EXPLORING SIMILARITIES AND VARIATIONS OF HUMAN MOBILITY PATTERNS IN THE CITY OF LONDON

P. Sulis[1],*E. Manley[1]

[1] Centre for Advanced Spatial Analysis, University College London, United Kingdom
(patrizia.sulis.14, ed.manley)@ucl.ac.uk

KEY WORDS: time-series, cluster analysis, smart card data, human mobility, spatiotemporal patterns

ABSTRACT:

The availability of new spatial data represents an unprecedented opportunity to better understand and plan cities. In particular, extensive data sets of human mobility data supply new information that can empower urbanism research to unveil how people use and visit urban places over time, overcoming traditional limitations related to the lack of large, detailed data sets. In this work, we explore patterns of similarities and spatial differences in human mobility flows in London, analysing their temporal variations in relation to the liveliness measured in a number of places. Using data sourced from the Oyster smart card and Twitter, we perform a time-series cluster analysis, exploring the similarity of temporal trends amongst places assigned to each cluster. Results suggest that differences in patterns appear to be related to the central and peripheral location of places, which present two or more temporal trends over the week. The type of transport network connecting the places (Tube, Railways, etc.) also appears to be a factor in determining significant differences. This work contributes to current urbanism research investigating the daily rhythms in cities. It also explores how to use mobility data to classify places according to their temporal features, with the aim of enhancing conventional analysis tools and integrating them with new quantitative information and methods.

1. INTRODUCTION

Understanding how urban space is used over time by people is a very relevant topic in urbanism and planning research: exploring the rhythms of the places and the presence of people in streets and public spaces can reveal how valuable a place is for its users, how well it is designed, how safe it is. It can also unveil useful information about the activities and the type of people using the space during the different times of the day. The large amount of human mobility data (i.e., sourced from mobile phones and social media) recently available to researchers represents, therefore, a new, relevant source of information for urbanism. It can be employed to describe in detail the human mobility patterns across the city or to measure the qualities of places from a mobility perspective. For example, mobility data can be used to quantitatively evaluate the liveliness of places, which is one of the most renowned characteristics for an urban place. The spatial information about human mobility may also be applied to classify places according to features different from morphology: metrics based on these data sets can integrate conventional indicators frequently used in urbanism analysis (i.e., land use, urban density), therefore empowering the process of urban analysis and planning.

In this paper, we present the result of the cluster analysis of a mobility data set, performed to explore the similarities and variation of human mobility patterns in London. The analysis investigates the temporal trends and spatial patterns in terms of presence of people for a number of places across the city. Using the values of urban vitality calculated in a previous work using the Oyster smart card and Twitter data sets, we perform the cluster analysis of time-series (each of them an array of values representing the hourly variation of human mobility flows) to explore the similarity of such places, depending on spatial features that are not directly derived from morphology. We employ a specific combination of distance and cluster technique (dynamic time warping dtw and $hdbscan$) in order to dedicate particular attention to the continuity of flows over time, rather than the simple magnitude of flows interesting a place during specific hours in the day (as it happens, for instance, at peak hours in central, working areas). We believe that including the newly available mobility data in urbanism research can improve its ability to understand urban phenomena. New data and analysis tools can be integrated with conventional tools usually used in urbanism and urban planning and used to validate well-known concepts, to unveil new phenomena to urbanism and planning, and can be employed to quantitatively measure new characteristics of places to be used together with established metrics.

The structure of the paper is the following: Section 1 introduces the study, Section 2 presents a short overview of previous work, Section 3 describes the data sets used, Section 4 describes the methodology. Section 5 illustrates the results, Section 6 discusses relevance and limitation of the work, and Section 7 concludes the paper.

2. PREVIOUS WORK

Human mobility data has been not used extensively in past urbanism research, a possible reason being the lack of detailed data, the difficulties in collecting it for wider areas, the inadequacy of data to capture the urban phenomena in the correct way because collected for other purposes (i.e., transportation). Amongst previous research, the work of Jan Gehl (Gehl, 2011, Gehl and Svarre, 2013) represents a very relevant example of quantitative data about human mobility used for urbanism purpose. He developed a detailed methodology for collecting information about how public spaces are used: he extensively employed the manual collection of data, both for qualitative and quantitative information.

The deluge of new spatial data recently available (Batty, 2013, Kitchin, 2014) may overcome previous limitations, and in fact

*Corresponding author

represents an unprecedented opportunity for urbanism research to exploit new detailed information and explore spatial features from a different, collateral perspective. The newly available data sets make possible to extract quantitative information about spatiotemporal patterns of human mobility in urban space for longer time periods and for wider areas in the city (Ratti et al., 2006, Girardin et al., 2008, Hasan et al., 2012). Previous studies which appear to be particularly relevant for urbanism research explore the possibility of highlighting spatial patterns of preferences according to the different activities in the neighbourhoods (Calabrese et al., 2010), and detecting the actual boundaries experienced by inhabitants during their daily trips and rhythms in their neighbourhood (Cranshaw et al., 2012). These two examples show how the extensiveness and the fine granularity of new spatial data sets can empower urbanism tools in exploring traditional and topics and unveiling spatial phenomena previously unreachable due to the lack of information.

It may also be possible to validate well-known concepts used in urbanism theory in a quantitative way, according to the available information about human mobility and the presence of people in a place, for example measuring its vitality, which is one of the most renowned and relevant characteristics for an urban place (Jacobs, 1961). Regarding urban vitality and presence of people in places, some previous works used human mobility data to validate Jane Jacobs' idea of a relation between morphological diversity and liveliness of urban space (De Nadai et al., 2016, Sung et al., 2013). Following these new attempts, (Sulis et al., 2018) calculated the liveliness of places in London according to Jacobs' belief that the continuity and diversity of flows and presence of people in streets were an essential feature for a place to be recognised as vital. This work used mobility data as a proxy for calculating the vitality of a place in relation to the diversity and variation of flows over time. Three dynamic attributes, calculated at different temporal intervals, contribute to the final values of vitality for a number of locations in London.

Starting from this work, we developed our analysis to explore the similarity of temporal patterns of vitality for different places in London. One of the purposes of this analysis is to later use the results as a supplementary metric for classifying places according to spatial features that go beyond the morphology of places.

3. DATA DESCRIPTION

In this work, we use values of the metric of urban vitality calculated as in (Sulis et al., 2018). We measured the variations of human mobility flows as *diversity* (Dv), through three temporal indices calculated at different temporal intervals (daily, weekly, hourly, see Equation1). We use two data sets containing information about human mobility in London. One data set contains records about the use of a number of public transport modes (Tube, Overground, Railways and buses). The other data set is sourced from the Twitter API and contains information about the location and time stamp of tweets sent within the Greater London area.

$$Dv = \alpha_1 * intensity(I) + \alpha_2 * variability(V) + \alpha_3 * consistency(C) \quad (1)$$

The first data set contains information recorded through smart cards. We selected this type of data because it can reliably represent human mobility through the records of individual journeys (Roth et al., 2011, Munizaga and Palma, 2012, Zhong et al., 2014) and it can provide a good estimation of the density of human activity in cities (Zhong et al., 2016, Reades et al., 2016). This transport data set is constituted of a large collection of non-aggregated records. Each record represents a single event (i.e., a bus journey), and contains information about a specific spatial location and a time stamp. The richness and the fine granularity of the data make possible to measure and analyse spatiotemporal variations of urban flows at several intervals and to reach a detailed understanding of the variety of spatial dynamics and patterns in different areas of the city. This data set is provided by Transport for London (TfL), the authority responsible for the various means of public transport in Greater London. It consists of approximately one month of records (collected between January and February 2014) of the Oyster card, the smart card needed to get access to the public transport network. The card is used for entering and exiting through the ticket barriers of the rail network (Underground, Overground, DLR etc.), and for boarding the buses. Each recorded transaction represents an individual journey through the transport network, it is identified by a unique ID code and contains a significant amount of information about the journey.

For the rail network, the details recorded for each transaction and used for this analysis are:

1. a unique ID code identifying the journey
2. a code identifying the transportation mode of the journey
3. a code and a name identifying the station
4. a code identifying if the user is entering or exiting the station
5. a code identifying the day of the journey
6. a time stamp for the entrance or exit of the station, identifying the start or the end of the journey

For the buses, the Oyster card is required only when boarding. The details recorded for each transaction and used for this analysis are:

1. a unique ID code identifying the journey
2. a code identifying the transportation mode of the journey
3. a code and a name identifying the bus stop
4. a code identifying the day of the journey
5. a time stamp for the boarding of the bus, identifying the start of the journey

The limitations of the data set are related to the size of the sample and the spatial distribution of the information recorded. Although the data set is very rich in details, it has constraints in time extension due to TfL privacy policy. Regarding seasonal variation, according to related work using data from the same source (Reades et al., 2016) they do not significantly influence the data trends, therefore we are confident with the reliability of the data set used. About the spatial distribution of the records in the data set, we decided to concentrate our analysis within localised areas of approximately 400 metres around the stations of the rail network (Tube and Railways) to obtain spatially detailed results rather than aggregated ones for larger. It is important to notice that this spatial constraint is common to other types of mobility data (i.e., data sourced from mobile phones): it is intrinsically linked to the locations of sensors recording the information (in this case, ticket barriers at stations).

To obtain the values of the vitality metric, we then combined the indices of diversity with weights (obtained through a regression model) that represent the density of human activity for each area in the study, as in Equation 2.

$$tweet\ count \sim \alpha_1 * intensity(I) \\ +\alpha_2 * variability(V) + \alpha_3 * consistency(C) \quad (2)$$

As in previous work (Traag et al., 2011, Hawelka et al., 2014), we chose Twitter data as a measure of this density. We collected the Twitter data through their public API for three months (January-March 2016, a time frame similar to the one used in the Oyster card), and we used the total amount of tweets sent for each area of our study. The final values of vitality used in the following cluster analysis are therefore a combination of two types of mobility data. Each value represents the liveliness of each area of study according to the variation of human flows in time. Liveliness is quantitatively estimated by measuring how continuous flows and the presence of people in a place are. In this way, we have a measure not only of the magnitude of the flows but also of their diversity in that particular area for a specific point in time (daily, weekly etc.).

4. METHODOLOGY

In this study, we are interested in exploring the similarity of patterns and spatial differences in human mobility flows amongst a number of areas in London. In particular, we are interested to understand if it is possible to recognise similarities in the variation of temporal trends about the presence of people in places, similarities that can be later used to classify such places.

To evaluate similarity, we employ the hourly variation of the vitality metric to perform a clustering analysis of time-series amongst the areas of our study. The cluster technique ($hdbscan$) and the distance metric (dynamic time warping, dtw) employed in the analysis are selected according to the type and format of our data (time-series, each of one a 1-D array of hourly values of vitality) and the purpose of our work, which is evaluating the similarity in the variation of the temporal trends rather than the mere variation of magnitude of flows, as our interest lays mainly on understanding temporal patterns of the flows and the continuity of the presence of people in places.

The first part of the analysis uses the vitality values of an average week day and weekend day. The average hourly trends (an array of 20-length vectors) for all areas in the data set are clustered, and areas are assigned to clusters and visualised in maps in order to evaluate the spatial distribution of the labels. The visualisation also helps to evaluate the coherence of the performance of the $hdbscan$ clustering algorithm in accordance with the empirical knowledge of the urban dynamics in the areas. The second part of the analysis employs instead daily, non-aggregated hourly time-series (no average values). The algorithm clusters the temporal trend of each area for all days in the data set. This analysis shows how the areas are assigned to different clusters across several days and make possible to observe the variation of temporal trends over time, which is also helpful to identify routines and anomalies in the weekly variation of the use of that place.

4.1 Selection of distance and algorithm

Measuring similarity amongst time-series can be a complex matter because the focus must be on the similarity of trends and not merely on magnitude of values. Therefore, the choice of the metrics employed might not be straightforward. In our analysis, we needed to select a combination of distance and cluster algorithm suitable for our type of data and the objective of our work: evaluating the similarity of multiple temporal trends (time-series) that represent a spatial phenomenon actually happening in urban places.

The literature suggests a number of possibilities for time-series cluster analysis (Tan et al., 2005, Warren Liao, 2005, Esling and Agon, 2012, Aghabozorgi et al., 2015). One of this is using a metric called Dynamic Time Warping (Kruskal and Liberman, 1983, Bundy and Wallen, 1984, Minnaar, 2014).

The dynamic time warping distance [1] finds the optimal non-linear alignment between two time-series (in this case a matrix of vectors). It can be used as a measure for distance and similarity: the optimal path defines the 'distance' between two given sequences and the similarity between 2-D time-series numpy arrays. The input is an array of vectors (our hourly values of vitality, in the form of a matrix of normalised pair distance, the output is a superior diagonal matrix.

Cosine dissimilarity (calculated as $1 - cosinesimilarity$) is another metric suggested for the time-series cluster analysis (Perone, 2013). Similarly to dtw, the judgement is on the orientation of temporal trends (similarity in patterns and trends) and not on the mere value of vitality. It measures whether two vectors are pointing in the same direction (therefore are similar), where vectors here are our time-series of vitality values.

Both the metrics (dtw and $cosine$) evaluate the distance between the time-series objects based on the temporal trends so that the attention is on the similarity of mobility patterns and not only on the intensity of flows over time. To do that, the matrices obtained when calculating the distances are normalised dividing each row of the matrix for the maximum value of the row, which represents the maximum hourly value of vitality for each day and area. In this way, the distances are calculated based on the similarity of temporal trends, and not on the magnitude of vitality, coherently with the understanding of urban vitality we considered in this work.

We applied both the distance metrics with various cluster algorithms that we considered appropriated for our data set, and we then evaluated the results obtained in the cluster analysis. Since we choose a distance that is specific for time-series analysis (i.e., no Euclidean distance), we also selected a number of unsupervised learning techniques that are appropriated for these distances: in this case, density-based spatial clustering applications with noise. Amongst the advantages of these techniques, they do not assume that the clusters have spheric shape, they remove noise points from the data set, they present a non-flat geometry and an uneven cluster size. Disadvantages are mainly represented by the fact that they require a fine-tuning of parameters, i.e. a good combination of $MinPts$ and e for DBSCAN. We selected the most appropriate algorithm for our data set through a comparison of different clustering methods using a similar geometry: DBSCAN, $hdbscan$, Mean-shift. After a number of attempts, we

[1] the Python package used is available at https://pypi.python.org/pypi/dtaidistance/0.1.4

selected *hdbscan* as it performs better with our data set. In particular, *hdbscan* (Campello et al., 2013, McInnes et al., 2017) is more robust than DBSCAN, and returns an acceptable cluster result with little or no parameter tuning (primary parameters = minimum cluster size): this was also ideal for the preliminary analysis we performed at the early stages of this work.

Results obtained using dynamic time warping and cosine dissimilarity distance with the *hdbscan* algorithm are comparable and very similar in the assignment of average and daily trends to each cluster. However, in this specific case and for this specific data set, we think that dynamic time warping is preferable because final results include a higher number of clusters and fewer objects labelled as noise. In our case, this makes possible to develop a more detailed analysis of the trends, especially in relation to their spatial distribution, which is a central point in our work.

We based the selection of the algorithm on:

1. the optimal number of clusters: too many clusters are confusing for the evaluation of results, too few may neglect interesting niche trends (for both average and daily values)
2. the spatial distribution of labels in the city and the comparison with empirical knowledge based on spatial flows and rhythms (i.e., tube and train stations belong mainly to different labels)

From our evaluation, the combination of *dtw* metric $+hdbscan$ appears to be the most appropriate method to be applied to this dataset. The joint evaluation of spatial distribution, temporal trends and list of areas for each cluster appear to be coherent with our empirical understanding of the urban dynamics in the city of London.

4.2 Reassignment of noise objects

As previously mentioned, the *hdbscan* algorithm does not assign all objects of the data set to the clusters: some of them are labelled as 'noise' if the algorithm cannot assign them to any label because of the parameters set in advance. The objective is therefore to find an appropriate balance between the number of clusters obtained (not too many for the purpose of results evaluation) and the number of objects labelled as noise. Once we found the most appropriate number of clusters for our data set, we then considered to re-assign some of the objects labelled as 'noise', in order to recover at least a number of objects that are close to the clusters. We built a function that considers the distance between each object labelled as 'noise' and the set of objects already assigned to each of the clusters.

This function follows different steps:

1. it calculates the distance amongst the object to assign and each object already belonging to each cluster;
2. it calculates the average distance amongst the 'noise' object and the different clusters;
3. it assign the object to the cluster with the smaller average distance;
4. in order not to lose precision, it is possible to set a threshold for a minimum average distance required for the object to be re-assigned.

We considered different parameters when evaluating the distance to reassign the objects (min, mean, max distance). Results show similar outcomes, therefore we choose the mean distance to further develop the analysis.

One limitation to consider in the re-assignment of 'noise' object is the loss of precision (after all, some data was not assigned by the algorithm). Setting a threshold in the re-assignment function (point 4) can be a possible solution. In this case, part of the 'noise' objects is not re-assigned.

number of cluster	number of areas
0	13
1	22
2	32
3	30
4	57
5	74
6	123
7	111

Table 1. number of areas per cluster

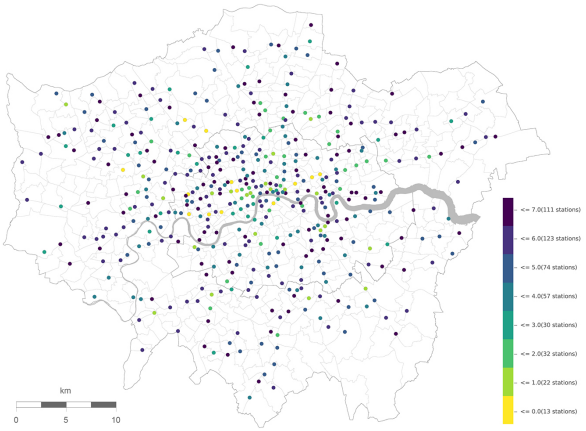

Figure 1. Cluster assignment for an average week day

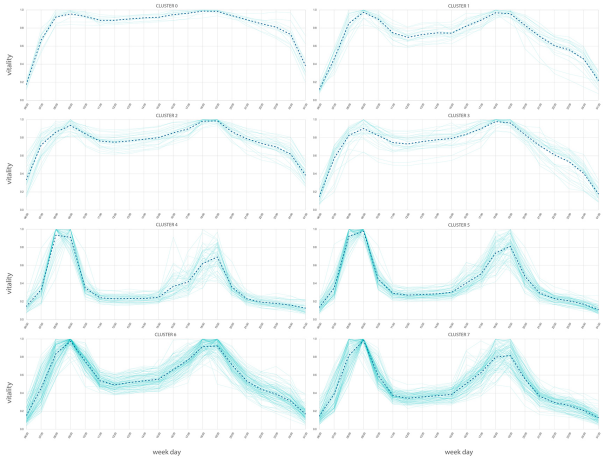

Figure 2. Temporal trends for clusters for an average week day

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-W11-51-2018 | © Authors 2018. CC BY 4.0 License.

5. RESULTS

5.1 Cluster analysis of average days values

In this section, we present the main results of the cluster analysis. The spatial distribution of the labels shows a wide variety in the temporal trends of areas during the week (Figure 1), whereas the weekend seems to indicate more heterogeneity amongst central areas and peripheral areas. Temporal profiles (Figure 2) show two main groups representing a clear two-peak profile (following the commuting rhythms, see profiles below) and a more continuous profile (which represent a more lively place interested by flows possibly attracted by a variety of activities over time, see profiles above).

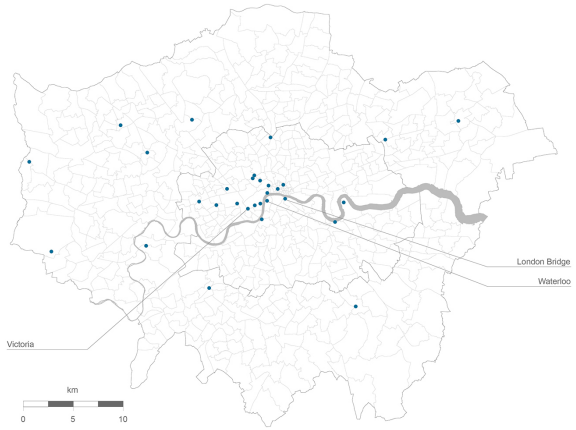

Figure 3. Cluster number 3: central, touristic areas

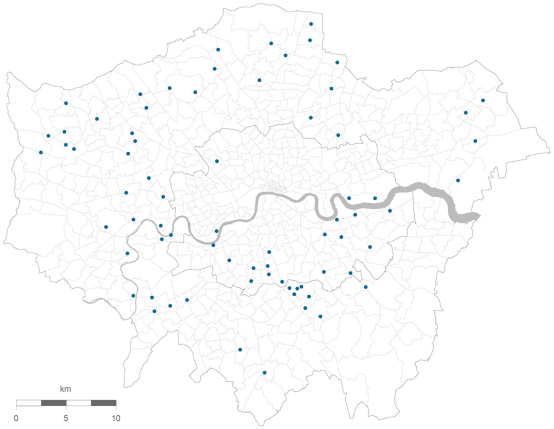

Figure 4. Cluster number 5: peripheral, residential areas

Weekend trends show more noise in the results, possibly because of the greater heterogeneity of rhythms, generally less related to routine behaviours (as commuting) happening during the week. Moreover, weekend profiles do not show the two peaks profile so distinctly, rather almost a 'bell' shape around the central hours of the day, and a long tail in the late hours, an indicator of night activities.

A detailed analysis of the spatial distribution of the labels in the city makes possible to visualise interesting results suggesting how different areas act similarly depending on their locations or the type of transport network that serves the area. If we compare two different clusters, we can observe how cluster 3 (Figure 3) presents a slightly two peak profile, with the afternoon and the evening showing a constant flow and presence of people in the areas, and a softer decline in the flows towards the night. Places belonging to this cluster show an uneven spatial distribution, with many areas within the inner London boundary. However, looking at the names of the areas, we can observe that the majority of them share a common trait: they are areas that attract touristic flows and includes Heathrow Airport and three main stations in central London (London Bridge, Victoria, Waterloo). Cluster 5 (Figure 4) instead shows a very defined two-peak behaviour, with the morning peak at a higher value of vitality and a neater profile, indicator that the morning peak time coincides for the majority of the commuters of the areas. The evening peak instead appears fuzzier and shows many irregularities.

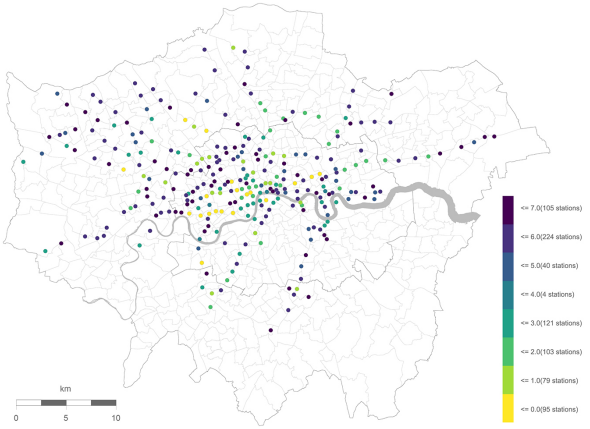

Figure 5. Cluster assignment for areas served by Tube

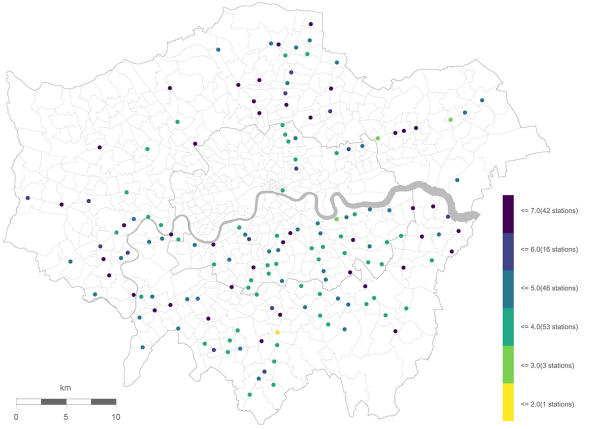

Figure 6. Cluster assignment for areas served by trains

This can have a number of reasons: people may finish work at different times, or they delayed their return journey attending to other activities (shopping, pub or other leisure activities after work). The night tail may suggest that areas in this cluster are mainly residential. And in fact looking at the spatial distribution of the cluster label, it shows mainly areas outside the inner London boundary, in the periphery where residential areas are common.

Another interesting example is represented by the spatial distri-

bution of the labels selected according to the type of transport network that serves the various areas of the city. Observing the maps showing the areas served by Tube stations and Railways stations, two main differences appear. The first one is mostly geographical, with the Tube areas and the Railways area neatly divided in a North-South line by the river Thames. This first geographical difference might also have an influence in the second one, related to the temporal patterns of mobility flows in the areas, and inferred by the cluster assignment. Whereas areas served by Tube stations (Figure 5) show a heterogeneity in the cluster assignment, with many places showing a continuous temporal trend across the day, areas served by Railway lines (Figure 6) show a distinct prevalence of the two-peak behaviour, possibly driven by the prevalence of mobility flows related to commuting from the residential areas towards the city centre (at least during the week days).

5.2 Cluster analysis of non-aggregated daily values

After analysing the results obtained for the average week and weekend days, we now show the results of cluster analysis of all daily time-series for the several days in our data set. The objective of this analysis is to observe the consistency of label assignment of each area across the different days of the week. We can observe a distinct difference from Monday to Saturday (Figure 7). This difference is even more visible when observing the distribution of labels assigned across the days (Sundays were excluded from analysis because many data were missing). During the week days, the majority of the areas are assigned to a label characterised by a two-peak profile, which is somehow expected given the routine of week rhythms, driven by commuting for work, schools and similar activities. During the weekend, results are more varied and show that the label assignment almost split in two, with many areas still assigned to the previous two-peak profile but with a majority now assigned to a label that show the typical weekend profile with a continuous, 'bell' shape. In both the cases (week and weekend days), although a clear majority is assigned to only one label, the interesting part of the analysis is represented by the results outside the main label, that show a clear variety of behaviours happening besides the main common one (future work will focus on this aspect).

label changes	number of areas
1	80
2	125
3	147
4	104
5	41
6	21
7	5
8	5

Table 2. number of areas per label changes

Following these results, we were interested in exploring how many labels each area was assigned to during several days, in order to understand the variability or regularity of the temporal trends of each area (Figure 8). The majority of areas shows a change from 2 to 3 labels over the week, possibly related to the week-weekend alternation. Some areas are assigned every day to the same label, others change four or more labels during the same time. Furthermore, there are areas that show the very same identical label patterns, which means they behave in the very same way across the days, although they do not share spatial proximity. Identical label patterns are most common for areas with 1 and 2 label change, whereas there are no common patterns for 4 and more label change. Besides the identical patterns, we also explored the similarity of label patterns for each group of areas that share the same number of label change. Therefore, we calculated the covariance and correlation amongst the time-series of the areas of each group, to explore if the areas changed towards the same labels in each change. Preliminary results are not definitive, therefore this aspect will be the focus of future work in this research.

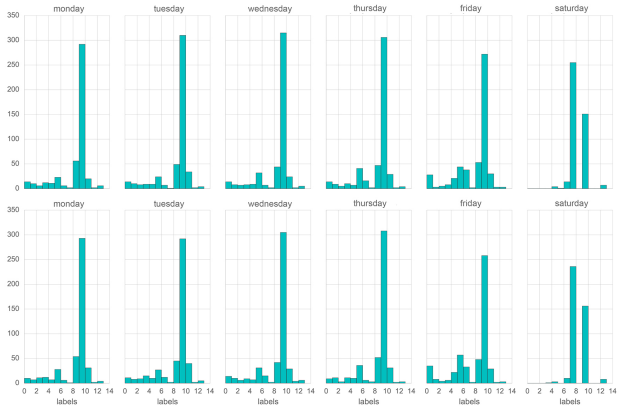

Figure 7. Number of areas assigned to each cluster

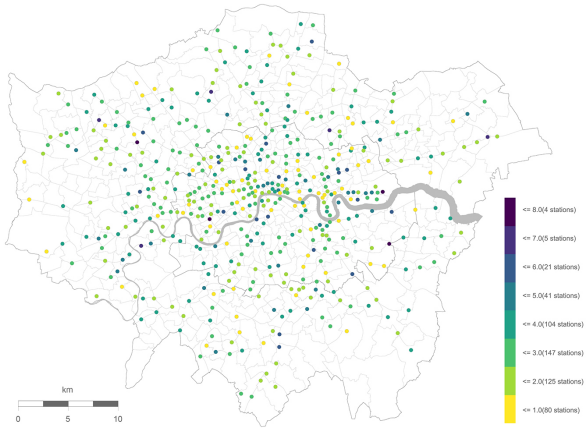

Figure 8. Number of label changes for each area

6. DISCUSSION

As mentioned in the introduction, our main interest in this study is to unveil similarities and differences in the spatiotemporal patterns of human mobility and presence in places for various areas across London. Results show that similarities exist, also amongst places that not necessarily share spatial proximity. One of the possible applications of our work is to use these outcomes to define a quantitative indicator to classify places according to features that measure human mobility and daily rhythms in the urban space. This classification can be used in the process of analysis and planning of places in cities. In this case, our indicator is obtained from data about the actual presence of people in places and not, as sometimes happened in past works, deduced from the

place morphology. Furthermore, this additional metrics can be used together with other spatial features (topology, morphology etc.) to compose a taxonomy of places according to a specific combination of multiple indicators. For example, a place can be identified as a generator of human mobility when a continuous high level of vitality is present in correspondence of a variety of activities that are spatially diffused in a certain area, a variety that attracts diverse people (i.e., inhabitants and strangers) during different times of the day. This is only an example of how such a combined metric can contribute to a more comprehensive classification of urban places. Future work will focus on this possible application to urbanism research. More generally, our work contributes to the exploiting of new spatial data available, especially data about human mobility at such a high resolution, that can empower urbanism research in exploring well-known concepts, validating them and unveiling new elements that are relevant to the spatial analysis and the consequent planning of cities.

Inevitably, there are also some limitations in this work, mainly related to the data set and the methodology used:

1. the original data set presents a bias about the spatial distribution of the locations where the data are recorded, depending on the location of the sensors. Furthermore, despite its extensiveness, the data set presents some missing records that led to exclude some of the days from the analysis: this happened for some of the weekend days, which also happen to be the most interesting to analyse because of the wider variety of behaviours in comparison to the week routines;
2. the fact that the $hdbscan$ algorithm labels some of the objects as 'noise' inevitably means that the analysis cannot be completed for all the objects in the data set. The reassignment technique that we used does not completely resolve this issue;
3. a process of validation with other types of data would be useful, also to evaluate how well the methodology is capturing the phenomena. We did a preliminary validation by observing the patterns emerging from the analysis and recognising the coherence of their spatial distribution with our empirical knowledge of the city. However, in future work, we aim at comparing the distribution of the labels representing a specific temporal pattern with established urban indicators, such as land use or the spatial distribution of activities, for which we have collected data using different sources, including open-source database such as OpenStreetMap. This validation may clarify if and how the similarity in terms of the liveliness of the place is related to the spatial distribution of specific activities (i.e. leisure activities that attract a variety of people over time).

7. CONCLUSION

In this paper, we present the results of the data analysis performed to explore the similarities and variation of human mobility patterns in London. The study aims at understanding how places across the city present similar temporal trends and spatial patterns in terms of presence of people in places. The analysis was performed through an unsupervised learning technique that clustered time-series representing the hourly values of the urban vitality of several areas across the city. Results show that the similarity appears to be driven by flows and phenomena characteristic of the places (as tourism), by the spatial locations of the places (city centre and periphery), by the type of transport network that serves the areas (Tube/Railway network). Results also show that the areas tend to present different temporal trends (and are therefore assigned to different clusters) across several days, meaning that the patterns of human mobility and presence in places vary over time. We will further investigate these differences in the temporal patterns of places. We believe that this type of work can be useful to urbanism research to empower analysis and in particular a classification of places in cities based on spatial features that not necessarily are derived from the morphology of places. It can also be relevant to planners and policy makers in the process of planning new parts of cities, or regenerating old ones through the introduction of new policies (i.e., pedestrian areas) or activities that can encourage urban vitality. Future work will focus on the integration of these results with other conventional urban data, also using them as a method of validation for the cluster analysis results (i.e., land use and activity distribution).

ACKNOWLEDGEMENTS

The authors would like to thank Transport for London for providing the data set used in this work.

REFERENCES

Aghabozorgi, S., Seyed Shirkhorshidi, A. and Ying Wah, T., 2015. Time-series clustering A decade review. *Information Systems* 53, pp. 16–38.

Batty, M., 2013. Big data, smart cities and city planning. *Dialogues in Human Geography* 3(3), pp. 274–279.

Bundy, A. and Wallen, L., 1984. Dynamic Time Warping. In: A. Bundy and L. Wallen (eds), *Catalogue of Artificial Intelligence Tools*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 32–33.

Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liu, L. and Ratti, C., 2010. The geography of taste: analyzing cell-phone mobility and social events. In: *Pervasive computing*, Springer, pp. 22–37.

Campello, R. J. G. B., Moulavi, D. and Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In: *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 160–172.

Cranshaw, J., Schwartz, R., Hong, J. I. and Sadeh, N. M., 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In: *ICWSM*.

De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D. and Lepri, B., 2016. The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective. *arXiv:1603.04012 [physics]*. arXiv: 1603.04012.

Esling, P. and Agon, C., 2012. Time-series Data Mining. *ACM Comput. Surv.* 45(1), pp. 12:1–12:34.

Gehl, J., 2011. *Life Between Buildings: Using Public Space*. Island Press. Google-Books-ID: X707aiCq6T8C.

Gehl, J. and Svarre, B., 2013. *How to Study Public Life*. Island Press. Google-Books-ID: DUGiAQAAQBAJ.

Girardin, F., Calabrese, F., Fiore, F., Ratti, C. and Blat, J., 2008. Digital Footprinting: Uncovering Tourists with User-Generated Content. *IEEE Pervasive Computing* 7(4), pp. 36–43.

Hasan, S., Schneider, C. M., Ukkusuri, S. V. and Gonzlez, M. C., 2012. Spatiotemporal patterns of urban human mobility.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. and Ratti, C., 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41(3), pp. 260–271.

Jacobs, J., 1961. *The life and death of great American cities*. New York: Random House.

Kitchin, R., 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kruskal, J. and Liberman, M., 1983. *The Symmetric Time-Warping Problem: From Continuous to Discrete.* Addison-Wesley.

McInnes, L., Healy, J. and Astels, S., 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2(11), pp. 205.

Minnaar, A., 2014. Time Series Classification and Clustering with Python.

Munizaga, M. A. and Palma, C., 2012. Estimation of a disaggregate multimodal public transport OriginDestination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies* 24, pp. 9–18.

Perone, C., 2013. Machine Learning : Cosine Similarity for Vector Space Models.

Ratti, C., Williams, S., Frenchman, D. and Pulselli, R. M., 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning b Planning and Design* 33(5), pp. 727.

Reades, J., Zhong, C., Manley, E. D., Milton, R. and Batty, M., 2016. Finding pearls in London's oysters. *Built Environment* 42(3), pp. 365–381.

Roth, C., Kang, S. M., Batty, M., Barthlemy, M. and Perc, M., 2011. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE*.

Sulis, P., Manley, E., Zhong, C. and Batty, M., 2018. Using mobility data as proxy for measuring urban vitality. *Journal of Spatial Information Science*.

Sung, H.-G., Go, D.-H. and Choi, C. G., 2013. Evidence of Jacobss street life in the great Seoul city: Identifying the association of physical environment with walking activity on streets. *Cities* 35, pp. 164–173.

Tan, P.-N., Steinbach, M. and Kumar, V., 2005. *Introduction to data mining*. Pearson Addison Wesley, Boston. OCLC: 58729322.

Traag, V., Browet, A., Calabrese, F. and Morlot, F., 2011. Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pp. 625–628.

Warren Liao, T., 2005. Clustering of time series dataa survey. *Pattern Recognition* 38(11), pp. 1857–1874.

Zhong, C., Arisona, S. M., Huang, X., Batty, M. and Schmitt, G., 2014. Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science* pp. 1–22.

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F. and others, 2016. Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data. *PLoS ONE* 11(2), pp. e0149222.