Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 64 (2022) 76–83

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Scientific Conference "The Science and Development of Transport – Znanost i razvitak prometa – ZIRP 2022"

Methodology for public transport mode detection using telecom big data sets: case study in Croatia

Krešimir Vidović[a,*], Petar Čolić[a], Saša Vojvodić[a], Anamarija Blavicki[a]

[a]*Ericsson Nikola Tesla, Krapinska 45, 10000, Zagreb, Croatia*

Abstract

Determining the number of passengers using public transport services is a challenging and time-consuming task that relies either on manual observations (e.g. manual counting passengers in vehicles or at stations) or the application of technical solutions (using data from automatic fare collection system (AFC) or automatic passenger counters (APC), which is characterized either by the provision of an incomplete picture (AFC) or by a solution which in practice is installed in a small number of vehicles if any (APC). The new approach which uses anonymized telecom-originated big data sets and data science principles can be used as a smart data driven approach for determining the use of public transport. Anonymized telecom big data sets represent "digital breadcrumbs" that people leave while moving through the city. When paired with additional data sets (e.g. public transport timetables, location of public transport stations, information on public transport lines, etc.), it can be used for modal split detection. In this paper, a new methodological approach is proposed that uses anonymized telecom big data sets and a statistical modelling approach to identify possible public transport trips among all other trips. This methodology has been tested in a case study in the City of Rijeka and validated using ground truth data obtained from traditional sources.

Peer-review under responsibility of the scientific committee of the International Scientific Conference „The Science and Development of Transport - Znanost i razvitak prometa –ZIRP2022

Keywords: telecom big data; data science; public tranport utilisation; modal split.

* Corresponding author.
E-mail address: kresimir.vidovic@ericsson.com

Peer-review under responsibility of the scientific committee of the International Scientific Conference „The Science and Development of Transport - Znanost i razvitak prometa –ZIRP2022
10.1016/j.trpro.2022.09.010

1. Introduction

Anonymized telecom big data sets are established as a recent data source used for various types of analytics in transportation and mobility planning. These data sets are characterized by a large sample (dependent on the mobile operator's market share) and have been recognized as a fast, reliable and cost-effective solution for accurate and easily repeatable data collection and analysis. Since this data represents information on the approximate position of population in time only, additional analytical techniques are required to try to understand and reconstruct the population movement (population trips). Based on the characteristics of these movements (time of occurrence, approximate distance, speed, duration, etc.), additional context can be assigned to these movements, such as purpose, transport mode, etc. One of the most challenging tasks in the analytics is the detection of transport mode. Identification of transport mode usage (modal split) using this kind of analytics is perceived as a possible solution for real and reliable determination of modal split since other methods are highly complex and, in general, rely on either survey or manual observations. This task is particularly challenging in urban environments. Most of the urban mobility measures targeting greener and safer mobility are addressing the increase of use and optimization of public transport services. However, in order to properly identify the real utilization of public transport services, the exact number of passengers has to be determined. This paper proposes a new methodological approach for determining the transport mode for population trips by using statistical modelling in urban areas. The paper is focused on defining the methodology for identifying the number of passengers in public transport bus service in urban environment but can be applicable to any type of public transport that uses lines and stops. This methodology has been applied in a case study in Croatia and validated using ground truth data.

Paper is structured as follows. Following the Introduction, section Current research analyses existing experience of usage of anonymized big data set in mobility management and in transportation engineering. Thirds section drafts the proposed methodology, and fourth section presents the results of methodology application and validation using data gathered from City of Rijeka. The paper ends with conclusion and description of further steps.

2. Current research

Anonymized telecom big data sets are an emerging data source that has been subject to research for the past two decades. The researchers are trying to find use for these types of analytics in various business domains, from transport, marketing, public services, and public health.

Anonymized telecom big data sets are usually not easy to acquire, and authors and researchers either might have access to them as a part of various "big data challenges and hackathons" organized by the telecom network operator or might purchase them from the operator for a specific analytical purpose. The collection and anonymization of data is a very time-consuming and sensitive process since all operators have to be compliant with strict data privacy procedures (including GDPR). From an operational and technical perspective, operators must have human and technical resources for the collection and anonymization of this data. Furthermore, not all operators are suitable as a data source. In order to keep up with the requirements on the representativeness of the sample, operators with heterogenous types of users (some operators are targeting the youth, some are targeting business users, etc.) and operators with significant market share should be candidates. Besides the selection of operator, those performing analytics should also be aware that there are several data sources within the operator that might be used. The first and most common one in the early phases of using anonymized telecom big data sets is a source called "Charging Data Record" database, which consists of data on the user's telecommunication activities that are stored (for a longer period of time) and used for billing purposes. The other data source, which is usually collected on demand, is called the "signaling data source" and it includes information on the signaling data exchanged between a mobile phone terminal and a mobile network. The advantage of this data type is that it contains information collected even if the user is not in a particular service (e.g. if the mobile phone is turned on but is not engaged in data transfer or voice call).

Therefore, the main advantage of these data sources is that in anonymized ways it collects information on population movement in time, and all analytical use cases that are based on analytics of population migration can benefit from this. For example, in marketing, this type of research is used to measure the real number of people attending an event, to determine the number of people passing by a store or a billboard, to identify the optimum position of a bank branch etc., Jia et al. (2016); Wassouf et al. (2020). In public health domain, people movement

analytics is used for the potential tracking of the spread of epidemics or to approximate the number of people in a certain area for the optimization of medical service, Grantz et al. (2020); Valdano et al. (2021). In public service domain, this approach can be used for performing a periodic census or for reform of state (re)organization, Novak et al. (2013); Reades et al. (2007). In tourism, big data analytics is to determine the number of tourists and its segmentation per country of origin, for determination and prediction of real touristic demand, etc., Li et al. (2018); Song & Liu (2017). However, the authors have, in general, mostly applied this approach in the domain of transport and mobility in general. The most common use case in mobility and transportation is determination of origin destination matrices, Bachir et al. (2019); Calabrese & Lorenzo (2011); Filić et al. (2016), identification of user trajectories, Chen et al. (2014); Geo et al. (2014), identification of traffic flow parameters Galloni et al. (2018), for smart mobility toolboxes Šoštarić et al. (2020), creation of national transport model, Friso (2020), for estimation of urban mobility, Vidović et al. (2017), for enhancement of traffic safety by identifying drivers using mobile phone services while driving, Čolić et al. (2022), and, finally, identification of transport mode, Bachir et al. (2019); Chin et al. (2019).

The general conclusion is that this data source is promising and that its advantages outweigh the disadvantages. The main advantages are manifested in the size of the population included in the research (this can be 40 or more percent of the entire population, based on the operator's market share), in how simple it is to collect and preprocess the data (it is counted in hours or days as opposed to a census or a survey that might last for weeks or months), and in the fact that the process can be easily repeated. The main disadvantage is in the inaccuracy of user position approximation, which in urban environment might be characterized with a position error of several hundred meters and in a complex process of data acquisition. Positioning error is a problem expected to be solved by evolution since all new generations of mobile telecommunication services (2G-5G) have significantly increased the user positioning accuracy.

3. Methodology

Methodology for the trip mode detection is based on the application of statistical modelling. However, prior to the selection of the most appropriate method, several approaches have been tested in order to identify the most appropriate one.

The first approach included the observation of only one particular bus line. Based on the information about the stops, all trips that have started and finished in the vicinity (a 300-meter radius) of some of the bus stops of those lines are extracted. These trips are then filtered according to speed (to match the public transport speed). Moreover, all "visited" locations should be within the defined distance from the bus line axes or accompanied stops. The main disadvantage of this method is that it includes only trips that have been performed by one line and does not detect trips that have included transfers (two or more lines).

The second approach has included observance of several (or all) bus lines, where all the conditions presented in the previous approach are also applicable. The main issue in this approach is that there is a certain number of stops within the city and a possible number of trips that fulfill combinations of origin, and destination from available stops exceeds the real number of trips performed by public transport, and a number of trips performed by other modes of transport (cars during rush hours, cycling trips, etc.) can be misidentified as public transport trips.

The next approach was the construction of an algorithm that aims to monitor the locations of passengers and thus determine whether they used a public transport service or not. The process includes identification of the stations for the specified line. This data as well as the trips are then analyzed using the algorithm (the output of the algorithm are trips detected around two or more consecutive stations). The shortcoming of this approach is the large variability of results depending on parameters such as station radius and speed filter.

Another approach uses time clustering. The goal is to make clusters according to the timestamp around the stations and to try to identify a "jump in the number of users" around the time scheduled for the arrival of the bus at that station. This method is not sufficient when there are other stops within the selected radius around the observed station. The reason for this is that the jump may indicate activity around some of other public transport stops.

Finally, the approach used in this research is based on Bayesian statistics and Bayesian model. Bayesian statistics is a theory in the field of statistics based on the Bayesian interpretation of probability where probability expresses a degree of belief in an event. A Bayesian model is a statistical model where you use probability to represent all

uncertainty within the model, both the uncertainty regarding the output but also the uncertainty regarding the input (aka parameters) to the model Puga et al. (2015). The proposed methodology is presented in Fig. 1.

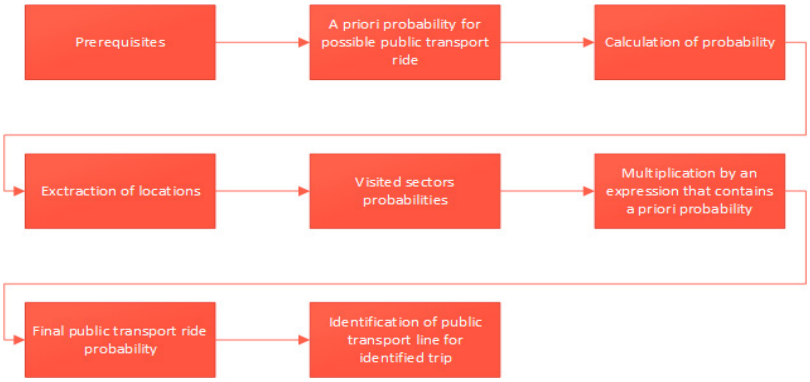

Fig. 1. Methodology for trip mode detection

In order to perform the analysis, there are prerequisites that have to be fulfilled.

Step 1: The first prerequisite is the existence of a user migration database for a particular area and particular timeframe, which was obtained from anonymized telecom big data set analytics and in accordance with data privacy principles. This database contains all user migrations (trips), with the details regarding the origin and destination of the trip, including visiting locations along the trip, trip distance, speed, duration, time of occurrence, etc. The database was obtained from operator with significant market share and preprocessed and processed in order to present insight on migration patterns. The methodology for performing such analytics is described in Šoštarić et al. (2021). Furthermore, the area of interest has to be divided in additional sectors using spatial decomposition method such as Voronoi diagram. For the area of interest additional data sets are required. These data sets include a data set on the road network (for road public transport), a data set on all available public transport lines in the area of interest (that are part of the analysis) and a data set on all public transport stops in the area of interest.

Step 2: In the second step, a priori probability for possible public transport ride is being determined. A priori probability refers to the likelihood of an event occurring when there is a finite number of outcomes, and each is equally likely to occur. The outcomes in a priori probability are not impacted by the prior outcome. The result is a general probability that someone will take a public transport trip (before any calculations are made). This information should be gathered from an external source or should be extracted from urban mobility data (e.g. average number of bus trips / all trips in a specific day).

Step 3: In third, the calculation of probability is carried out for each sector. Probability is calculated using the following equation using data identified in the prerequisite phase of this process. Probability is calculated as quotient of sum of number of stops and lines in geographic sector and sum of number of stops, number of lines and road length in same sector (Probability = (Stops + Lines) / (Stops + Lines + Roads)).

Step 4: In the fourth step, the extraction of locations related to particular trip is performed. For every recorded trip, all locations (start, end, visited) are extracted in order to identify the start location, the end location and all visited locations.

Step 5: In this step the probabilities of visited sectors is calculated. The calculated visited sectors probabilities are multiplied, and result is the number that will be named “BP”. Then the probabilities that the ride was not public transport ride are also multiplied, because of normalization purposes and result value will be named “NBP”.

Step 6: In this step, values calculated in previous step, BP and NBP are additionally multiplied by an expression that contains a priori probability.

Step 7: The sum of BP and NBP, after multiplication from previous step is calculated, and divide BP by that sum to get final public transport ride probability for given trip.

Step 8: In this step the decision is made on which public transport line was used if trip is classified as a bus ride. A start location is taken with first visited location of each trip that was classified as public transport and searched for

public transport lines inside of circles of some radius around those locations. If there were more than one public transport line for given trip, one will be randomly chosen and assign it to that trip. Important thing to mention is that each public transport line had assigned weight to it, which means not all lines had the same probability when being randomly selected. Those weights are nothing but number of rides certain bus line does in one day.

4. Case study and validation

The lack of reliable ground truth data is a key issue for proper validation of this type of research. In general, data for validation can come from three sources. The first data source can be automatic passenger counting systems installed in public transport vehicles. This data source is reliable and accurate, and data can be mapped both to geographic location and transport line/station. A shortcoming of this data source is the fact that these systems are rarely installed in all public transport vehicles or are not installed at all. The second data source for validation can be data from an automatic fare collection system. An automatic fare collection system can provide valuable data regarding utilization of public transport vehicles/lines using data from validators. A shortcoming of this methodology is that public transport operators might require that passengers validate their ticket only when entering bus and not when exiting the bus, so this information might be incomplete. In absence of more precise data, the most usual data acquisition method for counting passengers in public transport is manual people counting. The observer either stands at the bus station and counts passenger entering or leaving the buses or is driving along the line and counts passenger in a bus/line. Any data source that relies on manual counting only might be less reliable or incomplete; however, currently this is the best available ground truth data source.

Validation is, therefore, based on the number of passengers (passenger counting) in the public transport, during which the lines that carry 5% of the total number of passengers in the public transport of the City of Rijeka were analyzed. Therefore, five lines (Lines 11, 12, 13, 26, 30) were selected that together carry of 5% of all passengers. Those lines were selected since they have different characteristics in terms of number of departures, environment (centar, urban – non centar, rural). Any other line could be selected as well in order to test methodology. Based on the obtained results, in the first step, 2018 data (data from the City of Rijeka) was validated. This data was precisely and in detail collected for the needs of Traffic Masterplan of Primorje-Gorski Kotar County, Lika-Senj County, and Istria County. After confirming the accuracy of this data with data collected by the field research in this research, the data was used for the validation of results obtained from statistical modelling. According to the data obtained by field research, a decrease in the number of transported passengers was found, but the ratio of the share of the number of transported passengers on individual lines remained similar. A decrease in the number of transported passengers is a result of the COVID-19 pandemic. The results of the share of counted passengers on individual lines from this project and in relation to the data from 2018 are shown in the Table 1.

Table 1. Comparative analysis – public transport, shares of passengers per line, field research

Line	Share of counted passengers per line	Share of passengers per line – data from the City of Rijeka
11	36.0%	39.4%
12	14.0%	29.2%
13	4.9%	4.1%
26	29.8%	20.7%
30	15.3%	6.7%
SUM	100.0%	100.0%

The obtained results show that the data mostly correlates. Significant anomalies are present on two lines (differences in the percentage of up to 50%). This can be explained by significant changes in the structure and number of passengers due to the global COVID-19 pandemic. In accordance with the determined similar shares, the shares of

transported passengers on bus lines were used for further validation, and according to the data from the City of Rijeka, shown in Fig. 2.

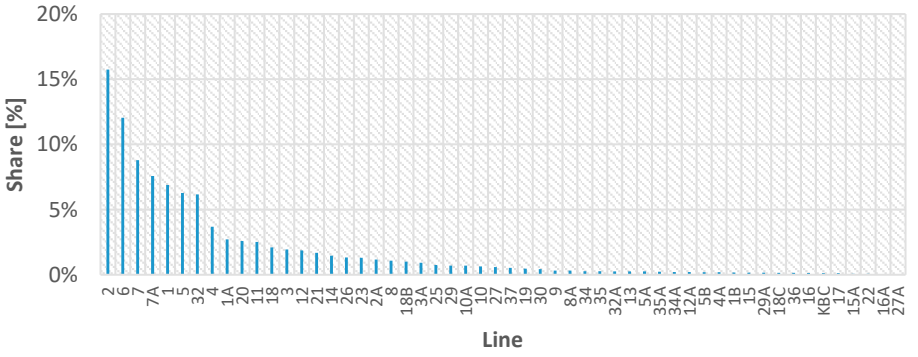

Fig. 2. Shares of transported passengers on bus lines

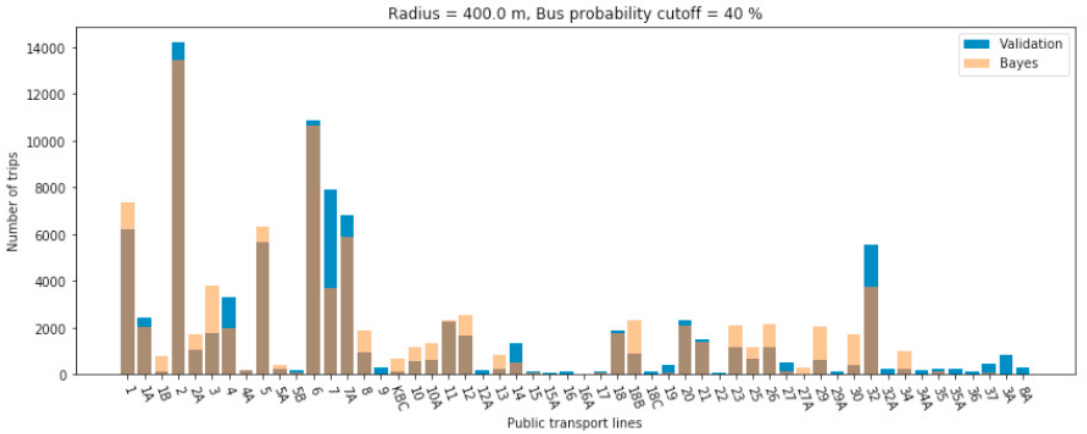

Fig. 3. Shares of transported passengers on bus lines

In accordance with the defined methodology for the detection of travel modes (based on Bayesian statistics and Bayesian model), the share of passengers according to public transport lines was determined. A comparative analysis of the obtained results showed a significant match between the shares on most of the lines. In Fig. 3, the number of trips per public transport lines for one day is presented. Blue bars represent validation data, and orange bars are data we obtained by using Bayes algorithm on a big data set. Brown bars present the “match” between those to values. The radius inside of which we look for those lines is 400.0m and we take all trips whose probability of being public transport is bigger than 40% as definite bus trips. Several rangers of probability values were tested during the model creation and validation and chosen treshold resulted in most appropriate match with validation data. As we can see, orange and blue bars overlap in most cases, which means that numbers are quite similar. Some of the best results were achieved for lines 2 (95% overlap), 6 (98% overlap), 7A (86% overlap) and 5 (90% overlap). In addition, the analysis of the data obtained by the survey has also shown similar results of the modal split in accordance with the big data and transport model. According to the survey, 73.1% represent motor vehicles, 16.9% public transport, while according to big data results, 78.5% represent motor vehicles and 21.1% public transport.

5. Conclusion

This paper has proposed the concept of the methodology for detection of public transport service users using anonymized telecom big data sets and statistical modelling. Statistical modelling was applied on the existing user migration database generated from telecom data, and new data sources were introduced in order to identify usage of public transport. These data sets included information of public transport lines, stations, and timetables. Several approaches have been tested, and finally the application of Bayesian statistics and Bayesian model have been selected and applied on the real data set from the City of Rijeka. For validation purposes, ground truth data set was acquired using a traditional process (manual passenger counting). The algorithm was proven successful for a number of lines, where its results correlated with the ground truth data up to 98 percent. Next steps will include fine tuning of the algorithm in order to increase its accuracy on a bigger number of lines with different characteristic (longitudinal lines, transversal lines, circular lines etc.). The algorithm will then also be applied in rural environment, where other challenges (regarding higher position inaccuracy) might degrade its performances.

References

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: Emerging Technologies*, *101*, 254–275. https://doi.org/10.1016/j.trc.2019.02.013

Calabrese, F., & Lorenzo, G. Di. (2011). Estimating Origin- Destination Flows using Mobile phone Location Data. *Cell*, *10*, 36–44. https://doi.org/10.1109/MPRV.2011.41

Chen, C., Bian, L., & Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, *46*, 326–337. https://doi.org/10.1016/j.trc.2014.07.001

Chin, K., Huang, H., Horn, C., Kasanicky, I., & Weibel, R. (2019). Inferring fine-grained transport modes from mobile phone cellular signaling data. *Computers, Environment and Urban Systems*, *77*. https://doi.org/10.1016/j.compenvurbsys.2019.101348

Čolić, P., Jakovljević, M., Vidović, K., & Šoštarić, M. (2022). Development of Methodology for Defining a Pattern of Drivers Mobile Phone Usage While Driving. *Sustainability*, *14*(3), 1681. https://doi.org/10.3390/su14031681

Filić, M., Filjar, R., & Vidović, K. (2016). Graphical presentation of Origin-Destination matrix in R statistical environment. *36. Skup o Prometnim Sustavima s Međunarodnim Sudjelovanjem Korema „Automatizacija u Prometu 2016“*.

Friso. (2020). Recent developments of big data in the Dutch national model – Study with mobile phone data. *International Journal of Technology, Policy and Management*, *20*(1), 54–69.

Galloni, A., Horváth, B., & Horváth, T. (2018). Real-time Monitoring of Hungarian Highway Traffic from Cell Phone Network Data. *ITAT 2018 Proceedings*, *2203*, 108–115.

Geo, T., Study, C., Mobile, E., & Data, P. (2014). *Reconstructing Trajectories from Sparse Call Detail Records*.

Grantz, K. H., Meredith, H. R., Cummings, D. A. T., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., Mehta, S., Solomon, S., Labrique, A., Kishore, N., Buckee, C. O., & Wesolowski, A. (2020). The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18190-5

Jia, Y., Chao, K., Cheng, X., Zhang, T., & Chen, W. (2016). Big data assisted human traffic forewarning in hot spot areas. *Signal and Information Processing, Networking and Computers - Proceedings of the 1st International Congress on Signal and Information Processing, Networking and Computers, ICSINC 2015*, 367–374. https://doi.org/10.1201/b21308-46

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323. https://doi.org/10.1016/j.tourman.2018.03.009

Novak, J., Ahas, R., Aasa, A., & Silm, S. (2013). Application of mobile phone location data in mapping of commuting patterns and functional regionalization: A pilot study of Estonia. *Journal of Maps*, *9*(1), 10–15. https://doi.org/10.1080/17445647.2012.762331

Puga, J. L., Krzywinski, M., & Altman, N. (2015). Corrigendum: Bayesian statistics. *Nature Methods*, *12*(11), 1098–1098. http://www.nature.com/doifinder/10.1038/nmeth1115-1098b

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular census: Explorations in Urban data collection. *IEEE Pervasive Computing*, *6*(3), 30–38. https://doi.org/10.1109/MPRV.2007.53

Song, H., & Liu, H. (2017). *Predicting Tourist Demand Using Big Data*. 13–29. https://doi.org/10.1007/978-3-319-44263-1_2

Šoštarić, M., Jakovljević, M., Lale, O., Vidović, K., & Vojvodić, S. (2020). Sustainable Urban Mobility Boost Smart Toolbox. *Proc of 1st International Conference Public Transport & Smart Mobility Innovative Solutions for Smart Urban Mobility*.

Šoštarić, M., Vidović, K., Jakovljević, M., & Lale, O. (2021). Data-driven Methodology for Sustainable Urban Mobility Assessment and Improvement. *Sustainability (Switzerland)*. https://www.bib.irb.hr/1138783

Valdano, E., Okano, J. T., Colizza, V., Mitonga, H. K., & Blower, S. (2021). Using mobile phone data to reveal risk flow networks underlying the HIV epidemic in Namibia. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-23051-w

Vidović, K., Mandžuka, S., & Brčić, D. (2017). Estimation of Urban Mobility using Public Mobile Network. *Proceedings of 59th International Symposium ELMAR-2017*, *2017-Septe*, 21–24. https://doi.org/10.23919/ELMAR.2017.8124426

Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00290-0