GLOBAL URBAN CITYSCAPE - UNSUPERVISED CLUSTERING EXPLORATION OF HUMAN ACTIVITY AND MOBILITY INFRASTRUCTURE

TANIA PAPASOTIRIOU[1] and STEPHAN CHALUP[2]
[1,2]*The University of Newcastle, Australia*
[1]*Soultana.papasotiriou@uon.edu.au*
[2]*stephan.chalup@newcastle.edu.au*

Abstract. It is widely accepted that cities cultivate innovation and are the engines of productivity. The identification of strengths and weaknesses will enchant social mobility providing equal opportunities for all. The study at hand investigates the relationship between social mobility and transportation planning in 1,860 central urban areas across the globe. Datamining processes combining open-sourced, automated, and crowdsourced information from four major pillars of social mobility (demographics, human activity, transport infrastructure, and environmental quality) are used to describe each location. Next, unsupervised clustering algorithms are used to analyse the extracted information, in order to identify similar characteristics and patterns among urban areas. The process, which comprises an objective framework for the analysis of urban environments, resulted in four major types of central areas, that represent similar patterns of human activity and transport infrastructure.

Keywords. Information retrieval; similarity measures; computer methodologies; unsupervised clustering; urban performance.

1. Introduction

In early 2020 the World Economic Forum performed a holistic assessment of 82 countries, classifying them according to their performance on critical aspects of social mobility (WEF, 2020). One of the main findings of this study is the need to develop global momentum to tackle the inequality that arises from new social mobility challenges. However, these challenges are not spatially homogenous as cities are complex and dynamic bodies of mixed personal activities scattered across various locations. Understanding the spatial variability of urban structures is paramount for operators and stakeholders to define effective management policies and services, such as sustainable and innovative transport planning. As cities around the world are now recognising the value of tracking their performance and progress, various instruments, indicators and targets are being used to benchmark, identify and promote sustainable transport policies.

In parallel, the rapid growth of data availability and the exponential development of digital platforms allow the detection of mobility patterns in urban

PROJECTIONS, Proceedings of the 26th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2021, Volume 2, 539-548.

areas with more detail than ever before. As these patterns are highly related to spatial interaction and land use (Rodrigue, 2020), capturing the heterogeneity of urban ecosystems by means of digital data is one of the most important lines of research within the field of Urban Computing. Numerous studies are seeking for solutions to capture the dynamic characteristics of urban settlements, and different technologies and techniques are used to gather and analyse urban data produced by static, mobile and crowdsensing sensors (Zheng, 2019). While spatiotemporal variations in human mobility, urban functional regions and city-wide human activity have been studied extensively (i.e. Becker et al., 2013; Caceres & Benitez, 2018; Zhang et al. 2020), important gaps remain. Researchers have pointed out that most case studies are often limited to local city contexts and are based on the use of isolated datasets (Regmi, 2020). Moreover, traditional and static methods do not apply to real settings, where various functional areas perform quite differently depending on time i.e. on working weekdays and weekends/holidays (Terroso Saenz, et al., 2020). New studies of urban analytics (Calafiore, et al., 2021) highlight the power of open data in a variety of contexts worldwide. However, the attempt to unpack contemporary mobility, diversity and similarities of human dynamics is once more restricted to small samples of cities and isolated data sources.

Building on these arguments, the study at hand is combining heterogeneous sources to interrogate the characteristics of urban areas with dynamic use of land, and high human activity. A clustering analysis is employed to identify similarities and analogies of urban settings, highlighting the relationship between human activity and transport infrastructure across the globe. Given the enormous amount and diversity of digital data produced from different aspects of urban life, this study proposes an objective framework that allows researchers to extract knowledge from open-source raw data, that can be updated dynamically. This framework can support decision-makers in evaluating and benchmarking the current state of any urban location, as well as tracking the impact of interventions in the urban environment.

The paper is structured as follows: The first part, focused on data selection, presents a series of indicators used to describe the human activity and transport infrastructure, that are available for any given location around the world. Moreover, it outlines the datamining process and discusses sources of information used in the study. The second part is dedicated to the clustering process. It demonstrates the use of t-SNE algorithm (Van der Maaten & Hinton, 2008) to identify and represent similarities among the selected entities, and the application of the Mean Shift algorithm (Cheng, 1995) to detect clusters that depict similar urban behaviours. Next, an overview of the resulting cluster outcomes is presented, followed by a discussion outlining the main conclusions.

2. Data collection

2.1. SELECTION OF URBAN INDICATORS

Several efforts have been made towards the establishment of social mobility and transport infrastructure indicators for urban areas (Arcadis, 2017; Chestnut

& Mason, 2019; Dixon et al., 2019). However, the indicator sets used in the mentioned studies vary widely, and consensus towards a widely accepted system that can be applied across cities is yet to be reached. Moreover, local regulations make the application of some assessment tools in certain regions of the globe questionable, or even invalid (Macedo et al., 2017).

In this study, we combine three types of data extracted from Open APIs or publicly available databases, specifically: directed (open-sourced and free licensed), automated (scraped or constructed), and volunteered (crowdsourced). Extraction of the data took place from 8/6/2020 until 15/6/2020. Table 1 is showing a summary of the selected indicators and their sources. The indicators correspond to major pillars of social mobility such as demographics, transport infrastructure, human activity, and environmental quality. The methods used to extract the data are outlined in the following.

Table 1. Summary of domains and indicators of the urban profile dataset.

Domain	Indicator	Attributes	Source	Type
Areas of interest	location	City , Country	(UN, 2019)	Directed
	centroid coordinates	Latitude, Longitude	(UN, 2019)	Directed
	centroid type	type of centroid	(OSM, 2020)	Directed
	isochrone radius	radius of urban unit	(OSM, 2020)	Automated
Demographics	population	Number of residents	(UN, 2019)	Directed
Human activity	POI / number of Venues	Arts & Entertainment, College & University, Event, Food, Nightlife, Spot Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport	(Foursquare, 20	Volunteered
	Walkability	Walk score	(Walkscore, 20	Automated
Transport infrastructure	Pois / Mobility	parking ,taxi hubs, public & transportation	(OSM, 2020)	Automated
	Network infrastructure	street length total, street length average, street segments, intersections, street density, intersection density	(OSM, 2020)	Automated
Environmental quality	Air quality	Air Quality Index (AQI)	(AERIS, 2020)	Automated

2.2. DESCRIPTION OF DATA MINING

Areas of interest: We consider 1,860 cities with a population that exceeds 300,000 residents from around the globe. The 2018 Revision of World Urbanisation Prospects (UN, 2019) contains an extensive database with details regarding the population and the geographical characteristics (latitude and longitude) of the centroid of each city. The description and type of each centroid (city centre or urban agglomeration centre) is provided from taginfo (OSM, 2020). Here we focus on the central locations of each city, defined as the area within 30 minutes walking distance (travel speed 4.5km/h) from the centroid. Each area corresponds to an "urban unit". For the calculation of the isochrone radius r_{iso} of each urban unit we have used Open Street Maps (OSM) (Haklay & Weber, 2008) combined with the OSMnx Python package (Boeing, 2017a, 2017b)

Transport infrastructure: The same sources that were employed to calculate the isochrone radius (OSM and OSMnx) were used to retrieve mobility Points of Interest (POIs) such as parking facilities, taxi stops and public transport stops within the urban units. Moreover, through the statistics modules of OSMnx we

have identified the geometrical characteristics of the road network of each unit such as total street length, average street length, number of street segments (S), and number of intersections (I). The last two attributes were used for the calculation of street density (D_s) and clean intersection density (D_i) as follows: $D_s = \frac{S}{A}$ and $D_i = \frac{I}{A}$ where A refers to the isochrone area and is calculated as $A = \pi \cdot r_{iso}^2$

Human activity: The application Foursquare (Foursquare, 2020) provides a global database of user content on venue data. This information is organized in a hierarchical order of multiple levels and the selection of the hierarchical level is relevant to the scale of this study. Through the application's API, we retrieved 10 high-level venue categories that describe popular human activities: Arts & Entertainment; College & University; Event; Food; Nightlife; Outdoors sports & Recreation; Professional & Other Places; Residence; Shop & Service; Travel & Transport. Moreover, to obtain a better understanding of the character of human activity, we used again open API's to identify the walk score (Walkscore, 2020) of the centroid of every location.

Environmental quality: To describe the environmental characteristics of each urban unit we extracted the Air Quality Index (AQI) for each centroid corresponding to 4 periods of a working day (every 6 hours) through the AERISweather API (AERIS, 2020). The final AQI score was the mean value of the extracted information.

2.3. DATA PROCESSING

Following data collection, the dataset was processed, and the gained information that used to calculate the street and intersection density and total air quality (i.e. street length, street segment count, intersection count, radius and AQI metrics) was excluded from the final selection. The dimension of the final dataset is (1860, 19) corresponding to 19 indicators that describe the use and behaviour of the 1,860 urban units.

Before proceeding to the next part of the analysis, we need to normalise our dataset and bring the different ranges of raw features to a common scale. This has been achieved by scaling each indicator to values between 0 and 1 via the sklearn MnMaxScaler function. The normalised valued X_{scl} for each indicator X is given as: $X_{scl} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$, where X_{max} and X_{min} is the maximum and minimum values of the indicator across the dataset, respectively.

It is worth noting that the abovementioned urban attribute dataset was compiled from publicly available sources. No imputation methods were used during or after data collection and no commercial sources were used for this analysis, except the AERIS Dataset, a trial version of which was sourced for the needs of this study.

3. Clustering process

Compilation of the dataset described above allowed interrogating the characteristics of urban units to detect similarities and patterns in human

activity and transport infrastructure.

To uncover sets of urban units with similar characteristics we employed an unsupervised clustering prosses of the abovementioned high dimensional space of 19 indicators. The process takes place in four steps (Algorithm 1): After pre-processing and data scaling, the high dimensional data are projected into a 2D map using the t-SNE algorithm, and then we apply Mean-shift clustering to identify clusters of similar patterns. Finally, we plot the clusters to overview and interpret the produced clusters. This procedure is described in the following algorithm.

```
*Algorithm 1: t—SNE and Mean shift clustering*
For each domain
Scale  data —-> X_scl
For perplexity in range (10, 50,10)
    Compute t—SNE (X_scl)
    Select the min KL divergence —-> t—SNE
    Plot t—SNE
For n_samples in range (10, 50,5)
    Compute bandwidth (n—samples) —-> bandwidth
    Means shift clustering (bandwidth,t—SNE)
    Estimate the number of clusters —-> cluster labels
    Plot clusters
Make Plots —-> boxplots and basemaps
```

3.1. DIMENSIONALITY REDUCTION

An algorithm capable of transforming high-dimension databases in two- or three-dimensional vectors that can be easily visualised in a scatter plot is the ‘t-Distributed Stochastic Neighbour Embedding’ or t-SNE for short (Van der Maaten & Hinton, 2008). Generally, t-SNE is employed to project high-dimensional data to a low-dimensional two-dimensional XY space, while preserving much of the local structure. This is achieved by minimising the Kullback-Leibler divergence between the original distribution and probability distribution of points in the low-dimensional map.

The low-dimensional projections are represented in such a way, that nearby points correspond to similar objects and distant points correspond to dissimilar objects. This facilitates visualising isolated cluster structures. The t-SNE algorithm due to its cost function can produce different results under different initializations. In the study at hand, we used the default algorithm parameters (refer to sklearn.manifold.TSNE) except of perplexity, which was set to 50 to gain a better understanding of global geometry. In addition, the number of iterations was set to 10000 (Wattenberg et al., 2016). The outcome of the clustering process is graphically depicted in Figure 1.

3.2. MEAN-SHIFT CLUSTERING

Although several clustering techniques can be applied to the resulting 2D feature space, we have selected the Mean-shift algorithm (Comaniciu & Meer, 2002) on the grounds of its lower computational requirements, its ability to find groups with

various shapes, and the low number of hyper-parameters that require fine-tuning. Mean-shift is capable of identifying core clusters that are generating strong stance and outliers with a small influence that will not belong to a cluster. The aim is to identify peaks of densities in the feature space using Gaussian kernels. Each point is iteratively shifted towards the mean of all the points within the kernel until all points converge to a local maximum of density nearby them, hence group into the same cluster. The radius of the kernel (bandwidth) determines the number of peaks detected from the algorithm and it can be estimated automatically using cross-validation in a probabilistic setting.

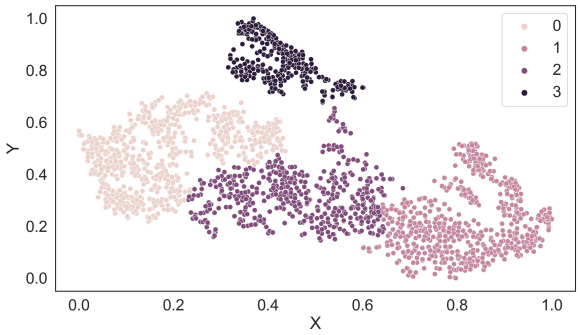

Figure 1. Visualization of the t-SNE two-dimensional feature space highlighting the clustering outcome of the Mean-shift algorithm.

For the application at hand, we have used the sklearn implementation, testing the default parameters for bandwidth estimation. To evaluate the stability of the Mean-shift outcome we used the Calinski-Harabasz Index (CH) (Caliński & Harabasz, 1974; Wang & Xu, 2019). Since there are no ground truth labels, the evaluation is performed using the model itself and higher CF scores respond to models with better-defined clusters (Table 2). Moreover, in Table 3 we compare the outcome with other clustering algorithms. A detailed description of the way these algorithms operate is outside the scope of this paper. The algorithm resulted in the identification of 4 clusters, no outliers were detected (Figure 1) and the produced labels were assigned to each urban unit.

Table 2. (Left) Evaluation of clustering stability of Mean shift outcomes with different bandwidth. Table 3: (Right) Comparison of Calinski-Harabasz Index between different clustering algorithms.

N_samples	Number of clusters	CH score
10	9	3280.80
15	7	3265.46
20	5	3262.57
25	4	3229.74
30	4	3314.54
35	4	3255.21
40	4	3308.85

Clustering algorithm	Number of clusters	CH score
Mean shift	4	3314.54
Birch	4	2989.87
K Means	4	3101.19
AGNES	4	947.19
DBSCAN	4	96.0222

4. Clustering overview

The clustering process discussed above resulted in 4 major types of urban units (Clusters) scatter across the globe, shown in Figure 2. Our initial assumption that urban units in each constellation will have similar stance reveals the characteristics of each group. Figure 2 also depicts the general behaviours regarding each feature. The characteristics of each Cluster are briefly outlined in the following.

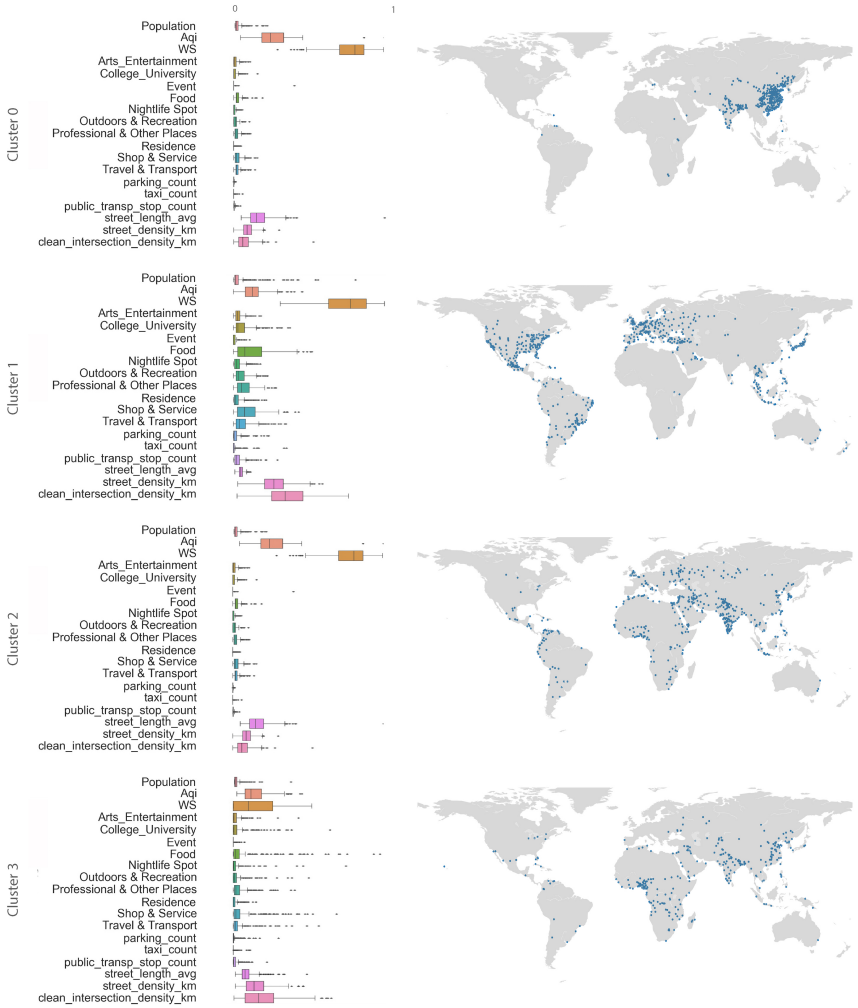

Figure 2. Distribution of clusters across the globe and boxplot of features.

Cluster 0 (number of urban units = 437): Represents cities with a minor number of venues within the urban units and minimum POIs of transport infrastructure. These units are walkable, with long streets, however, they exhibit low air quality.

Samples are mostly found in Eastern Asia, with China and India dominate the cluster.

Cluster 1 (number of urban units = 613): Vibrant urban units with an increased number of venues and substantial transport infrastructure belong to this cluster. The central areas of the cities are very walkable with good environmental conditions. The street network is dense with small segments and increased intersection density. Cities that belong in this cluster are scattered across the globe with strong presence in Europe, USA, Canada, Brazil, Japan and Australia.

Cluster 2 (number of urban units = 498): This cluster describes walkable units with moderate air quality and dense road networks that do not have high numbers of venues or advance transport infrastructure. Cities exhibiting such characteristics are found mostly in India, Russia, Iran and north-east parts of South America.

Cluster 3 (number of urban units = 312): The last cluster represents urban units with extremely dense and short street networks. The cities that belong to this cluster are barely walkable with poor air quality, and the existing transport infrastructure is limited. Most of these units are in China, India, Nigeria and Congo.

5. Discussion and conclusion

The concept of social mobility is widely considered as an indicator measuring countries' ability to provide equal opportunities for all. (OECD, 2011). This study, through an objective and data-driven process, explores heterogeneous, openly available, city-level information to identify similarities in urban behaviours that operate, according to Sorokin, as indicators for social openness(Sorokin, 1959).

The presented process revealed that 73 countries (Figure 2) have central urban units in more than two clusters. This fact demonstrates high levels of living standards inequalities between neighbouring cities. From those 9 countries - China, Colombia, India, Philippines, Saudi Arabia, South Africa, South Korea, Venezuela, Vietnam - with cities in all four clusters display extremely different living standards and urban behaviours in the selected indicators. In the same notion, the central urban units of 24 countries belong in 3 clusters revealing significant differences in human activity, transport infrastructure and environmental conditions with potential socio-economic deprivation. Some of these countries are Australia, Brazil, Canada, Italy, Qatar, Russia, Taiwan, and USA.

Clusters within countries	Number of countries
4	9
3	24
2	40
1	77

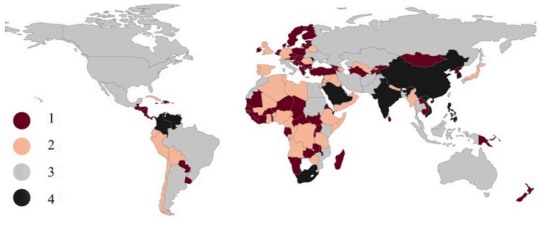

Figure 3. Choropleth map of different types of urban units within countries. .

This study has two limitations that call for further investigation. First the lack of open and accessible datasets that provide in a consistent manner city level information regarding economic factors, such as cost of living and purchasing power. As cities are searching for ways to improve livability and track their performance, this type of information is expected to become publicly available in a standarised form, for more and more cities. Second the demographic and geographic biases of crowdsourcing information retrieved from Foursquare. The application is usually referring to young and technology savvy users, and provides inadequate information for certain locations.

The described methodology demonstrates how we can extract - to a reasonable extent -reliable and up-to date city-level data for every urban location around the world, in a systematic and objective manner. The framework can support decision-makers in evaluating and benchmarking the current state of cities, tracking progress and performance of urban planning interventions and identifying best practices within cities of similar context . Moreover, the framework can be used to quickly identify differences and deprivation within urban settings of the same country, steering policy makers towards interventions that will support equal opportunities for all.

Acknowledgements

The first author would like to acknowledge the support from the University of Newcastle during her PhD studies. The authors would like to thank Sam Spurr, Nicholas Foulcher and Asad Abbas, for the fruitful discussions during the course of this research.

References

Aeris, [.: 2020, “AerisWeather API” . Available from <https://www.aerisweather.com/support/docs/api/> (accessed June 2020).

Arcadis, [.: 2017, “Arcadis Sustainable Cities Mobility Index” . Available from <https://www.arcadis.com/media/8/B/8/%7B8B887B3A-F4C4-40AB-AFFD-08382CC593E5%7DSustainable%20Cities%20Mobility%20Index.pdf> (accessed June 2020).

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A. and Volinsky, C.: 2013, Human mobility characterization from cellular network data, *Communications of the ACM*, **56**(ISSN 0001-0782), 74-82.

Boeing, G.: 2017a, OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, *Computers, Environment and Urban Systems*, **65**, 126-139.

Boeing, G.: 2017b, “Isochrone Maps with OSMnx + Python” . Available from <https://geoffboeing.com/2017/08/isochrone-maps-osmnx-python/> (accessed June 2020).

Caceres, N. and Benitez, F.G.: 2018, Supervised land use inference from mobility patterns, *Journal of Advanced Transportation*, **2018**(ISSN 0197-6729), 8710402.

Calafiore, A., Palmer, G., Comber, S., Arribas Bel, D. and Singleton, A.: 2021, A geographic data science framework for the functional and contextual analysis of human dynamics within global cities, *Computers, Environment and Urban Systems*, **8**(ISSN 0198-9715), 101539.

Caliński, T. and Harabasz, J.: 1974, A dendrite method for cluster analysis, *Communications in Statistics-theory and Methods*, **2**(ISSN 0090-3272), 1-27.

Cheng, Y.: 1995, Mean shift, mode seeking, and clustering, *IEEE transactions on pattern analysis and machine intelligence*, **17**(ISSN 0162-8828), 790-799.

Chestnut, J. and Mason, J.: 2019, “Indicators for Sustainable Mobility.” . Available from <https://www.itdp.org/publication/indicators-sustainable-mobility/> (accessed June 2020).

Comaniciu, D. and Meer, P.: 2002, Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on pattern analysis and machine intelligence*, **24**(ISSN 0162-8828), 603-619.

Dixon, s., Irshad, H., Pankratz, D.M. and Bornstein, J.: 2019, “The 2019 Deloitte City Mobility Index” . Available from <https://www2.deloitte.com/content/dam/Deloitte/br/Documents/consumer-business/City-Mobility-Index-2019.pdf> (accessed June 2020).

Foursquare, [.: 2020, “Foursquare Developer API” . Available from <https://developer.foursquare.com/.> (accessed June 2020).

Haklay, M. and Weber, P.: 2008, Openstreetmap: User-generated street maps, *IEEE Pervasive Computing*, **7**(ISSN 1536-1268), 12-18.

Van der Maaten, L. and Hinton, G.: 2008, Visualizing data using t-SNE, *Journal of machine learning research*, **9**, 2579-2605.

Macedo, J., Rodrigues, F. and Tavares, F.: 2017, Urban sustainability mobility assessment: indicators proposal, *Energy Procedia*, **134**(ISSN 1876-6102), 731-740.

OECD, (.F.E.C.-O.D.: 2011, *Divided we stand: Why inequality keeps rising*, OECD Publishing, Paris.

OSM, (.S.M.: 2020, “Taginfo” . Available from <https://taginfo.openstreetmap.org/keys/place#values> (accessed June 2020).

Regmi, M.B.: 2020, Measuring sustainability of urban mobility: A pilot study of Asian cities, *Case Studies on Transport Policy*, **8**(ISSN 2213-624X), 1224-1232.

Rodrigue, J.P.: 2020, *The geography of transport systems (5th ed.)*, Routledge.

Terroso Saenz, F., Muñoz, A. and Arcas, F.: 2020, Land□use dynamic discovery based on heterogeneous mobility sources, *International Journal of Intelligent Systems*, **36**(ISSN 0884-8173), 478-525..

Sorokin, P.: 1959, Social and cultural mobility, *New York*, **4**, 99-145.

UN, (.N.: 2018, “World urbanisation prospects” . Available from <https://population.un.org/wup/Download/Files/WUP2018-F22-Cities_Over_300K_Annual.xls> (accessed June 2020).

Walkscore, [.: 2020, “Walkscore API” . Available from <https://www.walkscore.com/professional/api.php> (accessed June 2020).

Wang, J., Huang, Z., Xu, H. and Kang, Z.: 2018, Clustering analysis of human behavior based on mobile phone sensor data, *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 64-68.

Wattenberg, M., Viégas, F. and Johnson, I.: 2016, How to use t-SNE effectively, *Distill*, **1**(ISSN 2476-0757), e2.

WEF, (.E.F.: 2020, “The Global Social Mobility Report 2020 Equality, Opportunity and a New Economic Imperative” . Available from <http://www3.weforum.org/docs/Global_Social_Mobility_Report.pdf> (accessed June 2020).

Zhang, X., Sun, Y., Zheng, A. and Wang, Y.: 2020, A New Approach to Refining Land Use Types: Predicting Point-of-Interest Categories Using Weibo Check-in Data, *ISPRS International Journal of Geo-Information*, **9**(ISSN 2220-9964), 124.

Zheng, Y.: 2019, *Urban Computing*, MIT Press..