Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 47 (2020) 417–424

Transportation Research Procedia

www.elsevier.com/locate/procedia

22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18-20 September 2019, Barcelona, Spain

Land use inference from mobility mobile phone data and household travel surveys

Noelia Cáceres[a], Francisco G. Benítez[a]*, Luis M. Romero[a]

[a]Transportation Engineering, School of Engineering, University of Sevilla, Camino de los Descubrimientos, Sevilla, Spain

Abstract

The mobility data derived from mobile phones may provide hints regarding land-use. Activity zones, be residential or productive, feed the global mobility once acting as origin and/or destination of trips. This research presents an approach to characterise the predominant activity of the sectors of a case of study, the metropolitan area of Malaga (Spain), using mobility patterns. The methodology is tested and compared with the socio-economical information provided by the Official General Statistics and Economic Information in order to quantify the reliability of the approach.

Keywords: Land-use; Mobile phone data; Household mobility survey; Trip matrices

1. Introduction

Most activities conducted by people over a region are repeated on daily basis, and are affected by the structure of the region. This is the reason the land use zoning and their socioeconomic characteristics have been customarily regarded as explanatory features of the daily mobility. For transport optimisation and planning most of the effort has been channeled to collect data of weekday trips to portray people's mobility of a generic working day. In some areas were tourism is relevant, the characterisation and quantification of people displacements have attracted the attention of transport planners, and mobility surveys have included questions related to it. In this last case, surveys are quite dissimilar to classical household, or phone-based, data collecting. Due to the fact that surveys are costly, they are conducted in widely spaced time intervals, being this the main reason why available mobility data get frequently outdated.

* Corresponding author. Tel.: +34-954-487-315
*E-mail address:*benitez@us.es

10.1016/j.trpro.2020.03.117

Since the last two decades, the pervasive use of mobile telephony has offered a huge amount of data to be exploited in order to identify mobility patterns of users, see Pan et al. (2006), Bonnel et al. (2018), Bachir (2019), or Bachir et al. (2019).

Besides, the published literature is very profused regarding prior studies using mobile phone data to infer land-use using multiplicity of approaches and methodologies, see for example Isaacman et al. (2011), Phithakkitnukoon et al. (2010), Cui (2018), or Tang et al. (2019).

In a previous work, Caceres et al. (2018), land uses were identified by a) combining in a simultaneous way the pattern of trips generated and attracted by each zone, and b) accepting the observed fact that a multiplicity of mobility patterns coexists in multi-activity zones. This work presents advances based on exploiting the information provided by two mobility data sources, which supplement each other, to reach higher reliable origin-destination matrices. The main goal pursued focuses on the interest for automatic detection of preeminent land use and secondary activities for the purpose of planning in advance on urban, transport infrastructures and traffic regulations. The hypothesis is (preliminary) backed by the outcomes derived from the empirical analysis carried out in the area of study. Further exhaustive investigations are ongoing to corroborate the methodology and findings.

2. Empirical setting

2.1. Study area

The study area corresponds to the urban agglomeration of Malaga, Spain. It consists of the city of Malaga, with 570,000 inhabitants, and fourteen surrounding municipalities. The population of the agglomeration is around one million inhabitants on 1400 km^2 divided into 178 transport zones (TZ), 128 of them corresponds to the city. The main land-use distribution shares are almost 1/3 for residential, 1/3 for mixed uses, and the last third is equally balanced between industrial, commercial and general services. The transport network is modelled by 178 centroids, 1601 regular nodes and 3939 links. The macro-zoning consists of 46 larger zones (MZ), defined by aggregating TZs under a multiplicity of socioeconomic criteria.

2.2. Trip matrices data and comparative analysis

The most exhaustive and current quantitative mobility information available, of the case of study, corresponds to OD matrices derived from two sources: household travel surveys (HTS) and mobile phone data (MPD). The last household travel survey was conducted in 2014 based on trip patterns and travel choices made by residents, which was expanded using census, socioeconomic and employment indexes. The estimated OD matrices pictured an average winter working day over the area of study.

The MPD come from one telecommunication provider operating in Spain; the MPD-based matrices were created based on events generated by users of the mobile network (anonymising the footprints). The events consisted of active phone calls, text messages, as well as passive interactions regarding displacements between cells and static periodic updates. The raw data were finally extrapolated using similar cited indexes to provide insight representative of the total population. The data captured corresponds to two weeks in February 2015 covering the same area of study, and it frames the average weekday between all pair of zones during each hour of the day.

A comparative analysis regarding the trip data provided by the two types of mobility matrices (HTS, MPD) was conducted. Both matrices are limited to data collected from information related to persons over 18 years old, to comply with legal issues.

2.3. Trip distribution

This section explores the trip data as a function of the hour of the day and the travelled distance (shortest network distance between the origin and destination centroids). From the mobile data perspective, the detection of movements of crowds is strongly subject to the number of events generated by phones as they communicate with the

network. The more events are generated the more footprints are available to infer the trip. In this regard, a longer duration increases the possibilities of a call or message event, or even the above-mentioned passive events.

Likewise, a longer trip distance also offers more opportunities in generating events (e.g. due to movement events created when a user changes from one group of cells to another). When trips are made in less time (because they imply shorter distances or are made at faster speeds), mobile phones leave fewer footprints of their "approximate" locations during their movement. The consequences of these aspects cannot be ignored when using MPD. Focusing on the travelled distance (based on shortest network distance between the origin and destination centroids), Fig. 1a reveals that for medium- and long-distances, normally made by motorised modes, the travel rates are very similar from both sources (MPD in blue and HTS in red). But for distances less than 2.5 km, a significant reduction of the MPD rates is appreciated. This suggests that mobile data tends to underreport short-distance trips (less than 2.5 km and/or 30 minutes), as it is showed in Fig. 1b. In this respect, the modes of transport related to short-distance trips are usually non-motorised, primarily walking, as stated by Asadi-Shekari et al. 2013, although cycling also occurs. Research in literature shows that the average 'walking' speed is around 5 kilometers per hour (km/h) while that of cycling is around 10-12 km/h, see Fishman et al. (2013) and Sieg (2016), depending on factors such as user's age, gender or even surface condition. Crossing short-distance trips with low travel speeds, a movement on the scale of neighborhood in cities is obtained; this is difficult to be detected with the spatial resolution offered by mobile technology (strongly dependent on the granularity of mobile network). Moreover, in terms of travel time, many of those trips involve less than 15-30 minutes, a reduced time window in which the generation of mobile events is less likely. Fig. 1a also displays the travel rates derived from HTS only considering trips with a duration (reported by respondents) greater than 15 minutes. These travel rates (bars in green) are certainly close to the MPD rates (bars in blue), especially in the context of short-distances. This confirms the problematic about the sparse representation of mobile events to infer trips in such time window. Trips implying longer distances made at faster speeds may also take 15-30 minutes; so this issue may also arise in medium- and long-distances. However, in this case it is less pronounced since there are other events (e.g. changes from one group of cells to another) to infer trips consistently.

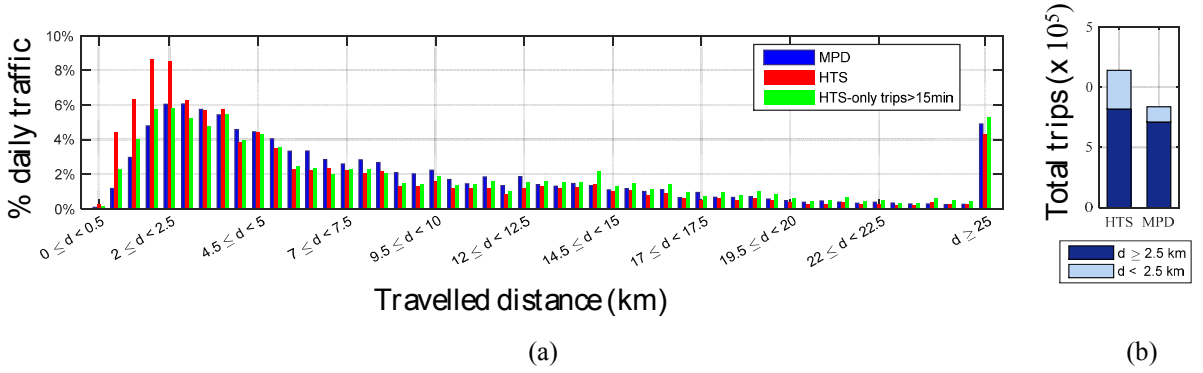


Fig. 1. Trip distribution from both sources (MPD vs. HTS): (a) Percentage of daily trips classified by travelled distance ranges (in kilometers, km); (b) Total daily trips are coloured as function of the travelled distance.

Regarding the trip distribution by time of the day (Fig.2), both sources distinguish that most of the trips during a workday are usually concentrated between 07:00 and 21:00 hours, and there are similar distinguishable characteristics in both curves. The most intense time period of the day occurs in the morning (between 07:00–09:00 hours), when the majority of people commute to work, school or business (mandatory type); but the distribution also shows a remarkable peak in the afternoon (due to double-shift work which produces pronounced peak periods around 13:00–15:00). It is worth noting that in those peaks the travel rates derived from mobile data are less intense than from survey data; in contrast, the rates are higher than surveys in off-peak periods and during the evening. The reason of such deviation may be explained by the nature of many trips. Based on survey data, the majority of trips during the peak-period in the morning and in the afternoon are short-distance; many of them typically associated with the trip chaining (e.g. intermediate short-time activities). In addition, for purposes involving regular (or mandatory) activity, made primarily during the peak periods, survey-based approaches also tend to overestimate trip

rates, see Greaves (2000) and Stopher et al. (2003). This fact, together with the underreported short-trips from mobile data source, can explain the notable deviation between HTS rates and MPD rates showed in Fig 2 during the peak-periods. In contrast, trips in off-peak periods are mainly dominated by discretionary purposes; the same occurs in the evening, when many of the trips are linked, apart from returning home, with some other non-regular purposes like shopping, entertainment, and other recreation trips. In survey-based approaches, the more spontaneous or discretionary trips are severely underestimated, see Greaves (2000) and Stopher et al. (2003) or underreported, see Badoe and Steuart (2002). Individuals may forget to report discretionary trips that are typically associated with trip chaining, Bricka and Bhat (2006); so that it is difficult to make any inference on this kind of trips. This issue is overcome by MPD, as data is passively collected, without being affected by either non-responses or human errors.

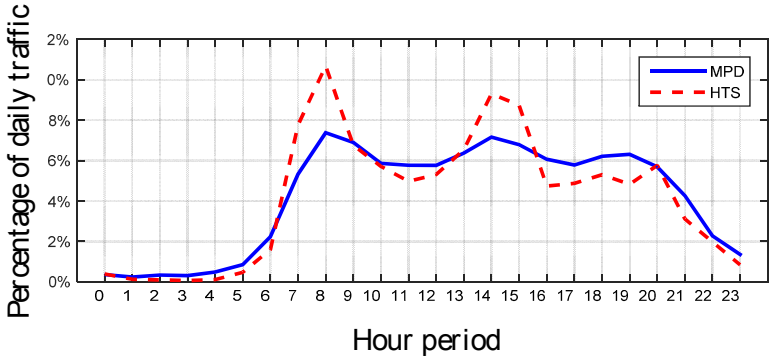


Fig. 2. Trip distribution from both sources (MPD vs. HTS) with the percentage of daily trips by the hour period associated to the departure time of the trip ('8 Hour period' refers to 08:00-08:59).

2.4. HTS-MPD data fusion

The use of two mobility data sources, with their own respective peculiarities, screens out some of the main disadvantages of each one. It is stated that mobility estimation requires large sample size of observations to infer valuable estimates. In this study approximately 3% of population (30.000 persons) in the study area was interviewed to derive HTS-OD matrices. From inferring MPD-OD matrices approximately 200.000 mobiles handsets were exploited. Though, in this last case, some assumptions made might affect the full reliability of the captured information (i.e. the sample was not affected by statistical sampling; just one mobile operator was involved; some socio-economic population segments are over/under represented; non-uniform phone activity), some relevant advantages were implied (i.e. size of the sample exceeds the survey-based HTS; the sample includes non-resident visitors within the area). The advantages of using both sources in a cooperative manner allow to take advantages of the main functionalities of both sets: mobile data are collected passively, thus non-response and non-reported information regarding the actual trips made are minimised; besides, the timeliness nature of the data collected diminishes non-reporting trips (a non-negligible share in HTS). In addition, HTS brings in information regarding socioeconomic characteristics of transport system users, the modes used for the displacements and the stages of a trip, difficult to be obtained by other means.

The data fusion approach applied herein had the objective of inferring OD trip matrices from both data sources, MPD and HTS, in order to maximise the representativeness of the mobility in the region of study. The final OD matrices bring in a) the representativeness of the MPD regarding total trips generated and attracted by the different zones of the area of study, diminishing the lack of reliability regarding information associated to short trips (either on distance and/or duration); and b) the higher accuracy of the HTS in relation to trip distribution by distance.

The inference schema follows an entropy-maximisation and information-minimisation problem:

$$\min_{g_{ij}} \quad \sum_{i,j} g_{ij}\left(\log\left(\frac{g_{ij}}{m_{ij}}\right)-1\right)$$

$$s.t. \quad \sum_{ij\in Bin(b)} g_{ij} = P_b \cdot T \qquad \forall b \in \{1,...,|B|\} \qquad \text{(a)} \tag{1}$$

$$\sum_{i\in I}\sum_{j\in J} g_{ij} = \sum_{i\in I}\sum_{j\in J} m_{ij} \qquad \forall I,J \in \{1,...,|MZ|\} \qquad \text{(b)}$$

where g_{ij} stands for the estimated number of trips from transport zone i to j, m_{ij} the number of trips between transport zones i and j from the daily MPD-based matrix. The final estimated matrix g_{ij} will be forced to have a similar structure as the prior MPD-based matrix m_{ij}, keeping the OD relations provided by MPD. Additional information is included in the problem, to adjust the deviation of the observed data related to short trips detected in MPD, using the total number of trips and its distribution by distance provided by the HTS-based matrix.

The uppercase indices denote macro-level zones, and lowercases stand for transport zone. Restrictions (a) in (0) impose that the resulting matrix must fulfil the trip distribution by distance provided by the HTS-based matrix. Each OD-pair is classified in a discrete number of intervals, $|B|$; P_b is the proportion of trips in the distance range identified by bin b, and T is the total number of trips, magnitudes provided by the HTS-based matrix. To control the distortion at each OD pair, of the estimated matrix with respect to the prior one, during the estimation procedure, an additional set of constraints (b) is imposed, whose purpose is to bound the number of trips at macro-zone level (more accurate at lower granularity). This scheme forces trips at macro-zone level are maintained during the optimisation. The analytical solution of problem (1) is:

$$g_{ij} = \alpha_{b(ij)} \cdot \beta_{IJ(ij)} \cdot m_{ij} \tag{2}$$

where the dimension of vectors of coefficients α and β are respectively $|B|$ and $|MZ|$. The solution obtained preserves the original structure of m_{ij}, but the total number of trips T is distributed among the total number of OD pairs detected in the MPD, which is much higher than those captured by HTS. The problem (1) can be solved by means of a classical iterative proportional fitting procedure (Furness 1965; Salter 1989). The estimated matrix G preserves the information trend brought in by the prior matrix regarding the number of trips contained in OD pairs at traffic zone level ($R_p = 0.91$); resulting in a Pearson's coefficient at macro-zone level of value 1. By this scheme, short trips have increased their reliability in the estimated matrix with regard to the prior one. Fig. 3a depicts the trip distribution versus distance. It is worth noting that the estimated matrix (G) shows similar trip rates to those from HTS source; Fig. 3b also reflects this comparison.

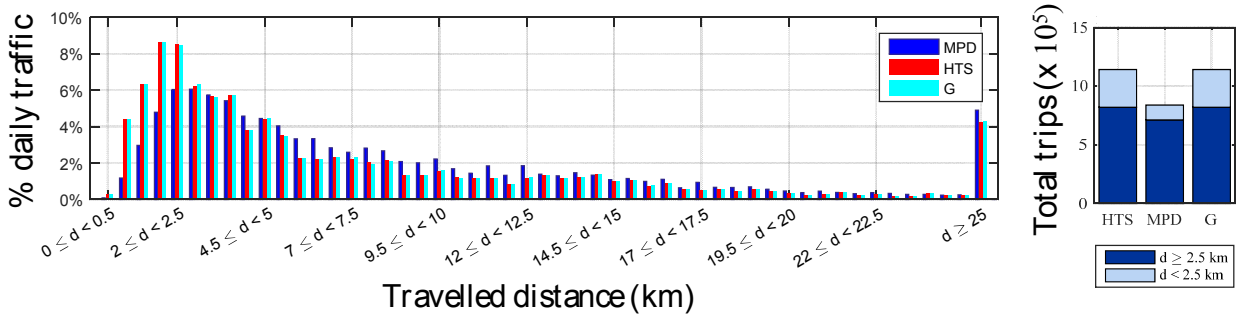


Fig. 3. Trip distribution from the estimated matrix (G) and the two sources (MPD and HTS-based matrix): (a) percentage of daily trips by travelled distance; (b) total daily trips.

3. Mobility patterns and land-use inference methodology

The purpose of the tested methodology is to identify and classify the transport zones according to the land-use based on the combined HTS-MPD mobility information. The mobility data has provided the total number of trips originated and terminated in each zone at a particular hour period, for weekdays, differing from the standard-stated trip production-generation and attraction scheme. Later, 24-element vectors corresponding to trip distribution, originated O_k^d or terminated D_k^d, by each zone k and day-type are derived to be used as object for clustering. The CH index, Calinski and Harabasz (1974), in conjunction with the pairwise Euclidean distance and the Dunn's index, Dunn (1973), for choosing the best clustering outcome are used.

Fig. 4 presents the average weekday pattern in each group (thick line in green) for originating trips (a) where three distinct groups are identified, and terminated trips (b) with other three clusters; lines in black reflect the profiles of all zones classified in each group (identified by the number of zones N). Subscripts on grouping identification G_{sn} stand for set membership s and clustering n. The shape of the patterns characterises diverse behaviours encountered in the area of study. Regarding data associated to trips originated from zones, three distinct clusters are identified: i) G11 where two clear separated peaks are identified at lunch-time and evening-time, corresponding to business-related zones (work, education, services, etc.) where people are involved in split shifts or half-day activities with home-return trips for lunch and at the end of the working period; ii) G12 pattern with three smooth separated peaks suggesting a mixture between business and commercial activities; and iii) G13 pattern with one very pronounced peak earlier in the morning, a clear identification of residential zones where trips are originated. In relation to trip terminated, other three clusters are inferred: iv) G21 with a sharp peak in the early morning which can be easily associated to business-related zones; v) G22 with three blunt peaks characterising areas of mixed activities (commercial/industrial); and vi) G23 with two more pronounced peaks, at lunch and evening time, typical of residential areas.

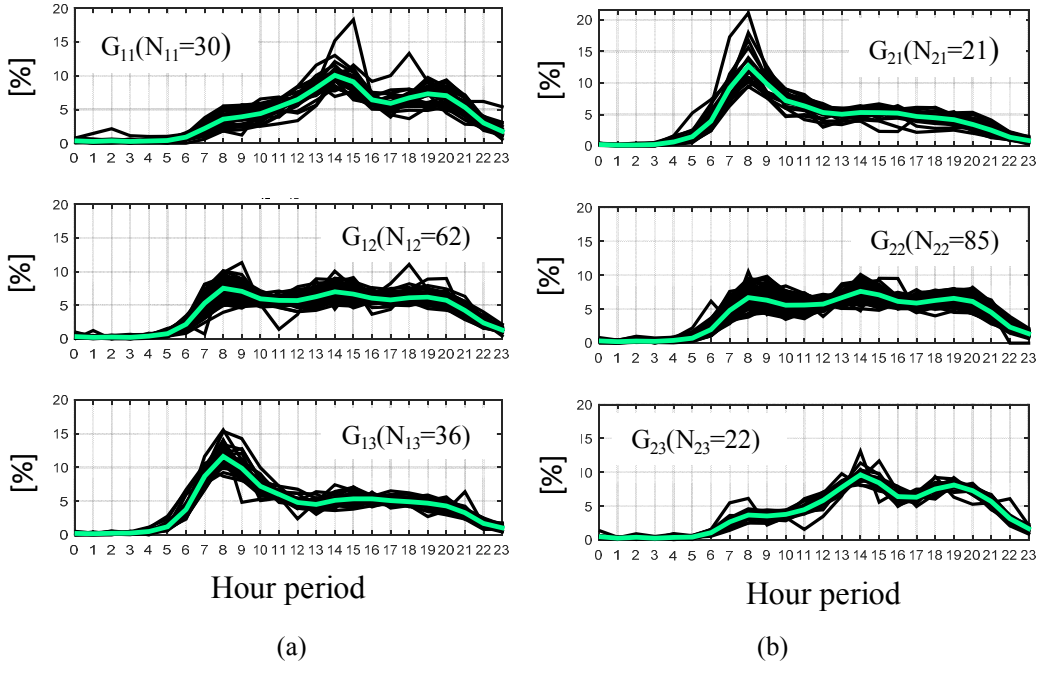
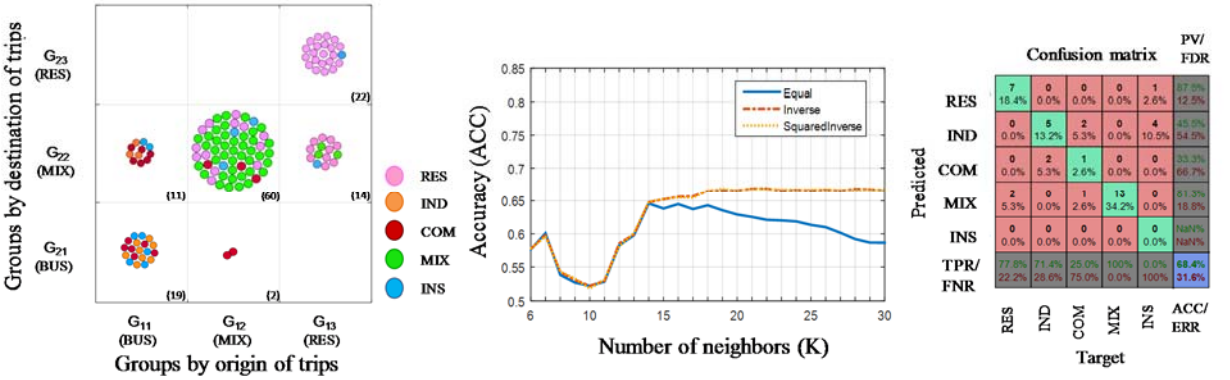


Fig. 4. Weekday patterns: (a) departure trip distribution from zones and (b) arrival trip distribution from zones (versus initial/ending time).

According to an intuitive land-use classification, based on the pattern previously inferred, a mapping can be generated where those originating-trip groups are presented versus terminating-trip groups, depicted in Fig. 5a. Each dot corresponds to a zone common to both groups; the colour identifies the actual land use (business: BUS, mixed:

MIX, residential: RES). Each square box includes the number of implied zones. Although most zones are correctly classified according to the originating-trip and terminating-trip patterns, as they are the cases of residential zones (RES) and mixed-activity zones (MIX), there are a set of them that shift land-use classification from the originating/terminating trip perspectives.

In order to improve the classification success of mixed activities zones a k-nearest neighbors algorithm (KNN) approach is applied using a training data set of 90 zones whose land-use is known a priori. The model is previously trained to generate the classification rules using a test set of 38 zones, with balanced categories (i.e. residential, commercial, industrial and mix). The rules are constructed based on assigning the category of an element according to the category of the closest neighbours. The contemplated parameters are: the number of neighbours (between 6 to 30), distance measure (cosine metric), and distance weighting function (equal, inverse and squared inverse). A multiplicity of data-splitting runs, into calibrating and testing sets, has been used to generate the fitted models and to assess the testing data. The predictions of the results were averaged over the split runs. Fig. 5b presents the evolution of the accuracy (ACC) as a function of the number of neighbours (K), reaching a 68%. The performance of the classifier model is presented in Fig. 5c based on the confusion matrix or contingency table, see Stehman (1997), for the case of square inverse weighting function and a number of 18 neighbours. This table total the zones correctly/incorrectly predicted by the classification models. The columns are associated to the actual category of the data (target). The rows reflect the predictions made by the model. The diagonal elements count the correct classifications for each category, and the off-diagonal elements show the errors made by the model. Each cell presents the percentage of the total test size. Other indicators are the *true positive rate* (*TPR*): proportion of positive cases correctly identified, the *positive predictive value* (*PPV*): proportion of the predicted positive cases correctly identified, the *error rate* (*ERR*), *false negative rate* (*FNR*) and *false discovery rate* (*FDR*). The accuracy for the selected test set is 68.4% The results for each of the categories yield high *PPV/TPR* rates, though the worst results in classification in terms of *PPV/TPR* occur for institutional class (INS). In this last category, none of the possible zones have been correctly classified, and the model is not able to classify a zone properly. A plausible explanation might arise from the fact that the wrong classified zones involved general services like healthcare/education/leisure facilities in connivance/proximity with residential areas. Despite this, the accuracy of this approach is quite satisfactory for the rest of categories, taking into account that it is only based on patterns of trips originated and terminated in zones. The best rates for *ACC/PPV/TPV* are achieved for the major categories of land uses in the studied area: RES and MIX. This conclusion is important attending to the interest in urban and transport planning.



(a) (b) (c)

Fig. 5. (a) Classification of zones according to originating and terminating trips; (b) Accuracy levels as a function of the number of neighbours: (c) Confusion matrix for KNN classifiers using the testing data set.

4. Conclusions

The methodology described here should be helpful in assessing land use classification in an urban environment. It is composed of two techniques; the first one allows inferring a more reliable mobility data, by using a fusion mapping between the two most relevant sources of information regarding OD matrices; this balances the advantages

and disadvantages brought in by each data collecting technique, household trip surveys (HTS) and mobile phone data (MPD). In particular, it is worth to underline the under-reported information provided by MPD regarding short trips versus the higher representativeness of OD pairs and sample coverage, and the opposite features of under-reported OD areas drawn by HTS versus higher reliability of short trip duration/time distribution. The second technique identifies the predominant activity of the urban zoning using mobility patterns, derived from the fused mobility data, and a supervised classification approach as the KNN.

The analysis has been conducted in the urban agglomeration of Malaga (Spain). The outcomes, though limited to just one empirical case and not exhaustive enough, could led to greater improvement in the application of land use characterisation.

Acknowledgements

The authors would like to thank the following institutions: Public Works Agency and Regional Ministry of Public Works of the Regional Government of Andalusia, Malaga Area Metropolitan Transport Consortium, and Malaga City Council, for the offered help during this research. Finally, N. Caceres acknowledges the funding provided by the Spanish Ministry of Economy and Competitiveness through the Torres Quevedo Programme (PTQ-13-06428).

References

Asadi-Shekari, Z., Moeinaddini, M., Shah, M.Z., 2013. Non-motorised level of service: Addressing challenges in pedestrian and bicycle level of service. Transport Reviews, 33, 2, pp. 166-194.

Bachir, D. Estimating urban mobility with mobile network geolocation data mining. PhD Thesis, Universite Paris-Saclay, 2019.

Bachir, D., Khodabandelou, G., Gauthier, V, El Yacoubi, M., Puchinger, J., 2019. Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transportation Research Part C, 101, pp. 254–275.

Badoe, D. A., Steuart, G. N., 2002. Impact of interviewing by proxy in travel survey conducted by telephone. J. Adv. Transp., 36, 1, pp. 43–62.

Bonnel, P., Fekih, M., Smoreda, Z., 2018. Origin-destination estimation using mobile network probe data. Transportation Research Procedia 32, pp. 69–81.

Bricka, S., Bhat, C. R., 2006. Comparative analysis of global positioning system-based and travel survey-based data. Transp. Res. Rec., 1972, pp. 9–20.

Caceres, N., Benitez, F.G., 2018. Supervised land use inference from mobility patterns. J. Adv. Transp. 2018, ID 8710402.

Calinski, R. B., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics, 3, pp.1–27.

Cui, Y. Augmenting household travel survey and travel behavior analysis using large-scale social media data and smartphone GPS data. PhD Dissertation. University at Buffalo, USA, 2018.

Dunn, J.C., 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3, pp.32–57.

Fishman, E., Washington, S., Haworth, N., 2013. Bike share: A synthesis of the literature. Transport Reviews, 33, 2, pp. 148-165.

Furness, K.P., 1965. Time function iteration. Traffic Engng. and Control 7, pp. 458–460.

Greaves, S. P. Simulation Household Travel Survey Data for Metropolitan Areas. Ph.D. Dissertation, Dep. Civil and Environmental Engineering, Louisiana State University, Baton Rouge, Louisiana, 2000.

Isaacman, C., Becker, R., C´aceres, R., Kobourov, S.,Martonosi, M., Rowland, J. M., Varshavsky, A., 2011. Identifying important places in people's lives from cellular network data. Pervasive Computing. K. Lyons, J. Hightower, and E.M. Huang (Eds.). Springer-Verlag, 2011, pp. 133-151.

Pan, C., Lu, J., Di, S., Ran, B., 2006. Cellular-based data-extracting method for trip distribution. In Transportation Research Record: Journal of the Transportation Research Board, No.1945, pp. 33-39.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C., 2010. Activity-aware map: identifying human daily activity pattern using mobile phone data. Human Behavior Understanding (Ed. Salah A. A., Gevers T., Sebe N., Vinciarelli A). Springer LNCS 6219.

Salter, R. J. Highway Traffic Analysis and Design. MacMillan, London, 1989.

Sieg, G., 2016. Costs and benefits of a bicycle helmet law for Germany. Transportation, 43, 5, pp. 935-949.

Stehman, S. V., 1997. Estimating standard errors of accuracy assessment statistics under cluster sampling. Remote Sensing of Environment, 60, pp. 258– 269.

Stopher, P. R., Wilmot, C. G., Stecher, C., Alsnih, R. Standards for household travel surveys—some proposed ideas. 10th International Conference on Travel Behavior Research in Lucerne, 2003.

Tang, L., Gao, J., Ren, C., Zhang, X., Yang, X., Kan, Z., 2019. Detecting and evaluating urban clusters with spatiotemporal big data. Sensors, 19, 461.