Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 8710402, 12 pages
https://doi.org/10.1155/2018/8710402

Research Article

Supervised Land Use Inference from Mobility Patterns

Noelia Caceres [1] and Francisco G. Benitez [2]

[1] *Transportation Engineering Unit, AICIA, Camino de los Descubrimientos, s/n, 41092 Seville, Spain*
[2] *Transportation Engineering, Faculty of Engineering, University of Seville, Camino de los Descubrimientos, s/n, 41092 Seville, Spain*

Correspondence should be addressed to Noelia Caceres; ncaceres@etsi.us.es

Received 29 November 2017; Revised 19 March 2018; Accepted 11 April 2018; Published 21 May 2018

Academic Editor: Giulio E. Cantarella

This paper addresses the relationship between land use and mobility patterns. Since each particular zone directly feeds the global mobility once acting as origin of trips and others as destination, both roles are simultaneously used for predicting land uses. Specifically this investigation uses mobility data derived from mobile phones, a technology that emerges as a useful, quick data source on people's daily mobility, collected during two weeks over the urban area of Malaga (Spain). This allows exploring the relevance of integrating weekday-weekend trip information to better determine the category of land use. First, this work classifies patterns on trips originated and terminated in each zone into groups by means of a clustering approach. Based on identifiable relationships between activity and times when travel peaks appear, a preliminary categorization of uses is provided. Then, both grouping results are used as input variables in a K-nearest neighbors (KNN) classification model to determine the exact land use. The KNN method assumes that the category of an object must be similar to the category of the closest neighbors. After training the models, the findings reveal that this approach provides a precise land use categorization, yielding the best accuracy results for the major categories of land uses in the studied area. Moreover, as a result, the weekend data certainly contributes to finding more precise land uses as those obtained by just weekday data. In particular, the percentage of correctly predicted categories using both weekday and weekend is around 80%, while just weekday data reach 67%. The comparison with actual land uses also demonstrates that this approach is able to provide useful information, identifying zones with a specific clear dominant use (residential, industrial, and commercial), as well as multiactivity zones (mixed). This fact is especially useful in the context of urban environments where multiple activities coexist.

1. Introduction

Generally, people perform different activities throughout the day over a region. Many of these activities are repeated on daily basis, producing recognizable patterns in time. Mobility is also closely linked with the structure of cities, which have to serve a variety of human needs (housing, working, shopping, leisure, and other activities). This land use planning affects travel behavior (e.g., dense and mixed-use environments tend to produce short trip lengths); hence land use and mobility are indirectly related. Besides, most weekly trips are associated with home-work commuting, primarily concentrated on weekdays (Monday–Friday). Nevertheless, weekend traffic has increased over time and, in some areas, it is viewed as equaling or even surpassing weekday traffic. Moreover, weekend travel behavior is expected to be substantially dissimilar from weekday due to differences in spatial and temporal constraints. With this in mind, policy makers and planners should also consider weekend trip information for defining zoning policies and strategies. One of the main difficulties found to address this issue is that travel data used in such studies come from surveys in which people are asked to describe their travel behavior on an average day. Since surveys are costly and time consuming, besides the abovementioned major concentration of home-work commuting trips, many travel studies only collect information about weekday behavior and ignore weekend days. Nowadays, new technologies offer effective options for collecting trip data in an efficient and quick way. In particular, the pervasive use of mobile phones has made this technology emerge as a promising alternative in travel behavior studies [1–4]; and this idea has inspired intensive research to conduct the analysis of weekday and weekend mobility [1, 5–7]. Results have demonstrated that this technology provides information about mobility patterns with larger sample size, higher update frequency, wider coverage, and more reduced cost and time for data collection. Although it is not exempt from shortcomings due to its temporal and spatial resolution [8, 9], mobile phone data can be regarded as a reasonable source for synthesizing mobility over a region. This paper aims at exploring and understanding travel patterns using trips derived from mobile technology, with the final purpose of determining prospective land uses. Knowing more about the relationship between land use and mobility patterns can help planners improve mobility prediction models or redefine zoning regulations. Hence several works have investigated the way phone data can also reveal details on land use by clustering techniques [10–13], but also by eigendecomposition [14], kernel density estimation [15], relational Markov networks [7], or even supervised-learning techniques [16]. This study extends the effort in examining the functional relationship between land uses and mobility patterns but, unlike other works in the literature, it is based on combining the information derived from patterns of trips originated and terminated at a given zone. In people mobility, the choice of making trips is affected not only by the role of the attracting zone but also by the generating one. For instance, work-related destinations have a very substantial impact on attracting trips, regardless of the typology of the trip origin. Hence, both types of patterns are jointly explored in this work for better determining of actual land use. In addition to proposing an alternative for automatically detecting patterns by origin or destination zone, this work makes other contributions. Most of the works in the literature provide a categorization ranging from two different types (e.g., residential or nonresidential) to a wide variety of categories (e.g., residential, commercial, industrial, and parks). It is reasonable to assume that patterns in residential areas differ from those in business areas. However, since the exclusive use of this variation may raise questions when areas are of the same use but with different intensity of activities, this work allows discerning zones under this circumstance. Moreover, the societal dynamics of cities is not often easy to be geographically separated; for instance, European cities are traditionally compact, with a dense historical core where different functions of use occur (e.g., residential buildings with commercial and civic uses on ground floor). As a consequence, different mobility patterns can coexist in areas with multiple activities. Apart from the common types for residential and business uses, this study also deals with the detection of that kind of multiactivity zones. These findings are demonstrated by comparing results with actual land uses that supplements the zoning system over a real area of study.

The paper is organized as follows: Section 2 describes the studied area and the data source, as well as the method used for automatically detecting meaningful groups in originating and terminating patterns and then for inferring land uses. Section 3 explores the discovered patterns and presents the results, highlighting the significance of using weekday-weekend data for better land use discovering. Finally, conclusions and future work are given.

2. Data and Methods

2.1. Data. The studied area is the city of Malaga, located in the South of Spain on the Mediterranean coast, with a population of 570,000 inhabitants; but its whole metropolitan area has about one million inhabitants including surrounding municipalities. The studied city is divided into several geographic areas known as traffic zones, defining the zoning system. A traffic zone is the unit of geography most commonly used in conventional transportation planning models to divide the planning region into areas of relatively homogeneous land use and demographic characteristics. These zones follow census geography boundaries, consisting of one or more census blocks, block groups, or census tracts. According to the zoning system, the city is divided into 128 traffic zones, for which their actual land uses are known. In particular, these categories are mainly (34.4%) residential-RES and (35.1%) mixed-MIX uses, which are primarily residential areas blending a mix of compatible activities necessary to meet the needs of the population (including services, education, offices, and commercial activities). The other types are 9.4% industrial-IND (e.g., light to heavy industrial facilities and industrial parks, limited commercial and office uses or even certain business such as the airport or the central railway station); 11.7% commercial-COM (e.g., office buildings, shopping centers, and retail establishments); and 9.4% institutional-INS (e.g., public/semipublic uses like educational, health, and community services). These land uses will be used as ground truth data to train and test the models presented in Section 3.

Given that the study pursues identifying the specific land use that represents the activity in these traffic zones, the travel data to be used have to be expressed according to the zoning system. These data can be derived from traditional methods (e.g., surveys or census) or from innovative approaches such as mobile technology, which presents the advantage of capturing weekday and weekend data under equal conditions. This work focuses on this last group of data, based on aggregated and anonymized mobile traces collected and processed by an operator with the largest market share (around 40%) in the studied area. These traces have been derived from events generated by active interactions (e.g., when users make/receive calls or text messages), as well as additional events occurring in the background (idle status), without user's participation. These passive interactions are related to signaling: losing/regaining mobile signal, "alive" records when phones are on but not having created any other events for a sustained period of time (in the order of a few hours), and records notifying the entry in a new location area, defined as a group of adjacent cells. This allows increasing substantially the size of the data source. Each of these events is characterized by an encrypted user id, a timestamp when the event occurs, and a location referred to as a traffic zone. This is estimated by the operator using proprietary (undisclosed) algorithms based on the triangulation of cell tower signals. These events provide "footprints" regarding where people have been and when they were there, useful to extract trips. In this study, a trip is regarded as a one-way movement from a zone of origin to a zone of destination at a particular starting time. Since users are more likely to engage in an activity after a "stay," the first step is to identify which footprints are "stays." The trips occur between these "stays" locations (origin and end of trip); the rest are "passing" footprints created during user's movement. To identify such "stays" several works have developed different algorithms [1, 4, 7–9]. In this work, the identification is based on a time threshold in the subsequence of events. This threshold between consecutive events has been taken as a simple rule-of-thumb for identifying whether an event belongs to a possible new trip ($t_{\text{consecutive\_events}} \geq 30$ min) or to the same trip after a brief stop ($t_{\text{consecutive\_events}} < 30$ min). An event defines the end of a trip when the time difference with the next event is more than 30 minutes as long as they are distant enough in space (avoiding ping-pong effects which insert fake movements in the trajectory of users) to make a trip on the associated route. This end defines the beginning of the "stay" but also the origin for the next subsequence of events. Apart from checking whether events are frequently made in the neighboring group of towers to discard ping-pong records, the analysis of proximity of events in space and time with regard to the characteristics of transport network topology (route distance or time, resp.) also plays a key role in the "stay" detection procedure. For instance, the associated distance between two consecutive events and the difference between their timestamps have to be checked to ensure that they are compatible with the travel distance and travel time, respectively, for the possible routes. Trips are inferred once all existing events created by the sample of users are processed. The results are then expanded to account for the difference between this sample and the population in the studied area; this is done by determining users' home based on footprints from events generated at late night on weekdays (when people usually stay at home). Finally, trips are expanded based on census data taking into account the area where home is located. Trips are also hourly aggregated; that is, a trip is assigned to each hour period based on its starting time. Therefore, in order to comply with privacy regulations, trips are anonymized, aggregated, and expanded; so that it is not possible to associate data with users. Nevertheless, special attention must be paid to some concerns of this technology. The trip extraction is strongly subject to the sparsity of events: the more events are generated, the more footprints are available to infer the trip. Moreover, location estimation using mobile technology does not provide information about the exact position of users but general regions related to the service area of cell towers. This area varies depending on network granularity, from hundreds of meters (in urban areas) to tens of kilometers (in the countryside). Here, an extra spatial error is added since trips have to be expressed according to the zoning system. In this study, focusing on a dense urban area, the spatial resolution is claimed to be of the order of a few hundred of meters, varying from 200 to 300 meters depending on the density of cell towers. The event sparsity is enough to properly infer trips due to the use of active and passive events; in fact "alive" records provide periodic events when any others are generated for a sustained period of time (four hours in the context of this research work). Aware of these issues, the data source includes roughly 200,000 users generating more than 10 million of trips in the urban agglomeration of Malaga for two consecutive complete weeks in February 2015. Then, two separated datasets for the average weekday and weekend-day are available, containing the number of trips made between any pair of origin and destination (OD) zones every hour of the day.

2.2. Methodology. Everyday life is regarded as a sequence of activities performed by individuals at various places during a day, which shape mobility patterns over a region. Clustering is the process of classifying objects into groups (or clusters) so that the objects in the same group are more similar to each other than objects in other groups. Clustering starts with the choice of objects to be classified, which must synthesize conceptually the problematic studied. This study aims to reveal mobility patterns created by people when they perform different activities throughout the day, bearing in mind that the urban travel behavior is very complex in terms of OD-patterns. Therefore, assuming that the travel activity in each particular zone directly feeds the global mobility over a region (once acting as origin of trip and others as destination), the analysis deals separately with patterns regarding traffic originating and terminating in each zone by hour of day, both for weekdays and weekends. In this sense, it is necessary to clarify that this concept differs from the trip production and attraction scheme based on factors that generate and attract trips. Trip production is usually defined as the home end of a home based trip or the origin of a nonhome based trip, while trip attraction is defined as the nonhome end of a home based trip or the destination of a nonhome based trip. This study manages the total number of trips originated and terminated in each zone at a particular hour period, regardless of the nature of the zone where the trip starts or ends or the purpose of the trip. However, huge differences in the order of magnitude for traffic at different zones (even being of the same land use type) may affect the clustering using absolute units. To overcome this issue, different ways of normalization have been applied in a similar context [13, 16]. In this study, normalization over time is applied to express these variables in relative terms: as a ratio of total daily traffic. Therefore, the hourly distribution (in percentage) of traffic in each zone is used as object, resulting from the 24-element vector of trips originated (or terminated) at a given zone by hour of day normalized by the total daily trips originated (or terminated) in such a zone, particularly:

(i) O_k^{wd}: normalized vector of hourly trips originated in zone k on a weekday

(ii) O_k^{we}: normalized vector of hourly trips originated in zone k on a weekend-day

(iii) D_k^{wd}: normalized vector of hourly trips terminated in zone k on a weekday

(iv) D_k^{we}: normalized vector of hourly trips terminated in zone k on a weekend-day.

Next, the clustering classifies these objects based on their similarity. There are many ways to combine cases into groups; overviews of clustering procedures can be found in the literature [17]. One of them is the hierarchical clustering method, which basically forms groups by clustering cases into larger groups until all cases are members of a single group. The criteria for deciding groups are based on matrix of pairwise distances between the objects to be clustered. Several distance metrics are available to build this matrix with hierarchical clustering (Euclidean, Correlation, Cosine, etc.). To determine which clusters should be merged, the concept of closeness is defined by a specified rule in each of the existing agglomeration methods (Ward, Group-Average, etc.); hence the distance matrix after each merging is computed by a different formula based on the linkage method.

Finally, many clustering algorithms depend on the number of clusters; but this prespecified number is not known for this study. The determination of the optimal number of groups in a dataset is one of the main difficulties in cluster analysis. Although it is common to use criteria depending on the subjective judgement of planners, a variety of indices have been defined in the literature to externally decide the number that fits best a dataset, based on the evaluation of the clustering results. Several of such indices were studied [18], and Calinski and Harabasz's (CH) index was regarded as the most effective one in identifying the optimal number. The CH index evaluates the cluster validity based on the average between- and within-cluster sum of squares [19]. This CH index involves looking at the sum of squared distances within the partitions (well-defined clusters have a large between-cluster variance and a small within-cluster variance), but also taking account the number of clusters and number of observations. The larger the CH(k), the better the data clustering. The value of k, which strictly maximizes CH(k), is regarded as specifying the optimal number of clusters; here the evaluated values range from 2 to 10. For the calculation of the pairwise distance matrix, this study uses the Euclidean distance. The CH index is based on ANOVA technique; so it makes most sense to use it where the cluster analysis is in terms of Euclidean distances, whatever linkage methods can be used to group the objects. In particular, different linkage methods are explored: group-average, weighted-average, centroid, and Ward. Each of these methods may lead to different clustering results; hence, results have to be compared. In this study, the solution is finally selected by evaluating the "quality" of each clustering result based on the well-known Dunn's index (DI). Dunn's index [20] measures compact and well-separated clusters, where the maximum value represents the right partitioning (partition with the highest separation between clusters and less spread data in between clusters).

Once zones are classified, by their patterns of originating and terminating trips, the procedure for inferring land uses can be launched. For this purpose, this study uses the K Nearest Neighbors (KNN) classification model, one of the most popular and intuitive machine learning algorithms proposed [21]. According to this approach, the n input variables (also known as predictors or features) define an n-dimensional space where cases located near each other are said to be "neighbors." The KNN model classifies objects based on the categories of the most similar cases (the K nearest neighbors). When a new case is presented, its distance from each of the cases in the model is computed. The assignment of a category is based on the predominance of a particular category in this neighborhood. Then, the KNN method assumes that the category of an object must be similar to the category of the closest neighbors. In our problem, the category of land use for a zone should be similar to the category for other members of the resultant group. But, instead of using the resultant groups exclusively by originating trips or by terminating trips, both roles of zones are simultaneously considered. To bear this in mind, the KNN classifier uses the labels of resultant groups, both for originating and terminating trips, as input variables. The actual land use of the zones over a real area of study, extracted from the zoning regulations, is used as the output (response). The rules of the KNN classifier are created by a training set. In this study, the set of 128 zones is randomly split into two subsets, 90 used for training (70%) and the other for testing (30%). Then, the model is trained using the training set. The testing set is completely excluded from the training process and is used for independent assessment of the final models. The number of nearest neighbors to examine (K) is a parameter of the model. Others are the metric to measure the distance between a set of data and query points and the weighting function to determine the weight of the individual "votes" of the k-nearest neighbors. Several metrics can be used to measure the distances between the test data and each of the training data to choose the final classification output. However, the concept of similarity or distance for categorical data is not as simple as for continuous data. Since the input variables are categorical, cosine metric is the most appropriated one. Cosine similarity is a popular measure for text clustering [22, 23], which is the most similar case to label the resultant groups. This methodology is carried out separately for the case of weekday data and merged weekday-weekend data.

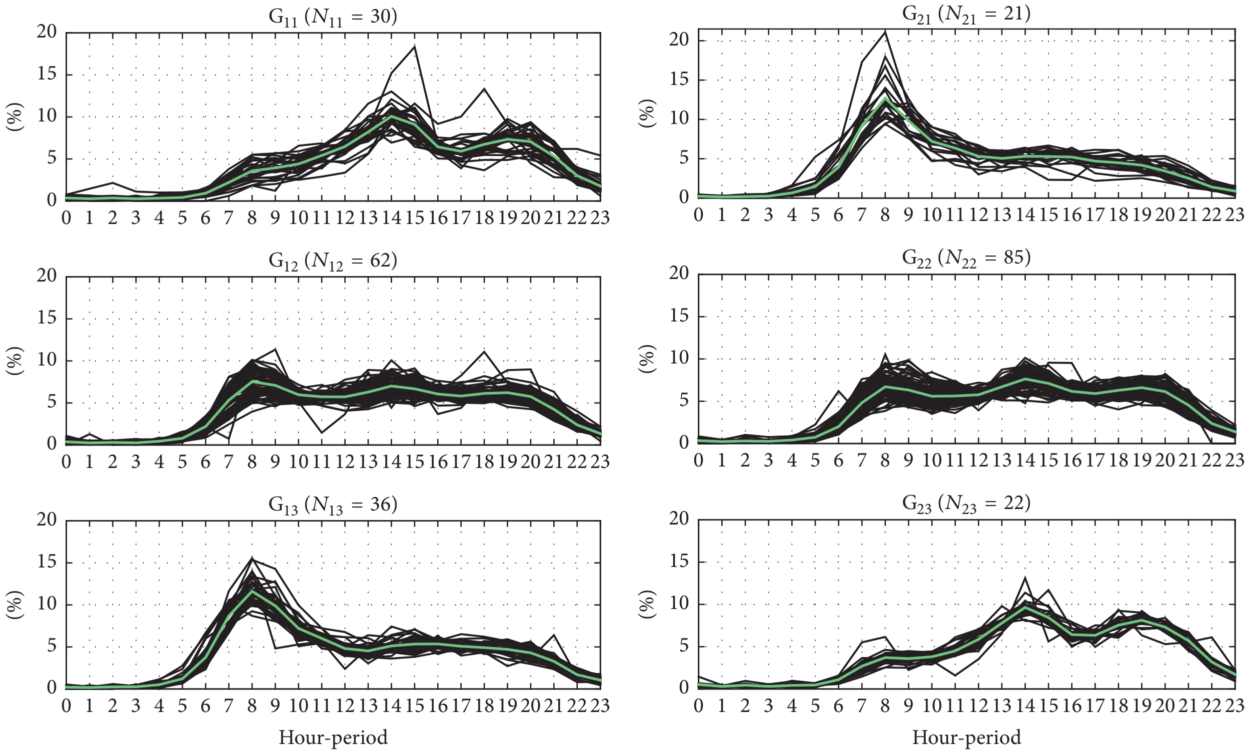
3. Analysis and Findings

3.1. Analysis of the Mobility Patterns. The first step conducted was to detect the possible patterns over the studied area, managing originated and terminated trips in a separate way, in order to identify their dependence with possible activities. The methodology is applied to four set of objects: normalized originated trips on a weekday ($O_k{}^{\text{wd}}$), normalized terminated trips on a weekday ($D_k{}^{\text{wd}}$), normalized originated trips on both weekday and weekend ($O_k{}^{\text{wd}}$, $O_k{}^{\text{we}}$), and normalized terminated trips both on weekday and weekend ($D_k{}^{\text{wd}}$, $D_k{}^{\text{we}}$). For each of these sets, Table 1 shows the number of clusters obtained by applying the approach described before, as well as Dunn's index obtained from each linkage method. Taking into account the fact that larger DI value means better cluster configuration, the clustering solution finally selected is marked with an asterisk. In general, the Ward method tends to produce more compact clusters than other methods, since it is aimed at minimizing the total within-cluster variance. It is remarkable that the number of patterns identified in the set of objects merging weekday and weekend data (Set 3 and 4) is higher than those just using weekday (Set 1 and 2), suggesting that the weekend can help to find more trends in travel.

As a first step, the time evolution of patterns for all cases is analyzed in order to find traceable relationships between activity and times when peaks appear. Focusing on the case of just using weekday data, Figure 1 presents the

Table 1: Number of clusters and Dunn's index (DI) obtained from each linkage method for the cases.

Linkage method	Set 1: (O_k^{wd})		Set 2: (D_k^{wd})		Set 3: (O_k^{wd}, O_k^{we})		Set 4: (D_k^{wd}, D_k^{we})	
	*N*clusters	DI	*N*clusters	DI	*N*clusters	DI	*N*clusters	DI
Average	3	0.1345	3	0.1676	8	0.2327	2	0.1984
Weighted	2	0.1174	2	0.1421	7	0.2365	5*	0.2967
Centroid	3	0.1248	3	0.1957	9	0.2555	4	0.2872
Ward	3*	0.1461	3*	0.1980	7*	0.2587	4	0.2946



(a)

(b)

Figure 1: Weekday patterns by the hour period associated with (a) departure time from a given zone and (b) arrival time at zone ("8" refers to 08:00–08:59).

abovementioned average pattern in each group (thick line in green) for originating trips (a) and terminated trips (b), as well as the profiles of all zones classified in such a group (lines in black) in order to see the grouping variability. For originating trips on a weekday (set 1), three different patterns are exhibited in Figure 1(a) (groups, G_{sn}, are identified by subscripts standing for set membership s, and the group number n resulting from clustering, N_{sn}, stands for the number of zones contributing to the corresponding group). The pattern of G_{11} is characterized by two peaks coinciding at lunch hours (around 2 p.m.) and in early evening (7 p.m. and 8 p.m.), when most people finish work. This behavior is typical for business-related zones (work, education, services, etc.). In Spain, school time is concentrated in half-day, which generates picking-up-children trips and home return trips for lunch. Moreover many people are employed in split shifts or even in part-time jobs, so a nonnegligible percentage returns home early in the afternoon, back to work, and return home in the evening. In the G_{12} pattern, there is not a clear, identifiable relationship between activity and times when peaks appear. In fact, the peaks occur both in the morning (trips starting around 8 a.m.) and during lunch hours (around 2 p.m.) and early evening (between 6 p.m. and 8 p.m.); that is, at times when people usually engage in different activities for different purposes. This suggests a mixture of activities in the same zone, for instance, residential areas blending a mix of activities to meet the needs of the population (including services, education, offices, and commercial activities). In the G_{13} pattern, there is a remarkable peak in the morning (nearly 8 a.m.), when people leave homes to start their business activity. This behavior, reflected in patterns from originating trips, normally occurs in zones dominated by residential buildings. In contrast, the reasoning behind patterns of terminating trips is totally the opposite, because they are based on destination zones of trips. Now the peaks shown in terminating-trip patterns

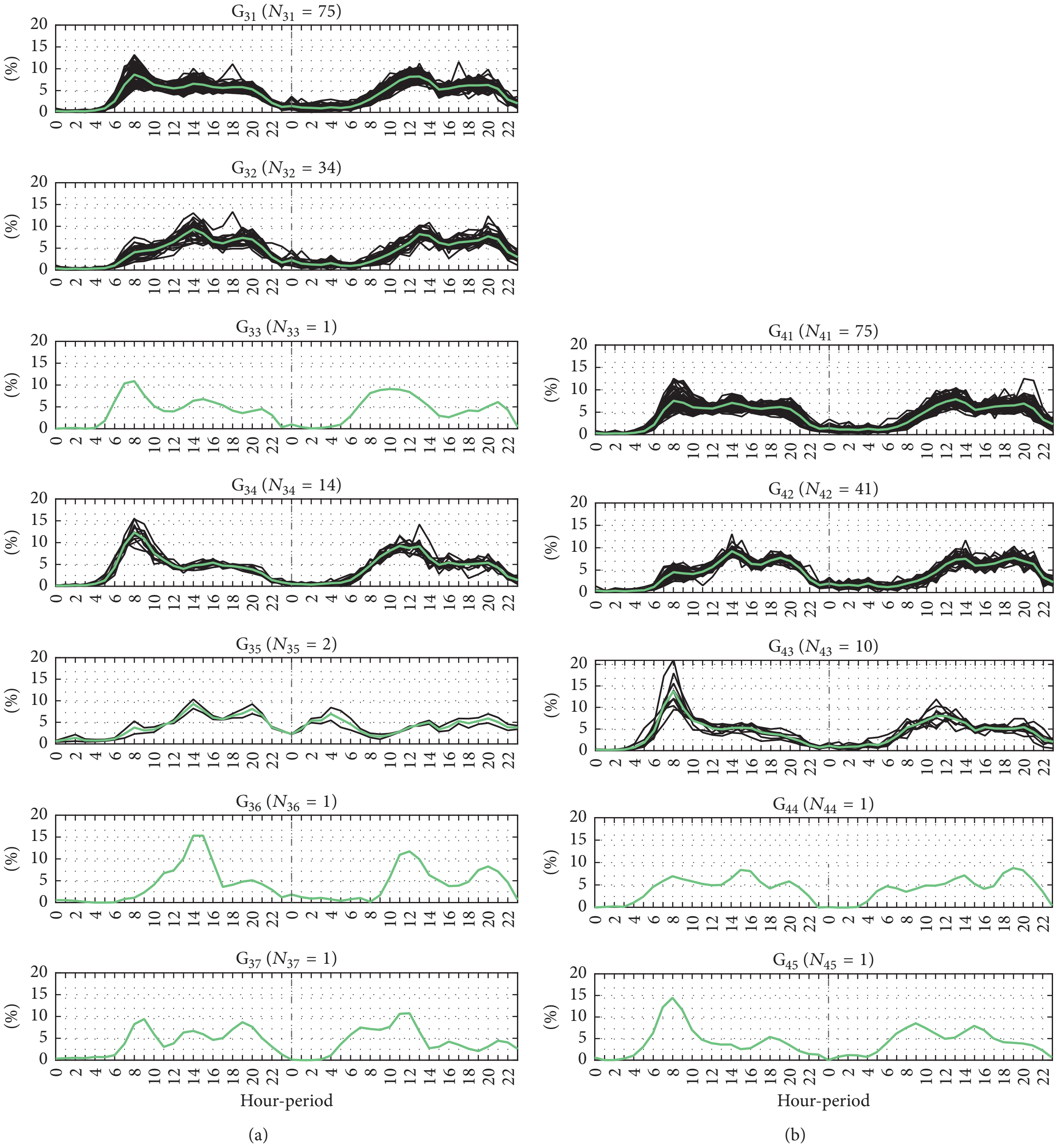


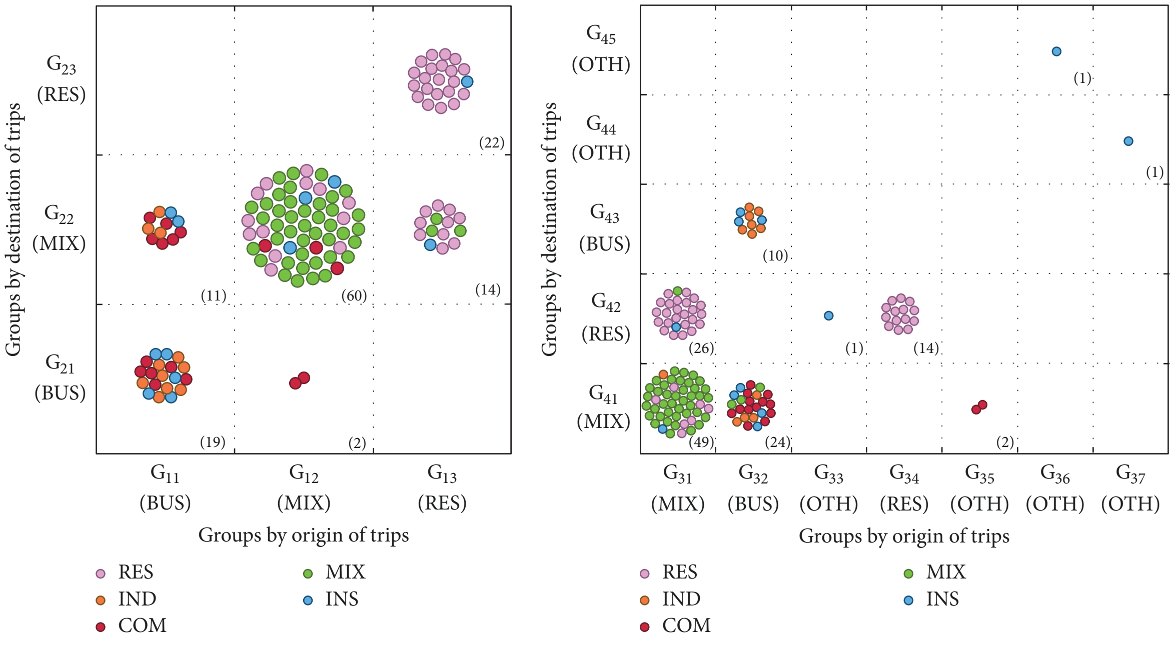
Figure 2: Weekday (left)-weekend (right) patterns by the hour period associated with (a) the departure time from a given zone and (b) the arrival time at a given zone.

(Figure 1(b)) are related to the arrival at a destination. In G$_{21}$ pattern, there is a remarkable peak in the morning related to the start of business hours, when people regularly go to work or to school. This clearly reflects the fact that these zones are more fitted to business areas (e.g., industrial, commercial, and institutional activities). The pattern for G$_{22}$ with three peaks (in the morning, during lunch hours and early evening) suggests areas composed of mixed activities, mainly residential but also industrial, commercial, and institutional areas. The pattern of terminating trips for G$_{23}$ has two remarkable peaks during lunch hours and early evening, most likely related to the hours for work-to-home return trips, clearly revealing residential areas as destination.

In the previous case, three groups have emerged from both originating and terminating sets. But the case of merging weekday and weekend data as a whole vector produces different results: seven and five groups for originating and terminating, respectively. In a similar way, Figure 2 presents the clustering results, with the average pattern in each group (in color thick line) as well as the profiles of all zones classified in such a group (in black), for originating trips (a), and terminating trips (b). Herein, although the number of resultant groups is higher than those using just weekday data, three of them concentrate on the majority of the zones, like in the case of weekday data. Focusing on trips originated by zone (Figure 2(a)), particularly on the groups concentering more zones (G_{31}, G_{32} and G_{34}), the distinctive peaks of their individual patterns are quite similar to the case of weekday data. In particular, G_{31} has three peaks in the morning (around 8 a.m.), during lunch hours (around 2 p.m.), and barely noted in early evening (between 6 p.m. and 8 p.m.). G_{32} has two peaks (one during lunch hours and other in early evening). And G_{34} has a remarkable peak in the morning. With a similar reasoning as the weekday case, zones in G_{31}, G_{32}, and G_{34} can be flagged as mixed, business, and residential related activities. In contrast, groups G_{33}, G_{35}, G_{36}, and G_{37} consist of zones with a particular travel activity (other), especially on the weekend, which makes them classified separately. For instance, G_{35}, like G_{32}, has two peaks as origin of trips on the weekday part (during lunch hours and in early evening), which correspond to the usual returning to home from offices, school, or business areas in general. Even both groups remain having these two peaks on the weekend, for time periods when people start to return home. This suggests that these zones are not only a place of work but also for shopping, services, or restaurant visits since work activity is significantly reduced for a huge number of people on weekends. However, G_{35} has an important travel component during late night hours on the weekend compared to G_{32}; hence they are classified in separate clusters. Looking at the two zones classified in G_{35}, they are in the city center. On workdays, the city center is actively used (as many homes, workplaces, and commercial services are situated there), while during the night no many movements are reported. In consequence, the patterns for G_{32} and G_{35} are quite similar focusing on the weekday part. But the weekend has a different pattern of movement in the city center. In particular, people visit the city center with a relatively high frequency during weekend evenings, and they stay long after midnight. Consequently, G_{35} has a typical nightlife peak of originating trips for people coming back home, which is not appreciated in G_{32} identified for business-related places. For the rest of groups with a particular travel activity (other), it is noted that these patterns belong to special zones: both G_{33} and G_{36} contain hospitals, while G_{37} is the main area of the university campus. Focusing on terminated trips by zone (Figure 2(b)), where peaks are related to the arrival at a destination, three groups also concentrate on the majority of zones (G_{41}, G_{42}, and G_{43}). Only two zones are classified in two separate groups (G_{44} and G_{45}). G_{41} has three peaks (in the morning, during lunch hours and early evening), G_{42} has two peaks (during lunch hours and in early evening), and G_{43} has a remarkable peak in the morning. With a similar reasoning as the weekday case, zones in groups G_{41}, G_{42}, and G_{43} can be flagged as mixed, residential, and business-related areas, respectively. G_{44} and G_{45} are characterized by specific travel activity unlike the other resultant groups, especially during the weekend. In particular, G_{44} contains the major share of the university campus and G_{45} a hospital, both also discerned in the previous case.

3.2. Identifying Land Uses. In previous subsection, a preliminary type of land use according to a general categorization (residential/business/mixed) is assigned to each group of patterns based on the relationship between activity and times when travel peaks appear, both for originating and terminating trips. Apart from these land uses, a different type (other) emerges from merging weekday and weekend data. Intuitively, it seems reasonable to envisage that zones classified into a group by their pattern of originating trips have similar category of land use. The same may be assumed for groups derived from patterns of terminating trips. But this preliminary land use identification based exclusively on originating or terminating trips may lead to confusion. Figure 3 depicts two scatterplots with the resultant groups by origin (represented on x-axis) and those by destination (represented on y-axis), both for weekday (a) and merged weekday-weekend data (b). Each circle represents a zone common to both groups, colored by the actual land use. The exact number is also indicated in brackets in the bottom right corner of the box (e.g., there are 19 zones common to both G_{21} and G_{11}). As it can be appreciated in Figure 3, there are certain zones that remain together by their patterns of originating and terminating trips. However, other zones classified into a group by originating trips do not remain together in the same group by their pattern of terminating trips. Therefore, a more exhaustive aggregation of zones is defined by combining both groups. In particular, it produces six new aggregations of zones using weekday data and nine using weekday and weekend data as a whole vector. In them, the land use associated with a specific aggregation can be more precisely identified.

By exploring the occurrence of actual land uses, represented by colors in Figure 3, some findings are revealed both for the case of weekday data (a) and merging weekday and weekend data (b). In both cases, it is appreciated that a zone having the same land use type both by origin and by destination can be certainly regarded as of this type. For instance, analyzing the occurrence of actual land use, zones flagged as residential both based on its originating trips and terminating trips is mainly composed of residential uses. Something similar occurs for zones flagged as business (e.g., industrial, commercial, and institutional activities) both by origin and by destination. This is, for instance, the case of the aggregation resultant from G_{32} and G_{43} in weekday-weekend data (Figure 3(b)), which comprises seven industrial areas inducing a noticeable effect on working activity in the studied area (e.g., major industrial parks) as well as three remarkable institutional spaces (e.g., some faculties/technical schools at the university). In contrast, zones marked as residential by origin but as mixed by destination suggest many different activities in a neighborhood more oriented to residential uses. Similarly, zones marked as business by origin and as mixed by destination (or vice versa) suggest that although they contain also homes, the main use of those areas deals with business activities. This reinforces the appropriateness of this



(a)

(b)

Figure 3: Aggregation of zones common to both groups derived by originating and terminating trips: (a) weekday data and (b) weekday-weekend data; the number of zones for each case is stated in parenthesis.

preliminary basic categorization, which leads to discerning zones with similar use but different intensity. Moreover, as might be expected, zones flagged as residential by origin are not regarded as business by destination (or vice versa), since they are incompatible uses based on the applied reasoning. However, this identification is not so direct for zones marked as "mixed," involving many different uses, or as "other" for special spaces, only revealed in the case of weekday-weekend data. Then, although this reasoning can discover useful information on land use, it is not enough to find a detailed understanding of the activity performed in the area, according to more disaggregated categorization schemes. In this sense, the inference procedure based on KNN classifier plays a key role to determine the category of land use. The KNN classification is based on the space defined by the input variables (in this case, the groups by originating and terminating trips), in a similar way as the visualization showed in Figure 3. The KNN method assumes that the category of an object must be similar to the category of the closest neighbors. Then, the category of land use for a zone should be similar to the category for other members of the resultant group. This assumption is verified by looking at Figure 3.

The KNN method aims to classify zones whose category is unknown given their respective distances to zones in a learning (training) set whose category is known a priori. In this study, the training set contains 90 zones randomly selected. Then, the model is trained to generate rules for classifying test data (38 zones) into the categories predetermined by the actual land use categorization (i.e., residential, industrial, commercial, mix, and institutional), used as ground truth. With KNN the three main parameters to be considered are the number of neighbors, distance measure, and distance weighting function. As explained in Section 2, cosine metric is used to calculate the distances between the testing set and all of the training data in order to identify its nearest neighbors and produce the classification output. The influence of these K nearest points in such output is specified by the selected distance weighting function: Equal (no weights), Inverse (weight is 1/distance), or Squared Inverse (weight is 1/distance^2). These three functions are evaluated in this study. Then the number of nearest neighbors is the main factor to be decided. It is possible to specify a finer or coarser classifier by deciding on the number of neighbors. Many neighbors can produce high accuracy but can be a very time-consuming process. A way to overcome this fact is by means of the distance weighting function, which makes the classifiers less sensitive to the chosen value of K. To check this issue, the number of neighbors K is evaluated in the range from 6 to 30. Moreover, in order to test the approach which does not depend on the used testing dataset, a Monte Carlo cross-validation has been also implemented [24]. According to this, the process of splitting the data into a calibrating set and a testing set is randomly repeated several times, generating different randomly partitions, always each case always appears in either the calibrating set or the testing set, but not in both. For each partition, the models are fitted using the corresponding calibrating data, and the predictive accuracy is assessed using the testing data. In particular, the accuracy is based on the percentage of correctly predicted class over all predictions. The results are then averaged over the splits; in this case, the partitioning has been repeated 100

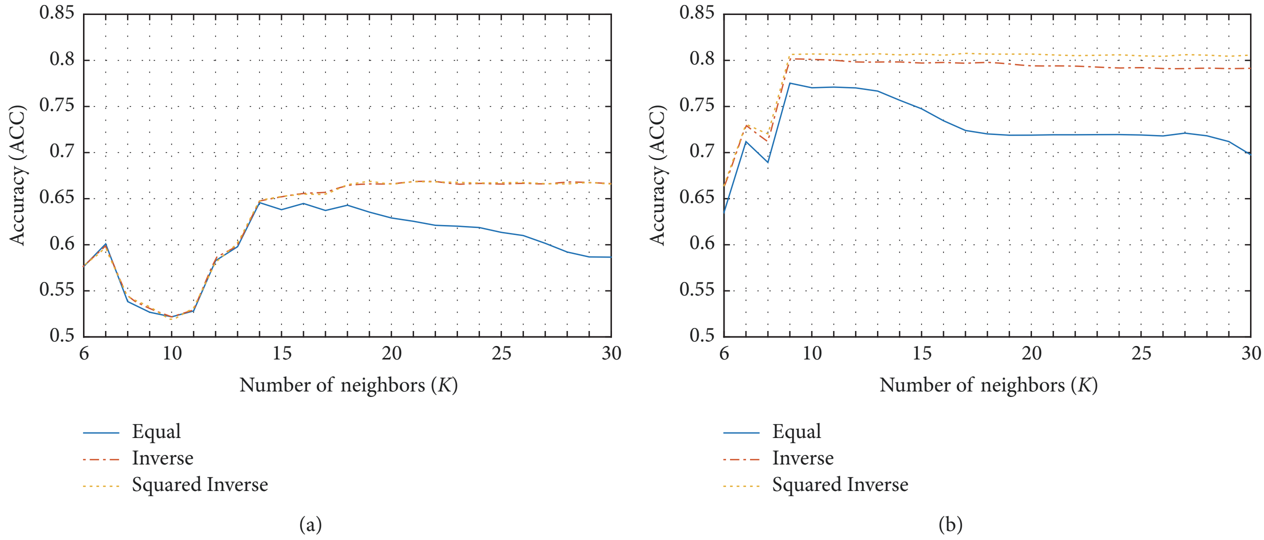


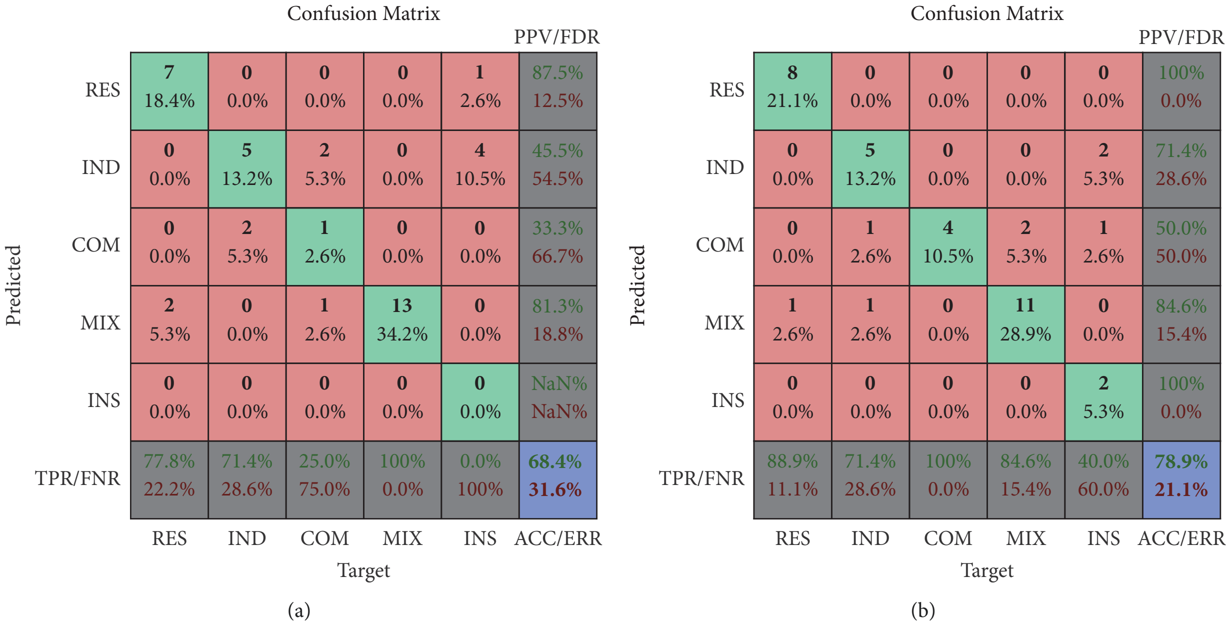
Figure 4: Evolution of accuracy levels as a function of the number of neighbors using distance weighting function Equal, Inverse, and Squared Inverse: (a) weekday data and (b) weekday-weekend data.

times. Figure 4 presents the evolution of the accuracy (ACC) as a function of the number of neighbors (K), for the case of weekday data (a) and merged weekday-weekend data (b).

As it is shown, the accuracy levels obtained after applying the KNN classifiers to merged weekday-weekend data are clearly greater than those for just weekday data. From a specific number of neighbors (e.g., $K = 18$ for just weekday, and $K = 10$ for weekday-weekend data) the ACC remains in the same order of magnitude when "Inverse" or "Squared Inverse" is used as distance weighting function. This is explained by the fact that the use of weights makes the classifiers less sensitive to the value of K. So, the distance weighting function is chosen to be Squared Inverse. Focusing on this case, using merged weekday and weekend data allowed for correct identification of around 80% of the land uses, while just weekday data allowed for reaching around 67%. Therefore, as a result, the weekend certainly contributes to find more precise land uses as those obtained by just weekday data.

Next, the performance of the classifier models for each category is explored in detail. For this purpose, one of the possible partitions of dataset is arbitrarily taken, and the trained models are applied to such test zones, both for weekday and merged for weekday-weekend data. The parameters for distance measure and distance weighting function are, respectively, "Cosine" and "Squared Inverse"; the number of neighbors are $K = 18$ for just weekday and $K = 10$ for weekday-weekend data. In this sense, the classification accuracy has been evaluated by the confusion matrix or error matrix [25]. This matrix counts the test zones correctly and incorrectly predicted by the classification models for the case of weekday data (Figure 5(a)) and merged weekday-weekend data (Figure 5(b)). The columns in the matrix correspond to the actual category of the data (target). The rows correspond to the predictions made by the model (predicted). Thus, the diagonal elements show the number of correct classifications made for each category, and the off-diagonal elements show the errors made by each model. Each cell also indicates the percentage of the total test size. Other measures derived from this matrix are the recall or *true positive rate* (TPR), which is the proportion of positive cases that were correctly identified, and the *precision* or *positive predictive value* (PPV), which is the proportion of the predicted positive cases that were correct. These metrics are depicted in green in the last row and column, as well as the respective complements (in red): *error rate* (ERR), *false negative rate* (FNR), and *false discovery rate* (FDR). For an ideally performing model TPR and PPV rates would be 100%.

As expected, the accuracy for the selected test set is around 68% in the case using just weekday data, while the accuracy using the classifier for the case of merged weekday and weekday data is clearly higher (79%). In general, the results for each of the categories yield high PPV/TPR rates. However, the confusion matrices also reveal that the worst results in classification in terms of PPV/TPR occur for institutional class (INS) for both cases. For just weekday, none of the possible zones has been correctly classified; in fact, the model is not able to classify a zone in such a category. In the case of using merged weekday and weekend data, the results are better but remain in a reduced rate of correct predictions. This may be explained by the nature of this category (INS), which comprises public/semipublic services to serve the needs of a population (e.g., educational, health, cultural, or other community-oriented uses). Some of them, like schools, hospital/healthcare facilities, or parks, are usually compatible with a residential environment situated within a neighborhood, but others are well-matched with working areas (e.g., university campus). In this last case, for instance, the times when travel peaks appear in zones covering the university campus perfectly match with a working zone like an industrial park. As it was appreciated in Section 3.1, this can be discerned with the use of



Confusion Matrix

Predicted	RES	IND	COM	MIX	INS	PPV/FDR
RES	7 18.4%	0 0.0%	0 0.0%	0 0.0%	1 2.6%	87.5% 12.5%
IND	0 0.0%	5 13.2%	2 5.3%	0 0.0%	4 10.5%	45.5% 54.5%
COM	0 0.0%	2 5.3%	1 2.6%	0 0.0%	0 0.0%	33.3% 66.7%
MIX	2 5.3%	0 0.0%	1 2.6%	13 34.2%	0 0.0%	81.3% 18.8%
INS	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
TPR/FNR	77.8% 22.2%	71.4% 28.6%	25.0% 75.0%	100% 0.0%	0.0% 100%	68.4% 31.6%
	RES	IND	COM	MIX	INS	ACC/ERR

Target

(a)

Confusion Matrix

Predicted	RES	IND	COM	MIX	INS	PPV/FDR
RES	8 21.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
IND	0 0.0%	5 13.2%	0 0.0%	0 0.0%	2 5.3%	71.4% 28.6%
COM	0 0.0%	1 2.6%	4 10.5%	2 5.3%	1 2.6%	50.0% 50.0%
MIX	1 2.6%	1 2.6%	0 0.0%	11 28.9%	0 0.0%	84.6% 15.4%
INS	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 5.3%	100% 0.0%
TPR/FNR	88.9% 11.1%	71.4% 28.6%	100% 0.0%	84.6% 15.4%	40.0% 60.0%	78.9% 21.1%
	RES	IND	COM	MIX	INS	ACC/ERR

Target

(b)

Figure 5: Confusion matrix for KNN classifiers using (a) weekday and (b) weekday-weekend testing data.

weekend data (where a set of zones with a particular travel activity was discovered), but not in the case of just weekday, and hence the poor results obtained for the corresponding classifier. Nonetheless, the ACC and PPV/TPR rates for the category INS can be improved by incorporating other zonal features in the inference procedure (e.g., population size). But, unlike traditional techniques that use variables of a physical/biological and socioeconomic nature to derive land uses, the proposed approach only focuses on mobility information. Despite this, the accuracy of this approach is quite satisfactory for the rest of categories, taking into account the fact that it is only based on patterns of trips originated and terminated in zones. The best rates for PPV/TPR are achieved for the major categories of land uses in the studied area: RES and MIX. This can be explained by the fact that, for both weekday data and merged weekday-weekend data, the results from clustering already allow us to distinguish these categories from the rest, since the corresponding aggregations of zones (Figure 3) are mainly composed of zones flagged by such categories. This is of paramount importance in urban environments, where premises are often used in addition to work and home (e.g., buildings with residential units above and commercial units on the ground floor). For the industrial category (IND), similar rates for TPR are obtained using weekday data and merged weekday-weekend data. However, a low PPV is reached using just weekday data (PPV = 45.5%), predicting in such category zones flagged not only as IND but also as INS. The reason of that is because both of them have similar patterns on weekday data in terms of travel peaks (the working hours are quite similar for work and study) and the clustering stage is not able to separate them. In contrast, these categories are better distinguished using weekday-weekend data, because the weekend introduces a distinct travel behavior to make them be better classified in a different group, getting a higher PPV (71.4%). Something similar occurs in the commercial category (COM) using just weekday data, for which the approach wrongly classifies zones in the target category (TPR = 25%). However, a substantial increment is appreciated using weekday-weekend data (TPR = 100%). Like the previous case, the difference in travel peaks associated with commercial activities on weekend patterns makes these zones well-separated from the clustering stage, getting better rates for PPV/TPV in the classification model. Therefore, the weekend data certainly contributes to better determine the category of land use for zones. In this sense, the approach presented in this work leads to discerning activities differentiating, for instance, from residential to mixed uses, or from industrial facilities to commercial areas.

4. Conclusion

Exploring mobility patterns generated by sequences of activities performed by individuals during a day can help planners identify how a particular zone is being used, with the final purpose of detecting land uses. With this in mind, this investigation uses mobility data derived from mobile events collected during two weeks over the urban area of Malaga (Spain). Using this source of information, this work proposes a clustering approach to automatically infer and classify patterns on trips originated and terminated in each zone in a separate way, founded on the idea that the choice of making trips is affected not only by the role of the origin zone but also by the destination. Based on the relationship between activity and times when travel peaks appear in patterns, a preliminary type of use according to a general categorization (residential/business/mixed) is assigned to each group of patterns, both for originating and for terminating trips. This reasoning leads to discerning zones with similar use but different intensity. Then, a KNN classification model is defined to determine the exact land use for zones, using the labels of resultant groups, for both originating and terminating trips, as input variables. It assumes that the category of land use for a zone should be similar to the category for other members of the resultant group. After training the models, the results are compared with the actual land uses of a testing set, demonstrating the potential for providing useful information on land use, yielding the best accuracy results for the major categories of land uses in the studied area. The findings reveal the relevance of integrating weekend trip information to better determine the category of land use for zones. In particular, the percentage of correctly predicted land uses using both weekday and weekend is around 80%, while just weekday data reach around 67%. Moreover, the proposed approach leads to identifying not only zones with a primary use (residential, industrial, or commercial) but also multiactivity zones (mixed). This is really appreciated for cities serving a mixture of human needs (housing, working, shopping, and other activities) in the same neighborhood. This is valuable not only for better understanding of the relationship between land use and mobility but also for determining more fitted land uses that supplement zoning regulations.

Other possible uses of the proposed approach focus on the detection of possible short-term changes on land uses motivated by nonroutine activities (e.g., itinerary exhibition, public concert). To identify land use changes, these types of nonrecurrent activities have to imply enough impact on travel behavior to make changes on the hourly patterns of trips originated and terminated in the corresponding zones. Thus, first of all, it is necessary to identify in which group the new pattern fits better by computing the "similarity" (using a quantifying criteria based, e.g., on a measure such as the Euclidean distance) between such a pattern and the average pattern in each group (thick line in green in Figures 1 and 2). These calculations have to be done for both originating and terminating trips in the corresponding zone. Once the groups are identified, the labels of resultant groups could be used as input variables in the trained KNN models. These models are trained based on the regular land uses of zones, so that they can be suitable to detect possible short-term changes impacting on mobility patterns. In case the groups of the new patterns remain the same, the labels used as input in the KNN models will be same as the previous condition; therefore, the same land use will be predicted. On the other hand, in case they change, a new land use will be predicted as a result. This kind of short-term changes is hardly detected by variables based on annual averages (e.g., socioeconomic data), generally used in traditional techniques for land use inference. Hence, the proposed approach (focused on just mobility patterns) deserves further research to detect short-term changes on land use. As a further research line, the inclusion of zonal features (such as population, number of employees, number of education centers, or housing balance) should be also investigated in order to improve the accuracy of the classifiers, especially for institutional uses. Further research based on this approach can contribute to a more detailed understanding of the activity performed in mixed areas, for instance, in order to explore the primary use in mixed zones (e.g., more oriented to business or residential uses).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the ERDF of European Union for its financial support via Project DIURMOVIL of the "Programa Operativo FEDER de Andalucia 2011–2015." They also thank the Public Works Agency and Regional Ministry of the Public Works and Housing of the Regional Government of Andalusia. Finally, Noelia Caceres acknowledges the funding provided by the Spanish Ministry of Economy and Competitiveness through the Torres Quevedo Programme (PTQ-13-06428).

References

[1] S. Çolak, L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González, "Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities," *Transportation Research Record*, vol. 2526, pp. 126–135, 2015.

[2] M. B. Rojas, E. Sadeghvaziri, and X. Jin, "Comprehensive review of travel behavior and mobility pattern studies that used mobile phone data," *Transportation Research Record*, vol. 2563, pp. 71–79, 2016.

[3] T. Wang, C. Chen, and J. Ma, "Mobile phone data as an alternative data source for travel behavior studies," in *Transportation Research Board 93rd Annual Meeting*, Transportation Research Board 93rd Annual Meeting, Washington, D.C, USA, 2014.

[4] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr., E. Frazzoli, and M. C. González, "A review of urban computing for mobile phone traces: Current methods, challenges and opportunities," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp 2013 - Held in Conjunction with the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013*, usa, August 2013.

[5] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 36–44, 2011.

[6] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: a mobile phone trace example," *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 301–313, 2013.

[7] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, "Discovering urban activity patterns in cell phone data," *Transportation*, vol. 42, no. 4, pp. 597–623, 2015.

[8] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin-destination trips by purpose and time of day inferred from mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 240–250, 2015.

[9] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, "The path most traveled: Travel demand estimation using big data resources," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 162–177, 2015.

[10] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Clustering daily patterns of human activities in the city," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 478–510, 2012.

[11] 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 *Proceedings of the 2011 19th International Conference on Geoinformatics, Geoinformatics 2011*, chn, June 2011.

[12] 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*, vol. 28, no. 9, pp. 1988–2007, 2014.

[13] M. G. Demissie, G. Correia, and C. Bento, "Analysis of the pattern and intensity of urban activities through aggregate cellphone usage," *Transportmetrica A: Transport Science*, vol. 11, no. 6, pp. 502–524, 2015.

[14] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: Segmenting space through digital signatures," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 78–84, 2010.

[15] X. Cheng and W. Li, "Analyzing human activity patterns using cellular phone data: a case study of Jinhe newtown in Shanghai, China," in *Proceedings of the Transportation Research Board 92rd Annual Meeting*, Washington, DC, USA, 2013.

[16] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proceedings of the the ACM SIGKDD International Workshop*, p. 1, Beijing, China, August 2012.

[17] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 16, no. 3, pp. 645–678, 2005.

[18] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[19] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, pp. 1–27, 1974.

[20] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[21] E. Fix and J. L. Hodges, *Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties*, Hodges. Discriminatory Analysis - Nonparametric Discrimination, USAF School of Aviation Medicine, 1951.

[22] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2011.

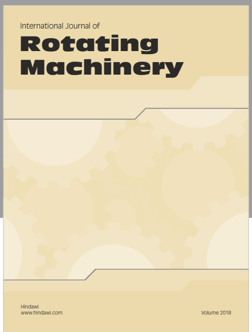
[23] S. Anitha Elavarasi and J. Akilandeswari, "Survey on Clustering Algorithm and Similarity Measure for Categorical Data," *ICTACT Journal on Soft Computing*, vol. 4, no. 2, pp. 715–722, 2014.

[24] J. Shao, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

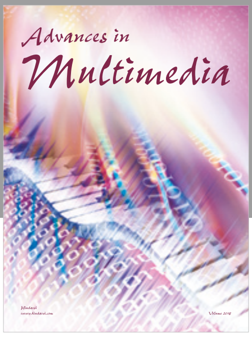
[25] R. Kohavi and F. Provost, "Special issue on applications of machine learning and the knowledge discovery process," *Mach. Learn*, vol. 30, no. 2, pp. 271–274, 1998.

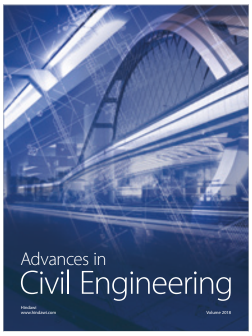
Journal of
Engineering

The Scientific
World Journal

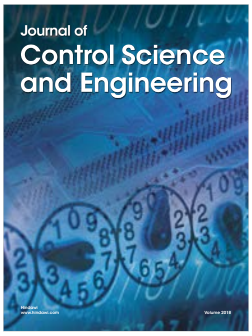

International Journal of
Rotating
Machinery

Journal of
Sensors

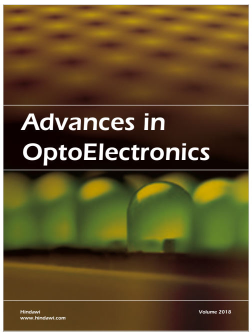

Advances in
Multimedia


Advances in
Civil Engineering

Journal of
Robotics

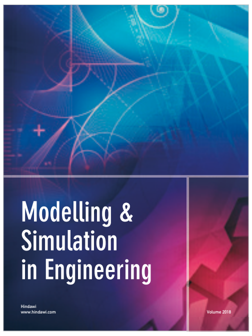

Journal of
Control Science
and Engineering

Journal of
Electrical and Computer
Engineering

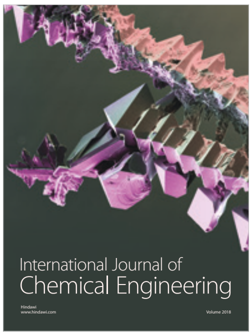

Advances in
OptoElectronics

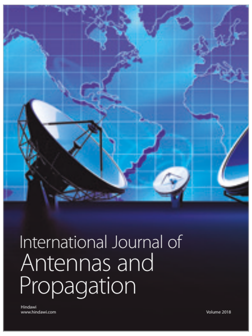
VLSI Design

International Journal of
Navigation and
Observation

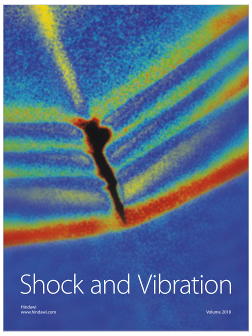

Modelling &
Simulation
in Engineering

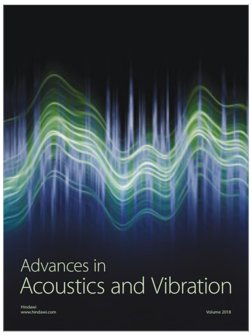
International Journal of
Aerospace
Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and
Propagation

Active and Passive
Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration