	4	0.740	0.654	0.772	0.763	0.775
	5	0.675	0.641	0.711	0.713	0.681
	6	0.671	0.644	0.662	0.638	0.684
	7	0.663	0.581	0.653	0.621	0.623
Full grid week	2	**0.921**	**0.920**	**0.915**	**0.913**	**0.933**
	3	0.897	0.876	0.895	0.890	0.878
	4	0.823	0.766	0.781	0.822	0.782
	5	0.751	0.672	0.747	0.744	0.749
	6	0.732	0.654	0.747	0.740	0.708
	7	0.717	0.617	0.706	0.708	0.706
Full grid day	2	**0.924**	**0.930**	**0.916**	**0.916**	**0.934**
	3	0.903	0.885	0.903	0.896	0.892
	4	0.820	0.800	0.794	0.822	0.853
	5	0.769	0.737	0.765	0.765	0.760
	6	0.742	0.684	0.768	0.767	0.741
	7	0.738	0.677	0.736	0.740	0.732

5. Conclusions and Future Research Lines

In this paper, we have presented a new supervised approach for meaningful place identification using mobility data provided by the CDRs stored by the operator Telecom Italia of the Milan mobile phone network as an alternative to classical unsupervised clustering techniques. In contrast to cluster analysis, our proposed method does not group or classify mobility data based on their general characteristics, regardless of whether the characteristics make sense for the purpose of classification. Instead, SAMPLID, our proposed method, learns a classification function based on prior knowledge provided to the classifier. The application of this method for identifying home and work places allowed us to evaluate the performance of the new proposed approach in classifying random cells, cells grouped in a grid, or a combination of both. Additionally, our results aided us in determining whether it is more appropriate to use data from a working day or a standard week for the problem of interest and in identifying what type of activity recorded by the CDRs is the most promising.

The best results for our proposed method with kNN as a supervised classifier (SAMPLID-kNN) were obtained by using a 20×20 subgrid with outgoing call data (*callout*) for one working day and $k = 10$. For all types of CDRs, the best results were obtained with the 20×20 subgrid, indicating that the model performs better when provided with information from nearby cells with a close spatial relationship, establishing neighboring relations that define clear transitions between cells of different types. The use of random cells with diverse spatial relationships does not seem to provide any advantage. Regarding the type of CDR used, although the best results were obtained using *callout*, there were no significant differences between the different data types, as all types record a common population pattern to a greater or lesser extent. The value of k also does not seem to have much importance, most likely because of the homogenizing effect of spatial aggregation, which assigns the same values to cells covered by the same RBS and smoothens the differences between cells covered by multiple RBSs. Moreover, although using data from one working day instead of one week does not lead to a significant improvement in the results, this approach allows us to identify the same patterns of population behavior from less information.

As expected, the best results for clustering algorithms were also obtained with *callout* data for one working day based on the 20×20 subgrid and k-medoids, which are more robust to noise and outliers than k-means. However, these results did not exceed those obtained with the model proposed in this paper.

Our results show that the k-medoids clustering algorithms are less sensitive than our model to transitions between different types of cells and have more difficulty in classifying home places far from the city center and work places in the city center.

Although the intrinsic characteristics of CDRs make it difficult to achieve greater precision in the identification of home and work places, this study has demonstrated the benefits of SAMPLID, our proposed supervised approach. Despite the effort required to correctly identify a significant number of patterns that enable any supervised classifier to learn effectively, we have demonstrated the potential of this approach in addressing the complex problem of meaningful place identification.

The large amount of information stored by mobile phone network operators about the activity of mobile phones in CDRs makes them a useful tool for identifying meaningful places frequently visited by individuals. On the other hand, CDRs are limited in location accuracy because they record positions only at the granularity of a radio base station. However, the increase in the density of the RBS network is proportional to the population density it serves, which helps keep the bias within acceptable margins for identifying meaningful places such as home and work.

As a final conclusion, we can affirm that the results obtained in this work using the supervised approach proposed by SAMPLID are sufficiently good to consider its application in any population analysis traditionally addressed using unsupervised clustering techniques, especially when dealing with large amounts of unlabeled data, such as CDRs. In this study, we have demonstrated that labeling a small representative portion of the

data can lead to substantial improvements by enabling the application of supervised learning techniques.

These promising results motivate the beginning of a new line of research aimed at identifying meaningful locations through a supervised approach using knowledge extracted from the mobility data provided by mobile phones. Clearly, the performance of the supervised approach proposed in this work could be improved by using more precise positioning methods, which are not always available in the necessary quantity and breadth.

Among the most interesting future research directions is the study of different sampling methods that allow us to select a set of cells or representative territory areas that provide us with a good knowledge base for training the classifier. Equally interesting would be the study of the characteristics that the used classifier should have and a comparative study of the most promising alternatives.

One of the problems we face when classifying a cell into a single category is that it may contain several classes of places or have shared uses. Thus, we believe that addressing the problem of meaningful place identification from a multi-label perspective is one of the most promising and innovative lines of research, given the intrinsic characteristics of the problem. In this context, novel knowledge selection methods or instance selection [39] and optimization of the used classifier [40] can also be applied to improve the performance.

Author Contributions: Conceptualization, Manuel Mendoza-Hurtado and Domingo Ortiz-Boyer; methodology, Manuel Mendoza-Hurtado and Domingo Ortiz-Boyer; software, Manuel Mendoza-Hurtado, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo; validation, Manuel Mendoza-Hurtado, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo; formal analysis, Manuel Mendoza-Hurtado and Domingo Ortiz-Boyer; investigation, Manuel Mendoza-Hurtado and Domingo Ortiz-Boyer; resources, Manuel Mendoza-Hurtado, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo; data curation, Manuel Mendoza-Hurtado and Domingo Ortiz-Boyer; writing—original draft preparation, Manuel Mendoza-Hurtado and Domingo Ortiz-Boyer; writing—review and editing, Manuel Mendoza-Hurtado, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo; visualization, Manuel Mendoza-Hurtado, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo; supervision, Manuel Mendoza-Hurtado, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo; project administration, Manuel Mendoza-Hurtado, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo; funding acquisition, Domingo Ortiz-Boyer and Juan A. Romero-del-Castillo. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Spanish Ministry of Science and Innovation by grant number PID2022-141869NB-I00 and the University of Cordoba (FPU-UCO-2020) for D. Mendoza-Hurtado.

Data Availability Statement: The data that support the findings of this study are available from BigDataChallenge contest at: https://dataverse.harvard.edu/dataverse/bigdatachallenge accessed on 14 August 2024. Restrictions apply to the availability of these data, which were used under license for this study. Additional data and code used are available at the figshare platform: https://figshare.com/s/073d3cbd472002f33ee7 accessed on 14 August 2024.

Conflicts of Interest: The authors declare no conflicts of interest.The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CDR	Call Detail Record
RBS	Radio Base Station
SMS	Short Messaging Service
kNN	*k*-Nearest Neighbors
SAMPLID	Supervised Approach for Meaningful Place Identification

Notations

T	a dataset
n	the number of instances
x_i	the feature vector of the i-th cell (or instance) in the dataset
y_i	the class label associated with the i-th cell x_i
X	the feature space
Y	the label (or classes) space
c	the number of classes
m	the number of features (or variables)
h	a classifier
x'_i	an unlabeled feature
y'_i	the class label predicted with the i-th cell x'_i
$S_i(t)$	the number of interactions or records in grid square i at time t
t	the time at which the interactions or records are counted
v	the coverage area or region
RBS_c_map	coverage map of the Radio Base Station
$R_v(t)$	the number of records in coverage area v at time t
A_v	the area of the coverage region v
$S'_i(t)$	the scaled number of records in grid square i at time t
κ	a constant to obscure the true number of records
k	the number of neighbors for the k-nearest neighbors algorithm

References

1. Fiandrino, C.; Zhang, C.; Patras, P.; Banchs, A.; Widmer, J. A Machine-Learning-Based Framework for Optimizing the Operation of Future Networks. *IEEE Commun. Mag.* **2020**, *58*, 20–25. [CrossRef]
2. Chen, H.; Song, X.; Xu, C.; Zhang, X. Using Mobile Phone Data to Examine Point-of-Interest Urban Mobility. *J. Urban Technol.* **2020**, *27*, 43–58. [CrossRef]
3. Anniki, P.; Siiri, S.; Rein, A. The Relationship between Social Networks and Spatial Mobility: A Mobile-Phone-Based Study in Estonia. *J. Urban Technol.* **2018**, *25*, 7–25. [CrossRef]
4. Quercia, D.; Di Lorenzo, G.; Calabrese, F.; Ratti, C. Mobile Phones and Outdoor Advertising: Measurable Advertising. *IEEE Pervasive Comput.* **2011**, *10*, 28–36. [CrossRef]
5. Ferrari, L.; Mamei, M.; Colonna, M. Discovering events in the city via mobile network analysis. *J. Ambient Intell. Humaniz. Comput.* **2012**, *5*, 265–277. [CrossRef]
6. Wang, Z.; Yue, Y.; He, B.; Nie, K.; Tu, W.; Du, Q.; Li, Q. A Bayesian spatio-temporal model to analyzing the stability of patterns of population distribution in an urban space using mobile phone data. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 116–134. [CrossRef]
7. Calabrese, F.; Di Lorenzo, G.; Liu, L.; Ratti, C. Estimating Origin Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Comput.* **2011**, *10*, 36–44. [CrossRef]
8. Ahas, R.; Aasa, A.; Mark, Ü.; Pae, T.; Kull, A. Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tour. Manag.* **2007**, *28*, 898–910. [CrossRef]
9. Kang, J.H.; Welbourne, W.; Stewart, B.; Borriello, G. Extracting Places from Traces of Locations. *SIGMOBILE Mob. Comput. Commun. Rev.* **2005**, *9*, 58–68. [CrossRef]
10. Zhuang, C.; Yuan, N.J.; Song, R.; Xie, X.; Ma, Q. Understanding People Lifestyles: Construction of Urban Movement Knowledge Graph from GPS Trajectory. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, Melbourne, Australia, 19–25 August 2017; pp. 3616–3623. [CrossRef]
11. Kim, D.H.; Hightower, J.; Govindan, R.; Estrin, D. Discovering semantically meaningful places from pervasive RF-beacons. In Proceedings of the UbiComp'09, New York, NY, USA, 30 September–3 October 2009; pp. 21–30.
12. Ahas, R.; Silm, S.; Järv, O.; Saluveer, E.; Tiru, M. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *J. Urban Technol.* **2010**, *17*, 3–27. [CrossRef]
13. Frias-Martinez, V.; Virseda, J.; Rubio, A.; Frias-Martinez, E. Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data. In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, New York, NY, USA, 3–16 December 2010; ICTD'10. [CrossRef]
14. Duan, Z.; Liu, L.; Wang, S. MobilePulse: Dynamic profiling of land use pattern and OD matrix estimation from 10 million individual cell phone records in Shanghai. In Proceedings of the 2011 19th International Conference on Geoinformatics, Shanghai, China, 24–26 June 2011; pp. 1–6. [CrossRef]
15. 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 Proceedings of the Pervasive Computing, San Francisco, CA, USA, 12–15 June 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 133–151. [CrossRef]

16. Pintér, G.; Felde, I. Commuting Analysis of the Budapest Metropolitan Area Using Mobile Network Data. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 466. [CrossRef]
17. 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. [CrossRef]
18. Susnea, I.; Dumitriu, L.; Talmaciu, M.; Pecheanu, E.; Munteanu, D. Unobtrusive Monitoring the Daily Activity Routine of Elderly People Living Alone, with Low-Cost Binary Sensors. *Sensors* **2019**, *19*, 2264. [CrossRef]
19. Nurmi, P.; Koolwaaij, J. Identifying meaningful locations. In Proceedings of the 2006 Third Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services, San Jose, CA, USA, 17–21 July 2006. [CrossRef]
20. Singh, S.; Singh Gill, N. Analysis and Study of K-Means Clustering Algorithm. *Int. J. Eng. Res. Technol. (IJERT)* **2013**, *2*, 2546–2551.
21. 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. Its Appl.* **2013**, *392*, 1459–1473. [CrossRef]
22. Raja, M.; Exler, A.; Hemminki, S.; Konomi, S.; Sigg, S.; Inoue, S. Towards pervasive geospatial affect perception. *GeoInformatica* **2018**, *22*, 1–27. [CrossRef]
23. Kumar, N.; Kumar, H. A fuzzy clustering technique for enhancing the convergence performance by using improved Fuzzy c-means and Particle Swarm Optimization algorithms. *Data Knowl. Eng.* **2022**, *140*, 102050. [CrossRef]
24. Park, H.S.; Jun, C.H. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **2009**, *36*, 3336–3341. [CrossRef]
25. Kononenko, I.; Kukar, M. Chapter 12—Cluster Analysis. In *Machine Learning and Data Mining*; Kononenko, I., Kukar, M., Eds.; Woodhead Publishing: Sawston, UK, 2007; pp. 321–358. [CrossRef]
26. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
27. Barlacchi, G.; De Nadai, M.; Larcher, R.; Casella, A.; Chitic, C.; Torrisi, G.; Antonelli, F.; Vespignani, A.; Pentland, A.; Lepri, B.; et al. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci. Data* **2015**, *2*, 1–15. [CrossRef]
28. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
29. García-Pedrajas, N.; Ortiz-Boyer, D. Boosting k-nearest neighbor classifier by means of input space projection. *Expert Syst. Appl.* **2009**, *36*, 10570–10582. [CrossRef]
30. García-Pedrajas, N.; del Castillo, J.A.R.; Ortiz-Boyer, D. A cooperative coevolutionary algorithm for instance selection for instance-based learning. *Mach. Learn.* **2009**, *78*, 381–420. [CrossRef]
31. Blondel, V.D.; Esch, M.; Chan, C.; Clerot, F.; Deville, P.; Huens, E.; Morlot, F.; Smoreda, Z.; Ziemlicki, C. Data for Development: The D4D Challenge on Mobile Phone Data. *arXiv* **2013**, arXiv:1210.0137.
32. de Montjoye, Y.A.; Smoreda, Z.; Trinquart, R.; Ziemlicki, C.; Blondel, V.D. D4D-Senegal: The Second Mobile Phone Data for Development Challenge. *arXiv* **2014**, arXiv:1407.4885.
33. Juszczak, P.; Tax, D.M.J.; Duin, R.P.W. Feature Scaling in Support Vector Data Description 2002. Available online: http://rduin.nl/papers/asci_02_occ.pdf (accessed on 14 August 2024).
34. Butler, H.; Daly, M.; Doyle, A.; Gillies, S.; Schaub, T.; Schaub, T. *The GeoJSON Format*; Request for Comments 7946; RFC Editor: Wilmington, DE, USA, 2016. [CrossRef]
35. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
36. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley: Hoboken, NJ, USA, 1990.
37. Santafé, G.; Inza, I.; Lozano, J. Dealing with the evaluation of supervised classification algorithms. *Artif. Intell. Rev.* **2015**, *44*, 467–508. [CrossRef]
38. Sawilowsky, S. New Effect Size Rules of Thumb. *J. Mod. Appl. Stat. Methods* **2009**, *8*, 597–599. [CrossRef]
39. del Castillo, J.A.R.; Ortiz-Boyer, D.; García-Pedrajas, N. Instance selection for multi-label learning based on a scalable evolutionary algorithm. In Proceedings of the 2021 International Conference on Data Mining Workshops (ICDMW), Virtual Conference, 7–10 December 2021; pp. 843–851. [CrossRef]
40. del Castillo, J.A.R.; Mendoza-Hurtado, M.; Ortiz-Boyer, D.; García-Pedrajas, N. Local-based k values for multi-label k-nearest neighbors rule. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105487. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.