




Characterization of Individual Mobility for Non-routine Scenarios from Crowd Sensing and Clustered Data

Inês Cunha¹, João Simões², Ana Alves^{2,3} , Rui Gomes³,
and Anabela Ribeiro¹

¹ Centre for Territory, Transports and Environment, University of Coimbra,
Polo II, 3030-788 Coimbra, Portugal

inesfariadacunha@gmail.com, anabela@dec.uc.pt

² Coimbra Institute of Engineering, Polytechnic Institute of Coimbra,
3030-199 Coimbra, Portugal
a21220126@alunos.isec.pt

³ Centre for Informatics and Systems, University of Coimbra, Polo II,
3030-290 Coimbra, Portugal
{ana, ruig}@dei.uc.pt

Abstract. Demand for leisure activities has increased due to some reasons such as increasing wealth, ageing populations and changing lifestyles, however, the efficiency of public transport system relies on solid demand levels and well-established mobility patterns and, so, providing quality public transportation is extremely expensive in low, variable and unpredictable demand scenarios, as it is the case of non-routine trips. Better prediction estimations about the trip purpose helps to anticipate the transport demand and consequently improve its planning. This paper addresses the contribution in comparing the traditional approach of considering municipality division to study such trips against a proposed approach based on clustering of dense concentration of services in the urban space. In our case, POIs (Points of Interest) collected from social networks (e.g. Foursquare) represent these services. These trips were associated with the territory using two different approaches: ‘municipalities’ and ‘clusters’ and then related with the likelihood of choosing a POI category (Points-of-Interest). The results obtained for both geographical approaches are then compared considering a multinomial model to check for differences in destination choice. The variables of distance travelled, travel time and whether the trip was made on a weekday or a weekend had a significant contribution in the choice of destination using municipalities approach. Using clusters approach, the results are similar but the accuracy is improved and due to more significant results to more categories of destinations, more conclusions can be drawn. These results lead us to believe that a cluster-based analysis using georeferenced data from social media can contribute significantly better than a territorial-based analysis to the study of non-routine mobility. We also contribute to the knowledge of patterns of this type of travel, a type of trips that is still poorly valued and difficult to study. Nevertheless, it would be worth a more extensive analysis, such as analysing more variables or even during a larger period.

Keywords: Urban mobility · Destination choice modelling · Clustering analysis

1 Introduction

Traditional transport supply design criteria require satisfying the highest demand hours, typically occurring during the work commute. However, in recent decades, the demand for leisure activities has increased due to some reasons such as increasing wealth, ageing populations and changing lifestyles, yet non-routine trips have only recently received more attention. Understanding personal travel patterns and modelling travel demand is essential to plan sustainable urban transportation systems to fulfil citizens' mobility needs. To do this effectively and timely, urban and transportation planners need a dynamic way to profile the movement of people and vehicles. Traditional survey methods are expensive, time consuming and give planners only a picture of what has happened.

In contrast, an emerging field of research uses the wide deployment of pervasive computing devices (cell phone, smart card, GPS devices and digital cameras) and transport system records (e.g. ticket validation counts; traffic counts) providing unprecedented digital footprints, telling where and when people are, for “urban sensing”. Moreover, the composition of social networks and human interactions is crucial not only for understanding social activities, but also for travel patterns. With millions of users around the world, transportation researchers have also realized the potential of Social Network Analysis (SNA) for travel demand modelling and analysis.

To date there is little published research, however, on how to realize this opportunity for the sector by capturing the views, needs and experiences of the travelling public in a timely and direct fashion through social media. All these mobility data, together with modern techniques for geo-processing, SNA, and data fusion, offer new possibilities for deriving activity destinations and allowing us to link cyber and physical activities through user interactions, in ways that can be usefully incorporated into models of land use and transportation interactions.

With data registered in mobile phones by users, using an App designated by SenseMyFEUP¹ [1], it was possible to characterize some aspects of the mobility of these users, as the destination selected for non-routine trips. The selection of this type of trips is related with the importance of recording some patterns that are not captured by the most usual mobility studies and that are important to characterize, in order to identify the gaps and to adapt the public transport offer, trying to decrease the use of the car, specially during the night-time. This type of information has been cross-referenced with social network data that records entries and preferences in places with night activities to characterize these patterns as accurately as possible.

It was possible to see that most of the trips that were made had origin and destination in the same municipality, with a greater tendency on weekdays. During weekends, the tendency to move to other municipalities increases, pointing to the availability

¹ <https://sensemycity.up.pt/project/sensemyfeup/> (September 2019).

to travel more distances being higher on weekends compared to weekdays. In the Municipality of Porto, Portugal, there was a greater number of movements in parishes with a higher offer of events/POIs, showing the attractiveness created by them. Geographic proximity also showed an influence, demonstrated both by the low travel times and by the proportion of trips to neighbouring municipalities concerning remote municipalities.

Regarding the modes of transport used during non-routine hours, the car was the most used, confirming the great tendency to use this mode already presented in other previous studies based on regular trips in the Porto Metropolitan Area, followed by foot. It was also identified the importance that certain variables had in choosing the destination, using a multinomial logit regression. The variables of distance travelled the number of check-ins, travel mode and whether the trip was made on a weekday or a weekend had a significant contribution.

The collected dataset of trips and the results will be discussed further, and although allowing very interesting results, and in order to obtain more reliable results, we have tried a new approach in this work using clustering analysis. With this, we identified agglomeration of POIs, i.e., the services that are offered in a concentrated way in the urban space. And, on the other hand, we also took into account the concentration of events that usually happen in certain areas of the city, i.e., looking for social events that are not related specifically to the importance of their venues but which are important enough to attract people. With this in mind, clusters of important areas were computed to represent the origin or the destination of a given trip, rather than the traditional municipality used in most of studies. The results obtained for both geographical approaches are then compared, considering the multinomial model studied to check which approach gives the best results for non-routine travel analysis.

In order to promote a broader discussion about the influence of geographical units in the analysis of mobility patterns and spatial data clustering, we present a review of the literature in Sect. 2. In Sect. 3, we present the proposed approaches of the comparison study (territorial units versus cluster techniques). In Sect. 4, we present the results and some discussion about them. Finally, on Sect. 5, conclusions are drawn and next steps are proposed pointing the main inputs for future research.

2 Literature Review

2.1 Studies on Non-routine Mobility Patterns

In recent decades, the demand for leisure activities has increased due to some reasons such as increasing wealth, ageing populations and changing lifestyles. This increase in demand has led to higher levels of emissions and congestion and, therefore, to greater concern in research [2].

Daily travel has already been comprehensively studied. However, non-routine trips have only recently received more attention, despite their environmental impact. Unlike fixed trips (i.e., trips to work where the arrival time and the trip destination are pre-defined), leisure activities offer more optimization flexibility since the activity

destination and the arrival times of individuals can vary [3]. The study of these trips is the focus of this paper.

An important issue to realize in this type of travel is that the option of starting this type of activity is confined to certain times of the day due to restrictions imposed by other activities [4]. Regarding the frequency of leisure trips per week, Tarigan and Kitamura [5] realized that the number of cars at home has a positive impact on the number of activities per week related to shopping and recreational activities, but not so much in social contact activities. On the other hand, the more bicycles there are at home, the more trips are per week related to sports and nature activities. Individuals living in the suburbs tended less to participate in nature-related activities than those living in the city. They also found that factors such as gender, age, life-cycle stage, vehicle ownership, and residence location have a significant influence on leisure travel patterns.

Another study focusing on leisure trips characteristics was that carried out by Sánchez et al. [6]. They found that men perform more leisure trips regardless of age, women jaunt average distances, and also that leisure trips are made closer to home regardless of the size and type of town. A more specific study was carried out by Sener et al. [7] where they analysed physical leisure activities. They concluded that socio-economic, demographic, domestic, employment-related and environmental variables affect the choice of physical recreation activity, relative to location, time of day, the day of the week and social context. Moreover, a study was conducted to understand whether there were leisure travel patterns differences between people living in an urban district and people living in a small city in a suburb. Große et al. [8] found out that the urban structure of a residential place influences the constitution of daily mobility styles and that there is a greater tendency to perform weekend trips and holidays by people living in an urban district.

2.2 Crowd Sensing for Mobility Characterization

In the field of data collection for transport studies, several studies on mobility patterns have already been done using data from mobile phones and social networks, confirming the various advantages as data providers. Some of these advantages were presented by Calabrese et al. [9], such as lower cost, larger sample size, higher update frequency, more spatial and temporal coverage and, also, provide unprecedented “digital footprints”. However, they have also identified some gaps, which could be filled by traditional methods. Examples of such gaps are the impossibility to access socio-economic and demographic characteristics due to privacy issues; the high probability that mobile phone users do not represent a random sample of the population tending to be biased, and the difficulty of using the database because of its format is not ready for modelling.

There are also some challenges in comparing different datasets, even if they are related. Anda et al. [10] pointed out that the main ones are the different harvest periods and different spatial units. However, they also have the advantage of using different human mobility sensors and complementary datasets, such as travel diary data, to complement the importance of the data fusion approach.

In the field of transport planning, one of the advantages of these digital footprints is helping to reduce underreporting of trips, which is a common phenomenon in travel surveys. Thomas et al. [11] showed this conclusion by carrying out a mobile application study in the Netherlands for one month, registering departure and arrival times, origins and destinations, and travel reasons for a group of 615 volunteers. To compare automatically detected and reported trips, volunteers also had to participate in a requested web-based recall survey and answered additional questions. Most trips were identified without any apparent defects in the length or duration of the trip, and modes of transport were classified correctly in more than 80% of these trips.

Digital footprints come also from the Location-Based Social Networking (LBSN) such as “Foursquare”, “Gowalla”, “Google Latitude” and “Facebook Places”. These social networks are an indispensable source of voluntary geographic information [12]. While providing data from different age groups and site categories, geo-tagged social network data can complement regular household survey data to validate findings of mobility and urban structure and to reveal new insights about nature and human behaviour [13].

2.3 Clustering Techniques Applied to Urban Space

In [14], the authors identify seven groups of data analysis techniques usually used to understand urban mobility: visualization, statistics, logical heuristics, graph theory, optimization, clustering and classification.

Regarding clustering, a study implemented in the PublicSense mobile crowd sensing framework [15] in the city of Xi’an, in Japan, proposes the use of complaints platform data to understand the current problems of the city (mobility, constructions, among others), and thus supporting the decision making with respect of management of city resources to solve problems. This study applies clustering, namely K-means, to identify those areas that contained more complaints, most related to mobility problems, and therefore require more attention. This same algorithm is used by Quadri et al. [16] to identify three (K) classes of places visited by each individual person from Call Detail Record data to identify non-routine trips: Mostly Visited Places, Occasionally Visited Places and Exceptionally Visited Places.

The NSE (National Science Experiment) project [14] aims to analyse and understand the mobility of young students, namely the determination of patterns in the means of transport used, through the exclusive collection of access points extracted from the Wi-Fi network. The identification of POIs and trips is performed through the application of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. In another study [17] this same algorithm was used to group visited GPS locations visited by elderly people according to different profiles of trips.

Gan et al. [18] propose a framework of identifying urban mobility patterns and urban dynamics from a spatiotemporal perspective and pointing out the linkages between mobility and land cover/land use (LCLU). For this, they used data from the AFC (Automatic Fare Collection) system of Nanjing metro and applied the K-means clustering algorithm to isolate seven types of stations (depending of they are used to return home, go to work, etc.). Next, a multinomial logit regression was estimated to explore the relationship between each distinctive cluster and local land use characteristics.

2.4 Comparative Analysis

After the collection and analysis of the mentioned studies, it was verified that some have already been done considering non-routine trips (leisure, social, tourism, etc.). However, as this type of studies has only recently been given more relevance, they still have some gaps, such as: dispersion in calculation methods, data sources, time intervals and scopes, the accuracy and completeness of the measurement are compromised by survey designs, socio-psychological variables are often not controlled, and more [19].

Complementary, the use of digital footprints presents many advantages for the characterization of the mobility compared to traditional methods. However, several studies have found that for mobility research, the best way to obtain a database with quality is to combine information from different available data sets, from traditional sources like mobility surveys, to more recent ones, like mobile phones data [10, 11, 13].

Such digital footprints represent large volume of geospatial and crowd sensed data that brings a new perspective of more up-to-date mobility studies. Unsupervised techniques, as clustering, can deal with this type and massive quantity of data, which are not labelled and are produced in real-time fashion. To the best of our knowledge, none of the scientific works found in the literature proposed a comparative study to the characterization of non-routine trips using territorial divisions against density-based clusters. Our study proposes to develop a methodology that contributes to the improvement of this knowledge. In the methodology proposed, we also exploit the use of silhouette score as an internal and external measure to identify the best configuration values for each density-based clustering analysed.

3 Proposed Approach – Methodology

As already mentioned, the area defined for the study corresponds to the city of Porto and its surroundings, more specifically the Porto Metropolitan Area. What led us to this choice was the fact that there is large-scale data on social networks, both at points of interest and events. To detect and collect movements of people, the SenseMyFEUP [1] app was used to collect trip data from volunteers in April 2016, targeting students and personnel of the Faculty of Engineering of the University of Porto. Participants were asked to install the corresponding application on their smartphones that automatically collected mobility data during trips and launched questionnaires enquiring participants on the used transportation mode.

Some characteristics of the resulting dataset, such as errors in raw data and user engagement analysis, were published in [20, 21]. This dataset was made available by their authors in a given Data Access Agreement, and after trip segmentation and anonymization. It comprises data from 239 unique participants from the Faculty of Engineering of Porto. In total, 23600 trips were gathered over 6300 h of commuting time. Since this is a study focused on non-routine movements, we studied every day of that month only from 7 p.m. to 7 a.m., except for the weekends for which the full day was analysed. After this filtering, we reached a total of 2246 non-routine trips. Since the study hours depend on the day in study, the sample was divided into two distinct groups: weekdays and weekend days.

Foursquare² is a social network with the main objective to register the users' visits to a Point of Interest (check-ins). The user can classify the place and add pictures or comments. This data source has many POIs and provides an Application Programming Interface (API) to collect data in an automated way. These POIs are classified in categories that forms a taxonomy with ten top-level categories: *Arts & Entertainment*, *College & University*, *Event*, *Food*, *Nightlife Spot*, *Outdoors & Recreation*, *Professional & Other Places*, *Residence*, *Shop & Service* and *Travel & Transport*.

Infoporto³ is a website that gathers Porto city events, encompassing a portal to disseminate information about the Porto Region. It allows the consultation and research of events to take place in that region. It has a strong connection with Facebook Events in the way that every event advertised has its link to this social network. This service is available free of charge to the public. We developed a script in python to collect events dated April 2016.

3.1 Global Approach

The method developed comprehends four phases:

1. Data collected from SenseMyFEUP dataset and from social networks (Foursquare and InfoPorto);
2. Clustering all events and POIs collected in two different layers considering their density.
3. Data processing, making two similar samples: (a) one based on territorial division (2934 trips); (b) one based on clusters techniques (2091 trips, after excluding those that are not having a cluster of POIs as destination), both for the month of April 2016 and considering the trips regarding leisure purpose (inferred by the closest popular POI top-level category) from the following Foursquare categories: *Arts & Entertainment*; *Event*; *Food*; *Nightlife Spot*, and *Outdoors & Recreation*;
4. Analysis and discussion of the obtained results by using multinomial logistic regression for the destination choice model estimation and comparing the data obtained with the two samples;

The modelling approach is aimed at identifying the most determining factors in the choice of a given location. To support this analysis, it was used the Statistical Package for the Social Sciences (SPSS) software⁴.

3.2 Territorial Division

OD matrices were obtained by crossing information from the trips and the events and POIs, the latter only considering those with more than 500 check-ins (the most popular ones). At this stage, the trips were treated at the scale of the segments. With an area of influence defined as a circle of radius equal to 500 m and centre in the places of the

² <https://developer.foursquare.com/> (July 2019).

³ <https://www.infoporto.pt/en> (July 2019).

⁴ <https://www.ibm.com/analytics/spss-statistics-software> (July 2019).

events/POIs, the determination of the destinations of the trips was made. Thus, in the case of these matrices, the origins correspond to the beginning of the trip and the destination to the municipality where the event or the POI took place, or vice versa.

3.3 Stages of Clustering

The methodology applied aimed to identify the main areas, which are most sought after in terms of non-routine destinations in weekdays and weekend days, and an attempt to infer the possible purposes of these trips, which led SenseMyFeup users to travel to a particular destination. It was intended that the implemented methodology could answer questions such as which areas of Porto are most sought after and at what hours, which correspond to the period outside the daily life, i.e. during the week after 18h30 to 07h00 of the next day and the weekend. Also, the clustering process could help understanding which areas of the metropolitan area offer more services and which can therefore be considered as regions of high attractiveness.

The first step was to select points of interest in the city of Porto, which could be related to the trips made by users. In addition, events in the city were also selected that may be related to possible trips as well.

In order to identify the regions with the highest concentrations of points of interest and events, we chose to use the clustering algorithms HDBSCAN (Hierarchical DBSCAN), DBSCAN and OPTICS (Ordering Points to Identify the Clustering Structure). One of the main difficulties in using these three algorithms is the correct definition of the initial configuration, i.e. the minimum number of points and epsilon most indicated for the case study. To do this, we chose to use validation to determine the best configuration. A common approximation is based on the use of validation indexes [22].

The clustering assessment can be divided into three main categories: internal, external and relative. The external approach is based on prior knowledge of the data. In turn, internal indexes are used to measure cluster quality without external information. Relative approximations are used to compare different groupings or clusters.

Considering these approximations, and since the manual visualization of each possible configuration would be very painful and we do not have any prior knowledge of the data, it was decided to follow the use of an index that works for the both measures that can help us to identify the best configurations for each algorithm (HDBSCAN*, DBSCAN and OPTICS): relative and internal measures. These algorithms were selected considering that we need to group spatial data but there was no clue about the number of clusters (a pre requisite to other clustering algorithms, such as K-means).

In order to assess the clustering phase, from the algorithms above, we used the silhouette score index [22]. This index is based on the principle of maximum internal cohesion and maximum separation between clusters, that is, it measures how close the objects are to each other, and belonging to the same cluster and how far apart they are from objects of other clusters. The silhouette score values range from -1 to 1 . Values less than 0.25 mean that no substantial structure has been found, values ranging from 0.26 to 0.50 means that the structure found is weak and may be artificial, values ranging from 0.51 to 0.70 mean that a reasonable structure was found. Finally, values above 0.7 mean that a strong structure was found.

In order to achieve a larger volume of samples, we chose to select the clustering configurations that maximize the assessment index, the silhouette score, but which may also have resulted in more than 30 clusters for statistical significance. For each layer (POIs and events), the OPTICS and DBSCAN algorithms were run with epsilon varying between 0.0005 and 0.0024, in intervals of 0.0001. For each epsilon value, the minimum point value varied from 4 to 170.

In the case of HDSBCAN*, since it does not use epsilon, but just the minimum of points, the algorithm was run for each layer with this configuration parameter from 4 to 170. Table 1 presents the results of the assessment for the event layer, as an example of evaluation performed to select the best algorithm and configuration, respectively.

Table 1. Configuration obtained in the event layer, in the weekend days.

Algorithm	DBSCAN	OPTICS	HDBSCAN
Minimum points	4	4	4
Epsilon	0.001	0.0012	N/A
Resulting number of clusters	20	16	16
Silhouette score	0.85	0.83	0.74

3.4 Model Generation - Destination Choice Models

These models tried to answer the following question: The user's role, the mode of transport used, weekday vs weekend day, travel time, distance travelled and the number of check-ins in the event/POI may influence the destination choice? And, how can the geographic approach influence these relations?

Two models were tested: one model considered trips starting and ending in a parish/municipality of Porto and the other considered a cluster as destination.

The destination choice models considered several types of variables, including both trip-related characteristics, event/POI attributes, and individual socio-demographics characteristics.

Trip-related characteristics explored in our specifications included nominal variables such as the transport mode used for the trip; and, whether the trip was on a weekday or on a weekend day; and scale variables: the trip travel time; and, the distance travelled.

Event/POI attributes included two variables: number of check-ins in the event/POI and category, scale and nominal variables, respectively.

Individual socio-demographic characteristics included the only variable available that is the role of the user, a nominal variable.

We used the SPSS program to estimate the destination choice model under the two different geographic approaches, which statistical inference, namely at the level accuracy of the model, is guaranteed through the statistical tests included. The model was not identified through tests on the data, but it was previously identified by the researchers for being the most adequate with this kind of variables and when the outputs are of the discrete choice kind.

4 Experimental Results – Interpretation

4.1 Destination Choice Models (Multinomial Logistic Regression)

First Model- Territorial Division Sample

General

The results indicated a good fit of the final model. The model explained 49.0% (Nagelkerke R²) of the variance in the destination choice and correctly classified 82.4% of cases. In addition, the final model is significant at a significance level of $\alpha = 0.050\%$ which means that the full model statistically significantly predicts the dependent variable better than the intercept-only model.

In Table 2, we can observe that three of the six variables presented significance to compose the final model. The variables “distance”, “time spent” and “weekday vs weekend” had a significant contribution ($p < 0.05$), but not “check-ins”, “mode” and “role”.

Trip-Related Characteristics

The model indicates that “distance”, “time spent” and the variable representing the week phase, that is, whether the trip would have occurred during non-routine hours on a weekday or during the weekend, had significance in the choice of destination category type. It was concluded that it was more likely that a user chooses to go to an “Event” than to an “Arts & Entertainment” POI if the trip occurred on a weekday. Additionally, it could be stated that it is 9.959 times more likely that the user chooses to go to an “Event” than to an “Art Exposition/Market” if it was a weekday. Furthermore, the variable “time spent” was shown to be significant. It was found that that increasing the time spent in a trip was associated to a slight likelihood of the user choosing to go visit an “Arts & Entertainment” POI.

Event/POI Attributes

Besides the category, the only variable studied in the model related to the characteristics of the events/POIs consisted of the number of check-ins. This proved not to be significant in the choice of the destination type.

Socio-demographic Characteristics

The only variable related to the individual socio-demographic characteristics introduced in the model was the role. This proved not to be significant.

Second Model-Clusters Division Sample

General

This model only considered trips that culminate in a cluster. In addition, the cluster itself is considered as an independent variable to add to the multinomial logistic regression. This new event/POI attributes related characteristic intuitively represents not only a means to filter trips that are not so important in the sense they are not going to hotspots in the city, but at the same time create a categorical variable which groups trips with close destinations (to the same end cluster). Additionally, we also registered if the trip started in a cluster (origin cluster).

Table 2. Multinomial logistic regression with territorial division.

	Event			Food			Nightlife Spot			Outdoors & Recreation		
	B	Exp(B)	Sig.	B	Exp(B)	Sig.	B	Exp(B)	Sig.	B	Exp(B)	Sig.
<i>Distance</i>	.000	1.000	.005	.000	1.000	.010	.000	1.000	.112	.000	1.00	.212
<i>Time</i>	.000	1.000	.314	.001	1.001	.007	.000	1.000	.527	.000	1.00	.653
<i>Check-ins</i>	.000	1.000	.970	.000	1.000	.811	.000	1.000	.996	.000	1.00	.591
<i>Weekday vs Weekend (ref. Weekday)</i>												
Weekend	2.298	9.959	.000	.177	1.194	.664	.394	1.483	.498	.089	.375	.870
<i>Mode (ref. Sustainable modes)</i>												
Car	0.116	1.123	.785	-.507	602	.229	-.621	.537	.276	.730	2.075	.223
<i>Role (ref. Non-teaching personal)</i>												
Student 1 st cycle	-1.423	.241	.859	-1.188	.305	.889	.773	2.166	.956	-7.102	1.308E-10	.374
Student 2 nd cycle	-2.072	.126	.796	-1.694	.184	.842	-.323	.724	.982	-6.941	1.578E-10	.384
Student 3 rd cycle	.676	1.965	.935	1.259	3.523	.886	.497	1.644	.972	.631	1.862E-07	.939
Researcher	.494	1.638	.972	.172	1.188	.990	3.137	23.027	.866	6.673	3.388E-10	.646
Teacher	.629	1.876	.937	-1.130	.323	.894	.652	1.919	.963	-6.194	3.297E-10	.438

Reference category is “Art Exposition/Market”. Significance level: $p < 0.05$

The results indicated a good fit of the final model. The model explained 28.7% (Nagelkerke R2) of the variance in the destination choice and correctly classified 84.8% of cases. In addition, the final model is significant at a significance level of $\alpha = 0.050\%$ which means that the full model statistically significantly predicts the dependent variable better than the intercept-only model.

In Table 3, differently from the previous experiment, five of the eight variables presented significance to compose the final model. For the simplicity of the table, only when a cluster is considered significant is shown. As the previous model, the variables “distance”, “time spent” and “weekday vs weekend” had a significant contribution ($p < 0.05$), but not “check-ins”, “mode” and “role”. Additionally, the new variables “cluster destination” and “cluster origin” also contribute significantly to the model.

Trip-Related Characteristics

This model indicated the same conclusion for the variables “distance”, “time spent” and “weekday/weekend”, showing that they had significance in the choice of destination category type. Nevertheless, this time, there are more categories where these variables are significant. This model, besides the conclusions made previously, now concludes that it was more likely that a user chooses to go to a “Food” than to an “Arts & Entertainment” POI if the trip occurred on a weekday.

Event/POI Attributes

Besides the category, the variables studied in the model related to the characteristics of the events/POIs consisted of the number of check-ins, cluster origin and cluster destination. These last two new variables proved to be significant in the choice of destination type. It was concluded that it is 1.044 times more likely that the user chooses to go to a “Nightlife Spot” POI than to an “Art Exposition/Market” if the destination of the trip is a density-based cluster.

Socio-demographic Characteristics

The only variable related to the individual socio-demographic characteristics introduced in the model was the user role. This proved not to be significant even in this model.

4.2 Comparing Results

We can see that for small samples, the differences are not substantial. For easier analysis, we put side-by-side the main results obtained through the two approaches, including the significant independent variables (S) and non-significant (N) for both models, as well as the respective percentages of correctly classified cases (Table 4). According to Table 4, both models suggest that the only independent variables that contributed significantly were the day of the week, the travel distance, and the time spent traveling. However, there is a slight improvement in the percentage of cases correctly classified by cluster analysis than by territorial division analysis, and the fact that the new variables (destination and origin of the cluster) are also significant. We believe that this improvement is due to the fact that this analysis is more focused on the POI/event, once the way the data is collected in clusters technique is different than in the territorial division.

Table 3. Multinomial logistic regression with clusters.

	Event			Food			Nightlife Spot			Outdoors & Recreation		
	B	Exp(B)	Sig.	B	Exp(B)	Sig.	B	Exp(B)	Sig.	B	Exp(B)	Sig.
<i>Distance</i>	.000	1.000	.000	.000	1.000	.000	.000	1.000	.002	.000	1.00	.117
<i>Time</i>	.000	1.000	.277	.001	1.001	.015	.000	1.000	.802	.000	1.00	.574
<i>Check-ins</i>	.000	1.000	.988	.000	1.000	.702	.000	1.000	.687	.000	1.00	.503
<i>Weekday vs Weekenday (ref. Weekday)</i>												
Weekenday	2.535	12.620	.001	-1.132	.322	.049	-.446	.640	.656	.665	1.944	.460
<i>Mode (ref. Sustainable modes)</i>												
Car	-.505	.604	.360	-.989	.372	.076	.219	1.245	.776	.621	1.860	.474
<i>Role (ref. Non-teaching personal)</i>												
Student 1 st cycle	-2.341	.272	.735	-1.420	.242	.862	-.772	.462	.958	.519	1.680	.963
Student 2 nd cycle	-3.687	.096	.593	-2.296	.101	.778	-2.466	.085	.866	-.751	.472	.956
Student 3 rd cycle	-.168	.025	.983	1.387	4.003	.877	-.148	.862	.992	10.749	4.66E+04	.936
Researcher	-.010	.845	1.000	-.130	.878	.994	1.518	4.563	.949	17.556	4.21+E07	.286
Teacher	-.037	.990	.996	-.741	.476	.927	-.545	.580	.970	.939	2.577	.413
<i>Cluster destination</i>	.001	1.001	.899	.025	1.025	.002	.043	1.044	.000	.011	1.011	350
<i>Cluster origin</i>	-.001	.999	.856	-.003	.997	.670	.022	1.022	.012	-.015	.985	.243

This study was a first attempt to extract useful information from georeferenced data from social media to study non-routine mobility. In addition to allowing us to contribute to the knowledge of this type of travel by studying the factors that influence the choice of event/POI category, the results obtained lead us to believe that a further analysis with a larger sample size and not as biased towards the type of people who took the trips would be worthwhile, and we would most likely have even better results.

Table 4. Multinomial logistic regression: comparing territorial with clusters.

	Territorial division (2934 trips)	Clusters (2091 trips)
Role	N	N
Transport mode	N	N
Weekday vs Weekend	S	S
Travel distance	S	S
Time spent	S	S
Check-ins	N	N
Cluster destination	N.A.	S
Cluster origin	N.A.	S
Overall percentage	82.4%	84.8%

5 Conclusion and Future Work

This paper discusses two techniques for extracting information for the study of non-routine travel, namely to characterize their purpose. With the increasing availability of social network and crowd sensed data unsupervised techniques are required to detect the city's main attractive regions in the city during those periods in order to anticipate the demand for transportation. Social events and Points-of-Interest are a good enrichment of traditional studies of these mobility patterns. This work compares the territorial division commonly used in studies of mobility patterns with the automatic recognition of clusters of dense concentrations of POIs or events serving origin-destination region of those trips. In this way, only the trips that are destined to reach a pole of attractiveness are considered in the new approach.

Two models of multinomial logistic regression were constructed from data available in the Porto Metropolitan Area, considering trips on non-routine periods in weeks and on weekends in a period of one month. Some of the trip-related and event/POI variables used showed to be significant to these models, as opposed to a unique demographic variable used due to anonymity of data that proved not to be significant.

With the clustering technique, it was possible to improve the model that only used municipalities as divisions in the study area: both in the significance of the variables and in the final accuracy. However, there is room for improvement as more variables can be added to the model, such as the period of visit at the destination of a trip. This feature is only available after anonymization if the daily session ID created by the crowd sensing mobile application is used. The dataset of trips used was limited to only one month. As a future approach to use the same methodology applied to a bigger dataset, we can test if the silhouette score can handle large data and see other indexes that can be used to identify the best configuration to each clustering algorithm. The choice of the index is important because that can be used to improve the validation time.

References

1. Santos, P.M., et al.: PortoLivingLab: an IoT-based sensing platform for smart cities. *IEEE Internet Things J.* **5**(2), 523–532 (2018)
2. Grigolon, A.B., Kemperman, A.D.A.M., Timmermans, H.J.P.: Mixed multinomial logit model for out-of-home leisure activity choice. *Transp. Res. Rec. J. Transp. Res. Board* **2343**(1), 10–16 (2013)
3. Gkiotsalitis, K., Stathopoulos, A.: Joint leisure travel optimization with user-generated data via perceived utility maximization. *Transp. Res. Part C Emerg. Technol.* **68**, 532–548 (2016)
4. Steed, J.L., Bhat, C.R.: On modeling departure-time choice for home-based social/recreational and shopping trips. *Transp. Res. Rec. J. Transp. Res. Board* **1706**(1), 152–159 (2000)
5. Tarigan, A.K.M., Kitamura, R.: Week-to-week leisure trip frequency and its variability. *Transp. Res. Rec. J. Transp. Res. Board* **2135**(1), 43–51 (2010)
6. Sánchez, O., Isabel, M., González, E.M.: Travel patterns, regarding different activities: work, studies, household responsibilities and leisure. *Transp. Res. Procedia* **3**, 119–128 (2014)
7. Sener, I., Bhat, C., Pendyala, R.: When, where, how long, and with whom are individuals participating in physically active recreational episodes? *Transp. Lett.* **3**(3), 201–217 (2011)
8. Große, J., Olafsson, A.S., Carstensen, T.A., Fertner, C.: Exploring the role of daily ‘modality styles’ and urban structure in holidays and longer weekend trips: Travel behaviour of urban and peri-urban residents in Greater Copenhagen. *J. Transp. Geogr.* **69**, 138–149 (2018)
9. Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transp. Res. Part C Emerg. Technol.* **26**, 301–313 (2013)
10. Anda, C., Erath, A., Fourie, P.J.: Transport modelling in the age of big data. *Int. J. Urban Sci.* **21**(Suppl. 1), 19–42 (2017)
11. Thomas, T., Geurs, K.T., Koolwaaij, J., Bijlsma, M.: Automatic trip detection with the Dutch mobile mobility panel: towards reliable multiple-week trip registration for large samples. *J. Urban Technol.* **25**(2), 143–161 (2018)
12. Sun, Y.: Investigating ‘Locality’ of intra-urban spatial interactions in New York City using foursquare data. *ISPRS Int. J. Geo-Inf.* **5**(4), 43 (2016)
13. Huang, A., Gallegos, L., Lerman, K.: Travel analytics: understanding how destination choice and business clusters are connected based on social media data. *Transp. Res. Part C Emerg. Technol.* **77**, 245–256 (2017)
14. Zhou, Y., Lau, B.P.L., Yuen, C., Tuncer, B., Wilhelm, E.: Understanding urban human mobility through crowdsensed data. *IEEE Commun. Mag.* **56**(11), 52–59 (2018)
15. Zhang, J., Guo, B., Chen, H., Yu, Z., Tian, J., Chin, A.: Public sense: refined urban sensing and public facility management with crowdsourced data. In: 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th International Conference on Autonomic and Trusted Computing and 2015 IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), pp. 1407–1412 (2015)
16. Quadri, C., Zignani, M., Gaito, S., Rossi, G.P.: On non-routine places in urban human mobility. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 584–593 (2018)
17. Sumudu Hasala, M., et al.: Identifying points of interest for elderly in Singapore through mobile crowdsensing. In: Proceedings of the 6th International Conference on Smart Cities and Green ICT Systems, pp. 60–66 (2017)

18. Gan, Z., Yang, M., Feng, T., Timmermans, H.: Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations. *Transportation (AMST)* 1–22 (2018)
19. Czepkiewicz, M., Heinonen, J., Ottelin, J.: Why do urbanites travel more than do others? A review of associations between urban form and long-distance leisure travel. *Environ. Res. Lett.* **13**(7), 073001 (2018)
20. Rodrigues, J.G.P., Pereira, J.P., Aguiar, A.: Impact of crowdsourced data quality on travel pattern estimation. In: *Proceedings of the First ACM Workshop on Mobile Crowdsensing Systems and Applications - CrowdSenSys 2017*, pp. 38–43 (2017)
21. Rodrigues, J.G.P., Aguiar, A., Queiros, C.: Opportunistic mobile crowdsensing for gathering mobility information: lessons learned. In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1654–1660 (2016)
22. Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E.M.: Internal versus External cluster validation indexes. *Int. J.* **5**(1), 27–34 (2011)