ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

NEUMANN JÁNOS INFORMATIKAI KAR

DIPLOMAMUNKA

OE-NIK	Hallgató neve:	**Szabó Kerecsen**
2022	Hallgató törzskönyvi száma:	**T/008706/FI12904/N**

Óbudai Egyetem
Neumann János Informatikai Kar
Biomatika és Alkalmazott Mesterséges Intelligencia Intézet

DIPLOMAMUNKA
FELADATLAP

Hallgató neve: **Szabó Kerecsen**
Törzskönyvi száma: T/008706/FI12904/N
Neptun kódja: [redacted]

A diplomamunka címe:

Városi mobilitás jellemzése mobil hálózati adatok alapján
Characterisatrion of urban mobility based on mobile network data

Intézményi konzulens: Dr. habil. Felde Imre Gábor
Külső konzulens:

Beadási határidő: 2022. május 15.

A feladat

Készítsen informatikai adatfeldolgozó rendszert a mobiltelefon hálózatokból származó adatok elemzéséhez. A feldolgozó rendszer alkalmas a mozgási metrikák számítására, mozgási útvonalak és tartózkodási helyek meghatározására. Készítse el a feldolgozott és számított adatok vizualizációjára hivatott szoftver megoldást. Készítsen esettanulmányt és elemzést a rendelkezésre álló adatok felhasználásával egy nagy kulturális elemzésen résztvevők tartózkodására és mozgására vonatkozóan. Foglalja össze a fejlesztési eredményeit.
A diplomamunka angol nyelven készül.

A diplomamunkának tartalmaznia kell:

- a keretrendszer rendszerterve,
- a tér- és időbeli vonatkozással rendelkező adat kezelésének bemutatása,
- az adatvizualizáció bemutatását,
- az esettanulmány és elemzés módszertana,
- az esettanulmány eredményei,
- a fejlesztés összefoglalója.

ÓBUDAI EGYETEM Neumann János Informatikai Kar 1. Budapest

Ph.

……………………………
Dr. Eigner György
intézetigazgató

A diplomamunka elévülésének határideje: **2024. május 15.**
(OE TVSz 55.§ szerint)

A diplomamunkát beadásra alkalmasnak tartom:

………………..
külső konzulens

……………………………
intézményi konzulens

HALLGATÓI NYILATKOZAT

Alulírott hallgató kijelentem, hogy a diplomamunka saját munkám eredménye, a felhasznált szakirodalmat és eszközöket azonosíthatóan közöltem. Az elkészült diplomamunkámban található eredményeket az egyetem és a feladatot kiíró intézmény saját céljára térítés nélkül felhasználhatja.

Budapest, 2022. május 10.

Szabó
....
hallgató aláírása

ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Informatikai Kar

KONZULTÁCIÓS NAPLÓ

Hallgató neve: Neptun kód: Tagozat:

SZABÓ KERECSEN [redacted] NAPPALI

Telefon: Levelezési cím (pl.: lakcím):

[redacted]

Szakdolgozat / Diplomamunka[1] címe magyarul:

VÁROSI MOBILITÁS JELLEMZÉSE MOBIL HÁLÓZATI ADATOK ALAPJÁN

Szakdolgozat / Diplomamunka[2] címe angolul:

CHARACTERISATION OF URBAN MOBILITY BASED ON MOBILE NETWORK DATA

Intézményi konzulens: Külső konzulens:

DR. HABIL. FELDE IMRE

Kérjük, hogy az adatokat nyomtatott nagybetűkkel írja!

Alk.	Dátum	Tartalom	Aláírás
1.	2021.02.05.	Mobilitási metrikák áttekintése	
2.	2021.03.05.	Cella adatok és jelentésük	
3.	2021.04.05.	CDR adatok szűrése	
4.	2021.05.03.	Adatvizualizáció	

A Konzultációs naplót összesen 4 alkalommal, az egyes konzultációk alkalmával kell láttamoztatni bármelyik konzulenssel.

A hallgató a Szakdolgozat I. / Szakdolgozat II. (BSc) vagy Diplomamunka 1 / Diplomamunka 2 / Diplomamunka 3 / Diplomamunka 4[3] tantárgy követelményét teljesítette, beszámolóra / ~~védésre~~[4] bocsátható.

Budapest, 2021.05.14.

....
intézményi konzulens

[1] Megfelelő aláhúzandó!
[2] Megfelelő aláhúzandó!
[3] Megfelelő aláhúzandó!
[4] Megfelelő aláhúzandó!

KONZULTÁCIÓS NAPLÓ

Hallgató neve: Neptun kód: Tagozat:

SZABÓ KERECSEN [redacted] NAPPALI

Telefon: Levelezési cím (pl.: lakcím):

[redacted]

Szakdolgozat / Diplomamunka[1] címe magyarul:

VÁROSI MOBILITÁS JELLEMZÉSE MOBIL HÁLÓZATI ADATOK ALAPJÁN

Szakdolgozat / Diplomamunka[2] címe angolul:

CHARACTERISATION OF URBAN MOBILITY BASED ON MOBILE NETWORK DATA

Intézményi konzulens: Külső konzulens:

DR. HABIL. FELDE IMRE

Kérjük, hogy az adatokat nyomtatott nagybetűkkel írja!

Alk.	Dátum	Tartalom	Aláírás
1.	2021.09.10.	Esettanulmányok áttekintése	
2.	2021.10.01.	Esettanulmányok módszertana	
3.	2021.11.05.	Esettanulmányok eredményei	
4.	2021.12.03.	Esettanulmányok összefoglalója	

A Konzultációs naplót összesen 4 alkalommal, az egyes konzultációk alkalmával kell láttamoztatni bármelyik konzulenssel.

A hallgató a Szakdolgozat I. / Szakdolgozat II. (BSc) vagy Diplomamunka 1 / Diplomamunka 2 / Diplomamunka 3 / Diplomamunka 4[3] tantárgy követelményét teljesítette, beszámolóra / védésre[4] bocsátható.

Budapest, 2021.12.10.

....
intézményi konzulens

[1] Megfelelő aláhúzandó!
[2] Megfelelő aláhúzandó!
[3] Megfelelő aláhúzandó!
[4] Megfelelő aláhúzandó!

KONZULTÁCIÓS NAPLÓ

Hallgató neve: Neptun kód: Tagozat:

SZABÓ KERECSEN [redacted] NAPPALI

Telefon: Levelezési cím (pl.: lakcím):

[redacted]

Szakdolgozat / Diplomamunka[1] címe magyarul:

VÁROSI MOBILITÁS JELLEMZÉSE MOBIL HÁLÓZATI ADATOK ALAPJÁN

Szakdolgozat / Diplomamunka[2] címe angolul:

CHARACTERISATION OF URBAN MOBILITY BASED ON MOBILE NETWORK DATA

Intézményi konzulens: Külső konzulens:

DR. HABIL. FELDE IMRE

Kérjük, hogy az adatokat nyomtatott nagybetűkkel írja!

Alk.	Dátum	Tartalom	Aláírás
1.	2022.02.08.	Klaszterezési technikák áttekintése	
2.	2022.03.08.	Klaszterezési eredmények jelentése	
3.	2022.04.05.	Szoftvertesztelési megoldások	
4.	2022.05.10.	Fejlesztés összefoglalása és lezárása	

A Konzultációs naplót összesen 4 alkalommal, az egyes konzultációk alkalmával kell láttamoztatni bármelyik konzulenssel.

A hallgató a Szakdolgozat I. / Szakdolgozat II. (BSc) vagy Diplomamunka 1 / Diplomamunka 2 / Diplomamunka 3 / Diplomamunka 4[3] tantárgy követelményét teljesítette, beszámolóra / védésre[4] bocsátható.

Budapest, 2022.05.14.

....
intézményi konzulens

[1] Megfelelő aláhúzandó!
[2] Megfelelő aláhúzandó!
[3] Megfelelő aláhúzandó!
[4] Megfelelő aláhúzandó!

Contents

List of Figures

List of Tables

Abstract

Since the mid-2000s, mobile phones 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. Therefore, the spatiotemporal logs of our journeys and additional communication device information establish a promising model to evaluate human mobility customs and socioeconomic indicators.

This study aims to develop a mobile network data processing framework, analysing phone users' urban mobility and socioeconomic status. This application of call detail records is not yet thoroughly explored, and there are no out of the box solutions. Nevertheless, the study contributed to the research group's publications with an accepted submission to IEEE and widened the literature on mobile network data used to infer socioeconomic status indicators. The analysis will focus on Budapest, using data from August 2014, sourced through Vodafone Hungary.

First, the methodology for deriving cell-level mobility and socioeconomic status indicators is proposed. Second, the radius of gyration and entropy is calculated for every user in the database, and cell averages are determined to characterise human mobility in Budapest. After, the study proposes an approach to estimating socioeconomic status by the price and age of the subscribers' phones, obtained by fusing a mobile phone property database with call detail records. Finally, the mobile network cells are assigned to groups using an unsupervised machine learning algorithm based on their metrics.

The developed software framework is tested using implemented unit tests in the source code and defining case studies. Well-known parts of Budapest were selected, and their socioeconomic status indicators were examined and compared. The study results are three-fold: a data processing framework has been developed to derive human mobility and socioeconomic status indicators from mobile network data. Furthermore, two case study results were presented: a large social event was analysed based on the attendees' metrics. Finally, the Budapest cells were clustered using k-means algorithms with average mobility and socioeconomic indicators.

Result visualisations were done using maps of voronoi cells around the base stations. The research showed the differences in average socioeconomic indicator distribution between Buda and Pest during a large scale social event, using mobile phone ages and prices. Furthermore, the clustering results prove that homogeneous groups of cells formed around Budapest, with similar mobility and socioeconomic metrics, and the most significant differences between the city centre and the suburbs. The main conclusion shows that mobile network data enhanced with mobile phone properties offers a widely available and versatile tool for characterising urban mobility and socioeconomic status. Furthermore, the developed framework can be utilised to analyse call detail records to a large extent, and examples have been presented of use cases.

Kivonat

A 2000-es évek közepe óta a mobiltelefonok életünk elválaszthatatlan részévé váltak. A készülék és a mobilhálózat közötti állandó kommunikáció önkéntelenül is nyomot hagy tevékenységünkről. Ezért az utazásaink tér-időbeli naplói és a kommunikációs eszközök kiegészítő információi ígéretes modellt alkotnak az emberi mobilitási szokások és a társadalmi-gazdasági mutatók értékelésére.

E tanulmány célja egy mobilhálózati adatfeldolgozási keretrendszer kidolgozása és a telefonhasználók városi mobilitásának és társadalmi-gazdasági státuszának elemzése. A hívásadat-nyilvántartások ezen alkalmazása még nincs alaposan feltárva, és nincsenek kész megoldások. Mindazonáltal a tanulmány egy elfogadott IEEE-beadvánnyal hozzájárult a kutatócsoport publikációihoz, és bővítette a társadalmi-gazdasági státuszmutatók következtetésére használt mobilhálózati adatokkal kapcsolatos szakirodalmat. Az elemzés Budapestre összpontosít, a Vodafone Magyarországtól származó, 2014. augusztusi adatok felhasználásával.

Először a cellaszintű mobilitási és társadalmi-gazdasági státuszmutatók levezetésének módszertanát javasoljuk. Másodszor, az adatbázisban szereplő minden egyes felhasználóra kiszámítjuk a gyűrűs sugarat és az entrópiát, és cellánkénti átlagokat határozunk meg a budapesti emberi mobilitás jellemzésére. Ezt követően a tanulmány javaslatot tesz a társadalmi-gazdasági státusz becslésére az előfizetők telefonjának ára és kora alapján, amelyet a mobiltelefonok tulajdoni adatbázisa és a hívásadatokat tartalmazó nyilvántartások összevonásával kapunk. Végül a mobilhálózati cellákat egy nem felügyelt gépi tanulási algoritmus segítségével csoportokhoz rendeljük a metrikáik alapján.

A kifejlesztett szoftverkeretet a forráskódban implementált egységtesztekkel és esettanulmányok meghatározásával teszteljük. Budapest jól ismert városrészeit választottuk ki, és azok társadalmi-gazdasági státuszmutatóit vizsgáltuk és hasonlítottuk össze. A tanulmány eredményei hármasak: kifejlesztésre került egy adatfeldolgozó keretrendszer, amely a mobilhálózati adatokból emberi mobilitási és társadalmi-gazdasági státuszmutatókat származtat. Továbbá két felhasználási eset eredményeit mutattuk be: egy nagy társadalmi eseményt elemeztünk a résztvevők mérőszámai alapján. Végül a budapesti cellákat klaszterezték k-means algoritmusok segítségével az átlagos mobilitási és társadalmi-gazdasági mutatók alapján.

Az eredmények vizualizálása a bázisállomások körüli voronoi-cellák térképeinek segítségével történt. A kutatás a mobiltelefonok életkorának és árainak felhasználásával mutatta ki a különbséget az átlagos társadalmi-gazdasági mutatók eloszlásában Buda és Pest között egy nagyszabású társadalmi esemény során. Továbbá a klaszterezési eredmények azt bizonyítják, hogy Budapest körül homogén sejtcsoportok alakultak ki, hasonló mobilitási és társadalmi-gazdasági mutatókkal, és a legjelentősebb különbségek a belváros és a külvárosok között mutatkoznak. A legfőbb következtetés az, hogy a

mobiltelefonok tulajdonságaival kiegészített mobilhálózati adatok széles körben elérhető és sokoldalú eszközt kínálnak a városi mobilitás és a társadalmi-gazdasági státusz jellemzésére. Továbbá, a kifejlesztett keretrendszer nagymértékben felhasználható a hívásrészletezési nyilvántartások elemzésére, valamint példák kerültek bemutatásra a felhasználási esetekre.

1 Introduction

According to the United Nations' 2018 estimate, 55.3 per cent of the world's population has lived in urban areas, which will reach 60 per cent by 2030. This rapid growth puts enormous pressure on urban developers and inhabitants, making everyday life harder for people who aspire to move to these densely populated areas.

Segregation, public safety, ease of transportation, travel and cohabitation are critical factors in city management. Up until the outbreak of the COVID-19 pandemic, city development plans have been relatively straightforward. Promoting and updating public transportation, building bypasses around residential areas, bicycle-friendly roads, and increasing walkability have been the centre of aim for urban developers to build human-friendly spaces.

However, early 2020 drastically changed day-to-day life in almost every aspect of urban residents. In the days where a simple daily commute has become a public health risk, it is an ever more important task we better understand human dynamics and mobility in urban areas. Between numerous challenges of adapting to the unprecedentedly fast and radical changes in our everyday life, we have to face the problems of social distancing, changing the ways of our transportation, how and when we shop and designing our home office. Frustration is a growing problem, and dynamic adaptive changes could help our transportation system, more competent logistics, and traffic control. For these outcomes, we need information on how an urban area and its residents function and what are their mobility habits and characteristics.

Since the dawn of the Global System for Mobile Communications (GSM) – in 1991, mobile communication system speeds reached the Gbit/s magnitude from the Kbit/s limits. With the evolution of 5G networks, we can observe Moore's law in information and communications technology. Furthermore, the growing amount of available mobile network data and research shows excellent potential to uncover human mobility characteristics using location-enabled GSM data in epidemiology, sociology, and urban planning.

In the age of mobile telephones, when almost half of the world's population has a smartphone in their pocket, there is enormous potential in the generated mobile network data. While most of these devices are already equipped with GPS, it is mostly inactive, and only a handful of applications log mobile phone positions based on it. Even though smartphones have the prospect of connecting to the internet, the most common function is cellular calls to date. The continuous communication between mobile phones and cell towers leaves the system operator with enormous network data. There is valuable information hidden in this data on our lives, and it can be used anonymously for a better understanding of human mobility. GSM logs enable extensive data analysis because only a few service providers are usually available in a region. In addition, because nearly everyone has a mobile phone, large masses can be involved in mobile network data analyses.

This study aims to widen the collection of the works exploring the possibilities of mobile network data processing and create a new application for call detail record (CDR) analysis. The objective is to process a given data set using a self-designed framework to extract indicators of human mobility and find connections to other personal features. Furthermore, socioeconomic status indicators can help understand urban characteristics and be concluded from an individual's mobile phone properties. This possibility will also be investigated, along with mobile phone user clustering based on the analysed indicators and spatial representation.

The work is structured as follows. In Section 2, the fundamental applications of mobile network analysis, important works and terminologies will be introduced. Section 3 discusses the design considerations and specifications of the data processing framework. Section 4 describes the course of implementation and the experiences during the work phases. Section 5 features software testing solutions for CDR analysis. Section 6 presents the results of the study, while Section 7 includes the conclusions, future works and improvement potentials.

2 Literature Review

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. However, interpreting large amounts of call detail records into useful information requires numerous tools and expertise. In the last ten years, dozens of research groups have published several major research journals discussing different applications of mobile network data analysis.

The sudden and rapid development of information and communication technology led to the first research results on the topic from 2007-2008. Characterising human mobility based on mobile network data is a relatively new research area. In this section, the analysis will introduce the most important works and results of the last 15 years as well as the commonly used mobility indicators used in the literature.

2.1 Call Detail Records

Using call detail records spanning over 52 weeks, accumulated over two-week-long periods, Gonzalez *et al.* analysed 100,000 randomly selected individuals' movements over half a year in Europe. They characterised basic human mobility patterns and discovered that most individuals travel only short distances and just a few moves over hundreds of kilometres and approximated the probability density function of travel distances with a truncated power-law [1].

Candia *et al.* carried out a comprehensive study on the mean collective behaviour of individuals and examined how space and time deviations can be described using known percolation theory tools. They also proved that the interevent time between consecutive calls is heavy-tailed, which agrees with previous studies on other human activities [2].

Pappalardo *et al.* reported the existence of two distinct profiles in human mobility: returners, which limit the majority of their movement to a few locations and explorers, whose mobility cannot be described in this way. They developed new models that capture their observational findings to support this theory. These results show a distinct role in phenomena spreading, which correlates with our dynamics and social interactions [3].

From the early 2010s, visual analytics approaches for exploring spatiotemporal urban data have been popular methods [4], [5], [6]. Senaratne *et al.* introduced a novel concept for pattern exploration and matrix representations of cellular network data. By extracting movement trajectories from mobile internet usage data, similarity measurements between users' spatial and temporal displacements, as well as home and work classifications, they described the urban dynamics of Santiago, Chile [7].

2.2 Urban Mobility

A typical application of CDR processing is the large social event detection and estimating the attendance during mass gatherings [8, 9, 10, 11, 12].

In the past ten years, many studies have achieved large-scale urban sensing and explored the mobility customs of individuals. Gender differences, however, have mostly been considered insignificant [13] or comparatively small [14]. On the contrary, Gauvin *et al.* showed that women visit fewer locations than men and spend more time at their top locations in Santiago, Chile. They were also able to relate mobility inequalities to several sociodemographic factors that may potentially explain the gender bias [15]. Aledavood *et al.* also showed that there are gender differences in call duration patterns across the day: women's calls are usually longer–especially during the evening and at night, concentrating those calls to a few individuals [16].

2.2.1 Radius of Gyration

Radius of gyration has been widely used to quantify the spatial dispersion of a person's daily activities in previous studies [1], [17]. Because we calculate it from raw data, it is referred as a low-level mobility indicator. Given a device's records as a list of location (l_i) and time(t_i) $\{(l_1|t_1), (l_2|t_2), ..., (l_n|t_n)\}$ the radius of gyration (r_g) can be defined as seen in Equation 1, where L is the set of locations, r_{cm} is the centre of mass of these locations and n_i is the number of visits at the location.

$$r_g = \sqrt{\frac{1}{N}\sum_{i \in L} n_i(r_i - r_{cm})^2} \tag{1}$$

In this calculation, after acquiring the corresponding locations of a user, the distance between these points r_i and the centre of mass r_{cm} was determined using the haversine formula.

2.2.2 K-Radius of Gyration

Similar to Section 2.2.1, k-radius of gyration is calculated using only the k most frequently visited locations of a user, defined in Equation 2, where L_k are the list of locations, $r_{cm}^{(k)}$ is the centre of mass calculated on these locations.

$$r_g^{(k)} = \sqrt{\frac{1}{N_k}\sum_{i \in L_k} n_i(r_i - r_{cm}^{(k)})^2} \tag{2}$$

Thus, $r_g^{(k)}$ represents the mobility range of a user, limited to the k-th most frequently visited location. Pappalardo *et al.* introduced the dichotomy of k-returners and k-explorers in human mobility [3]. To understand the characterisation ability of k-radius of gyration, let us assume an individual's $r_g^{(2)} \simeq r_g$, then his overall travel pattern is dominated by the two most often visited positions; hence called a k-returner. Contrariwise, if $r_g^{(2)}$ is much smaller than the unconditional radius of gyration, $k = 2$ is not sufficient, and more locations must be considered to achieve accurate characterisation, and the user is labelled k-explorer.

2.2.3 Entropy

In statistical mechanics, introduced in the 1870s by Ludwig Boltzmann, entropy measures the degree to which the probability of a system is spread out over different possible microstates. This application shows the diversity of a cell phone user's mobility: a higher value means the user is more likely to make the next call at a different location. In Equation 3 L is the set of locations visited by the user, p_i is the probability of the individual being active at the location [13]. H is also called the mobility diversity of an individual [18].

$$H = -\sum_{i \in L} p_i \log p_i \tag{3}$$

2.3 Socioeconomic status analysis

This study builds upon previous works at John von Neumann Faculty of Informatics, Óbuda University. The data set of the 2014 State Foundation Day has already been analysed [19] in respect of the large social event. In contrast, in this work, the attendees' socioeconomic status (SES) will be studied.

Using mobile phone prices as socioeconomic indicators have been proved to work well by Sultan *et al.* in [20]. They identified areas 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.

With respect to socioeconomic status analysis, Pintér *et al.* evaluated football fans' socioeconomic status indicators using their mobile phone details in Budapest during the 2016 UEFA European Football Championship. During data preprocessing, they eliminated CDRs from Subscriber Identity Modules (SIMs), which did not operate in mobile phones using Type Allocation Code (TAC) databases [12]. In another work, they analysed subscribers' wake-up times and explained how it correlates with their socioeconomic status. They also showed that the phone prices in the TAC database might have depreciated [21].

In an earlier study, Pintér *et al.* evaluated the connection between individuals' financial status and their mobility customs. The authors used the radius of gyration, entropy and Euclidean distance between home and work locations as mobility indicators. They applied data fusion methods with average real estate prices to determine the influence of wealth on mobility customs [22].

2.4 Clustering

In recent years, the interest in machine learning has skyrocketed and resulted in hundreds of serviceable algorithms available through multiple programming languages. As a result, we can unveil properties and connections in large data sets that would not be possible with basic heuristics using advanced data analysis methods. The three main approaches to problem-solving are divided by the nature of their signal or feedback to the learning system.

During supervised learning, the computer is presented samples of the learning input data set and their desired output. Based on these examples, the goal is to construct a general rule that maps the inputs to outputs.

A reinforcement learning program interacts with a dynamic system in which it must perform a particular objective. As it moves around in its problem space, the algorithm is provided feedback that is convertible to rewards, which it needs to maximise.

Unsupervised learning methods are used to discover underlying information in large data sets, where no labels are attached to the input for the processing algorithm. Instead of responding to a feedback signal, the algorithms identify commonalities in the operative data set. When it is assumed that the data set can be partitioned into groups on some available features, clustering techniques are used. A *cluster* can roughly be defined as a set of observed objects that are more similar to each other than to objects in other groups. The process of finding these groups is called clustering and has several approaches. Choosing the suitable clustering algorithm depends on the features of the input data set, outliers and the data objects, as well as the desired cluster characteristics [23].

The most common categories of clustering algorithms are density-based, hierarchical and partitional clustering. Each category has its strengths and weaknesses, and knowing them is crucial to determining which is most appropriate for an application. Density-based clustering aims to construct groups based on low and high-density regions. As a result, this clustering method does not require the number of clusters to be specified. Hierarchical clustering uses a top-down or bottom-up approach to create a dendrogram, which describes the hierarchy of points. It is often necessary to determine the number of clusters in this case. Partitional clustering also requires a user-specified k number of clusters to group objects in a non-overlapping way. In other words, every object must be sorted into only one cluster, and every cluster must include at least one object [24].

Machine learning algorithms have been proven to work well in discovering underlying information in mobile network data.

Al-Zuabi *et al.* developed an end-to-end solution to predict the personal information of customers. They tested a number of classification methods for gender prediction and achieved an accuracy of 85.6% [25].

Lenormand *et al.* used mobile network data combined with entropy-based metrics to measure the attractiveness of areas that can be used as a proxy for complex socioeconomic indicators. They used k-means clustering algorithms to group locations based on their visitors' average entropy, the radius of attraction and the ratio between the number of visitors divided by the population [26].

Ghahramani *et al.* proposed an approach to use an optimised self-organising feature map for revealing the underlying pattern structure of a mobile network data set. They tested density-based, hierarchical and partitional clustering methods for spatiotemporal data characterisation. They concluded that there is no universal definition of what is a suitable clustering method. Each algorithm has a different approach to solving spatiotemporal data clustering [27].

Based on a study by Zhao *et al.*, another promising research field is the problem of predicting individual socioeconomic status based on mobile network data. They proposed different approaches to solving the task and presented a semi-supervised hypergraph-based solution [28].

3 Methodology

In this section, the data set used throughout this study is first introduced, and then the CDR analysis apparatus and proposed framework are detailed. Lastly, an introduction to the unsupervised machine learning algorithms for clustering is given.

3.1 Data Sources

Processing large amounts of information requires a deep understanding of the data, careful planning and preparation of the working framework. The most common sources of spatiotemporal data analysis include relational databases, XML, JSON, GeoJSON and comma-separated value/text files. An essential step of data analysis is visual inspection, which also helps explore the available data types and prepare for preprocessing based on the identified anomalies in the source.

A CDR data set usually contains a caller identification (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' personal details helps research more than only trajectories and cell densities.

The data set used for this research was obtained from Vodafone Hungary Ltd. The number of active SIMs was 11,540,058 in Hungary, of which Vodafone had an estimated 24 per cent of the market share in 2014 [29]. The mobile network data set contains anonymous logs of customers' calls and text messages in Budapest and Pest County, Hungary, over approximately 1800 km^2. This study focuses on Budapest, using its administrative boundaries with an area of 525.14 km^2 and a population of 1,744,665 as of the 1st of January 2014 [30].

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 and the second and third are supplementary information about cells and devices.

The CDRs in this data set consist of a timestamp, a hashed device 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 exact 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, gender, whether it is an individual or a business and the subscription type (prepaid or postpaid). Unfortunately, some age and gender information is missing due to privacy restrictions.

TIMESTAMP	DEVICE ID	CELL ID	TAC
2014-08-18 00:00:00	4000132	5271	35580305
2014-08-18 00:00:00	4000286	4457	86984500
2014-08-18 00:00:00	4000286	4460	86984500
2014-08-18 00:00:00	4000320	2354	35603505
2014-08-18 00:00:00	4000424	4154	01318300

Table 1: CDR sample

ID	CELL TXT ID	CELL	LAC	LATITUDE	LONGITUDE
1	10011G	10011	115	47.5109745595405	19.0552397124617
2	10011U	10011	215	47.5109745595405	19.0552397124617
3	10012G	10012	115	47.5109745595405	19.0552397124617
4	10012U	10012	215	47.5109745595405	19.0552397124617
5	10013G	10013	115	47.5109745595405	19.0552397124617

Table 2: Cell lookup table sample

ID	AGE	SEX	CUSTOMER TYPE	SUBSCRIPTION TYPE
1	42	MALE	CONSUMER	POSTPAID
2	38	MALE	CONSUMER	PREPAID
3	27	FEMALE	CONSUMER	PREPAID
4	64	FEMALE	CONSUMER	PREPAID
5	UNKNOWN	UNKNOWN	BUSINESS	POSTPAID

Table 3: Device lookup table sample

In Tables 1, 2 and 3 samples of the preprocessed data are presented. Age and gender can only be published if the customer consented to use their data for marketing and research purposes; otherwise, we only see basic information on their subscription.

Preliminary data filtering shows 350 cells are missing location information; consequently, these records and the corresponding CDRs are excluded from further analysis. Aggregating sightings show an amount of 1,556,951 distinct devices, of which 553,944 are missing age and gender information.

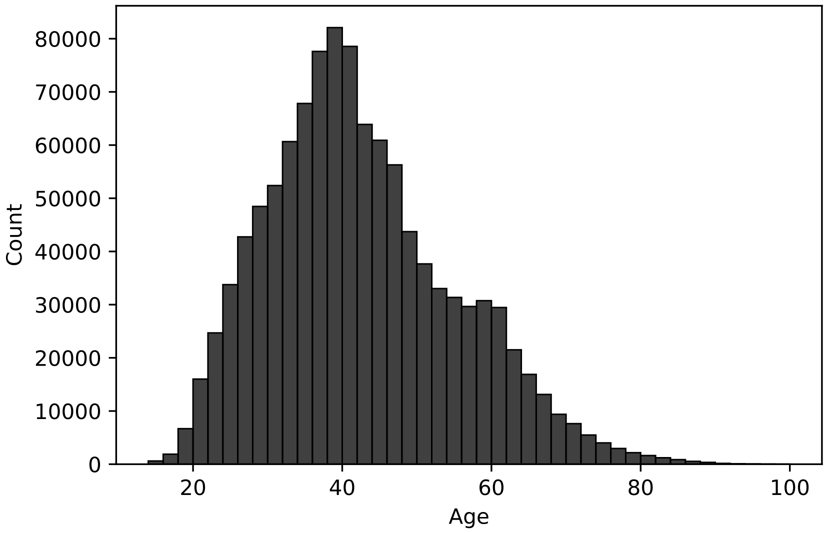
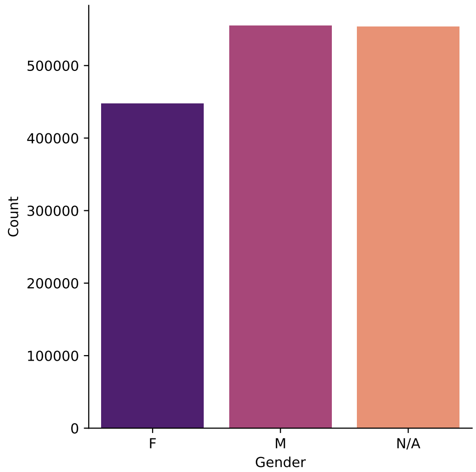


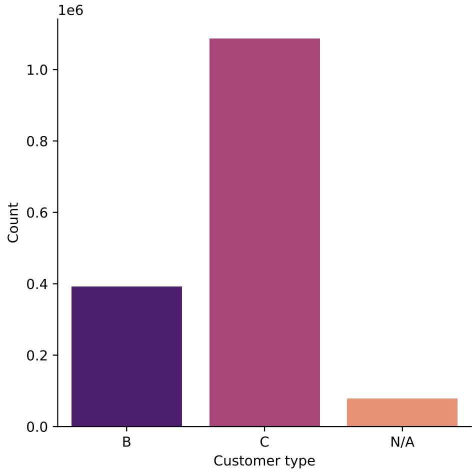
Figure 1: Age distribution among users.

Figure 1 shows age distribution using available data, Figures 2a, 2b and 2c show gender, customer and subscription type distributions, where *N/A* stands for the unknown values. Age distribution looks more or less like standard, normally distributed data. In contrast, the known gender distribution is roughly equal, although there is a high amount of unknown age and gender information. This might explain the low amount of young (18-20 years old) customers, or they still have their contracts under their parents' names.

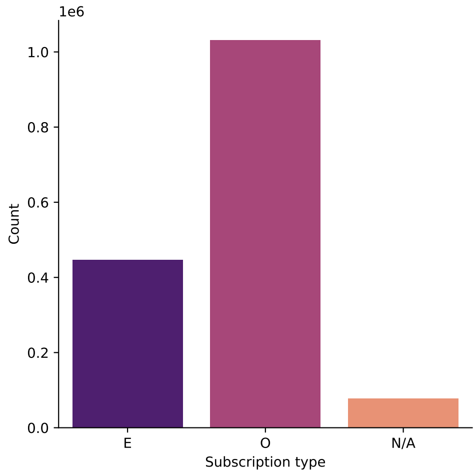
Hungary is known to have a *Baby Boomer* demographic cohort, which is visible in Figure 1, an increasing pattern around the age of 60. On the other hand, users around the age of 40 own most devices.



(a) Gender distribution



(b) Customer type distribution



(c) Subscription type distribution

Figure 2: Additional information on users.

Brand	Model	Year	Month	Smart	Price [€]	TAC
Apple	iPhone 3G	2008	7	True	90	01202400
BlackBerry	Passport	2014	6	True	170	35185206
LG	KP501 Cookie	2009	2	False	80	35701403
Nokia	6600	2003	10	True		35936500
Vertu	Constellation Quest	2010	12	True	6140	35379304

Table 4: TAC database sample

As a socioeconomic status 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 is from the data provided by 51Degrees fused [12] with the GSMArena database [31].

The source includes mobile phones' brand, model, release year, release month, logical value, whether it is a smartphone or feature phone, price in EUR and TAC as seen in Table 4. It is necessary to mention that the mobile phone prices might have depreciated [21]. For example, the database price of EUR 90 for the iPhone 3G can be misleading, as it is most likely the two-year contract price for the cheaper 8 Gigabyte (GB) version.

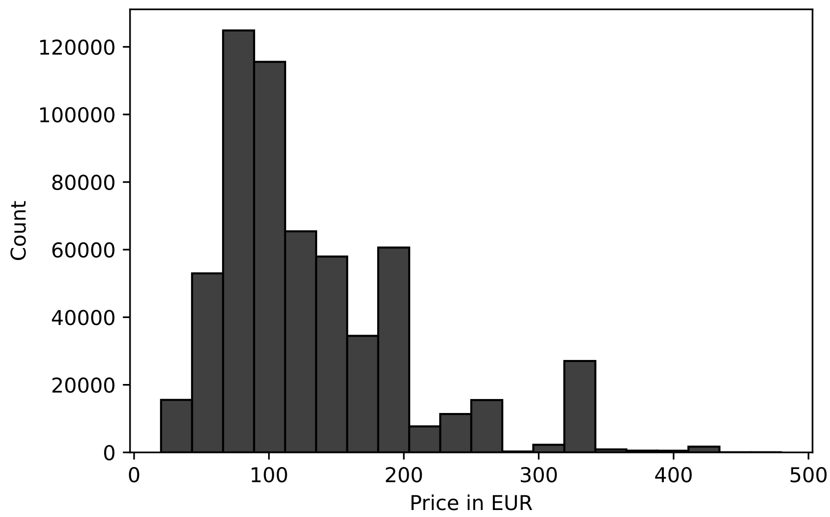


Figure 3: Mobile phone price distribution among users.

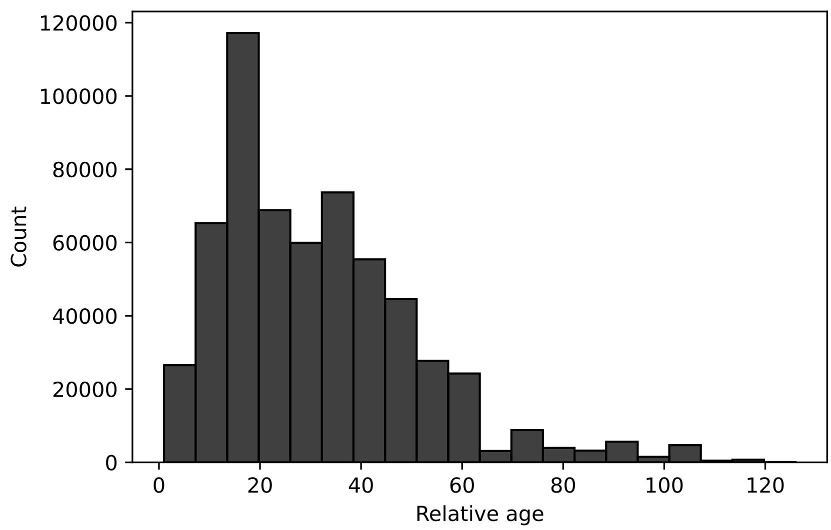


Figure 4: Mobile phone relative age distribution among users.

The users with the found mobile phone properties have a price and relative age distribution seen in Figures 3 and 4. In numerous cases, mobile phones are unknown, and this missing information reflects in the distribution. However, it is visible that the most common categories are the up to EUR 100 and 1-year-old phones. Also notable is that there are several phones in use up to ten years old, with over 20,000 at the age of five.

3.2 Data Preprocessing

Preliminary data set descriptions have been introduced in Tables 1, 2 and 3 in Section 3.1. The received data has already been cleared of unnecessary spaces, and long hash IDs have been swapped with incrementing integer values and uploaded to an SQLite database. This general-purpose storage solution was considered sufficient to run the proposed framework and has not been moved onto a more application-specific relational database.

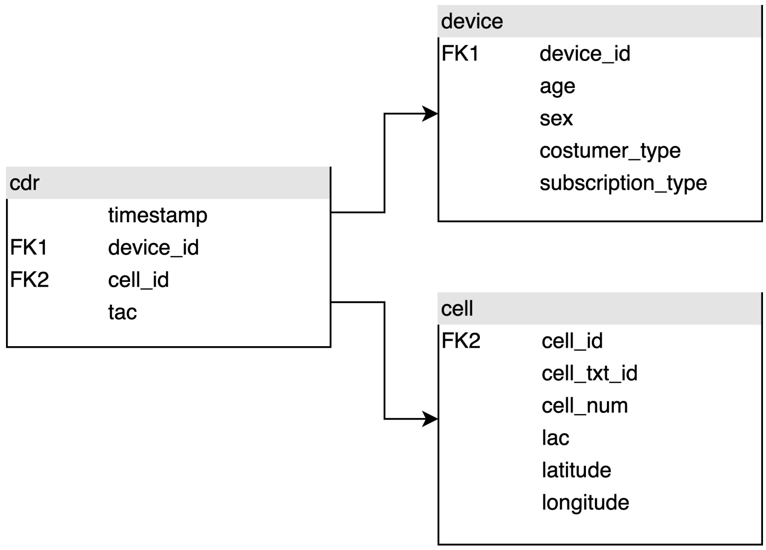


Figure 5: The relational database model containing CDRs and lookup tables.

As part of the preprocessing, cell IDs had to be modified. It was discovered that more than one cell shared the exact GPS location in the database. This was most likely the result of having both 2G, 3G and 4G transceivers facing different directions on the same base transceiver station. It is common practice for mobile network operators to have colocated cell towers for multiple cellular networks to save space and utility costs.

Unfortunately, we do not have information on the directions of the cells; hence the cells with the exact location have to be merged as in Equation 4. The analysis showed that anywhere from one to ten cells could be located on the same base station. This information is preferred to be used later, as it involves computations with fewer records, bearing in mind that after merging these cells, their IDs in CDRs must also be replaced with the new *base station ID.*

$$BS = \Big\{(lat|lon), (cell_1, cell_2, ..., cell_n)\Big\} \tag{4}$$

The database structure has not been modified during generating the new cell IDs so that the original data source is compatible with the new one. Private – foreign key connections have been created to show the relationship between the three tables as seen in Figure 5.

3.3 CDR Visualisation

After understanding the structure of the given mobile network data set, we have to establish a framework for evaluating the analysis results. The proposed solution is map-based data processing. We divide the recorded terrain into pieces and consider every person recorded in the area for a specific indicator calculation when aggregating the crowd. Defining the used division principle is an essential step toward spatial data analysis.

Displaying the merged cells, known as base station locations, on a map gives us a good representation of the geography. We can eliminate any outlier cell tower locations and perceive the cell sizes – moving out from the city centre, cell sizes increase, possibly from the lowering population density.

One way of dividing the area is using equal-sized two-dimensional rectangles. This would have the advantage of providing a homogeneous resolution. However, as seen in Figure 6, based on the structure of the city, the cell tower density highly varies based on the distance of the city centre. Therefore, choosing the right granularity would be impossible using the exact sizes in Budapest's inner city and outskirts.

Computing Voronoi diagrams under Euclidean metrics would give us the closest seed (cell tower) to any point of a plane. As a result, the process decomposes our map to regions called Voronoi cells, determining the smallest resolution achievable with this data set. An example of this representation is shown in Figure 6. 810 Voronoi polygons have been created from base transceiver station locations as the seeds marked by dots on the map.

It is visible in Figure 6 that the cell density in the city centre is much higher; in contrast, in the surrounding cities and villages, there are often only one or two cell towers per settlement. This is due to the different number of devices and cell tower loads. In this study, the analysis focuses on Budapest, using its administrative boundaries, marked on the map by the thick line.

The created voronoi polygons can be saved for later results visualisation and map creation. The most advantageous file format for storing geographical data is GeoJSON; it supports various geometry types, such as points, strings and polygons.

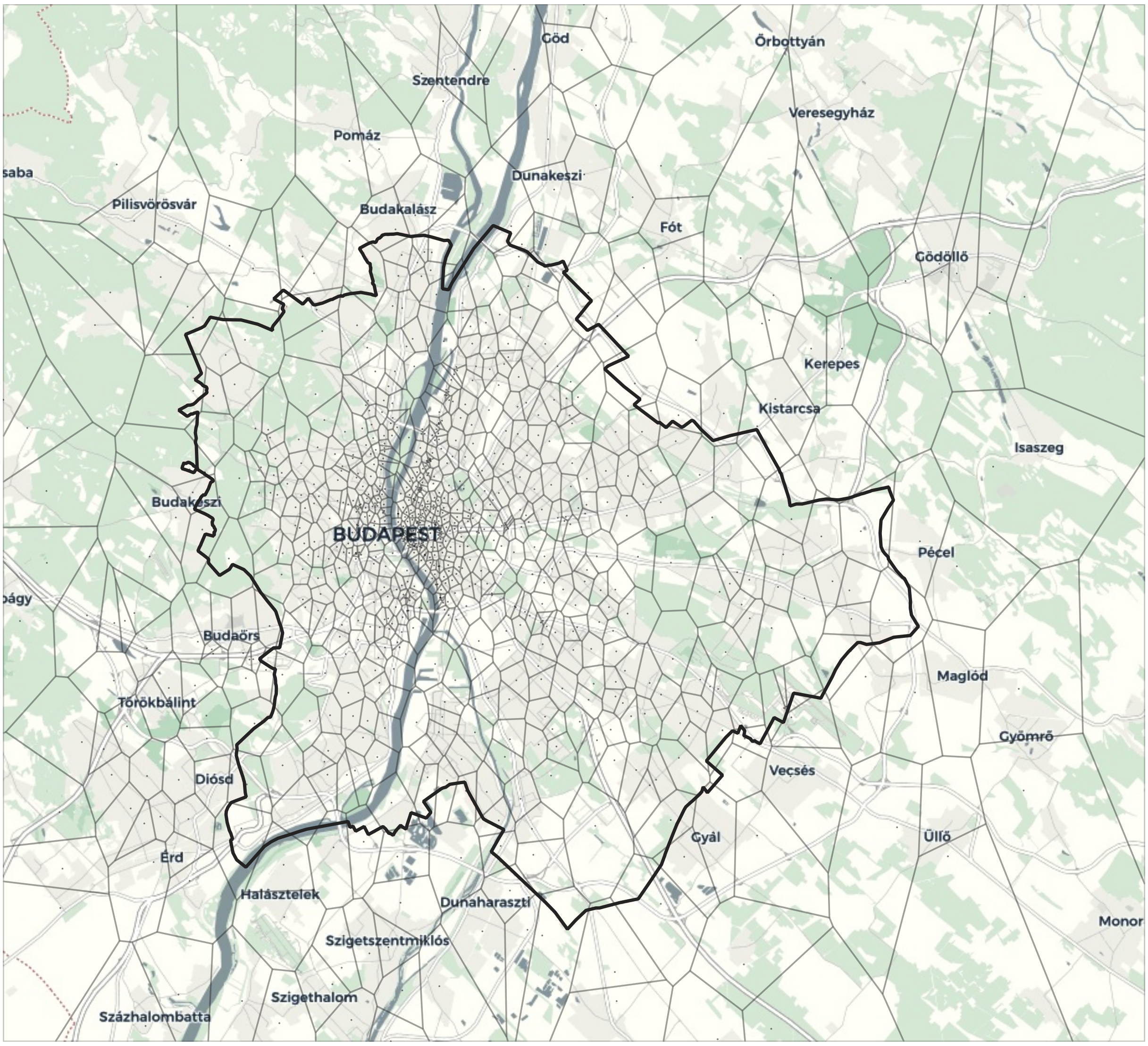


Figure 6: Voronoi polygons with cell towers as seeds in Budapest and its agglomeration.

3.4 Event Analysis

In this section, the methodology of the proposed large scale event analysis framework is presented. The aim is to produce spatial descriptors for the areas covering Budapest's Hungarian State Foundation Day celebrations. The call detail record analysis focuses on the city centre of Budapest during a large scale event in August 2014. The attendees of the main event of the Hungarian State Foundation Day are analysed based on their socioeconomic status. This section 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.

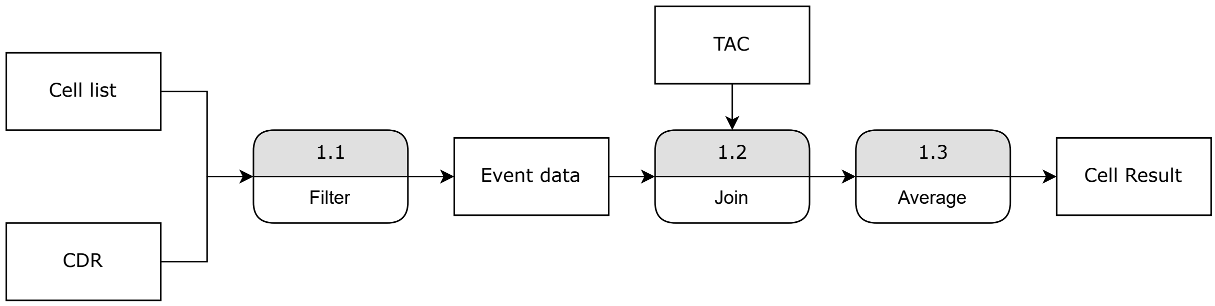


Figure 7: The large social event analysis framework.

The proposed framework is shown in 7 and works as follows: the cell list containing the cell towers that make most of the mobile phone transactions has to be selected and stored in a list. This can be done manually, knowing the terrain well or selecting the points of interest and the cells based on a predefined rule (e.g., x metre radius). Next, the CDR table from the database has to be filtered (Figure 7/1.1) using the cell list and time of the event, thus producing the event data output. This list contains all the call detail records that were most likely made by users at the time and place of the event. Next, the mobile phone price and relative age are used from the database introduced in Section 3.1 to estimate the socioeconomic status of attendees. Finally, the event data CDRs are joined on the records' TAC, selecting the corresponding phone property (Figure 7/1.2). At this point, the records without known TAC details can be eliminated. In some cases, where the TAC is known but the mobile phone property is not, records must also be eliminated from further analysis. An average SES indicator is used to generate a cell-by-cell descriptor, calculated in (Figure 7/1.3).

The final output of the proposed framework produces a list of cell IDs with their average socioeconomic status indicator. This list can be used to create maps and scatter plots to analyse the results and stored in a comma-separated value or JSON/GeoJSON file. Note that the framework can only function with one SES indicator at a time. However, as we only have two of these in this study, the calculations only have to run twice, with the appropriate mobile phone property configured 7/1.2).

3.5 Clustering

The proposed mobile network user clustering framework is presented in this section. Similar to Section 3.4, the aim is to produce a cell-by-cell descriptor, but in this case, the groups are created by an unsupervised clustering algorithm. In addition, different cases are determined by what input the machine learning algorithm is given to work with, but as a general rule: a list of cells and their mobility and socioeconomic status indicators.

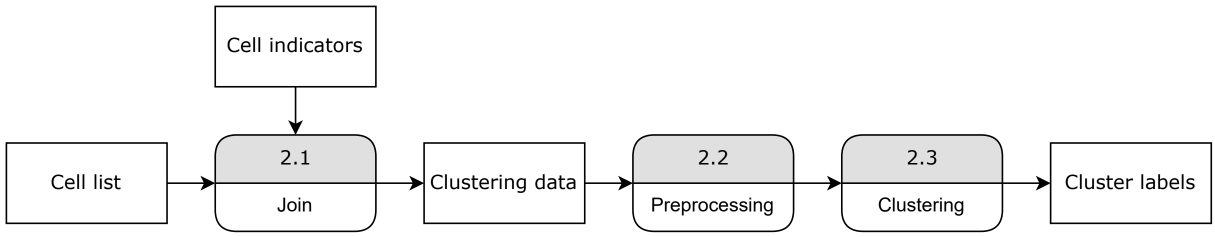


Figure 8: Cell indicator clustering framework.

The proposed framework is shown in Figure 8 and includes the following methods: the cell ID list must be prepared, where the analysis is to be concluded. The input cell list is then joined with the precomputed cell indicators (Figure 8/2.1). The output is the raw clustering data, which needs to be preprocessed (Figure 8/2.2) because the different indicators have different units of measurement. It is crucial to prepare input data for clustering algorithms, as they might result in inaccurate output without data set standardisation. The clustering algorithm (Figure 8/2.3) runs using the preprocessed data. It outputs the cluster labels, indicating which cells are more similar based on the given metrics than others. Finally, the analysis compares more clustering algorithms, chosen based on their characteristics and usage in related works (Section 2.4). The number of clusters must be determined beforehand using one or both of the following approaches: the elbow method or the silhouette coefficient.

The labels can be used to generate maps and scatter plots to analyse the results; hence they are saved in a separate file, noting what indicators were used in the process. Different indicator pairs need a separate configuration and run, but this solution is sufficient for this study, given the number of combinations.

4 Implementation

In this section, the study describes the implementation of previously proposed data analysis frameworks and the complementary calculations needed for the operation. First, the essential mobility indicators' computation is introduced, the large social event analysis, and lastly, the clustering implementation details are uncovered.

4.1 Mobility Indicators

In Section 2.2, the study introduced the methodology for deriving the most common mobility indicators used in the literature. These metrics are determined for every user based on the records they produced during the data accumulation period. In the case of this data set, the select queries on the database and calculations on the retrieved spatial data have to be applied to approximately 1.5 million individuals. The calculation would take a significant amount of time using a single thread. Therefore the analysis used the `multiprocessing` tool in Python, creating data-parallel threads to translate every user's call detail records into mobility indicators.

4.1.1 Radius of Gyration

The radius of gyration is one of the most frequently used low-level mobility indicators, introduced in Section 2.2.1. It represents the circle's radius where a specific user can usually be found. The implementation covers the computation for every user. The distribution is shown in Figure 9, with the only radius of gyrations greater than zero.

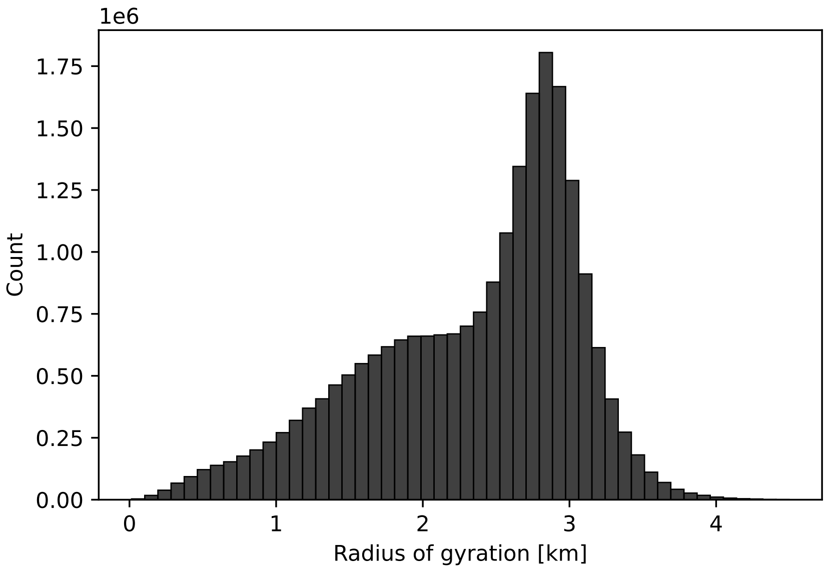


Figure 9: Radius of gyration distribution among mobile phone users.

4.1.2 Entropy

The mobility entropy of an individual gives the diversity of its travels between consecutive locations. Introduced in Section 2.2.3. The indicators calculation was completed during implementation, and the results are shown in Figure 10.

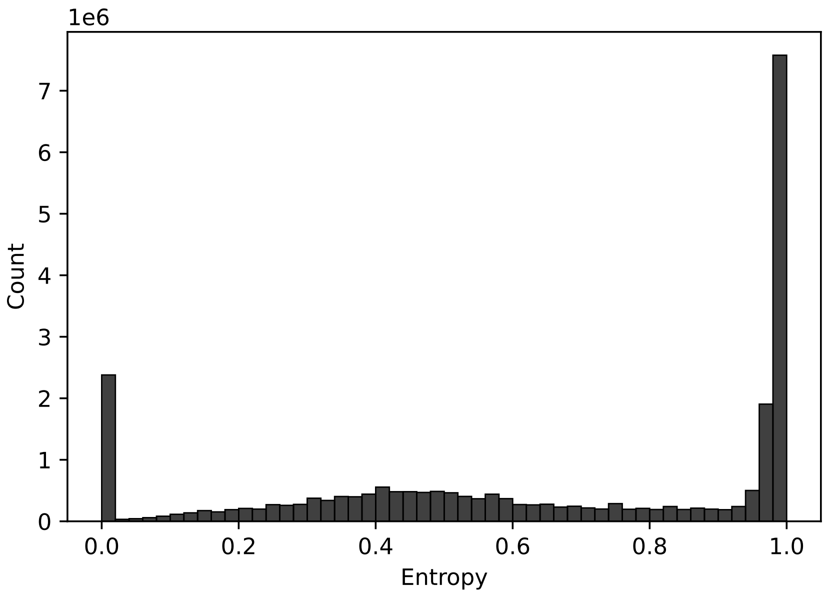


Figure 10: The mobility entropy distribution among mobile phone users.

It is visible that the most amount of users has an entropy close to one, meaning they almost always make phone calls from different locations. The second most amounts of users have an entropy of zero, meaning they always use their mobile phones in the same cell. The rest of the entropy groups peak around 0.5, meaning the users share their phone calls between two locations.

4.2 Event Analysis

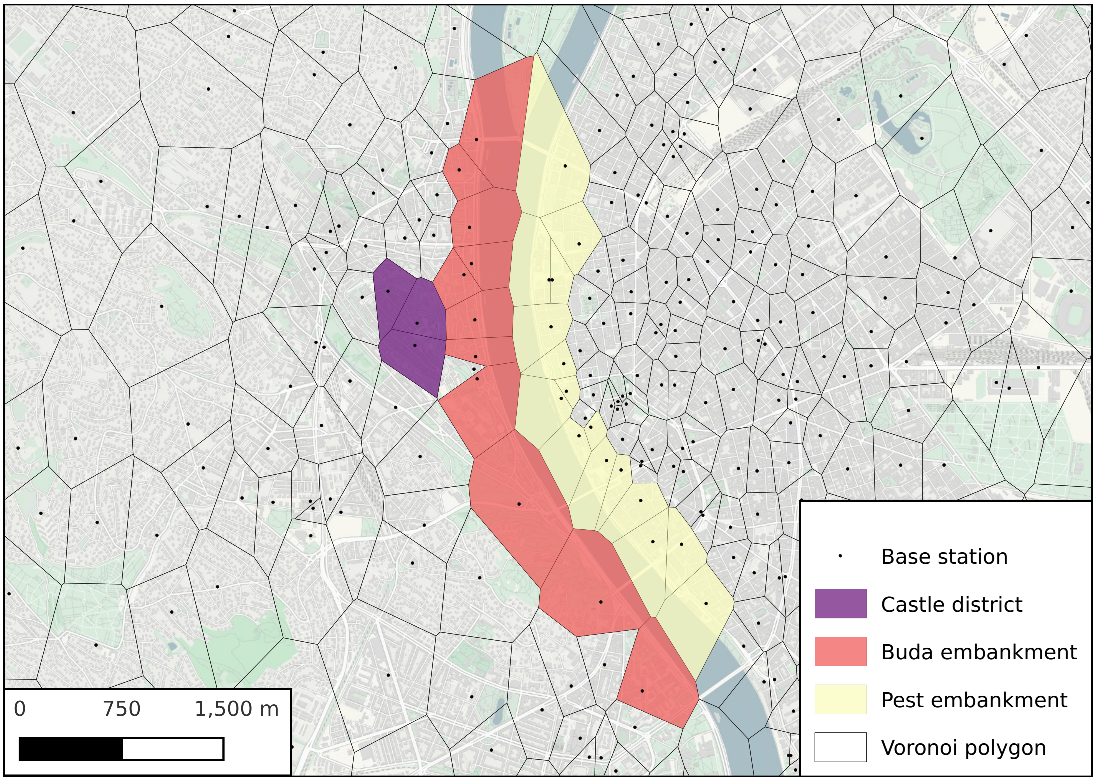


Figure 11: The selected cells for the large social event analysis.

A data processing framework has been developed to select the relevant CDRs with the associated SES indicators. Cells along the two sides of the river Danube are expected to be the first primary cells servicing the 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 11. Four cells were removed from the evaluation due to the meagre activity count of less than 500 during the event. Extracting the selected cell IDs helps filter and transform the large CDR database table to a more manageable, in-memory, pandas `DataFrame` structure [32]. The temporal filtration rule for the fireworks data is $\pm$ 30 mins to the actual event duration. This includes 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. In addition, 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).

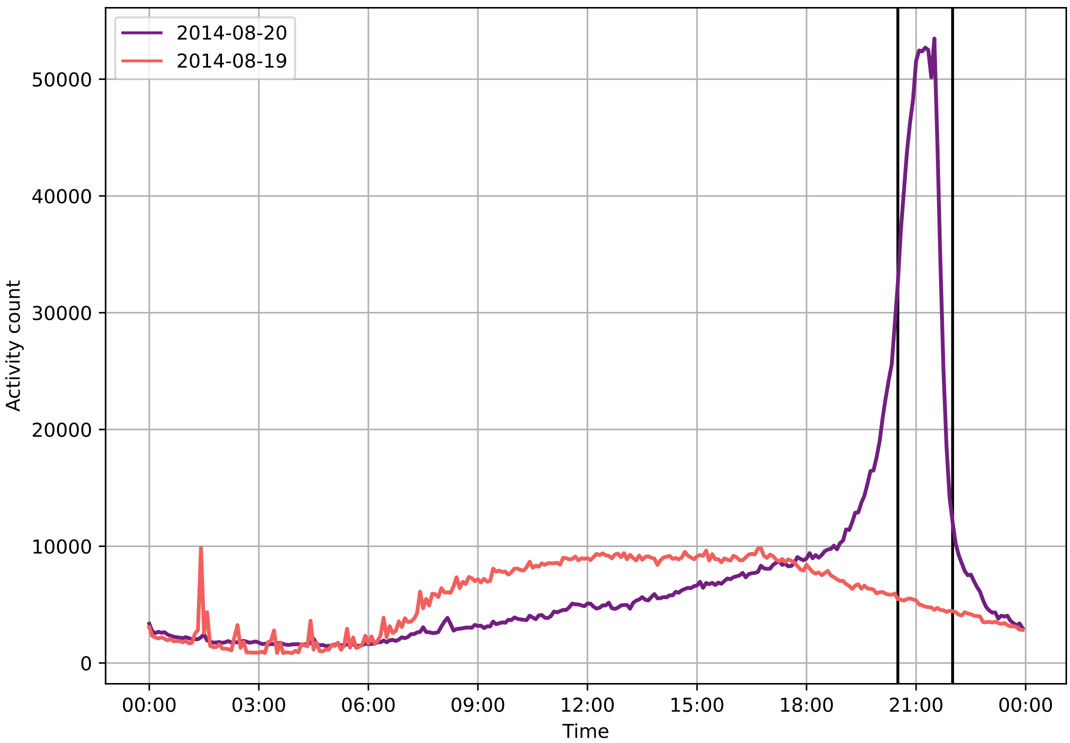


Figure 12: Daily cell activities in the studied area.

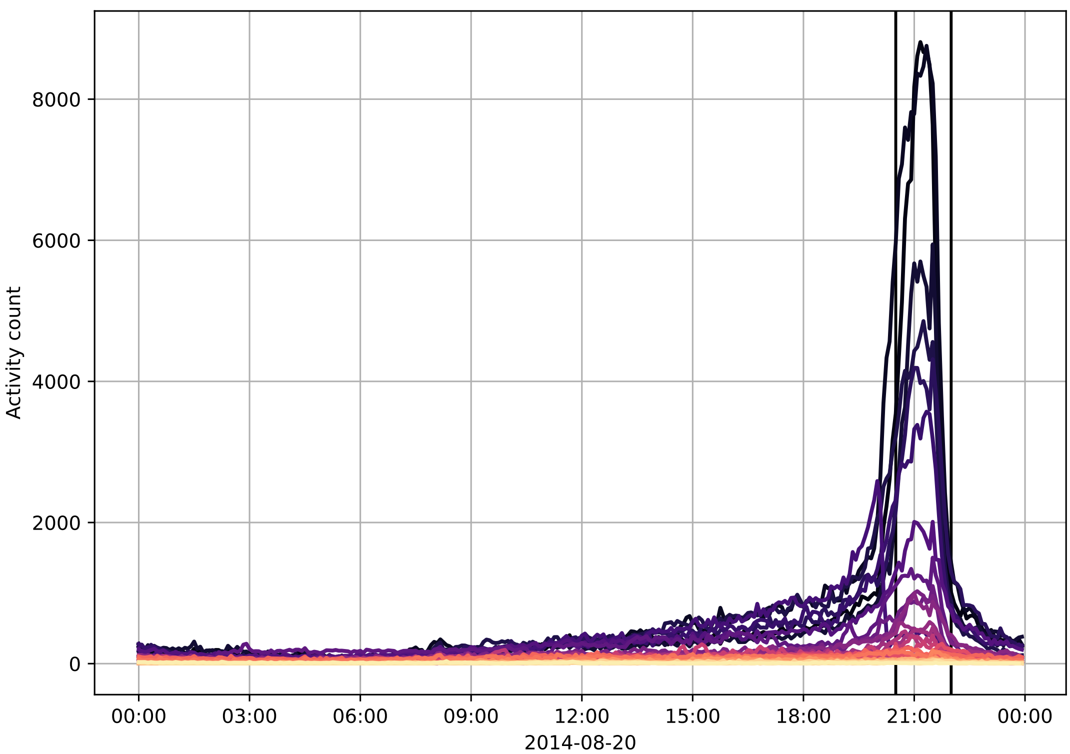


Figure 13: Daily cell activities between the studied cells.

4.3 Clustering

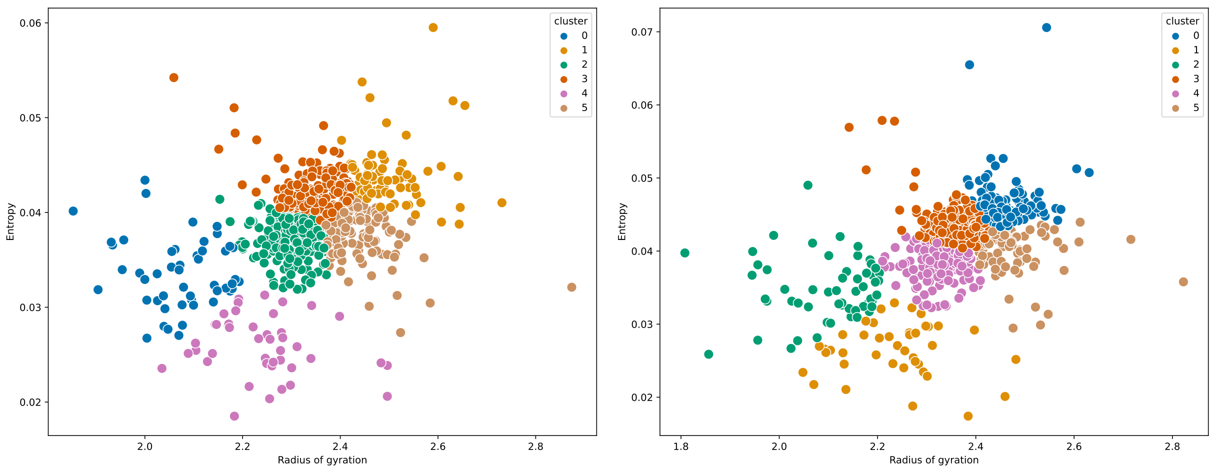
This section uses a conventional k-means algorithm to cluster mobile network cells based on mobile phone users' activity, mobility and socioeconomic status indicators. K-means clustering aims to find an optimal arrangement of objects, where, in a cluster, the sum of the squares of object distances to the centroid is a minimum.

As cell mobility indicators, the mean of user radius of gyration and entropy is used, calculated in Section 4.1. In addition, SES indicators from Section 3.4 are also used in the analysis.

Data preprocessing and clustering were executed using the `scikit-learn` [33] machine learning library in Python. It features several useful data scaling methods and the most common clustering algorithms. This work does not aim to create new clustering methods, as the currently available algorithms have been proven to work well on spatiotemporal data sets.

This analysis aims to discover human mobility and socioeconomic groups using historical data; hence clustering runtime does not need to perform well compared to live data streams. During the discussion of the results, runtimes are mentioned, if necessary, but they remain in the second – minute intervals.

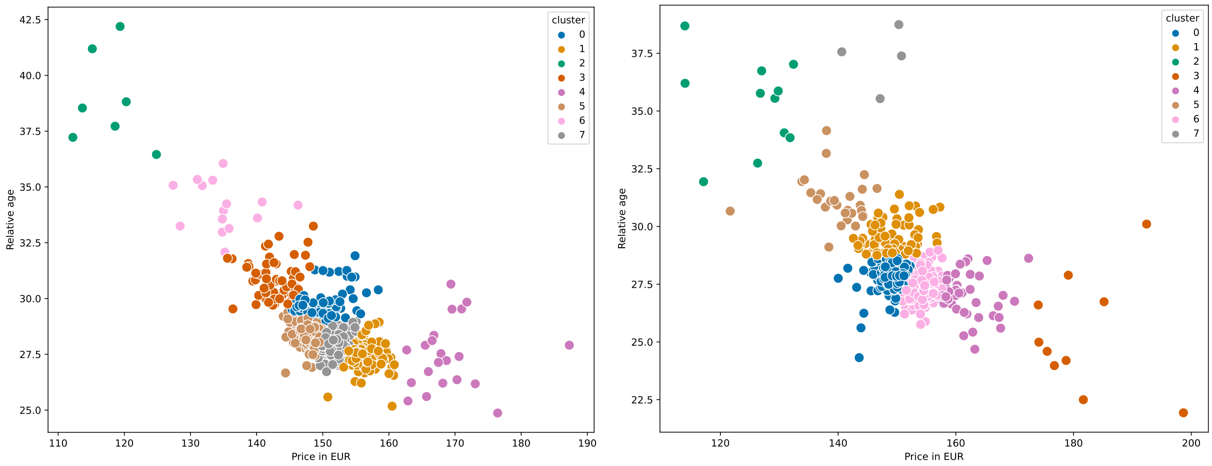
The study focuses on Budapest; hence the cell ID list had to be determined. Using the official city limits of Budapest as a vector layer in QGIS [34], the list could be exported after a clipping command on the previously created voronoi cell-map (Figure 6).



(a) (b)

Figure 14: Mobility clusters in Budapest on the 19^{th} (a) and 20^{th} (b) of August.

A parallel computing framework was created with n processes to generate the cell list with the corresponding mean values of the indicators used during clustering. Given the list of all users in the data set with their calculated mobility and socioeconomic indicators from Sections 4.1 & 4.2, the Budapest cell IDs were split into n equal pieces. Those were then passed to the parallel processing framework to determine the cell average indicator in the following manner.

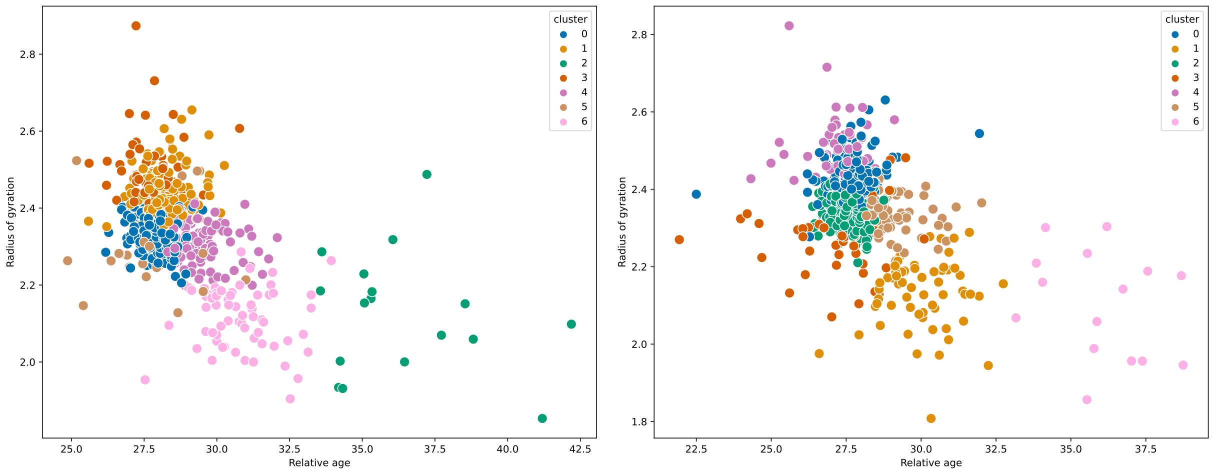


(a) (b)

Figure 15: Socioeconomic clusters in Budapest on the 19^{th} (a) and 20^{th} (b) of August.

First, an SQL query is executed on the CDR database via `sqalchemy` Python package [35], returning the unique device IDs from every cell, limited to the given time frame. Then, using the complete list of device IDs and the corresponding indicator, a join method is performed on the device ID. Finally, the cell mean indicator value can be determined on the indicator column. Concatenating the results of every cell in a process, then every cell sub-list of the n processes, the Budapest cell list result is given. Runtimes were around 1.5 minutes using 30 processes on 30 cores, with the bottleneck possibly being the SQLite queries for the device ID lists for every cell.

The above framework was applied to create mobility indicator lists of the average radius of gyration and entropy and socioeconomic indicator lists of average mobile phone price and relative device age. Clustering algorithms, however, require a unified scale in every dimension for the input data. Standardisation of a data set is a standard exercise before using many machine learning algorithms, as they might result in poor output if the indicators do not look more or less like standard normally distributed data [36].



(a) (b)

Figure 16: Mobility - Socioeconomic clusters in Budapest on the 19^{th} (a) and 20^{th} (b) of August.

The above-discussed indicators are set on different scales; thus, preprocessing is necessary. The `preprocessing` module of the `scikit-learn` Python library offers a `StandardScaler` method to standardise features of a data set by removing the mean and scaling to unit variance. This is the first phase of cluster creation in every case.

K-means clustering algorithm requires the number of clusters to be specified. This was determined by the elbow method [37], which involves running the k-means algorithm several times, incrementing k with each iteration and recording the sum of the squared estimate of errors (SSE). As one increases the number of clusters, SSE decreases, and as the curve bends, the elbow point represents a reasonable trade-off between inaccuracy and the number of clusters. As the next step towards clustering, this method was executed in every case to determine a minimum k.

For cluster centroid initialisation, the `k-means++` scheme was used, which determines them to be generally distant from each other, resulting in better distribution than random initialisation [38]. As clustering algorithm, the `sklearn.cluster.KMeans` class was used. The number of clusters and the initialisation method known, the computation time was seconds on the standardised cell mobility and socioeconomic status indicator data sets. Three clustering bases' were determined:

- Mobility – radius of gyration / entropy
- Socioeconomic status – mobile phone price / relative age
- Mobility & socioeconomic status – radius of gyration / entropy / mobile phone price / relative age

The above clustering methods are performed on the cell-level device ID aggregation on two separate days: the 19^{th} and 20^{th} of August, considering the former is a usual workday, while the latter is a national holiday with all day long celebrations around Budapest, introduced in Section 3.4.

The scatter plot of the clustering result labels are shown in Figures 14, 15 & 16. On the horizontal and vertical axes, two indicators are marked, the clusters are numbered from 0 to $k-1$ and the cells as points are coloured accordingly. In Figure 16, where four dimensions are present at clustering, for 2D representation, the radius of gyration and mobile phone relative age are used.

The cell IDs with their calculated cluster labels were saved for map representations in each case. The cropped Budapest voronoi cells around the cell towers can be coloured using these clusters, creating an appropriate visual representation of the spatial groups.

5 Testing

5.1 Unit Testing

Two unit tests were written to validate the mobility metrics computed from call detail records. As the framework was developed in Python, the `doctest` module was used [39], which executes code snippets that start with `>>>` and compares it against the expected output given in the following line.

```
def calc_ent(loc):
    """
    This function returns the entropy calculated from the latitude
    and longitude points in the arguments.
    >>> calc_ent([])
    nan
    >>> calc_ent([1])
    0.0
    >>> calc_ent([1,2,3,4])
    1.0
    >>> calc_ent([1,1,2,2])
    0.5
    """
```

```
def calc_rog(lat, lon):
    """
    This function returns the radius of gyration calculated from the
                                    latitude
    and longitude points in the arguments.
    >>> calc_rog([], [])
    nan
    >>> calc_rog([47.5], [19.05])
    0.0
    >>> calc_rog([47.5, 47.5], [19.05, 19.05])
    0.0
    >>> calc_rog([47.5, 47.5, 47.5], [19.05, 19.05])
    nan
    """
```

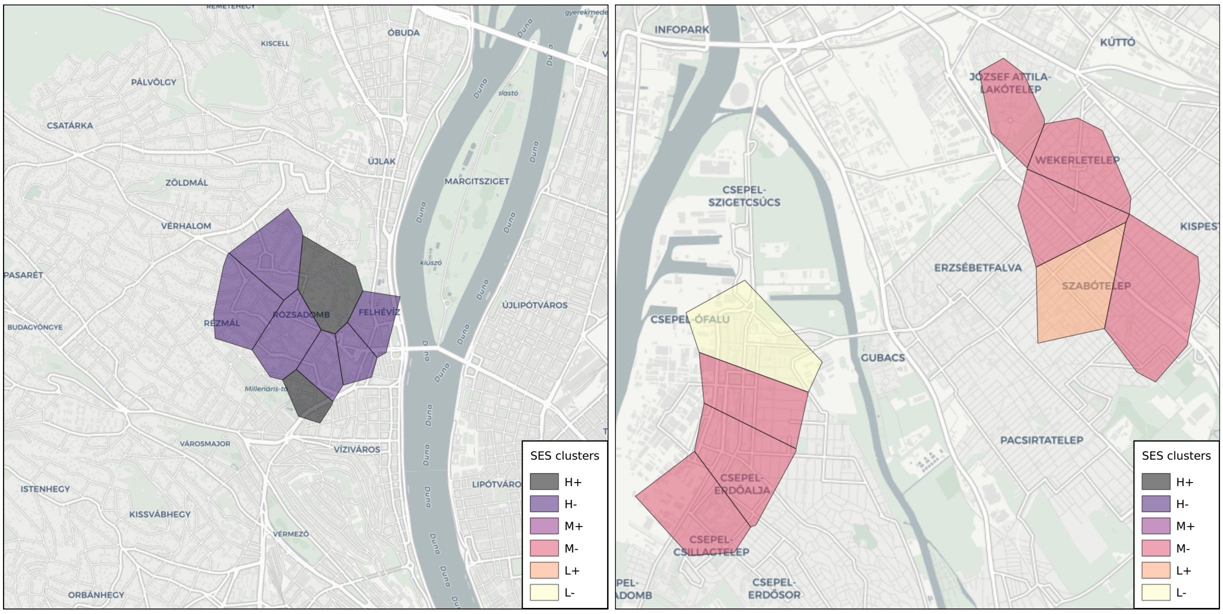
The mobile phone user entropy and radius of gyration indicator calculation algorithms passed all four of the tests; hence, the results are valid. The whole functions are published in Appendix A and B.

5.2 Case Study

The following case study is concluded to validate data integrity and the performance of the call detail record processing framework. Three well-known areas of Budapest are compared in terms of average socioeconomic status indicators. Rózsadomb is one of the most prestigious areas in Hungary, with exclusive villas, numerous embassies and one of the highest housing prices in the country.

Kispest includes one of Budapest's most prominent panel housing estates, built between the 1960s and the 1980s with over 10,000 flats. Csepel is located on the northern end of Csepel-sziget, the largest island on the river Danube in Hungary. The area is worth mentioning because more than 60 per cent of the 32,000 dwellings are housing estates, including Csepel-Csillagtelep, with a population of around 20,000.

The hypothesis is that Rózsadomb has higher average socioeconomic status indicators than Kispest and Csepel. Using the proposed clustering framework (Section 8), the area of Budapest is grouped into six categories. The clustering results of the three areas are visualised in Figures 17 and 18.



(a) (b)

Figure 17: The areas of the case study, Rózsadomb (a) and Csepel – Kispest(b).

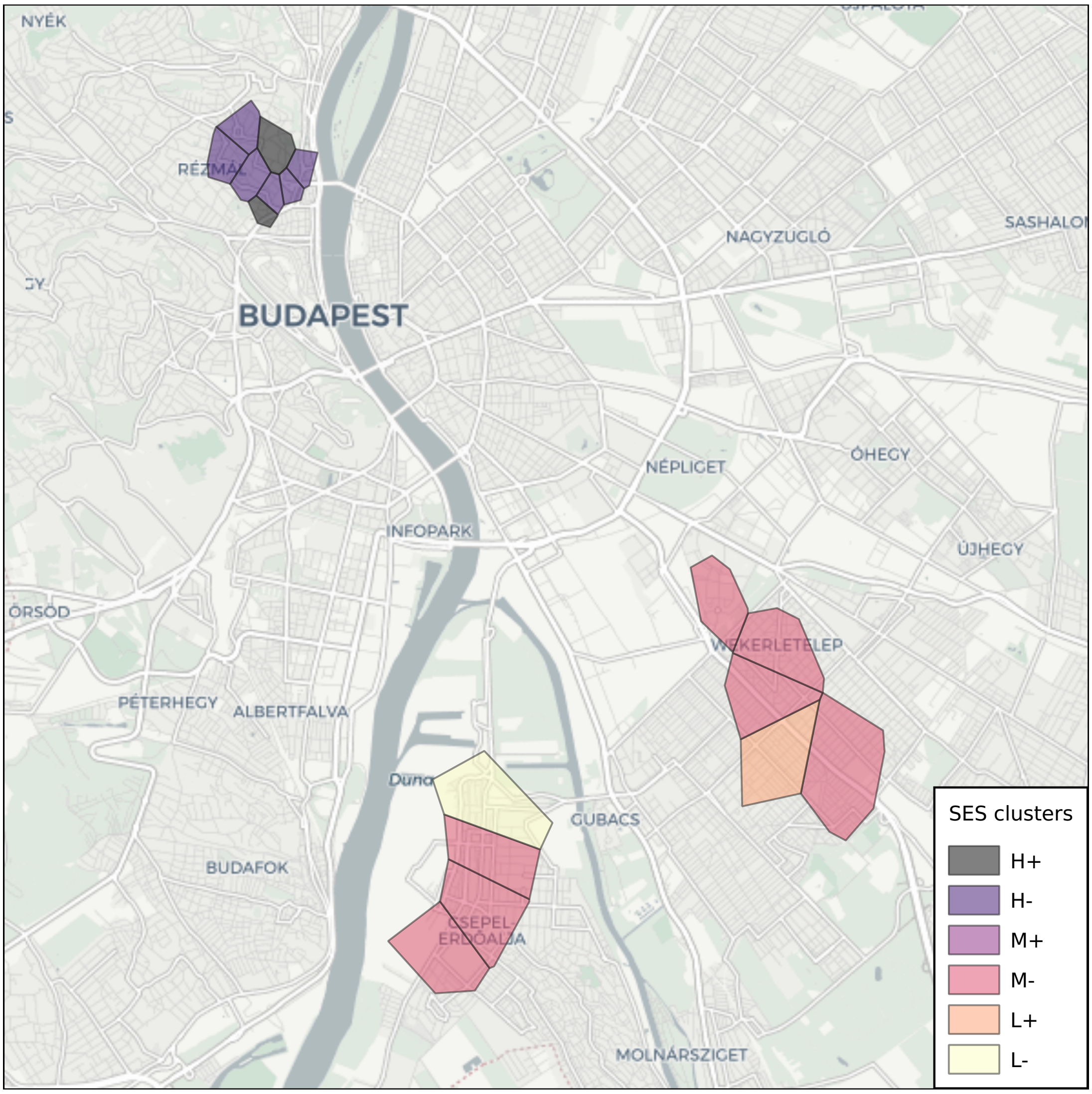


Figure 18: The case study areas compared

The results show significant differences between Rózsadomb and Csepel – Kispest. While the former primarily consists of H– and H+ socioeconomic status classes, the latter two are mainly M– and a few lower classes. This proves the hypothesis, stating that Rózsadomb has higher average socioeconomic status indicators than Kispest and Csepel. Based on the outcome of this case study, the mobile network data processing framework works correctly, and the clustering based on average mobile phone prices and ages reflects authenticity.

6 Results

In this section, the results of the study are introduced. First, the data processing framework development results are summarised, and then the event analysis and clustering outcomes are discussed.

6.1 Data Processing Framework

The aim of this study was to create a mobile network processing framework to infer user mobility and socioeconomic status indicators using call detail records and base station locations. The framework has been used throughout the large scale event analysis and clustering and was detailed in Sections 3.4 and 3.5.

The data processing framework was implemented in Python using several modules and libraries, detailed in the following sections. It consists of database queries, tabular data processing functions, and producing a cell list with their specific metrics. It was proven to work well with any spatiotemporal data from the source database, producing useful information for the event analysis and cell clustering. Furthermore, the results presented in Sections 6.2 and 6.3 are the products of this framework. The implementation allows easy modification possibilities to facilitate other call detail record databases or sources to infer socioeconomic status or other metrics.

6.2 Large Social Event Analysis

The large social event analysis literature is introduced in Section 2.2, the proposed methodology in Section 3.4 and the framework implementation in Section 4.2. The results discussed in this section have been submitted to the 16th IEEE International Symposium on Applied Computational Intelligence and Informatics.

St. Stephen's Day is celebrated in Hungary every year on the 20^{th} of August. Tens of thousands of people visit the capital for an all-day-long celebration and the main event, a 30-minute long firework 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 show a significant spike in cellular activity. Therefore, they were the primary subject of the event analysis.

The research focused on the socioeconomic status 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 Figures 12 and 13.

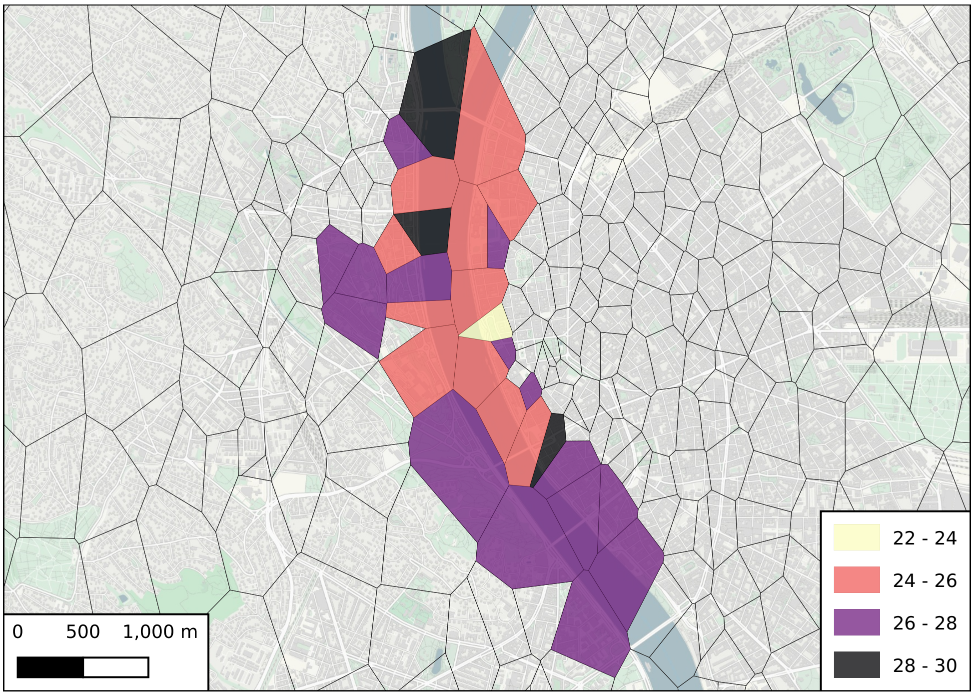


Figure 19: Average mobile phone relative age distribution by riverside cells among the large social event attendees.

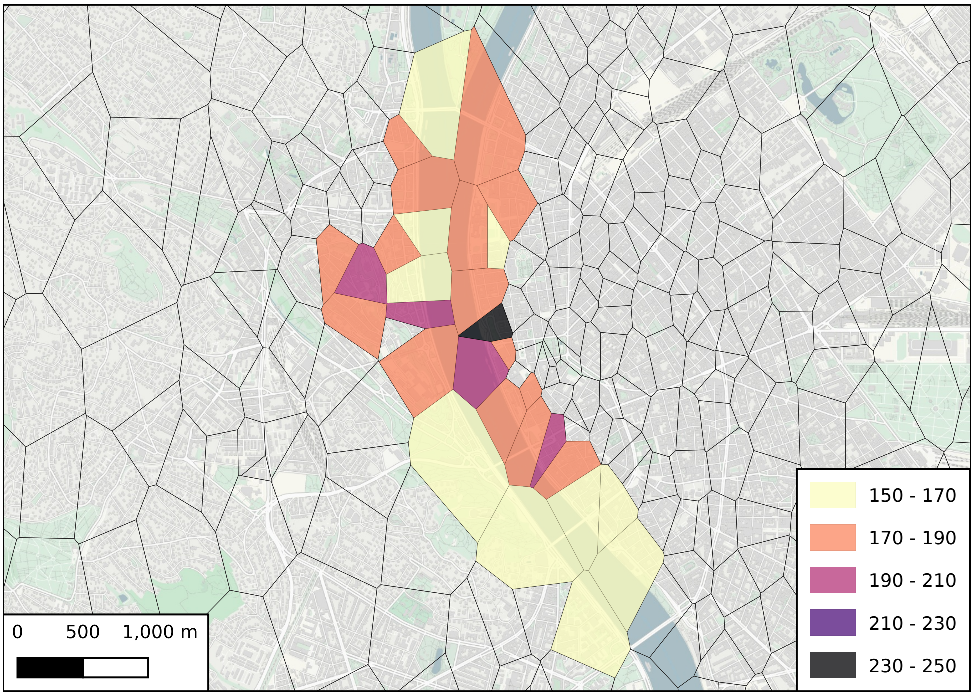


Figure 20: Average mobile phone price distribution by riverside cells among the large social event attendees.

A cell-by-cell average of mobile phone prices and relative ages was calculated for the SES indicator distribution. Figures 19 and 20 show 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.

The results demonstrate an opposite trend between mobile phone price and age. The scatter plot of the same data is shown in Figure 21, where the Pearson correlation coefficient $= -0.7329$. Data points are coloured based on their location in the city, but there are no visible groups based on socioeconomic status.

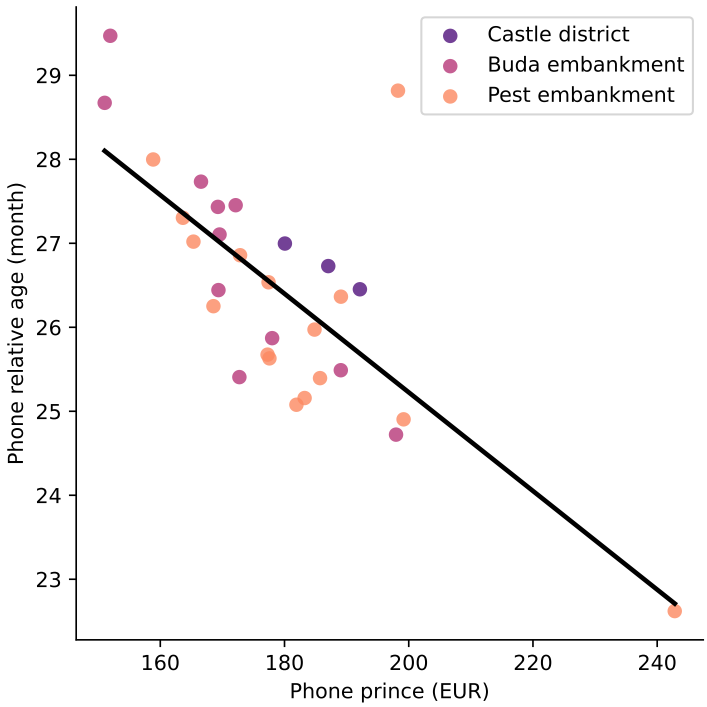


Figure 21: The correlation between average phone prices and ages in cells, where Pearson's $r = -0.7329$.

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.

This section 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.

It was expected that in Buda, where the housing prices are higher [22], more expensive and newer phones would generate the majority of the activity. Nonetheless, this study 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.

6.3 Clustering

In this section, the results of three clustering algorithms are presented. The literature is introduced in Section 2.4, the clustering methodology in Section 3.5 and the framework implementation in Section 4.3.

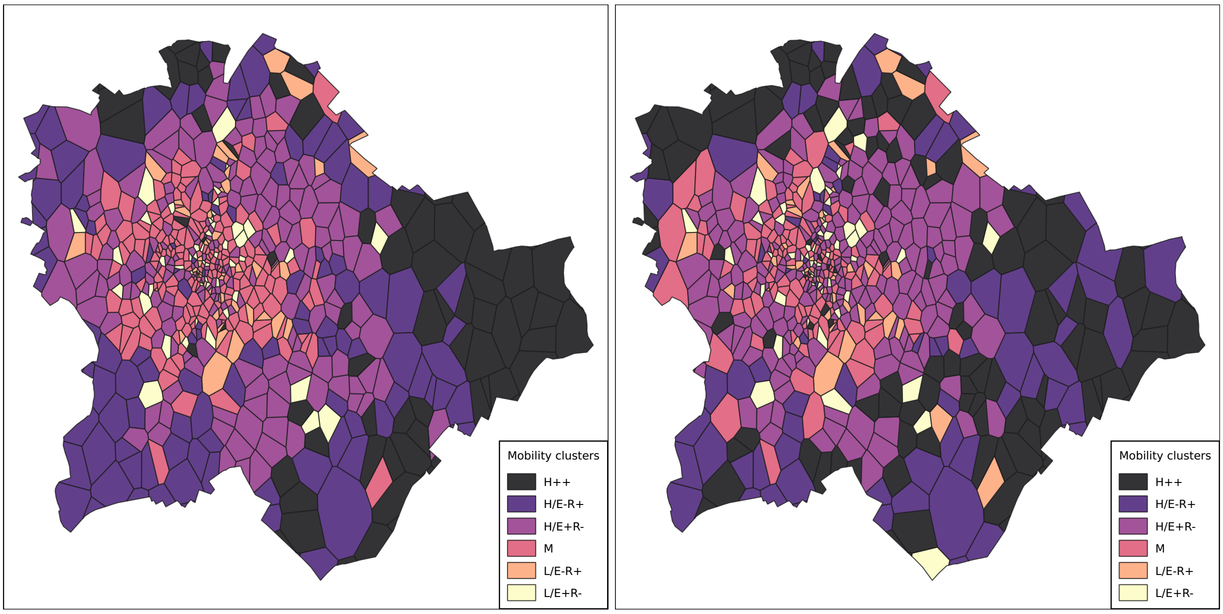
Six different input data set was prepared and processed using the developed clustering framework. Three sets of indicators were selected: mobility, SES, mobility and SES together; for two days: the 19^{th} and 20^{th} of August 2014. The two days are different because the former is a regular workday, and the latter is a national holiday, as detailed in Section 4.2.

The results are published in the form of maps of Budapest, where the voronoi cells are coloured by cluster cell centroid positions. Darker colours indicate higher values in the unit of measure of centroid dimensions.

6.3.1 Mobility Clusters

Mobile phone users' mobility customs were represented by their radius of gyration and entropy calculated over the whole period of the data collection, detailed in Section 3.5. Using the clustering framework and the list of cells with their average mobility metrics, six groups of cells were determined:

- H++ - High radius of gyration and high entropy.
- H/E–R+ - High metrics overall, lower entropy and higher radius of gyration.
- H/E+R– - High metrics overall, lower radius of gyration and higher entropy.
- M - Medium metrics.
- L/E–R+ - Low metrics overall, lower entropy and higher radius of gyration.
- L/E+R– - Low metrics overall, lower radius of gyration and higher entropy.



(a) (b)

Figure 22: Mobility clusters in Budapest on the 19^{th} (a) and 20^{th} (b) of August.

The visualised results are in Figure 22. It is visible that the outskirts of Budapest have higher mobility metrics on average. This is possibly the result of a significant amount of individuals commuting to the city centre to work regularly. The highest (H++) clusters are almost all on the eastern side of Pest – furthest away from the centre. The distribution on the 19^{th} is relatively homogeneous; however, on the 20^{th}, a large number of cells show higher average mobility metrics towards the city centre. This might result from the national holiday and people travelling from the outskirts and other settlements to the city centre for the celebratory events.

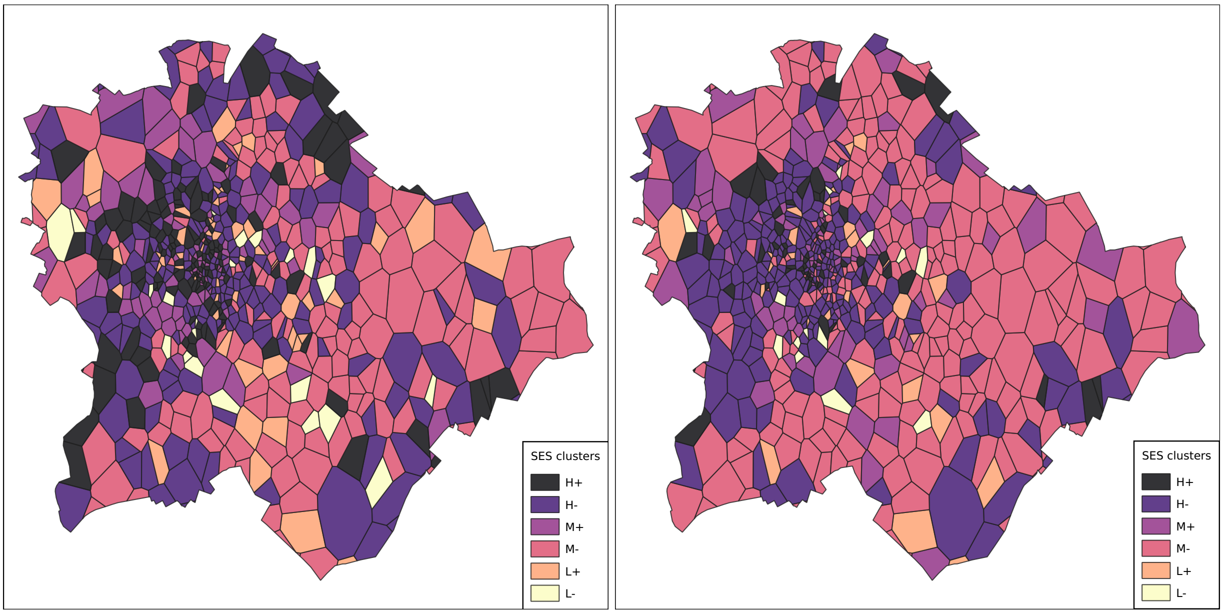
6.3.2 Socioeconomic Status Clusters

Mobile phone users' socioeconomic status was represented by their mobile phone's relative age and price detailed in Section 3.4. Using the clustering framework and the list of cells with their average SES metrics, six groups of cells were determined:

$$Cl_{ses}(k = 6) = \{H+, H-, M+, M-, L+, L-\} \tag{5}$$

where:

H+ is the highest, and L– is the lowest SES cluster.



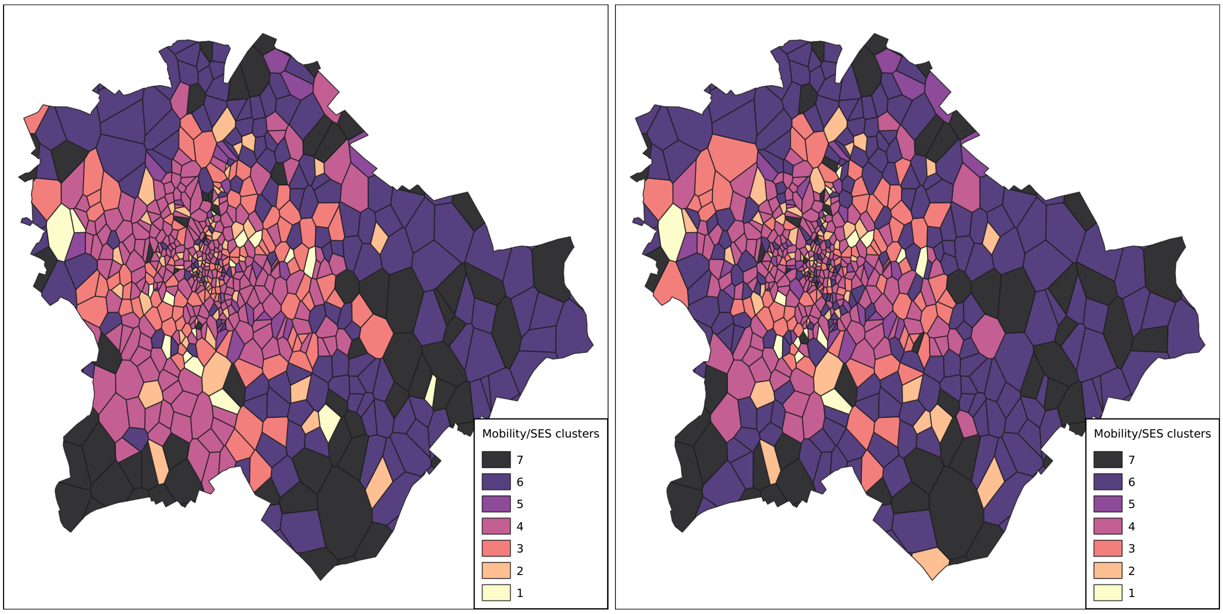
(a) (b)

Figure 23: Socioeconomic clusters in Budapest on the 19^{th} (a) and 20^{th} (b) of August.

The visualised results are in Figure 23. It is visible that the higher SES clusters are mainly in the city centre, central and south-west Buda, and north and south-east Pest. On the contrary, the lowest average SES cell clusters are east and south Pest and the outskirts. The group distribution is visibly lower on average but more homogeneous on the 20^{th}. This is possibly due to numerous regional and rural tourists visiting the State Foundation Day celebrations in Budapest or simply higher activity from individuals from groups of lower SES.

6.3.3 Mobility and Socioeconomic Clusters

Cells were clustered based on their average mobility and socioeconomic status indicators. These mixed groups were resolved using all four metrics used in Sections 6.3.1 and 6.3.2. Seven groups of cells were determined using the clustering framework and the list of cells with their average mixed metrics: $1-7$. The cell centroids exist in a four-dimensional space; hence, cluster labels were not given as in previous cases.



(a) (b)

Figure 24: Mobility – Socioeconomic clusters in Budapest on the 19^{th} (a) and 20^{th} (b) of August.

The visualised results are in Figure 24. It is visible that homogeneous groups of cells form on large areas. The outskirts and the centre are separate from each other. The spatial cluster distribution is more homogeneous on the 19^{th} than on the 20^{th} of August; however, there are only more minor differences between the two days, unlike in Sections 6.3.1 and 6.3.2. Changes in larger areas between the two observation periods are mainly visible in central and south Buda.

7 Conclusions and Future Works

In this study, the mobility customs and socioeconomic statuses of mobile phone users in Budapest were analysed using call detail records. The most common mobility indicators and mobile phone properties represented the users in a selected area. A mobile network data processing framework was developed to produce these metrics, and the methodology for a large scale event analysis was proposed. In addition, a clustering framework was introduced to connect socioeconomic status and mobility indicators, which is capable of grouping spatial cells based on one or more specified statistical values.

The developed frameworks were tested using software unit tests and case studies. The results were presented using maps of Budapest displaying voronoi cells around the base station locations. The large scale event analysis showed minor differences in socioeconomic status averages between Buda and Pest during the fireworks of 2014 State Foundation Day in Budapest. The implemented framework is capable of analysing different events.

The clustering methodology was proposed with a focus on the ability of reusability. Using different configurations, the framework can specify clusters based on metrics. In the analysis, three configurations were executed – mobility, socioeconomic status, and the two indicators together. Finally, the results were presented and discussed. Therefore, the conclusion shows that together with supplementary information, mobile network data can be used to build a robust spatiotemporal descriptor of an area.

The limitations of this study are the follows. First, the produced metrics on an area are only valid if the collected data is representative. Not enough call detail records result in outlier and false results. Moreover, the SES metrics used in this study are weak indicators of an individual's socioeconomic status. Having access to better indicators would result in a more accurate model.

Further advanced cell indicator calculations would also help create a more precise spatiotemporal model. In this study, net averages were used, but daily aggregations could help show more significant differences between cell groups. In addition, using inhabitant-based cell aggregation methods would eliminate the distorting effect of a subscriber being considered in an area wherever they live.

Appendices

A Radius of Gyration

```
def calc_rog(lat, lon):
    """
    This function returns the radius of gyration calculated from the
                                  latitude
    and longitude points in the arguments.
    >>> calc_rog([], [])
    nan
    >>> calc_rog([47.5], [19.05])
    0.0
    >>> calc_rog([47.5, 47.5], [19.05, 19.05])
    0.0
    >>> calc_rog([47.5, 47.5, 47.5], [19.05, 19.05])
    nan
    """

    if len(lat) < 1 or len(lon) < 1 or len(lat) != len(lon):
        return np.NaN

    center = [stat.mean(lat), stat.mean(lon)]

    dist = []
    for i in range(len(lat)):
        dist.append(haversine([lat[i], lon[i]], center))

    rg = math.sqrt(sum(dist)/len(dist))

    return rg
```

B Entropy

```
def calc_ent(loc):
  """
  This function returns the entropy calculated from the latitude
  and longitude points in the arguments.
  >>> calc_ent([])
  nan
  >>> calc_ent([1])
  0.0
  >>> calc_ent([1,2,3,4])
  1.0
  >>> calc_ent([1,1,2,2])
  0.5
  """

  if len(loc) < 1:
    return np.NaN

  if len(loc) == 1:
    return 0.0

  locList = []
  for l in loc:
    if not l in locList:
      locList.append(l)

  locNum = []
  for l in locList:
    n = loc.count(l)
    locNum.append(n)

  locSum = len(loc)

  locProb = []
  for i in range(len(locList)):
    locProb.append([locList[i], locNum[i]/locSum])

  s = 0
  for lp in locProb:
    s += lp[1]*math.log(lp[1])

  ent = -1 * (s/math.log(len(loc)))
  ent = round(ent, 5)

  return ent
```

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] Luca Pappalardo et al. „Returners and explorers dichotomy in human mobility". In: *Nature communications* 6.1 (2015), pp. 1–8.

[4] Natalia Andrienko and Gennady Andrienko. „Designing visual analytics methods for massive collections of movement data". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 42.2 (2007), pp. 117–138.

[5] Francesco Calabrese et al. „Real-time urban monitoring using cell phones: A case study in Rome". In: *IEEE transactions on intelligent transportation systems* 12.1 (2010), pp. 141–151.

[6] Günther Sagl, Martin Loidl, and Euro Beinat. „A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic". In: *ISPRS International Journal of Geo-Information* 1.3 (2012), pp. 256–271.

[7] Hansi Senaratne et al. „Urban mobility analysis with mobile network data: A visual analytics approach". In: *IEEE Transactions on Intelligent Transportation Systems* 19.5 (2017), pp. 1537–1546.

[8] Martin Wirz et al. „Probing crowd density through smartphones in city-scale mass gatherings". In: *EPJ Data Science* 2.1 (2013), pp. 1–24.

[9] Marco Mamei and Massimo Colonna. „Estimating attendance from cellular network data". In: *International Journal of Geographical Information Science* 30.7 (2016), pp. 1281–1301.

[10] 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.

[11] 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.

[12] 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.

[13] Chaoming Song et al. „Limits of predictability in human mobility". In: *Science* 327.5968 (2010), pp. 1018–1021.

[14] Ioanna Psylla et al. „The role of gender in social network organization". In: *PloS one* 12.12 (2017), e0189873.

[15] Laetitia Gauvin et al. „Gender gaps in urban mobility". In: *Humanities and Social Sciences Communications* 7.1 (2020), pp. 1–13.

[16] Talayeh Aledavood et al. „Daily rhythms in mobile telephone communication". In: *PloS one* 10.9 (2015), e0138098.

[17] Yang Xu et al. „Human mobility and socioeconomic status: Analysis of Singapore and Boston". In: *Computers, Environment and Urban Systems* 72 (2018), pp. 51–67.

[18] Luca Pappalardo et al. „An analytical framework to nowcast well-being using mobile phone data". In: *International Journal of Data Science and Analytics* 2.1 (2016), pp. 75–92. URL: `https://doi.org/10.1007/s41060-016-0013-2`.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[24] Vladimir Estivill-Castro. „Why so many clustering algorithms: a position paper". In: *ACM SIGKDD explorations newsletter* 4.1 (2002), pp. 65–75.

[25] Ibrahim Mousa Al-Zuabi, Assef Jafar, and Kadan Aljoumaa. „Predicting customer's gender and age depending on mobile phone data". In: *Journal of Big Data* 6.1 (2019), pp. 1–16.

[26] Maxime Lenormand et al. „Entropy as a measure of attractiveness and socioeconomic complexity in Rio de Janeiro metropolitan area". In: *Entropy* 22.3 (2020), p. 368.

[27] Mohammadhossein Ghahramani et al. „Spatio-Temporal Analysis of Mobile Phone Network based on Self-Organizing Feature Map". In: *IEEE Internet of Things Journal* (2021).

[28] Tao Zhao et al. „Predicting individual socioeconomic status from mobile phone data: a semi-supervised hypergraph-based factor graph approach". In: *International Journal of Data Science and Analytics* 9.3 (2020), pp. 361–372.

[29] Research summary for the National Media and Infocommunications Authority. *Electronic Communication Services Usage by Households and Individuals, 2014.* 2015. URL: `https://english.nmhh.hu/article/185603/Electronic_Communication_Services_Usage_by_Households_and_Individuals_2014`.

[30] Hungarian Central Statistical Office. *Gazetteer of Hungary 1st January, 2014.* 2014. URL: `https://www.ksh.hu/docs/hun/hnk/hnk_2014.pdf`.

[31] Mohit Sainani. *GSMArena Mobile Phone Devices.* Accessed on 02.03.2022. 2020. URL: `https://www.kaggle.com/msainani/gsmarena-mobile-devices`.

[32] The pandas development team. *pandas-dev/pandas: Pandas 1.4.2.* Version v1.4.2. Apr. 2022. DOI: `10.5281/zenodo.6408044`. URL: `https://doi.org/10.5281/zenodo.6408044`.

[33] F. Pedregosa et al. „Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[34] QGIS Development Team. *QGIS Geographic Information System.* QGIS Association. 2022. URL: `https://www.qgis.org`.

[35] Michael Bayer. „SQLAlchemy". In: *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks.* Ed. by Amy Brown and Greg Wilson. aosabook.org, 2012. URL: `http://aosabook.org/en/sqlalchemy.html`.

[36] Lars Buitinck et al. „API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning.* 2013, pp. 108–122.

[37] Ville Satopaa et al. „Finding a "kneedle" in a haystack: Detecting knee points in system behavior". In: *2011 31st international conference on distributed computing systems workshops.* IEEE. 2011, pp. 166–171.

[38] David Arthur and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding.* Tech. rep. Stanford, 2006.

[39] *Test interactive Python examples.* Accessed on 15.05.2022. 2022. URL: `https://docs.python.org/3/library/doctest`.