Evaluating the Socioeconomic Status of a Large Social Event Attendees

Kerecsen Szabó, Gergő Pintér, Imre Felde

Óbuda University, John von Neumann Faculty of Informatics, Budapest, Hungary
kerecsen.szabo@stud.uni-obuda.hu, pinter.gergo@nik.uni-obuda.hu, felde.imre@nik.uni-obuda.hu

Abstract—In this study, Call Detail Records (CDRs) from downtown Budapest were analysed, focusing on a large scale event in August 2014. The attendees of the main event of the Hungarian State Foundation Day have been analysed based on their socioeconomic status (SES). This paper proposes an approach to estimating SES by the price and age of the subscribers’ phones, obtained by fusing a mobile phone property database with the CDRs. We have found some tendencies between the attendees based on the location, from where they watched the fireworks. However, the results do not show significant differences in this geographical granularity.

Index Terms—mobile phone data, call detail records, type allocation code, data analysis, human mobility, social sensing, large social event, socioeconomic status

I. INTRODUCTION

Since the mid-2000s, mobile telephones have become an inseparable part of our lives. The constant communication between a device and the mobile network leaves an involuntary trace of our activities. The spatiotemporal logs of our journeys and additional communication device information establish a promising model to evaluate human mobility and socioeconomic customs.

For the last two decades, uncovering underlying information from mobile phone records has been a developing research field. Data scientists, spatial data analysts, physicists and applied mathematicians pay more and more attention to discovering cellular data. Interpreting large amounts of Call Detail Records (CDRs) into useful information requires various tools and expertise. In the last ten years, a number of studies have been published discussing different applications of mobile network data analysis. [1] [2]

St. Stephen’s Day is celebrated in Hungary every year on the 20th of August. Tens of thousands of people visit the capital for its all day long celebrations and the main event, a 30-minute long fireworks show. The main area of the event includes three bridges and the embankments on the Danube for approximately three kilometres. With great views from the Buda and Pest embankments and the Castle District, these areas should show a significant spike in cellular activity and will be the primary subject of the analysis.

This work aims to explore ways of using mobile network data fused with mobile phone properties to derive socioeconomic indicators, which can be used to characterise the SES of the attendees of a large scale event. The discovered SES distribution in the analysis may lead to a more in-depth understanding of urban life and large scale events. Furthermore, the method described in this paper can be applied to other mobile network and phone property data sets.

This paper is organised as follows: Section II summarises relevant literature. Section III describes the data set used in the analysis. Section IV details the methodology and the data processing features. Then, the study results are concluded in Section V, and Section VI discusses the conclusions and future work.

II. RELATED WORKS

A typical application of mobile network data processing is large scale social event detection and estimating the attendance during mass gatherings. [3, 4, 5, 6] This study builds upon previous works at John von Neumann Faculty of Informatics, Óbuda University. The 2014 State Foundation Day data set has already been analysed [7] regarding the large social event attendance. Whereas in this work, the socioeconomic status of the attendees is studied.

Using mobile phone prices as socioeconomic indicators have been proved to work well by Sultan et al. in [8]. They identified areas in Pakistan where more expensive phones appear more often using indicators of accessibility to services, infrastructure, hygiene and communication. Their model performed with an absolute Pearson’s correlation coefficient > 0.35 and p-value < 0.01.

In an earlier study, Pintér et al. evaluated the connection between individuals’ financial status and mobility customs. The authors used the radius of gyration, entropy and Euclidean distance between home and work locations as mobility indicators and applied data fusion methods with average real estate prices to determine the influence of wealth on mobility customs. [9]

Regarding socioeconomic status (SES) analysis, Pintér et al. evaluated football fans’ SES using their mobile phone details in Budapest during the 2016 UEFA European Football Championship. They eliminated CDRs from Subscriber Identity Modules (SIMs) during data preprocessing, which did not operate in mobile phones using Type Allocation Code (TAC) databases. [10] In another work, they analysed subscribers’ wake-up times and explained how it correlates with their socioeconomic status. The analysis demonstrated a strong positive connection between the two indicator. They also showed that the mobile phone prices in the TAC database might have depreciated. [11]
III. MOBILE NETWORK DATA

A CDR data set usually contains a caller ID, the cell tower it is connected to, its location, and a timestamp. Additional information on the purpose of communication, device type, and SIM holders’ details help investigate more than just trajectories and cell densities.

The data set used for this research was obtained by Vodafone Hungary Ltd. The number of active SIMs was 11,540,058 in Hungary, of which Vodafone had an estimated 22.7 per cent of the market share in June 2014 (Hungarian National Media and Infocommunications Authority). These CDRs contain anonymous logs of customers’ calls and text messages in Budapest, Hungary and its suburban areas, over approximately 525.14 km² (Hungarian Central Statistical Office).

The data set was collected between the 18th and 22nd of August 2014. A total amount of 191,528,883 records have been logged, between 8,890 cells. Three comma-separated value files have been acquired for the analysis in this research. The first one contained call data records, the second and third are supplementary information about cells and devices.

The CDRs in this data set consist of a timestamp, a hashed device identification (ID), a hashed cell ID and a type allocation code. TAC is the initial eight-digit segment of a device’s International Mobile Equipment Identity (IMEI), uniquely identifying a particular device. In this data set, CDRs are of active call record type, meaning a record was made when the user was making a call or sending/receiving a text message. Unfortunately, the data set does not contain information on cell switching, which would make more granulated data possible.

The supplementary cell lookup table contains cell IDs and positions as 2D coordinates in decimal degrees format. Cells at the same location were merged into base stations for further analysis, and the corresponding CDR’s cell ID values were updated. Uniting particular cells is necessary for more straightforward data analysis. Nevertheless, cell tower antennas planted at an exact location might face different directions, but we do not have this information.

The device table contains a hashed device ID, the customer’s age, sex, whether it is an individual or a business and the subscription type (prepaid or postpaid). Some age and sex information is missing due to privacy restrictions.

IV. METHODOLOGY

During data cleansing, unnecessary spaces have been removed from the end of the lines. Furthermore, device and cell IDs were hashed information, a fundamental step for user privacy. On the contrary, it has a low information density, thus have been replaced with incrementing integer values. The reassigning does not affect the contained information but is favourable to reducing data size. Due to the considerable amount of CDRs, the comma-separated value text files have been loaded into an SQLite database, with the scheme illustrated in Figure 1. To speed up data acquisitions, indices were created on timestamps, IDs and TACs.

Although latitude and longitude values are in string format, there is no need for the original precision of 13 decimal points. This would mean μm resolution, while anything above the sixth decimal place is useless in this application. The unnecessary information can be discarded to save space and increase query speeds.

As a socioeconomic status (SES) indicator, the analysis used relative mobile phone ages in months to the event (August 2014) and phone release prices in EUR. Information on resolving the TACs are from the data provided by 51Degrees fused [10] with the GSMArena database [12].

A data processing framework has been developed to acquire the relevant CDRs with the paired SES indicators. Cells along the two sides of the Danube are expected to be the primary cells servicing the firework attendees’ mobile phones on the embankments. These cell IDs were selected manually due to the uneven separation line on the river. Any other cells that might support the selected areas are determined by a 250 m radius around the main event area, as visible in Figure 2. Four cells were removed from the evaluation due to the insignificant activity count (< 500) during the event. Extracting the selected cell IDs help filter and transform the large CDR database table.

![Figure 1: The database structure showing CDR, device and cell tables with foreign key connections.](image1)

![Figure 2: The selected cells for the large social event analysis.](image2)
Figure 3: Cell activities showing the extra mobile network load due to the State Foundation Day celebrations on the river Danube embankments and the Castle District in Budapest. The temporal filtration rule for the fireworks data will be $\pm30$ mins to the actual event. This will include possible additional users who did attend the venues but did not use their phones during the half-hour show.

User data from the device database table can be joined on the CDRs’ device ID fields. This gives us the ability to analyse age and sex distributions in selected groups. The selected CDRs have been joined on the corresponding TAC values from the merged mobile phone property database for the SES indicator aggregation. For further analysis, relative ages of the appearing phones have been calculated from release dates and months to the event date (August 2014). The cell by cell SES indicator used in the analysis is derived from the arithmetic mean value of the cell’s indicator series.

V. RESULTS

The analysis focused on the SES indicator distribution among the State Foundation Day celebratory firework viewers in Budapest, on the banks of the river Danube and in the Castle District in August 2014. The time frame is 20:00 – 21:30, including half an hour before and after the 30-minute show, marked with vertical lines in Figure 3.

A cell-by-cell average of mobile phone prices and relative ages was calculated for the SES indicator distribution. Figure 4 shows the spatial distribution of these indicators using Voronoi polygons generated around the cell tower locations. Cells are coloured by the average SES indicators; the higher the value, the darker the colour.

Figure 4 demonstrates an opposite trend between mobile phone price and age. The scatter plot of the same data is shown in Figure 5, where the Pearson correlation coefficient is $-0.7329$. Data points are coloured based on their location in the city, but there are no visible groups based on SES. The expectation was that there would be a significant contrast between Buda and Pest in socioeconomic indicator...
distribution. However, there are only minor differences in this spatial resolution, from which it can be concluded that those interested in the event do not divide drastically into socioeconomic groups.

VI. CONCLUSIONS AND FUTURE WORK

This paper presented a concept of socioeconomic data analysis on a large social event using call detail records and mobile phone details. This work fits into current research tendencies, fusing mobile network generated mobility data with socioeconomic descriptors that make it possible to deduce socioeconomic status from anonymous Call Detail Records.

We expected that in Buda, where the housing prices are higher [9], more expensive and newer phones would generate the majority of the activity. Nonetheless, we found that slightly more expensive phones were active in Pest, but the difference, on average, was not substantial. The base station level aggregation may have partly caused this result, or the attendees might not have watched the fireworks isolated from each other based on social status.

As future work, the firework attendees could be grouped into visitors and homeowners in the activity areas. Calculating home positions on CDRs has already been demonstrated to be helpful in [7] and could make a difference in the conclusion of this study.

AUTHOR CONTRIBUTIONS

Methodology, K.S. and G.P.; Conceptualisation, G.P. and K.S.; Software – data processing, K.S.; Software – data preprocessing, G.P.; Validation, K.S. and G.P.; Visualisation, K.S.; Writing, K.S.; Supervision, G.P. and I.F.

ACKNOWLEDGEMENT

The authors would like to thank Vodafone Hungary and 51Degrees for providing the Call Detail Records and the Type Allocation Code database used in this study. Map tiles by CartoDB, under CC BY 3.0.

REFERENCES

[1] Marta C Gonzalez, Cesar A Hidalgo and Albert-Laszlo Barabasi. “Understanding individual human mobility patterns”. In: nature 453.7196 (2008), pp. 779–782.
[2] Julián Candia et al. “Uncovering individual and collective human dynamics from mobile phone records”. In: Journal of physics A: mathematical and theoretical 41.22 (2008), p. 224015.
[3] Martin Wirtz et al. “Probing crowd density through smartphones in city-scale mass gatherings”. In: EPJ Data Science 2.1 (2013), pp. 1–24.
[4] Marco Mamei and Massimo Colonna. “Estimating attendance from cellular network data”. In: International Journal of Geographical Information Science 30.7 (2016), pp. 1281–1301.
[5] Ian Barnett, Tarun Khanna and Jukka-Pekka Onnela. “Social and spatial clustering of people at humanity’s largest gathering”. In: PLOS ONE 11.6 (2016), pp. 1–12.
[6] Hendrik Hiir et al. “Impact of Natural and Social Events on Mobile Call Data Records – An Estonian Case Study”. In: Complex Networks and Their Applications VIII. Springer. 2019, pp. 415–426.
[7] Gergő Pintér et al. “Evaluation of mobile phone signals in urban environment during a large social event”. In: 2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE. 2018, pp. 247–250.
[8] Syed Fahad Sultan et al. “Mobile phone price as a proxy for socio-economic indicators”. In: Proceedings of the Seventh International Conference on Information and Communication Technologies and Development. 2015, pp. 1–4.
[9] Gergő Pintér and Imre Felde. “Evaluating the Effect of the Financial Status to the Mobility Customs”. In: ISPRS International Journal of Geo-Information 10.5 (2021), p. 328.
[10] Gergő Pintér and Imre Felde. “Analyzing the Behavior and Financial Status of Soccer Fans from a Mobile Phone Network Perspective: Euro 2016, a Case Study”. In: Information 12.11 (2021), p. 468.
[11] Gergő Pintér and Imre Felde. “Awakening City: Traces of the Circadian Rhythm within the Mobile Phone Network Data”. In: Information 13.3 (2022), p. 114.
[12] Mohit Sainani. GSMArena Mobile Phone Devices. Accessed on 02.03.2022. 2020. URL: https://www.kaggle.com/msainani/gsmarena-mobile-devices.