ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

PHD DISSERTATION

GERGŐ PINTÉR

# Analyzing the Mobility Customs of the Urban Population using Mobile Network Data

Supervisor
Dr. habil. Imre Gábor Felde

DOCTORAL SCHOOL OF
APPLIED INFORMATICS AND
APPLIED MATHEMATICS

BUDAPEST, 2022

**Members of the Defense Committee:**

Chair of the Committee:
Prof. Dr. habil. József Tar, DSc (Óbuda University)

Secretary:
Dr. habil. Sándor Szénási, PhD (Óbuda University)

Opponents:
Prof. Dr. Róbert Fullér, DSc (Széchenyi István University)
Prof. Dr. János Kertész, DSc, Member of the Hungarian Academy of Science (Central European University)

Members:
Dr. Tamás Tettamanti, PhD (Budapest University of Technology and Economics)
Dr. Péter Kádár, PhD (Óbuda University)
Dr. György Schuster, PhD (Óbuda University)

Replacement members:
Secretary: Prof. Dr. habil. Márta Takács, PhD (Óbuda University)
Members: Dr. Andrea Pődör, PhD (Óbuda University), Prof. Dr. András Molnár, PhD (Óbuda University)

**Members of the Comprehensive Examination Committee:**

Chair of the Committee: Prof. Dr. Róbert Fullér, DSc
Main Subject: Dr. Andrea Pődör, PhD
Secondary subjects: Prof. Dr. habil. Márta Takács, PhD; Dr. habil. Sándor Szénási, PhD

**Date of the Defense:**

July 7, 2022

# Declaration

I hereby state and declare that this Ph.D. thesis represents my own work and is the result of my own original research. I only used the sources listed in the references. All parts taken from other works, either as word for word citation or rewritten keeping the original meaning, have been unambiguously marked, and reference to the source was included.

Budapest, July 7, 2022

Gergő Pintér

Gergő Pintér

**Copyright**

**Colophon**

This document was typeset with the help of KOMA-Script and LaTeX using the kaobook class.

# Acknowledgement

I would like to express my gratitude Gábor Bognár, who practically established the mobile network data processing at Óbuda University. I have learned a lot from him in regard to how to process data of this order of magnitude. The design of my Call Detail Record processing framework has been influenced by his work. I also would like to thank my supervisor for his support.

Furthermore, I am grateful to the anonymous peer reviewers, who reviewed my papers and improved my research articles through their remarks and, indirectly, this dissertation as well.

Finally, I am deeply grateful to my family and friends for their love, support, and encouragement to finish this work.

A city is more than a place in space, it is a drama in time.

– Patrick Geddes

# Abstract

Mobile phones are now fundamental parts of our life; they are practically always with us, wherever we go, almost as if they were a part of our body. The continuous communication between a device and the mobile networks leaves traces of our whereabouts in the operator's system. In the last few decades, this information source has become a standard way of analyzing the characteristics of human mobility. During my research, I have analyzed Call Detail Records (CDR), covering Budapest, Hungary, and the surrounding settlements. The goal of my research was to develop a methodology and implement a data processing framework that can evaluate mobile network data. This framework should also be able to calculate mobility indicators and associate socioeconomic indicators with the subscribers. As the mobility patterns of the subscribers can be extracted from the CDR, the people can be distinguished based on their mobility customs. One of the main goals was to find a correlation between mobility and the socioeconomic status (SES) and infer SES from mobility.

Another goal was to analyze the commuting tendencies of the capital and its agglomeration. This objective required an algorithm to estimate the home and work locations of the subscribers. These two locations defined the direction of commuting. The further someone lives from their workplace, the longer the travel may last. Although several conditions affect the travel time, the time when people wake up can be a crucial indicator. The length of the workday is also connected to commuting. Mobile network data can be suitable for estimating these features.

I have utilized two features to characterize socioeconomic status, a direct and an indirect one. The indirect is the housing prices because the price-level of a neighborhood is used to infer the SES. When the home location of a subscriber is known, the typical housing price of that area can be associated with the subscriber. Considering if someone lives in a more expensive area, their socioeconomic status should be higher. A more direct feature would be the price of the subscriber's cellphone, which can be associated with the subscriber, even though the actual purchase price can be affected by multiple factors. I introduced a new method to analyze the correlation between these socioeconomic features and the widely used mobility indicators.

Based on mobile network data, I found that the wealth level in Budapest has a certain influence on travel customs. The real estate price of home locations does not have a remarkable effect on the number of visited places (Entropy) and Gyration, which suggests that the travel diversity of the inhabitants is not strongly dependent on their socioeconomic status.

On the other hand, the mean distance between home and work locations is significantly correlated to property values. Living in a high-priced neighborhood in Budapest requires working in more expensive regions and vice versa. The people living in cheaper regions have to take longer routes to their work sites. This relationship can be explained by the fact that the less expensive districts are located at the perimeter of the city, while most of the workplaces are in the wealthier regions. Analyzing the residents' real estate values at home and work locations, I found a strong positive relationship.

I also demonstrated that the mobile phone price is an expressive socioeconomic indicator. It is capable of clustering the areas of a city and distinguishing the subscribers by mobility customs.

A new indicator, called wake-up time, is introduced to describe a behavior of a group of subscribers, and it denoted the time when this group starts to use the mobile network in the mornings. A negative correlation was found between wake-up time and mobility indicators (Entropy, Radius of Gyration): People wake up earlier and travel more on workdays. It is quite the contrary on holidays. A correlation has been identified between the wake-up time and the socioeconomic status using socioeconomic classes derived from housing prices and cellphone prices. The subscribers living in less expensive apartments get up earlier, and this tendency holds true in respect of the mobile phone prices, as well: subscribers who own more expensive cell phones tend to get up later.

An application of my research would be to support sociological studies using the anonymized mobile network data as a frequent, countrywide, and cost-effective alternative to the questionnaire-based methods like the census. These results may help further analyze the city structures, both functionally and sociodemographicly. The socioeconomic findings of this research can also contribute to a better understanding of the social

structure of urban and rural environments if the analyses are performed at the country level. As city parts with early morning or late night activities may require different public transport services, the transportation infrastructure planning can also profit from this methodology. It could help urban development and — based on international examples — mobile network data can aid epidemic control. Business development could also benefit from the detailed insight of the neighborhood chronotypes, especially with the associated information on the home locations and the socioeconomic status of the subscribers.

# Kivonat

Napjainkra a mobiltelefon a mindennapi életünk alapvető részévé vált. Szinte mindenhová magunkkal visszük, mintha csak a testünk része lenne. A folyamatos kommunikáció a telefon és a mobilhálózat elemei között nyomokat hagy a tartózkodási helyünkről a szolgáltató rendszereiben. Az elmúlt évtizedekben ez az információforrás az emberi mozgáselemzés általános eszközévé vált. A kutatásom során a Budapesthez és a környező településekhez elérhető ún. hívásrészletező rekordokat (CDR) dolgoztam fel, amelyek az adott mobiltelefon-cellában létesített távközlési tranzakció (hívás, szöveges üzenet, vagy adatforgalom) különböző jellemzőit rögzíti, mint például az idő és a cella (ezen keresztül pedig a hely).

A kutatásom célja egy olyan adatfeldolgozó rendszer kidolgozása és megvalósítása, amely lehetővé teszi mobilhálózati adatok kiértékelését, mobilitási indikátorok kiszámítását, és társadalmi-gazdasági mutatók társítását az anonim előfizetőkhöz. Mivel a mobilitási jellemzők meghatározhatóak a mobilhálózati adatokból, az emberek csoportosíthatók szokásaik alapján. A cél összefüggést találni a mobilitás és a gazdasági jellemzők között, amelynek segítségével a vagyoni helyzet megbecsülhető a mobilhálózati adatokból származtatott mobilitási jellemzőkből.

A kutatás további céljai közé tartozott a főváros és az agglomeráció ingázásának vizsgálata. Ehhez meg kellett határozni az előfizetők lakó- és munkahelyét, amely meghatározza az ingázás irányát. Minél távolabb lakik valaki a munkahelyétől, valószínűleg annál több időt tölt utazással. A pontos utazási idő ugyan számos paraméter függvénye, azonban megfelelő indikátora lehet annak az ideje, amikor az emberek felkelnek. Emellett a munkanap hossza is összefüggésben áll az ingázással. A mobilhálózati adat pedig alkalmas lehet ezen értékek közelítésére.

A társadalmi-gazdasági mutató jellemzésére egy közvetett és egy közvetlenebb értéket használtam. Közvetett jellemző a lakások négyzetméterára, ugyanis a lakhely környékének átlagos lakásáraiból következtetek az előfizető vagyoni helyzetére, feltételezve, hogy aki egy drágább környéken lakik, az tehetősebb is. Mivel az előfizető lakóhelyét nagy biztonsággal sikerült meghatározni, a környék tipikus négyzetméterára az előfizetőhöz rendelhető. A közvetlenebb jellemző a telefon ára, ami szintén hozzárendelhető az előfizetőhöz a rendelkezésre álló adatok alapján, még ha a valós vételár több tényezőtől is függhet. Ezek kiértékeléséhez kidolgoztam egy új módszert, amely lehetővé teszi a széles körben használt mobilitási indikátorok és társadalmi-gazdasági jellemzők közötti összefüggések feltárását és vizsgálatát.

Mobilhálózati adatok alapján kimutattam, hogy a vagyoni helyzet befolyásolja a városon belüli utazási szokásokat. A lakóhelyek ingatlanárai nem hatnak kiemelten a meglátogatott helyek számára, ami azt erősíti, hogy az utazások változatossága nem függ számottevően vásárlóerőtől.

Ugyanakkor viszont a lakó- és a munkahely távolsága jelentős korrelációt mutat az ingatlanárakkal. Azoknak az embereknek, akik olcsóbb környéken élnek, jellemzően nagyobb távolságot kell megtenniük a munkahelyükig. Ez talán azzal magyarázható, hogy a kevésbé drága ingatlanok a város peremén helyezkednek el, míg a drágább régiókban található a legtöbb munkahely. A lakó- és munkahelyek környezetében jellemző ingatlanárak elemzése során szoros összefüggést találtam. Egy magasabb ingatlanárú környéken élő lakos magasabb ingatlanárú helyen is fog dolgozni, és mindez fordítva is igaz.

Szintén kimutattam, hogy a mobiltelefon ára erős indikátora lehet a vagyoni helyzetnek, különösen a telefon korával együtt. Nemcsak a város területeinek osztályozására alkalmas, hanem segítségével lehetővé válhat az előfizetők mobilitási szokásainak megkülönböztetése is.

Bevezettem egy új indikátort, az „ébredési időt”, amellyel az előfizetők egy csoportját jellemeztem, és azt jelöli, hogy ez a csoport reggelente mikor kezdi el használni a mobilhálózatot. Fordított korrelációt találtam az ébredési idő és a mobilitási mutatók (pl. entrópia, bolyongási kör sugara) között. Munkanapokon az emberek korábban kelnek föl és többet utaznak, munkaszüneti napokon épp ellenkezőleg.

Az ingatlan- és telefonárakból származtatott társadalmi-gazdasági osztályok alapján összefüggést találtam az ébredési idő és a vagyoni helyzet között. Az olcsóbb ingatlanárú környéken élő előfizetők korábban kelnek fel, és ez igaz az olcsóbb telefonnal rendelkezőkre is. Azok azonban, akik drágább helyen élnek vagy drágább telefonnal rendelkeznek, később is ráérnek felkelni.

A kutatásom egy felhasználási területe lehet a szociológiai kutatások támogatása anonim mobilhálózati adatokkal. Ez gyakori, az egész országra kiterjedő és költséghatékony alternatívát jelenthet a népszámláláshoz hasonló kérdőíves módszerekkel szemben. Ezek az eredmények segíthetik a városszerkezet mélyebb megértését mind funkcionális, mind szociodemográfiai szempontból. A kutatás társadalmi-gazdasági eredményei szintén hozzájárulhatnak a városi környezet társadalmi összetételének alaposabb megértéséhez. Továbbá országos lefedettségű adatokra kiterjesztve a vidéki környezet vizsgálatára is használható lehet. Mivel az egyes városrészek kora reggeli vagy késő esti élete eltérő tömegközlekedési szolgáltatást igényelnek, ezek az eredmények akár a tömegközlekedési rendszer fejlesztéséhez is hozzájárulhatnak. Segíthetik a várostervezést és – nemzetközi példák alapján – a mobilhálózati adat a járványok elleni védekezés egyik eszköze is lehet. Emellett a vállalkozásfejlesztés is profitálhat az egyes városrészekben megforduló emberek viselkedésének részletes ismeretéből, különösképpen a lakóhelyre és a becsült vásárlóerőre vonatkozó információkkal kiegészítve.

# Contents

# Abbreviations

**API** Application Programming Interface. 26

**CDR** Call Detail Record. vii, ix, 1, 2, 4, 6, 8–11, 13, 16, 18–21, 25–28, 30, 37, 40–42, 45–48, 50, 52, 65, 72
**CEU** Central European University. 28, 37, 38
**CPR** Control Plane Record. 4, 8
**CSV** Comma Separated Value. 25, 26

**GIS** Geographic Information System. 28
**GPS** Global Positioning System. 4, 27, 28

**HDA** Home Detection Algorithm. 8
**HUF** Hungarian forint. 21, 68, 76

**ICT** information and communication technology. 2, 40
**IMEI** International Mobile Equipment Identity. 13, 18

**KSH** Központi Statisztikai Hivatal, Hungarian Central Statistical Office. 15, 22, 23, 32, 40–42, 52, 75, 83

**MTI** Magyar Távirati Iroda, Hungarian news agency. 34

**OSM** OpenStreetMap. 22, 28, 88

**PCA** Principal Component Analysis. 69, 72–74, 82, 85, 86
**POI** Point of interest. 9, 12, 88

**SES** Social Economic Status. vii, 1, 2, 7, 9, 10, 29, 69, 71, 77–79, 81, 82
**SIM** Subscriber Identity Module. 7, 13–19, 25–27, 30, 31, 34, 35, 42, 50, 56, 65, 67, 69, 72, 74–76, 78, 82, 84
**SWC** sleep wake cycle. 1, 2, 11, 58

**TAC** Type Allocation Code. 7, 13, 14, 18–20, 26, 42, 78

**UMTS** Universal Mobile Telecommunications System. 13

**WGS 84** World Geodetic System, also known as EPSG:4326. 26
**WKT** Well-known text. 28

**XDR** eXtended Detail Record. 4, 8

# Introduction 1.

## 1.1. Background

Mobile phones are now fundamental parts of our life; they are practically always with us, wherever we go, almost as if they were a part of our body. The continuous communication between a device and the mobile networks leaves traces of our whereabouts in the operator's system. Via these devices, the mobile network can "sense" our movements, which is the basis of the "Smart City" concept [1].

In the last few decades, anonymized Call Detail Records (CDR) have become a common information source for analyzing the characteristics of human mobility. Numerous research has been published utilizing this source of data from all around the world, even from Hungary, e.g., [2–4]. The research focuses on Budapest and its agglomeration. CDRs contain the billed activities of the subscribers, providing information about the whereabouts of the population. Based on this massive information source, human mobility analysis is utilized in fields — among others — like social sensing, epidemiology, transportation engineering, urban planning, and sociology. Furthermore, the human sleep-wake cycle (SWC) is also studied by analyzing mobile phone network data.

The analysis of the human movement patterns based on the CDR data, which makes it possible to examine a large population cost-effectively, resulted in several discoveries about human dynamics. These works usually consider the population as a homogeneous group, and the classification is based on some mobility indicators [5]. The next step was adding external data sources to the mobile network data, extending the investigation to other population characteristics. Often, this external source is used to classify the data (e.g., by gender) and to analyze the classes using the previously introduced mobility indicators. Among others: social network data [6], transportation data [7], taxi trips [8], socioeconomic indicators (income, education rate, unemployment rate and deprivation index) [9], sale price of residential properties [10].

A part of this work also fits into the trends, using housing prices as a socioeconomic indicator. Housing price, however, is an indirect indicator for the SES, so a more direct indicator, the cellphone price, was also investigated. While Blumenstock et al. used the call history as a factor of socioeconomic status [11], Sultan et al. [12] applied mobile phone prices as a socioeconomic indicator and identified areas where more expensive phones appear more often. However, only manually collected phone prices were used, and the analysis was not performed on the subscriber level.

## 1.2. Research Goals

The goal of my research was to develop a methodology and implement a data processing framework that can evaluate mobile network data. This framework should also be able to calculate mobility indicators and associate socioeconomic indicators with the subscribers. As the mobility patterns of the subscribers can be extracted from the CDR, the people can be distinguished based on their mobility customs. One of the main goals was to find a correlation between mobility and socioeconomic status.

As the mobile network data does not contain information about the subscribers' income, the CDRs were required to be enriched with other data sources. My research focused on an indirect and a more direct feature in this regard. The indirect is the housing prices because the level of a neighborhood is used to infer the SES. When the home location of a subscriber is known, the typical housing price of that area can be associated with the subscriber. Considering if someone lives in a more expensive area, their socioeconomic status should be higher, as well.

A more direct feature would be the price of the subscriber's cellphone. As the actual purchase price of a mobile phone can depend on several factors, the age of the device may be used in line with the recommended retail price.

Another goal was to analyze the commuting tendencies of the capital and its agglomeration. Kiss et al. state that although commuting is an essential and common phenomenon, its measurement is occasional and inadequate [13]. Commuting is predominantly analyzed by the census, but that is performed only once in a decade; thus cannot follow sudden but permanent changes. They also stress that commuting should be examined frequently, and its methodology should be established [13].

As questioning the population is a slow, tedious and expensive task, it would be obvious to automate the process with the available info-communication technologies (ICT). In this research, the application of CDR processing was presented to examine commuting, mainly to Budapest, and the findings were validated by the results of studies that analyzed commuting using census.

Economic models distinguish city parts such as residential areas, industrial areas, business districts, and so on, but that is a relatively static, slowly evolving city layer. Mobile phone network data has the potential to describe the city structure via the inhabitants' mobility patterns. The next part of this research focuses on the effect of the SWC on the city structure. In this regard, it continued the commuting analysis, but the city structure was analyzed by the circadian rhythm of the people who live and work in a given area of Budapest.

Is it possible to cluster city areas by the time when the activity of the inhabitants, the workers, or the passers-by starts their activity in the morning or halts in the evening? Do city parts have "chronotypes"? Is there a structural or socioeconomic connection between the areas with the same "chronotype"? Can neighborhoods or districts be described by the terms "morningness" or "eveningness"? Another goal of this research was to answer these questions.

## 1.3. Meso-data

Is this work dealing with Big Data? No, not really, as one of the V's of Big Data [14], volume, does not apply here. Historical data has been processed for a limited area (the capital and its agglomeration) and a limited period (one month). Increasing the observation area and period would increase the volume of the data. However, a year of data for the whole country should still be effortlessly manageable with the current set of tools that are not extraordinary.

Could it be Big Data? Yes, it could, if, for example, the analytics should be updated in real-time or almost real-time, so the velocity would be an issue. Some current analyses could even take an hour or two, though there could be plenty of room for optimization in my queries and scripts.

What is it, if not Big Data? I like the term "meso-data" (and "meso-computing") that Matt Williams proposed in his essay [15] to describe the processing challenges of the order of magnitude of this data.

## 1.4. Data Science is like Origami

Origami[1] is the art of paper folding, where after a specific sequence of simple folding steps, a sheet of paper becomes a figurine of an animal or an object. Data science is like origami, as during the data processing, simple steps are applied one after another while it starts to shape and new information is born. As everyone has a sheet of paper, everyone has a pile of data nowadays.

The question is, who can fold that data into art?

[1] In the Japanese word, origami, *ori* means "folding", and *kami* means "paper" ("ka" changes to "ga" in the compound word).

Paper crane by Caro Asercion, CC BY 3.0

# Literature Review 2.

Understanding human movements and recognizing behavior patterns that occur during daily life in urban areas requires a systematic analysis of that human mobility. The evaluation of human travel is based on observations on the individual and group levels. In the last decades, several novel datasets, based on vehicular GPS and cellular network records or social media information, became available, which provided more accurate and sophisticated characterization of people's movements. This chapter provides a brief introduction to this field. Figure 2.1, shows the distribution of the referenced papers from the last two decades.

## 2.1. Mobile Network Data

The mobile phone network, during its operation, constantly communicates with the cell phones. This communication can be divided into two categories: (i) the passive, cell-switching communication that keeps the cell phones ready to use the mobile phone network at any time, and (ii) the active, billed usage of the mobile phone network, including phone calls, text messages or mobile internet usage. The Call Detail Records (CDR) collect the latter, containing information about the subscriber, the time of the activity, and the place (via the cell), where the activity occurred.

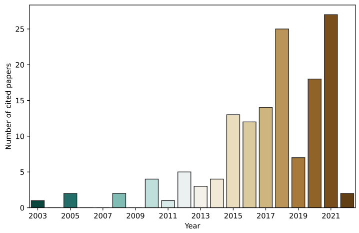


**Figure 2.1.:** Reference distribution of this work, from the last two decades.

Strictly speaking, Call Detail Record (CDR) contains only call information, and the expression eXtended Detail Record (XDR) is used to denote other types of billed communication (i.e., mobile internet traffic). As CDRs are generated only when a user makes or receives a call, its temporal resolution is low. In that sense, XDRs has a higher resolution since it is a mixture of human- and device-triggered communication [16]. An application can request data transfer without human interaction, downloading e-mails, for example. Note that, within this work, the expression CDR is also used to denote XDR, as the obtained data does not contain any information regarding the activity type (see Chapter 3 for details).

The so-called passive communication (also referred to as Control Plane Record (CPR)) is network-triggered at cell-switching, or, for example, when the network status of the cellphone is monitored [16]. Both the active and the passive communication have the same geographical resolution, as associated with the same base stations or cells that are supposed to be more or less constant. However, the temporal granularity of the passive communication can be significantly finer.

## 2.2. Mobility Indicators

There are some indicators that are widely used in the literature to characterize human mobility, like Radius of Gyration, Entropy, or the distance between Home and the Work locations (Section 4.6). These indicators are determined for every subscriber.

### 2.2.1. Radius of Gyration

The Radius of Gyration [17] defines a circle where an individual can usually be found. It was originally defined in (2.1), where $L$ is the set of locations visited by the individual, $r_{cm}$ is the center of mass of these locations, and $N$ is the total number of visits of time spent at these locations. The $r_i$ is the coordinate of location $i$, and $n_i$ is the number of visits or the time spent at location $i$ [5].

$$r_g = \sqrt{\frac{1}{N} \sum_{i \in L} n_i (r_i - r_{cm})^2} \tag{2.1}$$

### 2.2.2. K-Radius of Gyration

The K-radius of Gyration is calculated using only the $k$ most frequent locations of the individual [5], defined by (2.2). Pappalardo et al, used this approach to classify individuals by their mobility customs. Two classes named "Returners" and "Explorers" have been defined by the value of gyration. Returners are who spend most of their time between the $k$ most frequent locations ($r_g^{(2)} > r_g/2$), in contrast to the explorers whose activity area cannot be described with the $k$ most frequent locations ($r_g^{(2)} < r_g/2$) [10].

$$r_g^{(k)} = \sqrt{\frac{1}{N_k} \sum_{i=1}^{k} n_i (r_i - r_{cm}^{(k)})^2} \tag{2.2}$$

### 2.2.3. Number of Activity Locations

The number of activity locations (2.3) is simply the number of the visited cells during an observation period, defined by Xu et al. in [10].

$$A = |(s_1, s_2, ..., s_n)| \tag{2.3}$$

### 2.2.4. Mobility Entropy

The mobility entropy (or mobility diversity) of the visited locations characterizes the diversity of the individual's movements, defined as Equation (2.4), where $L$ is the set of locations visited by the individual, $l$ represents a single location, $p(l)$ is the probability of an individual being active at a location $l$ and $N$ is the total number of activities for an individual [18, 19].

$$MD = -\frac{\sum_{l \in L} p(l) \log p(l)}{\log N} \tag{2.4}$$

The generalization of the Mobility Entropy is the term, Travel Diversity [10]. Instead of the diversity of the locations, it determines the diversity of the travels between consecutive locations. It can be calculated to $k$-length transitions, where $k = 1$ gives back the location entropy. It is possible to consider the transitions with or without direction, contingent on whether the difference between $L_1 \longrightarrow L_2$ and $L_2 \longrightarrow L_1$ is important or not.

## 2.3. Social Sensing

CDR processing is often applied for large social event detection, such as football matches [20–24], concerts [25], sociopolitical events [23, 26] or mass protests [27]. When thousands of people are in the same place at the same time, they generate a significant "anomaly" in the data, whereas small groups usually do not stand out from the "noise". This is especially true when the passive, transparent communication between the mobile phone device and the cell is not included in the data, but only the active communication (voice calls, text messages, and data transfer) is recorded.

Wirz et al. estimated the crowd density during Lord Mayor's Show 2011 in London [28]. In [20–22] and [26], the authors examined the location of stadiums, where the football matches took place. Traag et al. [20] and Hiir et al. [26] also found that the mobile phone activity of the attendees decreased significantly. Xavier et al. compared the reported number of attendees of these events with the detected ones. Furletti et al. also analyzed sociopolitical events, football matches, and concerts in Rome [23]. Barnett et al. analyzed the attendees of Kumbh Mela, a 3-month-long Hindu festival [29]. Tourism is also studied via mobile network data; Xu et al. followed traveler groups within South Korean cities [30] and international travelers all around the country [31].

Qian et al. found that tourists tend to rest near close to their destinations or in the city center, using Weibo (Chinese microblog app) and Flickr (international photo-sharing service) [32].

## 2.4. COVID-19

The data used in this work predates the COVID-19 pandemic and solely focuses on the "normal" life of Budapest.

Epidemiology is used to be mentioned as a potential application of human mobility studies, with some applications like [33], but the COVID-19 pandemic prioritized its applications in Digital Epidemiology, as mobile phone network data can reflect the mobility changes caused by the imposed restrictions. The term "Digital Epidemiology" can be used when working with data that was not generated with the primary purpose of epidemiological studies [34], that had considerable applications even before the COVID-19 pandemic [35]. Mobile network data is also utilized to analyze human mobility during the COVID-19 pandemic and the effectiveness of the restrictions.

Willberg et al. identified a significant decrease in the population presence in the largest cities of Finland after the lockdown compared to a usual week [36]. Romanillos et al. reported similar results from the Madrid metropolitan area [37]. W. D. Lee et al. examined the effect of the SES on the mobility changes during the lockdown and found that the mobility of the wealthier subscribers decreased more significantly in England [38]. This is in good agreement with [39], where Yechezkel et al. found that poorer regions of Israel showed lower and slower compliance with the mobility restrictions. Khataee et al. compared the effect of the social distancing in several countries, using mobility data from Apple phones [40]. Bushman et al. [41], Gao et al. [42], Hu et al. [43] and Tokey [44] also analyzed effects of the stay-at-home distancing on the COVID-19 increase rate, in the US. Gao et al. found a negative correlation between stay-at-home distancing and COVID-19 increase rate [42]. Lucchini et al. studied the mobility changes during the pandemic in four states of the USA [45]. Bushman et al. analyzed the compliance to social distancing in the US using mobile phone data [41]. Yabe et al. found that one week into the state of emergency, human mobility behavior decreased by around 50%, resulting in a 70% reduction of social contacts in Tokyo [46]. This also confirms that mobile network data analysis is an efficient tool to monitor the effect of restrictions, just as Google Mobility data [47].

Still, these analyses might not be common enough. Oliver et al. asked: "Why is the use of mobile phone data not widespread, or a standard, in tackling epidemics?" [48]. This, however, is not within the scope of this study.

## 2.5. Selecting the Subscribers

Csáji et al. took into account subscribers who had at least ten activities during the observation period (15 months) [49]. Xu et al. chose to use those subscribers who had at least one activity record on at least half of the days during the observation period [10]. Pappalardo et al. discarded the subscribers who had only one location, and the individuals have at least half as many calls as hours are in the data set. Furthermore, the abnormally active (more than 300 calls per day) SIM cards are excluded [5]. I selected the SIM cards that have activity for at least 20 days (out of 30), the daily mean activity number is at least 40 on workdays and at least 20 on weekends, but not more than 1000. The upper limit is especially important to remove SIM cards that possibly operate in mobile broadband modems, for example. More details in Section 4.7. Filtering by activity is not necessarily sufficient to keep only individuals in the data set. Type Allocation Codes (TAC), on the other hand, can determine the type of the device and the exact model of a cell phone (Section 3.3).

## 2.6. Commuting

Identifying the home and work locations of a subscriber is a common and crucial part of the CDR processing, as a good portion of the people live their lives in an area that is determined by only their home and workplace [5, 50]. Since these locations fundamentally determine the people's mobility customs, the commuting trends can be analyzed between these locations. The commuting is studied within a city [50, 51], or between cities [52–54], and also examined by social network data, such as Twitter [55, 56].

Csáji et al. determined the subscribers' most common locations, and based on weekly calling patterns, identified the home and work locations [49]. Home locations showed a strong correlation with population statistics. Diao et al. applied a regression model to travel survey data to predict the activity type (e.g., home, work, or social) of the mobile phone location data by considering the temporal distributions of different activities [57]. Xu et al. determined the home locations and then applied a modified standard distance to measure the spread of each subscriber's activity space [58]. Pappalardo et al. used the Radius of Gyration (Section 2.2) to separate the subscribers based on their mobility customs and defined two classes: returners and explorers [5]. While in the case of returners, the radius of gyration is dominated by their movement between a few preferred locations, the explorers have a tendency to travel between a larger number of different locations. To demonstrate this dichotomy, they defined the k-radius of gyration, which refers to the gyration radius of the $k$ most frequent locations. The gyration radius of a two-returner is determined by the two most frequented locations, that is usually the home and work locations [5], so this method can also be used as a home detection algorithm. Pappalardo et al. [16] compared the estimated home locations of sixty-five subscribers with the known geographical coordinates of their residence location, using different types of mobile network data: CDR, eXtended Detail Record (XDR) and Control Plane Record (CPR). It has been found that XDRs should be preferred when performing home location detection.

Vanhoof et al. compared five different home detection algorithms (HDA), selecting the home cell by (i) the most activity, (ii) the most number of distinct days with phone activities, (iii) the most activities within a time interval (between 19:00 and 7:00), (iv) the most activities within a spatial perimeter, and (v) the combination of the temporal and spatial constraints [59].

Jiang et al. identified daily activity patterns (motifs, Figure 2.2), that can extend the home-work location based daily routine [50], the home locations were validated with census and household travel survey results.

Whereas Yin et al. separated the different types of activity (home, work, leisure, school) with chains of activity [60], providing different approaches for a similar purpose. Mamei et al. computed origin-destination flows with road network mapping and also validated the home location estimation with census data [53]. Zagatti et al. studied commuting in Haiti [52].

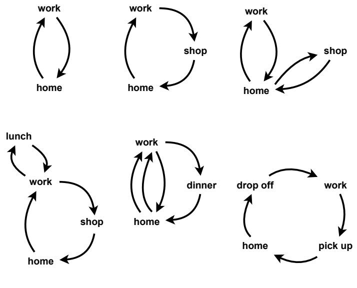


**Figure 2.2.:** Daily motifs, based on [50].

Dannemann et al. [61] partitioned the city of Santiago (Chile) into several communities and identified the socioeconomic composition of these communities based on the home-work trajectories.

### Connectivity

Traveling within a city or across cities is not necessarily bound to the home or work locations. Understanding the travel customs concerning recreational activities or tourism, for example, is also an important aspect of urban mobility.

Trasarti et al. used sequential pattern mining for identifying activity patterns in the CDR data and identified interconnections at the urban and national level [62]. Lee et al. analyzed commuting across ten Korean cities and determined the attractiveness of the cities based on which city attracted more commuters [63]. Fiadino et al. identified clusters where the visitors of Barcelona concentrate their activities and connections between the districts of Barcelona [51]. Using mobile network and POI data, Qian et al. determined that tourists tend to rest near sightseeing destinations and choose transportation hubs in the city center, furthermore connections between popular destinations were identified [32].

Ghahramani et al. investigated hotspots [64] and interaction flows between the parts of Macau and found that people tend to communicate within their close-proximity communities [65]. Fan et al. mapped the trajectories, extracted from CDRs to the road network, then the crowd flux was estimated [66]. Ni et al. extracted origin-destination information and confirmed positive connections between the population, key facilities (e.h., shopping malls or hospitals), transportation accessibility, and travel flows [67].

## 2.7. Gender Differences

Gauvin et al. revealed a gender gap in mobility, they found that women visit fewer locations than men, and spend their time less equally among those locations [68]. Goel et al. evaluated gender segregation by the social network interactions of the mobile network data, and found that Estonian speakers more likely interact with other Estonian speakers of the same gender [69]. Al-Zuabi et al. predicted the subscribers' age and gender based on mobile phone network activities [70]. Besides these examples, the other studies also attend to gender differences, like [71, 72].

## 2.8. Socioeconomic Status

The demographic metrics and Social Economic Status (SES) seem to have a significant relationship to individual travel behavior. Early studies aimed to investigate the correlation between the human travel characteristics and between SES [73, 74]. Utilizing mobility indicators, calculated from mobile network data, to infer SES is a current direction of mobility analysis, as the study of city structures (for example, Aung et al. classified land use types based on mobile network activity [75], Furno et al. fused GPS traces and mobile phone data [8]) led to the analyses of the socioeconomic structure of the population.

Cottineau and Vanhoof [19] developed a model to explore the relationship between mobile phone data and traditional socioeconomic information from the national census in French cities. Mobile phone indicators were estimated from six months of Call Detail Records, while census and administrative data are used to characterize the socioeconomic organization of French cities. The findings show that some mobile phone indicators relate significantly to different socioeconomic organization of cities. Pokhriyal et al. [76] used a computational framework to accurately predict the Global Multidimensional Poverty Index (MPI), in Senegal, based on environmental data and CDR. The methodology provides the accurate prediction of important dimensions of poverty: health, education, and standard of living. The estimations have been validated using deprivations calculated from the census.

Some investigations suggest that the mobile phone data can be used to predict individual SES [11] or regional socioeconomic characteristics [77]. Xu et al. [10] used an analytical framework on large-scale mobile phone and urban socioeconomic datasets to evaluate mobility patterns and SES. Six mobility indicators, housing prices, and income in Singapore and Boston have been used to analyze the socioeconomic classes. It was found that phone users who are generally wealthier tend to travel shorter distances in Singapore, but longer, in Boston. The research brought interesting findings but also showed that the relationship between mobility and socioeconomic status is worth investigating in other cities and countries as well.

Xu et al. [58] investigated the people's daily activities in Shenzhen, China, and identified so-called "north–south" differences in human activity, which findings are in good agreement with the socioeconomic divide in the city. Zhao et al. proposed a semi-supervised hypergraph-based factor graph model to predict individual SES, using data of Shanghai [78].

Castillo et al. calculated Human Development Index for locally available data for Ecuador to describe the socioeconomic status and used in comparison to their mobile phone based approach [79]. Barbosa et al. found significant differences in the average travel distance between the low and high-income groups in Brazil [54].

The different social classes live in different parts of a city, but CDRs also have been used to analyze gender, and minority segregation [69]. Cottineau et al. [19] explored the relationship between mobile phone data and traditional socioeconomic information from the national census in French cities. Barbosa et al. also found significant differences in the average travel distance between the low- and high income groups in Brazil [54].

Ucar et al. revealed a socioeconomic gap in mobile service consumption [80]. Vilella et al. found that education and age play news media consumption patterns in Chile, using a dataset that provides information about the visited websites [81].

While Blumenstock et al. used the call history as a factor of socioeconomic status [11], Sultan et al. [12] applied mobile phone prices as a socioeconomic indicator and identified areas where more expensive phones appear more often. However, only manually collected market prices were used. Beiró et al. examined the visitors of 16 malls in Santiago de Chile and found that people tend to choose a profile of malls more in line with their own socioeconomic status and the distance from their home [82].

Lenormand et al. utilized Entropy as a measure of attractiveness and socioeconomic complexity and found a positive and exponential relationship between income level and entropy, based on mobile network data at Rio de Janeiro Metropolitan Area [83]. Based on mobile network data, De Nadai et al. found that socioeconomic conditions, mobility, and physical characteristics of the neighborhood explain the emergence of crime [84].

Leo et al. investigated socioeconomic correlations between mobile phone communication records and anonymous bank transaction history over eight months. The latter contains daily debit/credit card purchases, their monthly loan measures, billing postal code, age, and gender [85]. They demonstrated that people of the same socioeconomic class are better connected and live closer to each other within their own class.

## 2.9. Sleep-wake cycle

The studies cited before mainly focus on the spatial distance between the home and work locations, as it is hard to estimate the travel time [86], using sporadic CDR data, though it has seasonal nature due to the human biorhythm. Moreover, human mobility is highly regular [17, 87], and the individual activity has a bursty characteristic [88]. Jo et al. found that by removing the circadian and weekly seasonality, the bursty nature of the human activity remains [89].

In the digital era, the human sleep wake cycle (SWC) is also studied using the info-communication systems, such as smartphones [90, 91], websites [81, 92, 93], social media [94] and call detail records (CDR) [71, 89, 95–99]. Cuttone et al. used screen-on events of smartphones to study the daily sleep periods [90], and Aledavood et al. examined the social network of different chronotypes, using the same data set [91]. Monsivais et al. identified yearly and seasonal patterns in calling activity and resting periods [97, 98]. Lotero et al. found a connection between temporal patterns and the socioeconomic status of the subscribers, namely, the wealthier wake up later [96]. Diao et al. found a difference in the daily activity between different districts of Boston [57].

Using Call Detail Records, Roy et al. found that chronotype is largely dependent on age, and younger subscribers are more likely to be evening people, but also found differences between women and men [72].

## 2.10. Visualization

Practically all the previously referenced publications use Voronoi tesselation for representing the cells of the mobile network. I also chose this method, but this does not mean that the Voronoi polygons generated around the base stations or the cell centroids give a perfect representation of the cell coverage. In the case of base stations that usually serve multiple antennas in different directions, the Voronoi polygons may be closer to the real coverage area, apart from the fact that some antennas may cover a much larger area than the others of the given base station. When the cell centroids are known (as in the case of the "April 2017" dataset of this work, Section 3.1), the Voronoi polygons are even less ideal approximation. On the other hand, close cells are merged during the data processing (e.g., in [51], or as I did). Especially after the merge, the Voronoi tesselation seemed the best available option.

Csáji et al. [49] and Ogulenko et al. [100] presented probabilistic positioning model as an alternative. Ricciato et al. also wrote about the issues of the Voronoi tesselation and compared different tesselation options, namely Naif Voronoi, Bi-layer Voronoi, and Proximus Voronoi [101]. Xu et al. applied Thiessen polygons to approximate the service areas of the base stations and converted the visitation frequency to hexagons [30].

When working with administrative boundaries or selecting an area to analyze, the cells can be associated by the centroid/base station coordinates or the Voronoi polygons (see Section 4.5). Lenormand et al. used an interesting approach approximating the administrative boundaries with the Voronoi polygons of the mobile network cells [83, Figure A2].

## 2.11. Social Media

There are studies that try to utilize data from social media as an alternative to, or along with [102], mobile network data. The before-mentioned directions are present in the social media based works. This section provides non-exhaustive examples. Galeazzi et al. geo-located Facebook data of 13 million users from France, Italy, and the UK and found that where Value Added per capita, and Population Density are high, resilience to mobility disruptions are higher [103]. Shepherd et al. analyzed the mobility trends in the United Kingdom (both domestic and international) during the pandemic, based on Facebook data [104]. They found differences in the mobility between central London compared to the rest of the UK, but Scotland, Wales, and Northern Ireland showed significant deviations from England.

Scholz et al. analyzed tourist flows in Austria [105]. Hawelka et al. evaluated global mobility patterns with geotagged tweets [106]. Jurdak et al. examined mobility via geotagged tweets posted in Australia [107], while Bokányi et al. analyzed commuting via Twitter in the US [56]. Using location data generated by the Foursquare app, Li et al. proposed a solution to forecast socioeconomic status by visiting Points of Interest (POI) [108].

# Data Sources 3.

In this chapter, the data sets used in this research are described. The "April 2017" (Section 3.1) and the "June 2016" (Section 3.2) CDR data sets are described, along with other sources like the estate price data (Section 3.4). Unless otherwise indicated, the CDR data denotes the "April 2017" data set through this work.

## 3.1. Vodafone April 2017

The CDR data were collected from Budapest, the capital of Hungary, and the surrounding county. Vodafone Hungary is one of the three mobile phone operators providing services in Hungary. The market share of the three big operators in Hungary has not changed significantly in the last few years. Vodafone Hungary had 25.5% in 2017 Q2 nationwide[109].

The communication between a cellular device and the mobile phone network can be divided into two categories: i) An administrative communication maintaining the connection with the service, for example, registration of the cell-switching, can be called passive communication. ii) When the device actively uses the network for voice calls, messages, or data transfer, that can be called active communication. The available data contains only active communication, which is sparser, so it cannot be used to track continuous movements.

The raw CDR data contains a long alphanumeric hash to identify the Subscriber Identity Module (SIM), a timestamp that was truncated to 10 seconds, and an ID of the cell, thus, a subscriber can be mapped to a geographic location in a given time. These are extended customer type (business, consumer), the subscription type (prepaid, postpaid), the age and gender of the subscriber, and the Type Allocation Code (TAC) of the device. The TAC is the first eight digits of the International Mobile Equipment Identity (IMEI) number that refers to the manufacturer, and the model of the device wherein the SIM card is active. These values are also present in every record, so, for example, the device changes can be tracked as well.

As for the cells, a separate table was provided with a cell ID, the geographic location of the cell centroid, the area of the cell, and the distance between the centroid and the base station. These values are an estimation based on a momentary state, especially with the UMTS (3G) cells due to their breathing mechanism, which can change the geographic size of the serving area for load balancing. The heavily loaded cells shrink, and the neighboring ones grow to compensate[110].

The rationale for this "wide" format may be that the subscriber data and the device can be changed within the observation period. This occurs about 3000 times during the observation period. The owner of the subscription can change its details and, of course, change the device if they bought a new mobile phone, for example. Subscriber and customer type were provided for every SIM, but age and gender were missing in many cases, presumably due to the privacy options requested by the subscriber.

The records include neither the type of the activity (voice call, message, data transfer) nor the direction (incoming, outgoing), and there was no data provided by the operator to resolve the TACs to manufacturer and model.

The “April 2017” CDR data set includes mobile phone network activity of the Vodafone users from Budapest (and the surrounding areas) in April 2017. This contains 955,035,169 activity records, from 1,629,275 SIM cards. Figure 3.1, shows the activity distribution between the activity categories of the SIM cards. Only 17.67% of all the SIM cards that have more than 1000 activity records provide the majority (75.48%) of the mobile phone activity during the observation period. Figure 3.2, shows the distribution of the SIM cards by the number of active days. Only about one-third (33.23%) of the SIM cards have activity on at least 21 different days. Despite the relatively large number of SIM cards present in the data set, most of them are not active enough to provide enough information about their mobility habits. For the exact selection criteria, see Section 4.7.

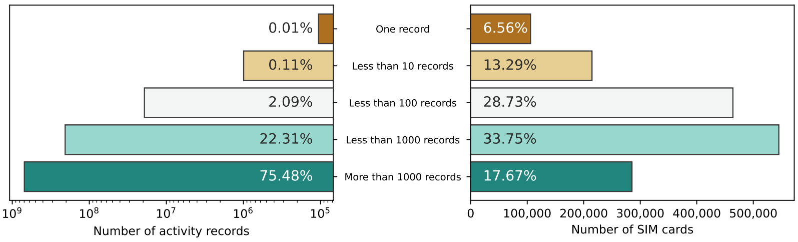


**Figure 3.1.:** SIM cards in the 2017-04 data set categorized by the number of activity records. The SIM cards with more than 1000 activity records (17.67% of the SIM cards) provide the majority (75.48%) of the activity.

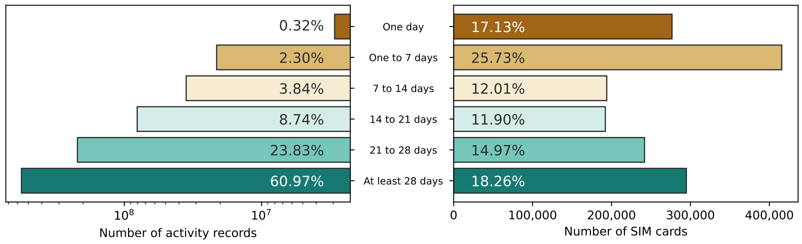


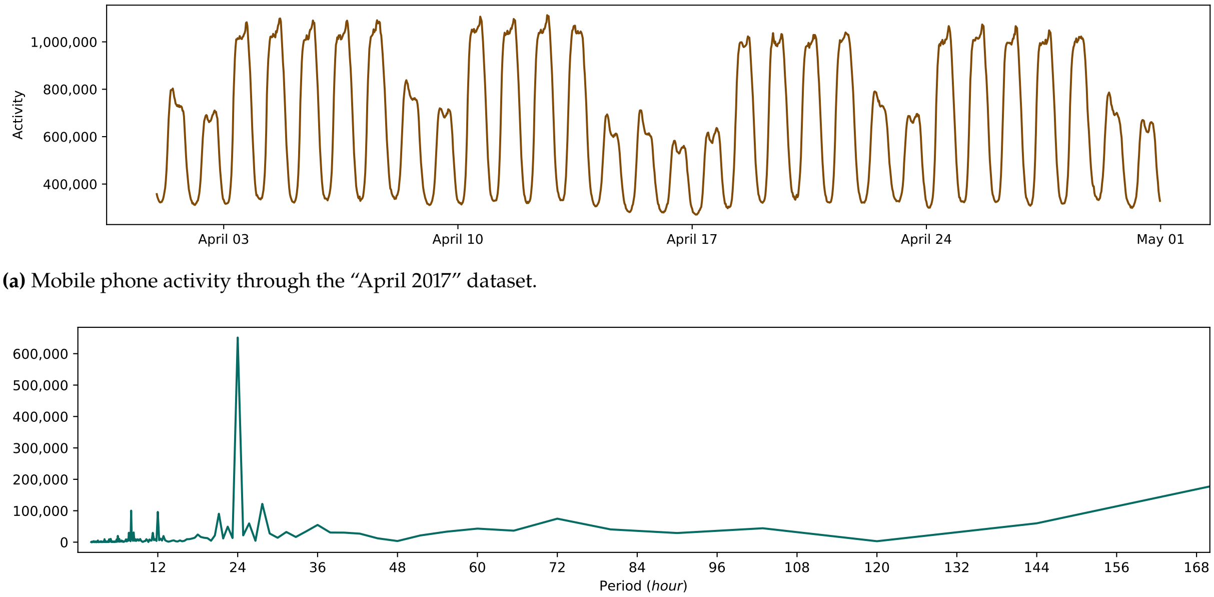
**Figure 3.2.:** SIM card distribution in the 2017-04 data set by the number of active days, in contrast of the number of records generated per category.

Figure 3.3a, displays the mobile phone activity as a time series for the “April 2017” dataset. It is quite regular, considering the 4-day weekend in the middle of the month due to Easter [1]. Figure 3.3b shows its Fourier decomposition to highlight the seasonality of the data. As expected, this 30-day dataset has a 24-hour periodicity.

### 3.1.1. Missing Properties

The subscriber type and subscription type were provided for all the SIM cards, whereas age and gender were frequently missing. For a group of SIM cards, some of these values have changed during the data set. These changes could be realistic as the subscription type, the subscriber (the ownership), or the subscription plan can be changed at will. Sometimes,

[1] From 2017, Good Friday is also a holiday in Hungary.



**(a)** Mobile phone activity through the "April 2017" dataset.

**(b)** Fourier decomposition.

**Figure 3.3.:** The mobile phone network activity (**a**) during the observation period of April 2017, and its Fourier decomposition (**b**).

the change affects the age or the sex as they become unknown or known when they were unknown before.

Let us take that someone changed their subscription plan, which is treated as a new contract between the subscriber and the operator, and this might have affected the personal data usability for marketing purposes. Either, someone just revoked that permission. Nevertheless, this affects about 2000 SIM cards out of about 1.6 million. In these cases, all the subscriber properties were set to unknown without trying to decide which value should be used.

### 3.1.2. Outliers

The age property contains some unrealistic values, like a more than 210 years old subscriber, and there were quite a few 115 years old subscribers. It is assumed that these values might be administrative errors. The number of at least 90-year-old people was 2926, and out of that, 554 were between 90 and 100, according to the data, which is unlikely. For comparison, 59,470 over 90 years old people lived in Hungary in 2017 KSH[111], which is 0.6% of the population.

## 3.2. Vodafone June 2016

Although this data set predates the formerly introduced one, I consider it secondary, as almost all of my work was performed on the "Vodafone April 2017" data set (Section 3.1). It is more recent and better in data quality. The collaboration between Vodafone Hungary and Óbuda University was an iterative process. Based on our remarks, they improved the provided data in a format and — if it was possible — in quality.

This section follows the structure of the previous one, describing the "Vodafone June 2016" data set.

The observation area of the CDR data was Budapest, the capital of Hungary, and the surrounding county. In 2016 Q2, the nationwide market share if Vodafone Hungary was 25.3% [109], the data format is similar as described in Section 3.1. This data set contains 2,291,246,932 records from 2,063,005 unique SIM cards. It is more than twice as many records for a practically same time interval and more than 26.6% increase regarding the unique SIM card number.

Although this data also covers one month, June, instead of April. It is a later period of the year that has some consequences: June, like an early summer month, is warmer with longer days, and there are usually more tourists in Budapest. Also, the school semesters end in June, which might be some impact on the mobility trends, and the 2016 UEFA European Football Championship took place during this observation period (for the Euro 2016 case study, see Section 5.1).

The main drawback of this data set is a large number of missing cell coordinates and possibly dirty data as regards the third Sunday of the month. Out of the 7239 cell IDs that appear in the data set, only 6268 has known geographic coordinates. 22.5% of the records (515,583,442) take place in unknown locations, which makes this data set highly unreliable when it comes to mobility analysis.

Moreover, 419 cells ceased to operate (did not have any activity) before the end of the month. Figure 3.4 shows the distribution of these cells during the month [2]. According to my knowledge, operators regularly adjust the network, including installing new cells and shutting down others. The problem is that the cell map does not associate temporal information with the geographic locations, so these changes are not represented. The numerous cells without geographic location might be the replacement cells of the removed ones or just temporary cells.

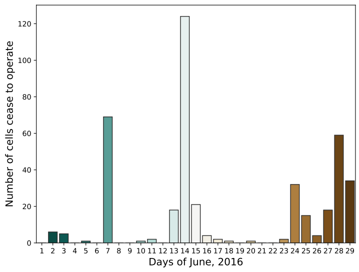


**Figure 3.4.:** Number of cells stopping to operate per day of June 2016

For example, Óbuda Island is a recreational area with marginal mobile phone activity during most of the year. But, it gives place the Sziget Festival, that had approximately 441,000 visitors in 2015, with a daily capacity of at most 90,000 visitors [112]. For this event, the operators install temporary cells to serve the massively increased demand of the mobile phone network capacity.

Figure 3.5, displays the mobile phone activity as a time series for the "June 2016" dataset. Unlike Figure 3.3a, it shows some considerable irregularities. Anomaly 1 was caused by the missing activity records on June 1, 2016. The data set contains about 16.5 million records for that day, which is only 18.5% of an average weekday. Interestingly, the obtained records are distributed across the whole day from numerous cells. So, it is not like a portion of cells or a period of the day was omitted. The majority of the assumed records are missing without apparent reason. Anomaly 2 is quite the opposite. June 12 contains an inexplicable activity surplus and is detailed later in this section. Anomaly 3, 5, 6 and 8 in Figure 3.3a are covered in Section 5.1. However, the reason for the sudden negative peaks denoted by numbers 7 and 9 is unknown. Just as in the case of anomaly 4.

[2] This phenomenon also affects the "April 2017" data set, but not to this extent, and almost every cell has a valid location, even if the adjustments are not documented in the received cell-map.

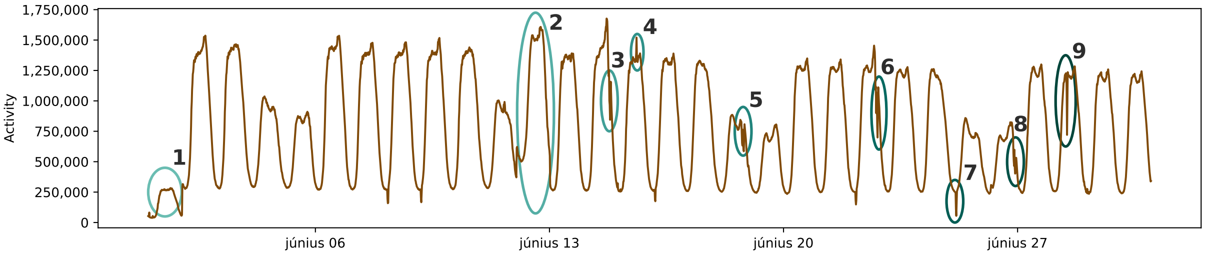


**Figure 3.5.:** Mobile phone activity through the "June 2016" dataset, with several anomalies.

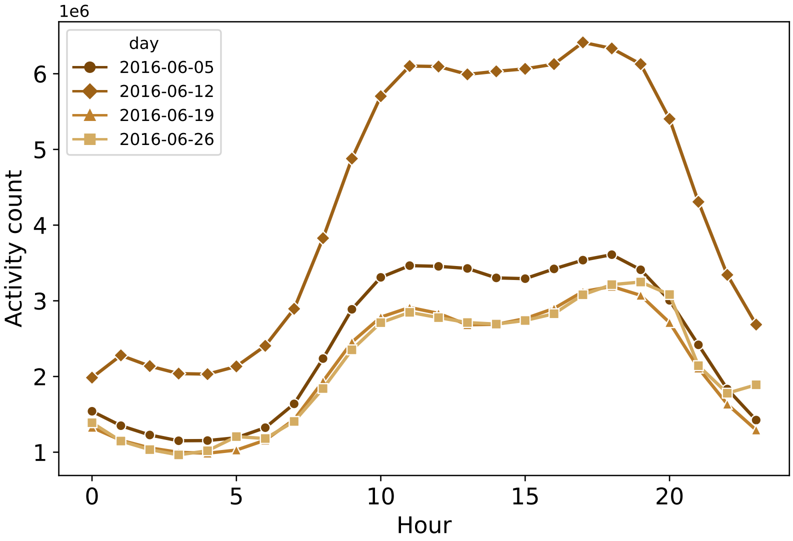


**Figure 3.6.:** Hourly aggregated activity on Sundays

Figure 3.6 shows the hourly aggregated activity on Sundays of June 2016. All four days have the same tendency. Between the last two Sundays of the month, there was hardly any difference. One notable difference is around 20:00, which may be explained in Section 5.1.4. On June 5, the activity levels are notably higher during the daytime, but this is within reason. However, on June 12, the activity was abnormally intense, almost twice as much as it would be expected, even during nighttime. I have no information about any events or circumstances that could have normally caused this activity surplus on June 12, 2016. It was not concentrated on a definite time interval or area of the city. At first, it could seem as if the records of that day had been duplicated, but there is no direct evidence of it in the data. As the surplus is unexplainable, I consider it a sign of dirty data.

Apart from the issues, the same descriptive analysis has been performed as on the "Vodafone April 2017" data set. Figure 3.7, shows the activity distribution between the activity categories of the SIM cards. The dominance of the last category, the SIM cards with more than 1000 activity records, is even more significant. This almost 27% of the SIM cards produce the more the 91% of the activity.

Figure 3.8, shows the SIM card distribution by the number of active days. Only the 34.59% of the SIM cards have activity on at least 21 different days. The ratio of the short-term (less than a week) present SIM cards is larger (more than 50%) than in the "Vodafone April 2017" data set. There are 241,824 SIM cards (11.72%) that appeared at least two days, but the

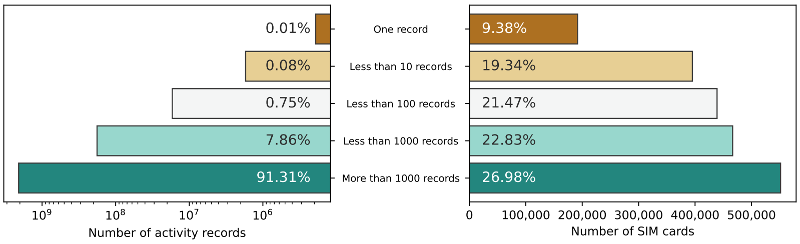


**Figure 3.7.:** SIM cards in the 2016-06 data set categorized by the number of activity records. The SIM cards with more than 1000 activity records (26.98% of the SIM cards) provide the majority (91.31%) of the activity.

difference between the first and the last activity is not more the seven days. High levels of tourism are usual during this part of the year.

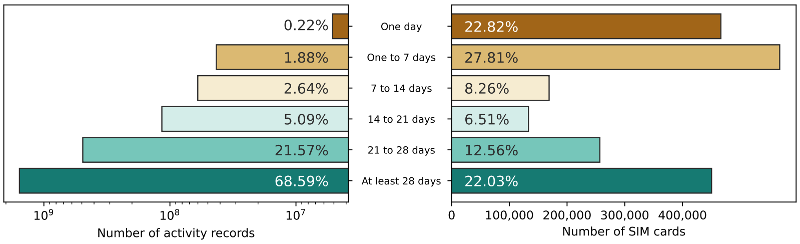


**Figure 3.8.:** SIM card distribution in the 2016-06 data set by the number of active days, in contrast of the number of records generated per category.

The subscriber property changes affected this data set as well. About 3000 SIM cards were affected out of about 2 million. Some very old subscribers are also present in this data set.

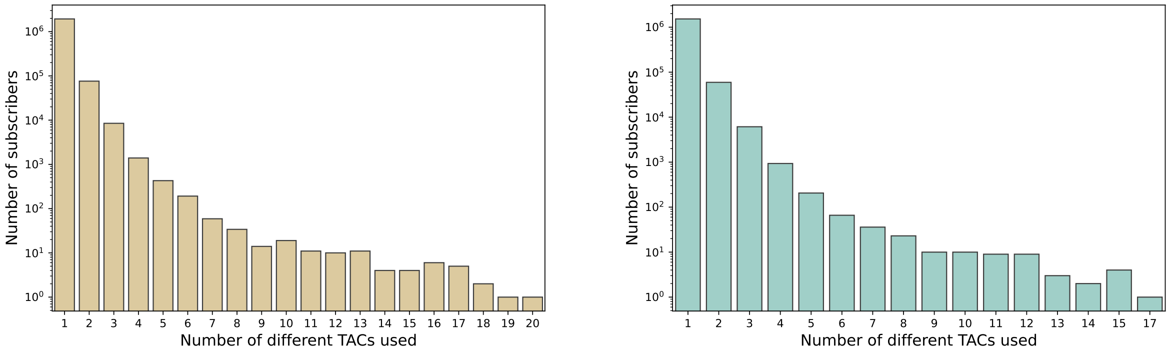
## 3.3. Device Types

Both CDR dataset contains Type Allocation Codes, that the first eight digits of the International Mobile Equipment Identity (IMEI) number, allocated by the GSM Association and uniquely identifies the mobile phone model, and every GSM capable device.

The TACs are provided for every record because a subscriber can change their device at any time. Naturally, most of the subscribers (95.71% in June 2016, and 95.8% in April 2017) used only one device during the whole observation period, but there were some subscribers, maybe mobile phone repair shops, who used multiple devices (see Figure 3.9). As a part of the data cleaning, the wide-format has been normalized. The CDR table contains only the SIM ID, the timestamp, and the cell ID. A table is formed from the subscriber and the subscription details, and another table tracking the subscribers' device changes.

### 3.3.1. Resolving Type Allocation Codes

To the best of my knowledge, there is no publicly available TAC database to resolve the TACs to manufacturer and model, although some vendors (e.g., Apple, Nokia) publish the TACs of their products. The exact model of the phone is required to know how recent and expensive a mobile phone is. Although this is not even enough to determine how much the cell phone cost for the subscriber as they could have bought it on sale or discount via the operator in exchange for signing an x-year contract. Still, the consumer price should designate the order of magnitude of the phone price.



**(a)** June 2016

**(b)** April 2017

**Figure 3.9.:** The number of different TACs used by the subscribers.

The dataset of TACs provided by *51Degrees* has been used, representing the model information with three columns: "HardwareVendor", "HardwareFamily" and "HardwareModel". The company mostly deals with smartphones that can browse the web, so the data set usually does not cover feature phones and other GSM-capable devices. Release date and inflated price columns were also included, but these were usually not known, making the data unsuitable to use on its own.

Although it cannot be separated by type, the CDR data contains not only call and text message records but data transfer as well. Furthermore, some SIM cards do not operate in phones but in other – often immobile – devices like a 3G router or a modem. 51Degrees managed to annotate several TACs as a modem or other not phone devices, which was extended by manual search on the most frequent TACs. There were 324,793 SIM cards that used only one device during the observation period and operated in a non-phone device.

### 3.3.2. Fusing Databases

For a more extensive mobile phone price database, a scarped GSMArena database [113] has been used. GSMArena[3] has a large and respectable database, that is also used in other studies [114, 115]. The concatenation of the brand and model fields of the GSMArena database could serve as an identifier for the database fusion. 51Degrees stores the hardware vendor, family, and model, where the hardware family often contains a marketing name (e.g., [Apple, iPhone 7, A1778]). As these fields were not always properly distinguished, their concatenation may contain duplications (e.g., [Microsoft, Nokia Lumia 820, Lumia 820]). So, for the 51Degrees records, three identifiers were built using the concatenation of fields (i) vendor + family, (ii) vendor + model, and (iii) vendor + family + model, and all the three versions were matched against the GSMArena records.

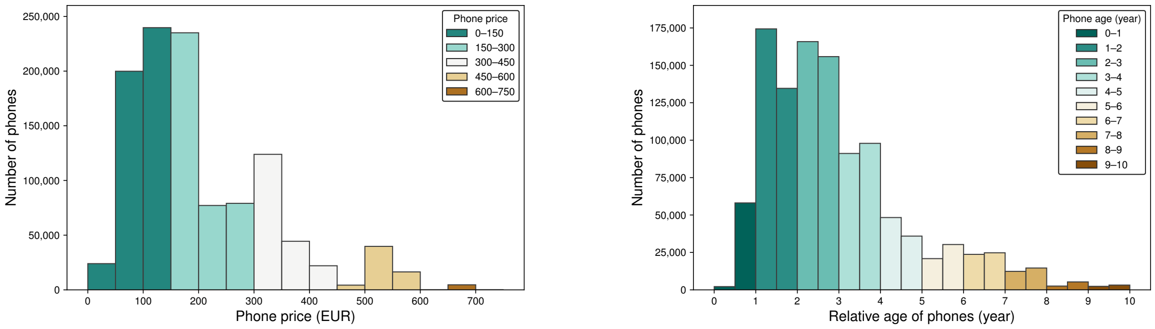
Another step of the data cleaning is to correct the name changes. For example, BlackBerries were manufactured by RIM (e.g., [RIM, BlackBerry Bold 9700, RCM71UW]), but later, the company name was changed to BlackBerry, and the database records are not always consistent in this matter. The same situation occurs due to the Nokia acquisition by Microsoft.

[3] `https://www.gsmarena.com/`

The simple string equality cannot be used due to writing distinction to match these composite identifiers, so the Fuzzy String match is applied using the FuzzyWuzzy 0.18 Python package, which uses Levenshtein Distance to calculate the differences between strings. This method was applied for all the three identifiers from the 51Degrees data set, and the duplicated matches (e.g., when the family and the model are the same) were removed. Mapping the GSMArena database to the 51Degrees adds phone price and release date information to the TACs, which can be merged with the CDRs.

From the GSMArena data, two indicators have been extracted: (i) the price of the phone (in EUR) and (ii) the relative age of the phone (in months). The phone price was left intact without taking into consideration the depreciation, and the relative age of the phone was calculated as the difference between the date of the CDR data set and the release date of the phone.

Figure 3.10 shows the distribution of the phone prices and relative ages within the April 2017 data set. The relative age has a nice distribution, showing that most cellphones are 1 to 3 years old. There were some new and very old phones still in use. The cellphone price distribution follows the relative ages. However, the number of expensive phones seems to be unrealistically low [4], so analysis has been performed on an expensive and well-known brand, the iPhone.



**(a)** **(b)**

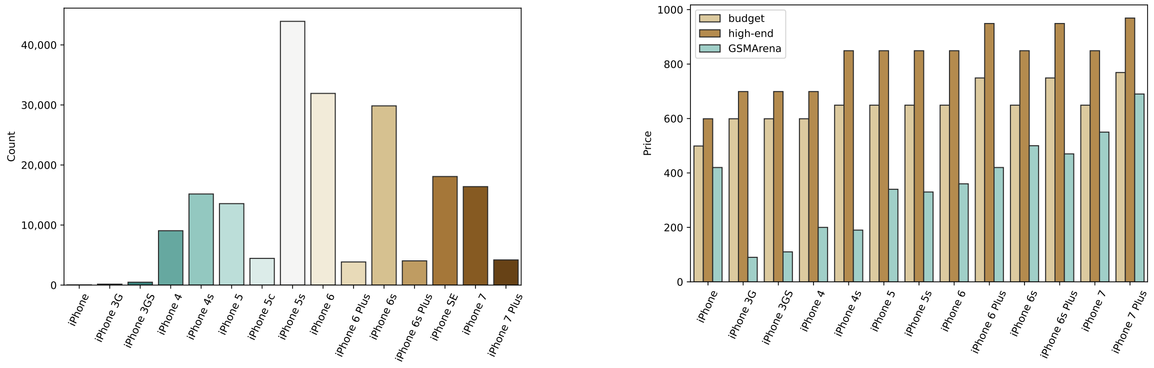
**Figure 3.10.:** Distribution of the mobile phone prices (**a**), and the mobile phone relative ages (**b**).

### 3.3.3. iPhones

As Apple iPhones are considered a status symbol [118], it makes them suitable to validate the phone price database [113]. Figure 3.11a shows the number of subscribers that exclusively use the certain iPhone models in the "April 2017" dataset. Using TAC values, it is not possible to distinguish the iPhone models based on specifications like storage. However, it is clear that the most expensive models ("Plus" versions) do not have a significant user base, in contrast with some older models like iPhone 4 and iPhone 5 series.

The launch prices of the iPhone models, released until April 2017, are obtained from [117]. Figure 3.11b, compares the two sources. As there are different versions of a certain model, a "budget" (with the lowest

[4] This was realized by an anonymous reviewer of my paper [116*].



**(a)** Model distribution

**(b)** Price comparison

**Figure 3.11.:** Based on the "April 2017" dataset, the different iPhone models in use are also displayed (**a**), and comparing Apple iPhone prices [117] with the GSMArena-based source [113] (**b**). Versions with the lowest amount of storage denoted by "budget", and versions with the most expensive versions categorized as "high-end".

amount of storage) and a "high-end" (the most expensive) version are also displayed. Although GSMArena price property is supposed to be a launch price, Figure 3.11b clearly shows that they are much lower than the original prices. Moreover, the older the phone is, the lower the available prices are, except for the first iPhone. Note that GSMArena prices are in EUR, whereas the ground truth prices are in USD, which cannot cause the difference. The results of this analysis imply that the phone prices might have depreciated.

## 3.4. Estate Price Data

Property estate price data was provided by the *ingatlan.com* estate selling website. The data contains slightly more than 60 thousand estate locations, floor spaces, and selling prices from the advertisements. The prices may not be the actual value that the buyer paid, but even if there was some bargaining, the order of magnitude should be reasonably accurate.

The data is from 2018, not from the same year as the CDR data. However, the price differences between the areas of Budapest have not changed significantly during those years, so it should be adequate to describe the average estate price of an area. The price of one square meter was calculated from the floor space and the selling price. In this way, the price level of two different estates in two very different parts of the city can be compared.

The data source contains slightly more than 85 thousand estate locations with floor spaces and selling prices. Figure 3.13, shows its distribution. Figure 3.12 shows the estate advertisements over Pest county, and the administrative border of Budapest is also displayed. The more expensive estate advertisements are represented both by color and larger markers.

Although 70.78% of the data points are within Budapest, there are some areas without property price samples, even in Budapest. Indicators are often aggregated by cells (Section 4.3), so it is crucial to know the average property price for an area covered by the given cell. Besides the cell-level aggregation, the property prices were aggregated on the suburb and district or settlement level. Budapest has more than 200 suburbs of

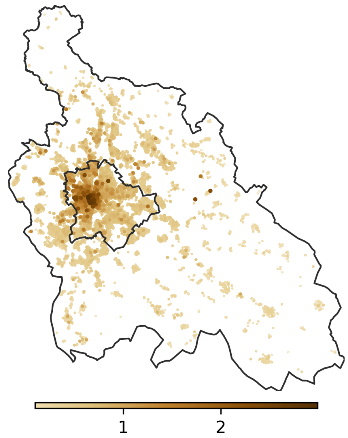


**Figure 3.12.:** Spatial distribution of the normalized real estate prices (million HUF).

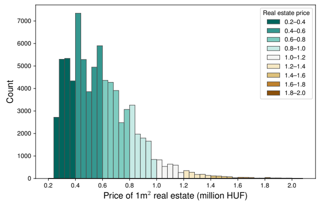


**Figure 3.13.:** Real estate price histogram from the *ingatlan.com* data source.

varying sizes, and the average property price of the suburbs are also determined. When a cell does not have a property price sample (or not enough), the average property price of the underlying suburb was used.

For every cell Voronoi polygons, it was determined how large part of the cell overlaps with the suburbs. The average property price of a cell is the weighted mean of the suburb property prices of the overlapping suburbs. This method significantly reduced the number of cells without estate price data and compensated for the extreme differences between the neighboring cells that might come from the advertisements. Figure 3.14 shows the result of the mean housing prices by the cell polygons.

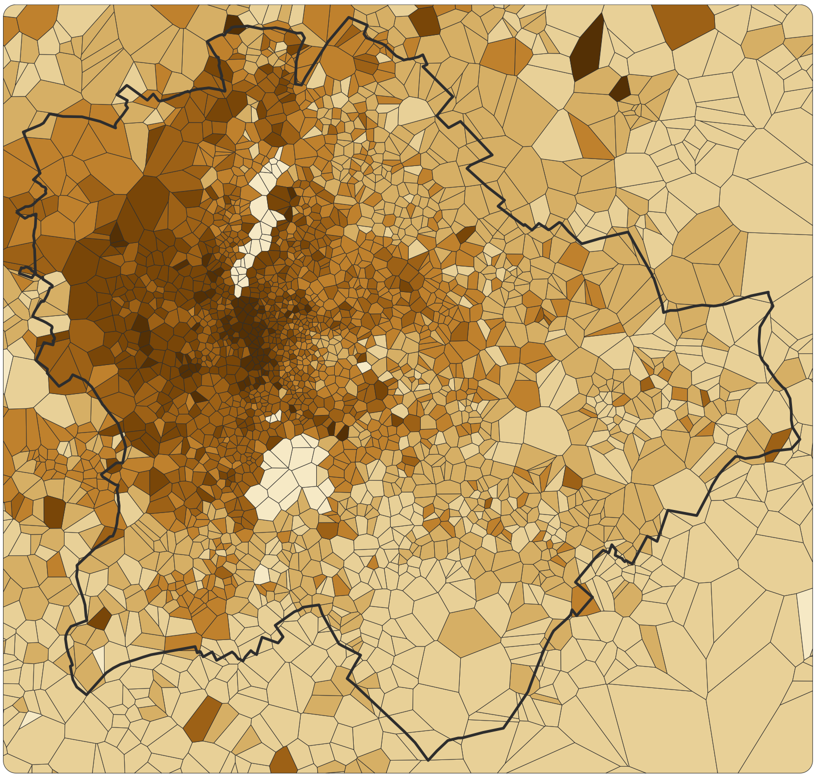


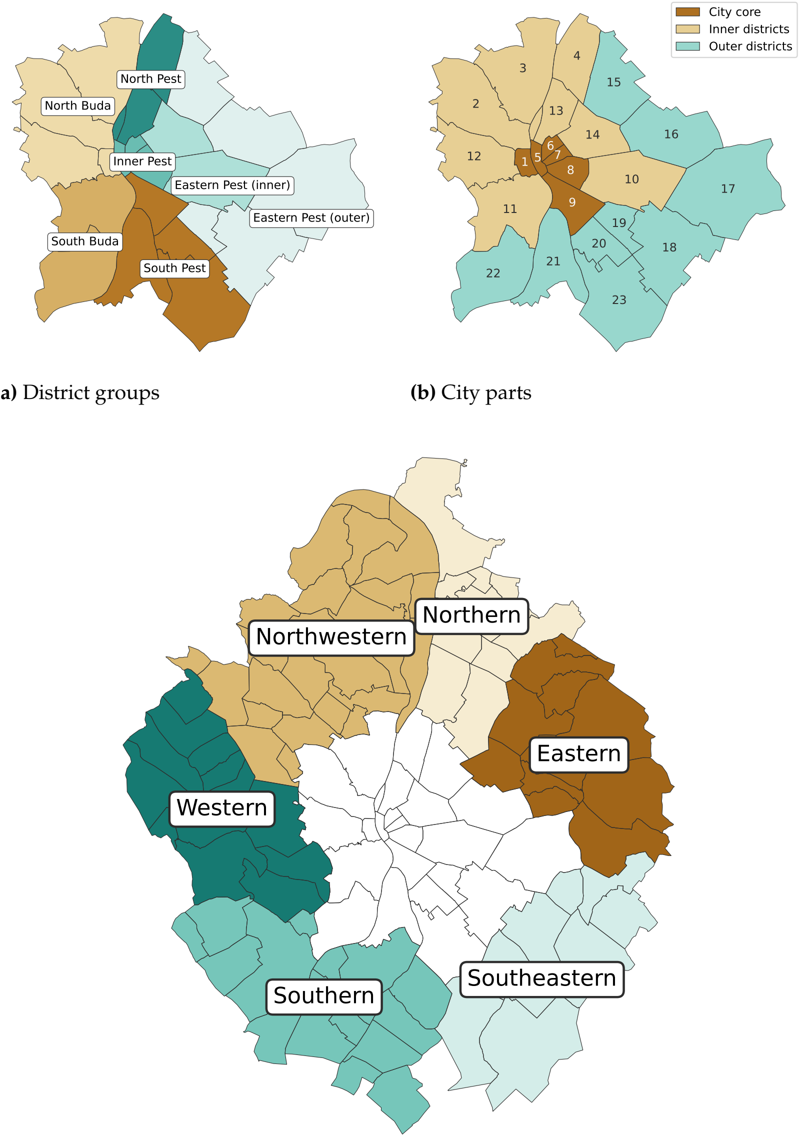
**Figure 3.14.:** Average real estate price per cell polygons from the *ingatlan.com* data source.

## 3.5. OpenStreetMap

OpenStreetMap (OSM) provides community-built map data about administrative boundaries (e.g., county, county, city, district), roads, railways, stations, and Points of Interest (e.g., museums, cafés) all over the world. I predominantly use OSM map data to visualize the mobile phone data in a spatial context.

Budapest is divided into 23 districts and more than 200 suburbs. Moreover, KSH groups the districts into three city parts (Figure 3.15b) and seven district groups (Figure 3.15a) as well. The agglomeration is divided into six sectors (Figure 3.15c). The administrative borders of these settlements, including the districts and suburbs of Budapest, are derived from OSM. Historically, District 21 (Csepel) is not part of either Buda or Pest, as it is located on the northern end of Csepel Island. Still, statistics used to classify as a Pest-side district (e.g. [119]).

Administratively, Margaret Island was part of the 13th district. Since July 2013, it has been directly under the control of the city, being a part of Budapest without belonging to any districts. However, some maps in this work still denote it as a part of District 13. Practically, the whole island is a recreational area covered with landscape parks.



**(a)** District groups

**(b)** City parts

**(c)** Sectors of the Budapest agglomeration

**Figure 3.15.:** District groups (**a**) and city parts (**b**) of Budapest, and the sectors of the Budapest agglomeration (**c**).

## 3.6. Other Data Sources

### Statistical Data

As a validation for the results (e.g., estimated population), statistical data is obtained from the Hungarian Central Statistical Office (KSH), using its interface of the spreadsheet sets, called *STADAT* [120].

### Astronomical Data

Also, as a validation, astronomical information (sunrise and sunset) has been obtained, for Budapest, from Visual Crossing, which collects global weather data [121]. This data was applied in Section 7.4 to compare with the calculated day lengths.

### Twitter

In Section 5.2.1, Twitter data is utilized that was obtained via its academic research access program [122], which provides access for non-commercial research purposes. To download historical tweets based on hashtags, the *twarc* software was used [123].

# Data Processing Framework 4.

A computational framework has been developed to process the CDR data, including preprocessing, cleaning, home and work location estimation, or calculating the mobility indicators (like Radius of Gyration and Entropy). This simplified process can be seen in Figure 4.1 and will be discussed further in this chapter.

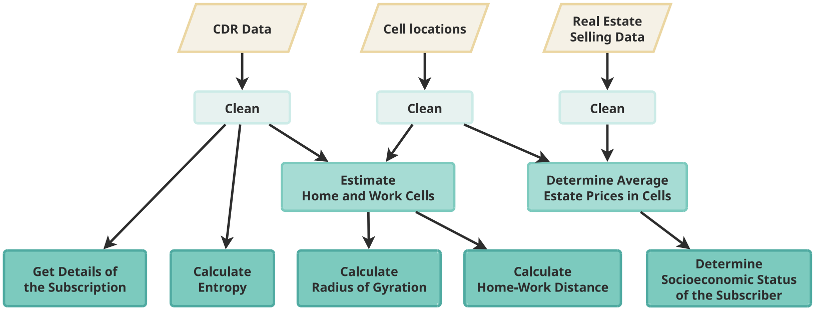

CDR Data
Cell locations
Real Estate Selling Data
Clean
Clean
Clean
Estimate Home and Work Cells
Determine Average Estate Prices in Cells
Get Details of the Subscription
Calculate Entropy
Calculate Radius of Gyration
Calculate Home-Work Distance
Determine Socioeconomic Status of the Subscriber

**Figure 4.1.:** The simplified processing the Mobile Phone data along with the Cell information and the Estate Price data used to determine the financial status of the subscribers.

## 4.1. Software Environment

I only used open-source software to build the data processing and visualization environment. The framework that practically is a set of scripts is primarily written in Python, using the following packages: Pandas, GeoPandas (Pandas for geospatial data) for the data manipulation, Matplotlib and Seaborn for plotting, NumPy, SciPy, Skikit-learn for numeric and scientific methods, NetworkX for graph/network analysis, OSMnx for downloading streets (and other objects) from OpenStreetMap as a NetworkX graph, just to mention the essential packages I have been relying on during my work.

For tasks, when the processing speed of the interpreted Python became a bottleneck, I have written small, task-specific programs in Go (in the beginning) or Nim (later). Small stream-processing tools have been written, using compiled languages, that read a large CSV file line by line, perform some operation on the records and write the result directly to the disc or accumulate in the memory. It worked well for these tasks because the activity records are independent of each other.

Otherwise, the usual procedure was to group the records by the SIM card ID and process a given subscriber's data on a single thread (for example, to calculate a mobility metric). Multiple subscribers' data was processed parallelly to speed up the computation.

All the CDR data is stored in a PostgreSQL database with an enabled PostGIS extension to add support for geospatial operations within the database. Spatial filtering can be faster this way, and only the minimum amount of records needs to be loaded into a (Geo)Pandas (Geo)DataFrame.

### 4.1.1. Visualization

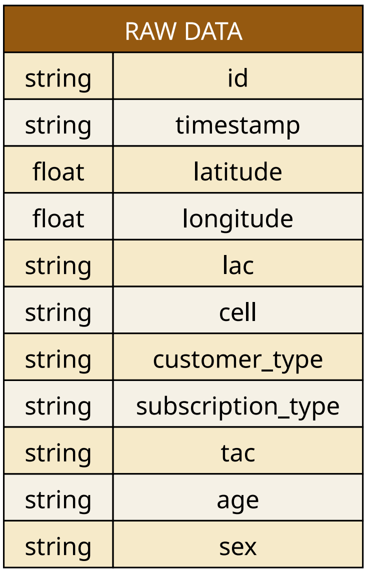
The visualization is a crucial step. Not just when the results are presented but during the analysis as well. One could argue that it is even more crucial to have a fast and reliable way to visualize the results during data analysis. Besides the descriptive statistics, the spatial distributions are also essential to validate a result.

Both Pandas and GeoPandas have APIs to plot (Geo)DataFrames using Matplotlib to have a quick view of the data, which can provide very fast feedback, but sometimes a more interactive tool is required. QGIS is an open-source Geographic Information System that can provide this. QGIS is an excellent software with an enormous number of features. I only used a small subset of its capabilities, usually just to visualize a GeoDataFrame dump or a GeoJSON with some linked data from a CSV or directly from the database.

## 4.2. Data Preparation

As the received data had a wide format (see Figure 4.2), they were normalized before importing into the database. The CDR table only contains the SIM ID, the timestamp, and the cell ID. a table has been introduced to store the SIM related properties (like subscription type, customer type, age, gender) and another to store the cell properties (cell centroid and base station coordinates).

Order numbers were used for SIM IDs in order of appearance instead of the long hash values to save disk space. The longitude and latitude values, provided in EPSG:4326 projection (also known as WGS 84) were rounded to 6 decimals because further decimals have no practical meaning in CDR positioning [1]. The received data uses property labels like “MALE”, “FEMALE” to indicate gender, “CONSUMER”, “BUSINESS” for customer type, “PREPAID”, “POSTPAID” for subscription type and the “UNKNOWN” string is used to denote unknown values. Again, these strings were shortened, and the unknown values were represented with the proper “NULL” value to save disc space.



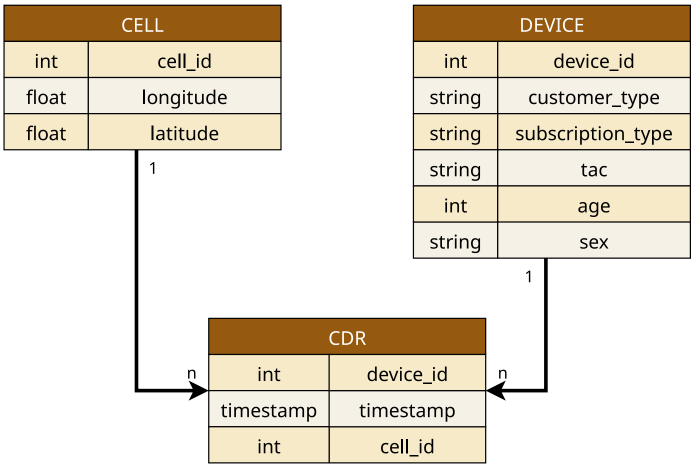
| RAW DATA | |
|---|---|
| string | id |
| string | timestamp |
| float | latitude |
| float | longitude |
| string | lac |
| string | cell |
| string | customer_type |
| string | subscription_type |
| string | tac |
| string | age |
| string | sex |

**Figure 4.2.:** The schema of the received data.

### 4.2.1. Normalization

The single table, received data has been normalized, and three new tables were formed (see Figure 4.3): i) cell, containing a cell ID and the coordinates of the cell centroid, ii) a device table, containing the device/SIM ID and information about the subscriber (age, sex), the subscription (consumer or business, prepaid or postpaid), and the Type Allocation Code (TAC) that can be used to identify the device where the SIM card is operating, iii) the CDR table, that serves as a link table to map a subscriber to a geographic location in a given time, (iv) TACs are loaded to another table after the dominant devices had been determined for the SIM cards, and (v), indexes were built for all the tables.

[1] The sixth decimal represents about 0.111 meter at the Equator. As the CDRs roughly provides a house-block level accuracy in downtown, the sixth decimal is more than enough in this use-case.



| CELL | |
|---|---|
| int | cell_id |
| float | longitude |
| float | latitude |

| DEVICE | |
|---|---|
| int | device_id |
| string | customer_type |
| string | subscription_type |
| string | tac |
| int | age |
| string | sex |

| CDR | |
|---|---|
| int | device_id |
| timestamp | timestamp |
| int | cell_id |

**Figure 4.3.:** Normalized data

## 4.3. Cell-Map Mapping

In most cases, the operator provided only the base stations' coordinates. Therefore, it is common practice [18, 49, 124–126] at CDR processing to use the base stations to map Call Detail Records to geographic locations and use Voronoi tessellation to estimate the covered area of the base station.

Vodafone Hungary provided the base station coordinates and the (estimated) cell centroids for the cells. With this, the position of the SIM cards is known with a finer granularity as if only the base station locations were known, as a base station can serve several cells.

The close cells within 100 m are merged using the DBSCAN algorithm of the Scikit-learn [127] Python package using the cell activity as weight, and the Voronoi tessellation is applied for merged cell centroids, similarly, as in [51].

It is possible to represent the merged cell by an imaginary cell centroid that could be the center of mass of the merged points, but I wanted to use an existing cell (centroid), the medoid, to represent the cluster. The number of the activity records of the cells was applied as a weight for the DBSCAN algorithm. This way, the most active or the most significant cell was chosen to represent multiple cells of an area.

After merging the cells, the 5412 cells were reduced to 3634. The merge affects mostly Budapest downtown since more mobile phone cells are applied in densely populated areas.

When mapping data with exact GPS coordinates to the cells, the Voronoi polygons were used to determine in which cell the given point belongs. As shown in Section 2.10, this method has some drawbacks, but without exact cell geometries, this is the only available solution. The estate price data (see Section 3.4) were mapped to cells via these Voronoi polygons. When there were no property price data in a cell, the average price of the underlying administrative area was used (e.g., suburb, district, or settlement).

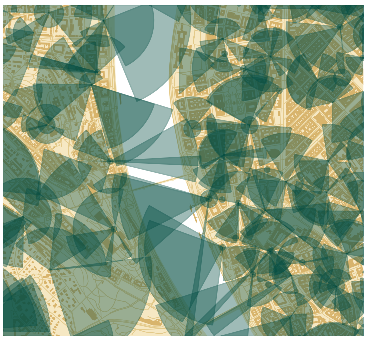


**Figure 4.4.:** Reconstructed cells.

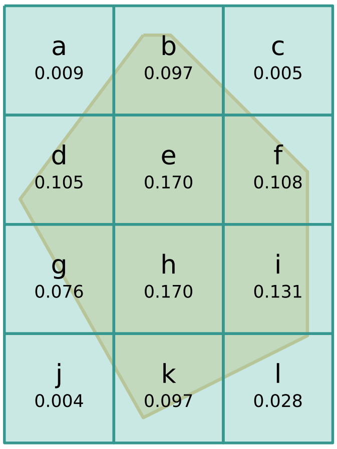
Gábor Bognár, a colleague of mine, tried to reconstruct the cell geometries from the received information, but his work resulted in unrealistic coverage (Figure 4.4). He suspected that the data was erroneous in respect of the geometries.

## 4.4. Rasterization of Locations

With an accurate geographic data source like GPS traces, it can be more natural to map locations to a spatial grid. With CDR data, the geographic locations are bound to the cell or a base station coordinates, and Voronoi tessellation is often applied, but what about the rasterization?

The problem is that it is not possible to know where a subscriber is within the cell. So, the maximal available accuracy is the cell. A finer resolution spatial raster can work well with point coordinates that can be classified to the rasters, but in the case of CDR, the classification can be only achieved with probabilities.

The intersection ratio with the cell polygons is calculated for every element of the grid (see Figure 4.5) that also serves as a probability of being in a given raster when the subscriber is in a given cell. Theoretically, this probability could be improved by removing rasters where people usually cannot be present. Using the example of Figure 4.7a, there are several cells that cover the river. When the subscriber is in one of those cells, they must be on the riverbank since being on a ship is usually not probable. OSM could provide landscape data for removing these rasters, but this functionality was not implemented in this framework.



| | | |
|---|---|---|
| a<br>0.009 | b<br>0.097 | c<br>0.005 |
| d<br>0.105 | e<br>0.170 | f<br>0.108 |
| g<br>0.076 | h<br>0.170 | i<br>0.131 |
| j<br>0.004 | k<br>0.097 | l<br>0.028 |

**Figure 4.5.:** Rasterization example with the coverage ratios.

When the probabilities are determined, I applied the Jefferson apportionment method (also known as the D'Hondt method or Hagenbach-Bischoff method) from the Python package, "voting", to distribute the subscribers of the cell between the rasters, using the probabilities as weight. It is feasible when only cardinality matters. For example, when one wants to plot a heatmap, but as the distribution cannot pay regard to the properties of the subscribers, it cannot be stated that subscribers in a given raster have certain mobility habits. On a cell level, this kind of statement can be made.

In [128], Galiana et al. also map Voronoi polygons to the grid but use a different approach for distributing subscribers between rasters.

## 4.5. Selecting an Area

In order to select an area for further investigation, a polygon has to be defined. The `geojson.io` provides an easy-to-use tool to define polygons over the map that can be downloaded in multiple GIS formats, such as GeoJSON or Well-known text (WKT). Figure 4.6, shows the observation area from the Central European University (CEU) demonstration case study (Section 5.2). It contains three different areas: (i) one from the Castle Garden to the Chain Bridge at the Buda side of the river, (ii) one in the middle from the bridge to the Kossuth Lajos square, covering the CEU buildings, and (iii) the parliament and the Kossuth Lajos square on the top of the map (Figure 5.8 highlights these buildings).

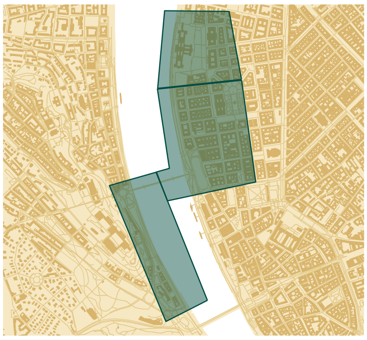
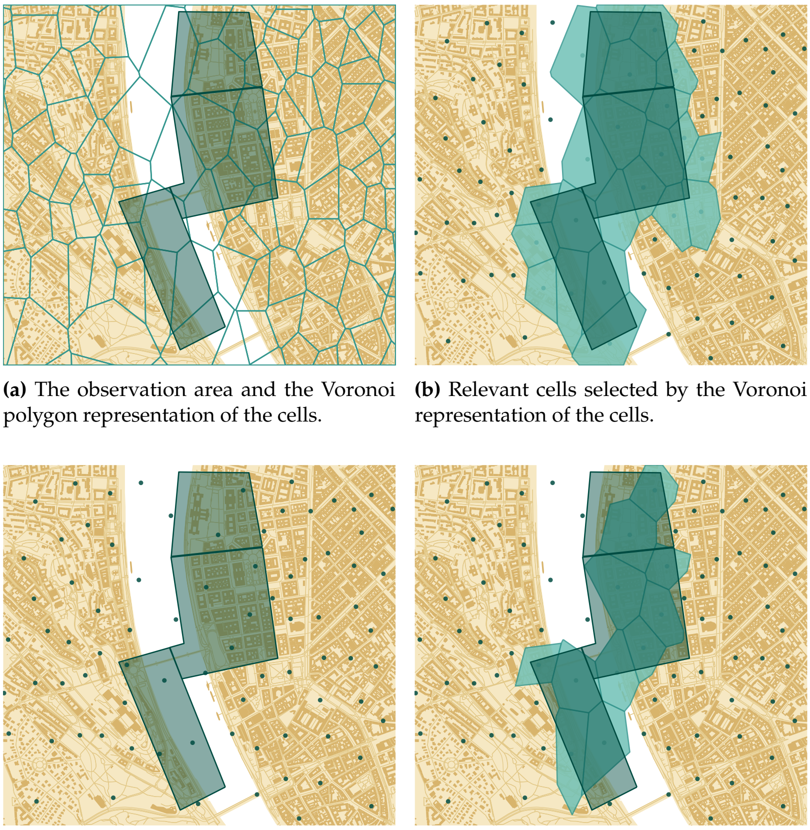


**Figure 4.6.:** Observation area from the CEU demonstration case study (Section 5.2).

The affected cells can be easily selected, using a polygon (or polygons) to define the observation area. The cells are represented with Voronoi polygons (Section 4.3) and the relevant cell selection can be performed in two ways: (i) selecting the cells Voronoi polygon intersects the selector polygon (Figure 4.7a and 4.7b), or (ii) selecting the cells which centroids intersect with the polygon (Figure 4.7c and 4.7d). The former method returns more cells, which can be more favorable in some cases, for example, when the target area is relatively small compared to the neighboring Voronoi polygons.



**(a)** The observation area and the Voronoi polygon representation of the cells.

**(b)** Relevant cells selected by the Voronoi representation of the cells.

**(c)** The observation area and the cell centroids.

**(d)** Relevant cells selected by the cell centroids.

**Figure 4.7.:** Comparison of the relevant cell selection methods.

## 4.6. Home and Work Locations

Most of the inhabitants in cities spend a significant time of a day at two locations: their homes and workplaces. In order to find the relationship between these most important locations and Social Economic Status (SES), first, the positions of these locations have to be determined. There are a few approaches used to find home locations via mobile phone data analysis [58, 129, 130].

The work location was determined as the most frequent cell where a device was present during working hours on workdays. Working hours were considered from 09:00 to 16:00. The home location was calculated as the most frequent cell where a device was present during the evening and the night on workdays (from 22:00 to 06:00) and all day on holidays. Although people do not always stay at home on the weekends, it is assumed that most of the activity is still generated from their home locations.

Most of the mobile phone activity occurs in the daytime (see Figure 4.8), which is associated with work activities. This may cause the home location determination to be inaccurate or even impossible for some devices.

This method assumes that everyone works during the daytime and rests in the evening. Although in 2017, 6.2% of the employed persons regularly worked at night in Hungary [131], the current version of the algorithm does not try to deal with the night-workers. Some of them might be identified as regular workers, but their work and home locations are mixed.

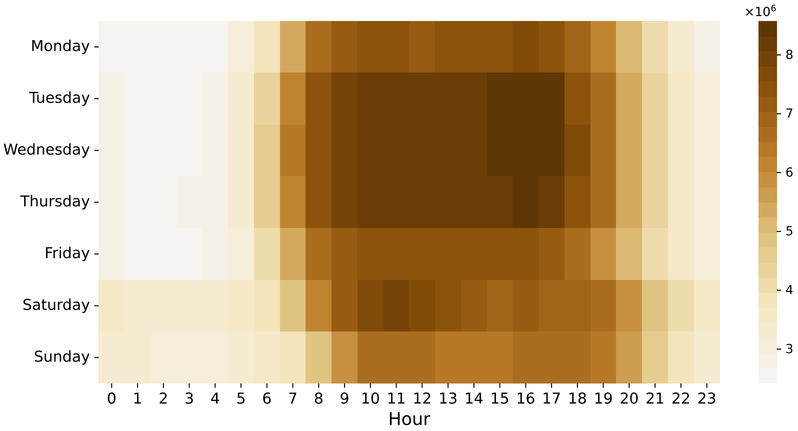


**Figure 4.8.:** The mobile phone activity distribution by days of week and hours based on the 2017 data set. Mondays and Fridays are not so bright as the other workdays, probably because of Easter Monday and Good Friday that are holidays in Hungary.

## 4.7. Selecting Active SIMs

As showed in Chapter 3, most of the SIM cards have very little activity records (Figure 3.1), that is not enough to provide adequate information about the subscribers' mobility customs. Thus, only those SIMs are considered active enough that had activity for at least 20 days, the average weekday activity is at least 40 records, and at least 20 on weekends. Additionally, the average activity of the SIM cannot be more than 1000 to filter out those SIM cards that possibly did not operate in a cell phone, but in a 3G modem, for example.

Note that activity can be either voice call, text message, or data transfer, also both incoming or outgoing. Hence, they cannot be distinguished in the dataset.

## 4.8. Calculating Indicators

As mentioned before, the data was stored in a PostgreSQL database, but the query engine of PostgreSQL was also utilized during the computation. Python is excellent for data science, as an interactive shell (also called read–eval–print loop, or REPL) provides a convenient environment for data analysis. However, the execution speed is not among the strong points of the interpreted languages. Considering that most of CDR processing happens at the subscriber level, and the subscribers are

During my research, the Scikit-mobility [132] Python package has been published that is capable of calculating well-known mobility indicators, such as Radius of Gyration or Entropy. However, by the time this was published, my own implementation was complete, tested, and well-optimized (within reasons), so I kept using my own software stack.

independent of each other, it seemed the best solution to partition the processing by subscribers. Then, the activity records of a single subscriber can be processed in a single thread even with a relatively slow interpreted language.

The *Pool* object, from the *multiprocessing* package, provides a convenient tool to parallelize the execution. A SIM ID is associated with a worker that executes a query to filter its activity records. Naturally, the query can contain spatial or temporal constraints as well. The database engine can efficiently select the activity records of a SIM card, utilizing the available many-core environment [2]. Then, the indicators were calculated per SIM card, and the partial results were collected and saved. The analysis and the visualization did not require considerable computing power and thus were often performed on a laptop.

[2] Lenovo x3650 M5 server, with 1024 GB of RAM, and two 18-core CPUs with HT, resulting in 72 logical cores. The nickname of this server was "naggykuttya" ("biggdogg").

# Social Sensing 5.

Social sensing is an application of the mobile network data (or social media) that utilizes humans as sensor carriers via wearable devices (like smartphones or smartwatches) in order to observe the situational awareness of the physical world [133]. I like to use the term, more specifically, when a social event or phenomenon is observed by the people's attendance. I reviewed applications, falling into this category, from the literature in Section 2.3 and 2.4. This chapter presents two case studies about the participants of large social events.

With my colleagues, we presented a conference paper [134*] about the fireworks display, the main event of the state foundation day, on August 20, 2014. It also comes under the category of social sensing.

## 5.1. Euro 2016

Football is one of the most popular sports worldwide, and European or World Championships, especially the finals, are among the most-watched sporting events. The Euro 2016 Final was watched by more than 20 million people in France [135], or the Germany vs. France semifinal was watched by almost 30 million people in Germany [135]. But what about Hungary?

According to the MTVA (Media Services and Support Trust Fund), which operates the television channel M4 Sport, the first Hungarian match was watched by about 1.734 million people, the second by about 1.976 million, and the third group match by about 2.318 million people*. With these ratings, the M4 Sport, turned out to be the most-watched television channel in Hungary during those days [137]. The whole participation of the Hungarian national football team was beyond expectations and raised interest, even among those who generally do not follow football matches. Nevertheless, is it possible to measure/correlate this interest with a mobile phone network?

### 5.1.1. Austria vs. Hungary

The first match against Austria (Figure 5.1) started at 18:00 on Tuesday, June 14, 2016. Before the match, the activity level was significantly higher than the weekday average and later decreased until the half-time. The activity level dropped to the average during the second half, which indicated that more people started to follow the match. Right after the Hungarian goals, two significant peaks occurred in the activity, indicating increased attention and the massive usage of mobile devices during the match.

As the data source cannot distinguish the mobile phone activities by type, it cannot be examined what kind of activities caused the peaks. The activity was supposed to be mostly data transfer or text messages, not phone calls. It simply does not seem to be lifelike to call someone during

* According to the Hungarian Central Statistical Office (KSH), the population of Hungary was about 9.83 million in 2016 [136].

the match just because of a goal, but sending a line of text via one of the popular instant messaging services seems feasible.

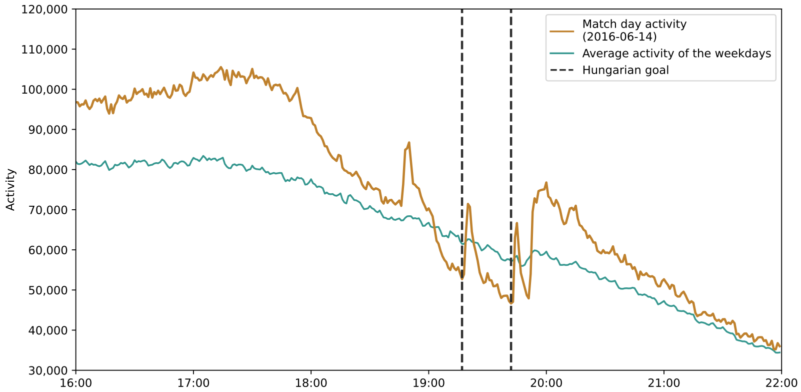


**Figure 5.1.:** Mobile phone activity during, and after the Austria–Hungary Euro 2016 match, in comparison with the average of the previous and the following two days.

### 5.1.2. Iceland vs. Hungary

The match against Iceland was played on Saturday, June 18, 2016. Figure 5.2 shows the mobile phone activity levels before, during and after the match. As the weekend activity is generally lower (see Figure 3.5), the average of the weekdays was used as a reference. The match began at 18:00, and from that point, the activity level was significantly below the average, except for the half-time break and the peak after the Hungarian goal. Interestingly, the Icelandic goal did not result in such a significant peak. Only a very moderate one can be seen in the time series.

Traag et al. [20] also found activity drop during a game, but in that case, the area of the stadium was analyzed, where the match was played, and there was no peak during the match.

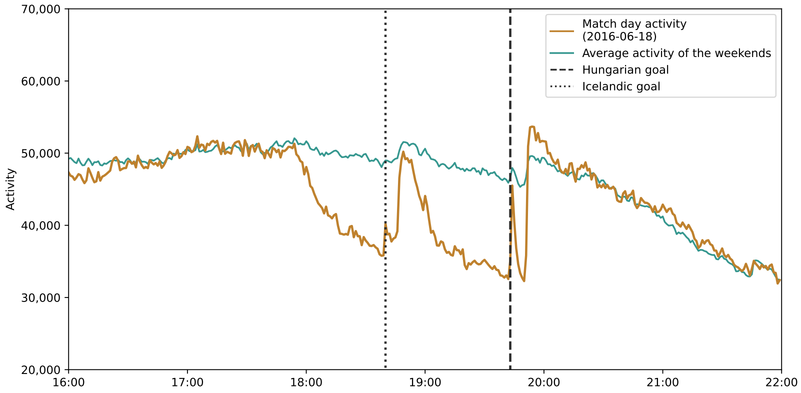


**Figure 5.2.:** Mobile phone activity during and after the Iceland–Hungary Euro 2016 match, in comparison with the average activity of the weekends.

### 5.1.3. Hungary vs. Portugal

On Wednesday, June 22, 2016, in the third match of the group stage of the 2016 UEFA European Football Championship, Hungary played a draw with Portugal. Both teams scored three goals, and with this result, Hungary won their group and qualified for the knockout phase. During the match, the mobile phone activity dropped below the average, but the goals against Portugal resulted in significant peaks, especially the first one (see Figure 5.3). On the other hand, the Portuguese equalizer goal(s) did not cause a significant mark in the activity. In the second half, the teams scored four goals in a relatively short time period, but only the Hungarian goals resulted in peaks. This observation suggests that the football fans had a notable influence on the mobile network traffic.

After the match, the activity level was over the average, which might represent the spontaneous festival in downtown Budapest. According to the MTI (Hungarian news agency), thousands of people celebrated in the streets, starting from the fan zones, mainly from Erzsébet square (Figure 5.5 a), Margaret Island (Figure 5.5 b) and Erzsébet square (Figure 5.5 c) direction Budapest Nyugati railway station. The celebrating crowd completely occupied the Grand Boulevard, and the public transportation was disrupted along the affected lines.

This social event is comparable to mass protests from a mobile phone network perspective. In Section 5.2.1 I have analyzed the mobile phone activity at the route of a mass protest. The activity of the cells was significantly high when the protesters passed through the cell. In this case, however, the affected area was smaller, and the sites along the Grand Boulevard were very busy at the same time after the game.

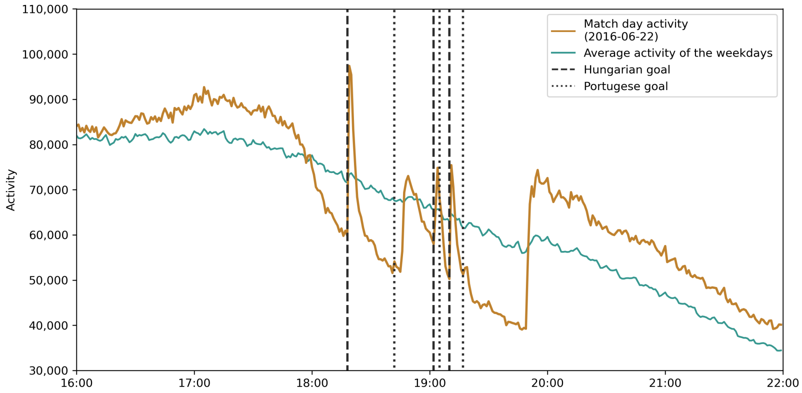


**Figure 5.3.:** Mobile phone activity during and after the Hungary–Portugal Euro 2016 match, in comparison with the average of the previous and the following two days.

The activities of the sites (multiple cells aggregated by the base stations) in Budapest downtown are illustrated in Figure 5.4. The highlighted site covers mostly Szabadság square (for the location, see Figure 5.5 a), where one of the main fan zones was set up with a big screen. The activity curve follows the trends of the whole data set (see Figure 5.3). There was high activity before the match, during half-time, and after the match for a short period. During the match, the activity decreased except for four not-so-significant peaks around the goals.

In the highlighted site, in Figure 5.4, almost 7 thousand SIM cards had been detected between 17:00 and 20:00. The data shows that 53.57% of the subscribers were between 20 and 50 years old, while 33.49% had no age data.

After the match, there was a significant increase in the activity in some other sites. These sites were (mostly) around the Grand Boulevard, where the fans marched and celebrated the advancement of the national football team to the knockout phase.

Figure 5.5 shows the spatial distribution of this social event using Voronoi polygons generated around the base stations locations. The polygons

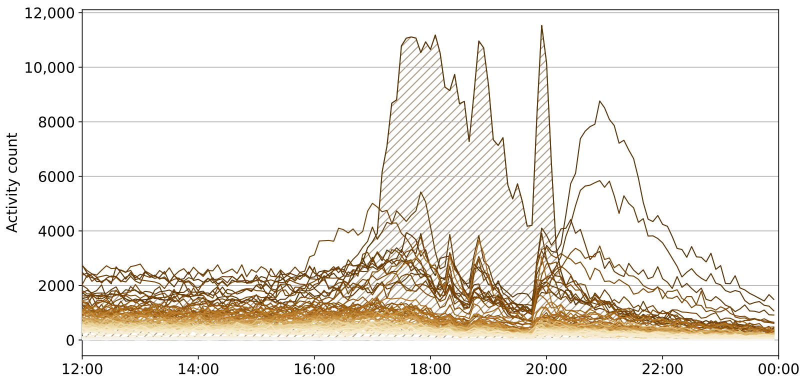


**Figure 5.4.:** Cell activities, in Budapest downtown, on the day of the Hungary vs. Portugal football match (June 22, 2016). The highlighted cell covers mostly the Szabadság square, where one of the main fan zone was set up.

are colored by the mobile phone network activity increase at 20:15, compared to the weekday activity average. For the comparison, the standard score [1] was determined for every base station with a 5-minute temporal aggregation. The darker colors indicate the higher activity surplus in an area. The figure also denotes the three main fan zones in the area, the routes of the fans by arrows, and the affected streets by thicker lines.

### 5.1.4. Hungary vs. Belgium

On Sunday, June 26, 2016, Hungary played the fourth and last Euro 2016 match against Belgium. Figure 5.6 shows the mobile phone network activity before, during, and after the match. During the match, the activity level was below the weekend average. The activity after the match was slightly higher than the average since the match ended late on Sunday when the activity average is usually low. This activity surplus may only indicate that the fans were leaving the fan zones and going home.

### 5.1.5. Homecoming

The Hungarian national football team returned to Budapest on June 27, 2016. A welcome event at the Heroes' Square has been held, where the football fans can greet the national football team. According to the M4 Sport television channel, approximately 20 thousand people attended the event [138]. Between 18:00 and 19:30, there were 4246 unique, non-phone SIM cards active on the site that covers the Heroes' Square. 3425 are known to use smartphones based on the operating system column of the GSMArena data set.

The cells of this base station cover a larger area, so not all of these subscribers attended the event, but on the other hand, it is not compulsory to use mobile phones during this event. Supposing that the mobile phone operator preferences among the attendees corresponded to the nationwide trends in 2016, there could even be about 17 thousand people, as the data provider had 25.3% market share [109].

Figure 5.7b shows a part of District 6 and the City Park with the Heroes' Square and the Voronoi polygons of the area are colored according to the Z-score values to indicate the mobile phone activity in the area at 18:35. The activity is considered low below −1, average between −1 and 1, high

[1] The standard score (or z-score) is defined as $z = \frac{x-\mu}{\sigma}$, where $\mu$ is the mean and $\sigma$ is the standard deviation.

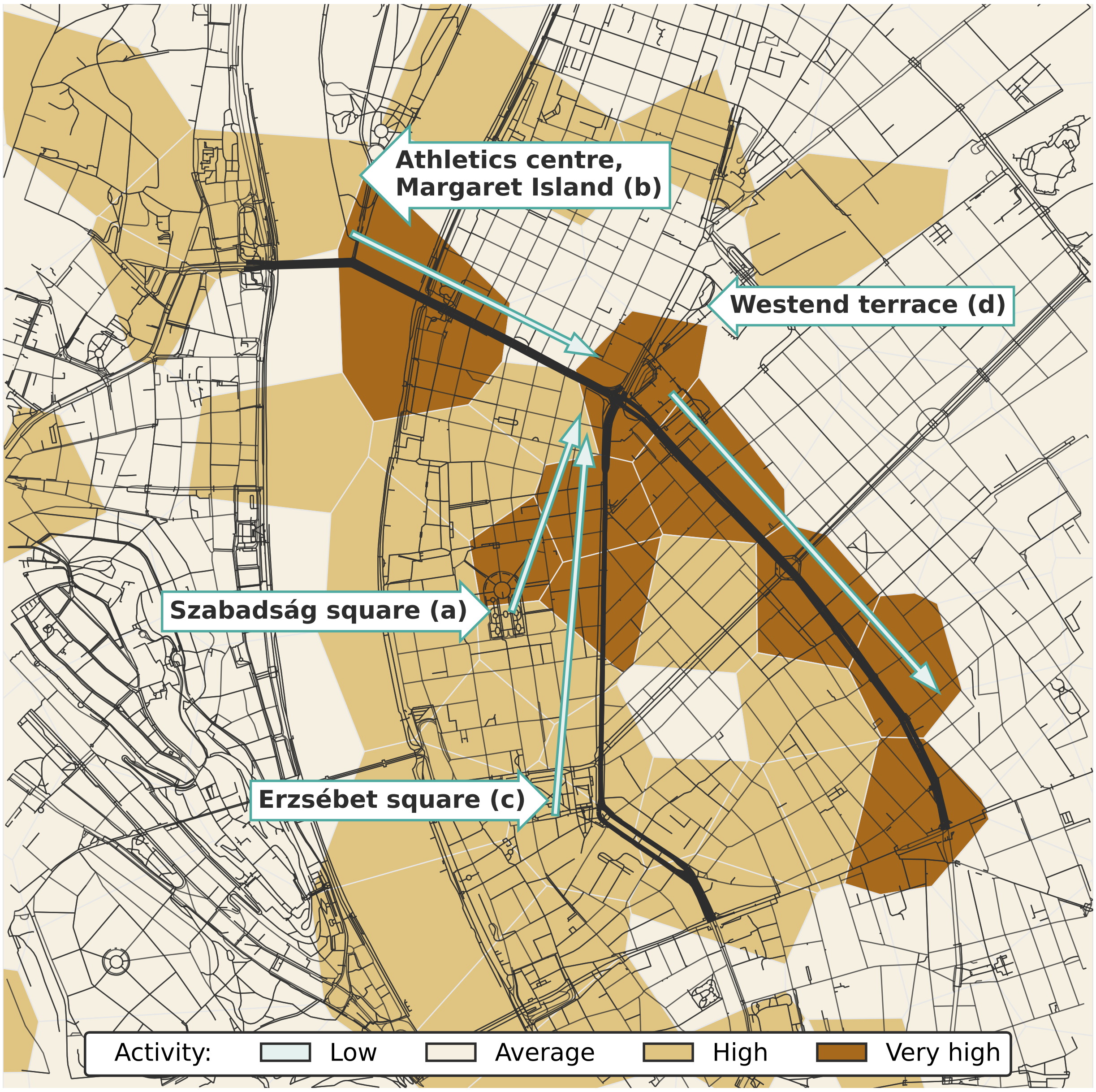


**Figure 5.5.:** After the Hungary vs. Portugal football match, the fans, delirious with joy, filled the streets. The arrows show their route from the main fanzones to along with the Grand Boulevard. Voronoi polygons colored by the mobile phone network activity at 20:20.

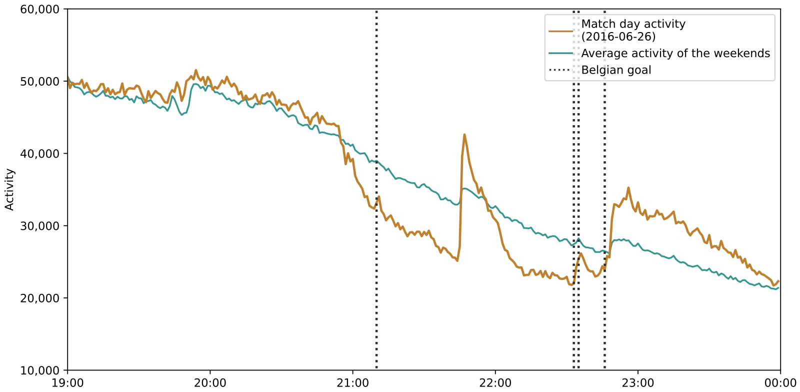
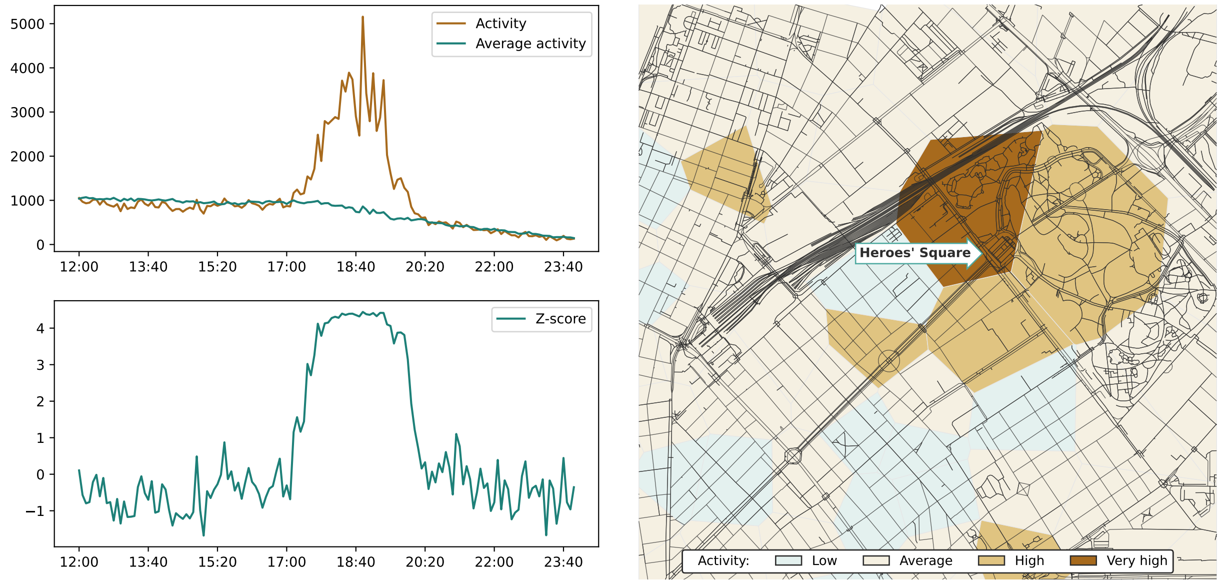


**Figure 5.6.:** Mobile phone activity during and after the Hungary–Belgium Euro 2016 match, in comparison with the average of the previous and the following two days.



**(a)** Activity and Z-score of the site, at Heroes' square.

**(b)** Spatial view of the activity at 18:35.

**Figure 5.7.:** Mobile phone network activity at Heroes' Square and its neighborhood, during the welcoming event of the Hungarian national football team.

between 1 and 2.5 and very high above 2.5. Figure 5.7a shows the mobile phone network activity (upper) and the Z-score (bottom) of the site, covering Heroes' Square. It is clear that the activity is significantly higher than the weekday average during the event, and the Z-score values also follow that.

## 5.2. Those Who Stood with CEU

The CDRs of this event have been utilized in the conference paper [139*] and its improved version [140*].

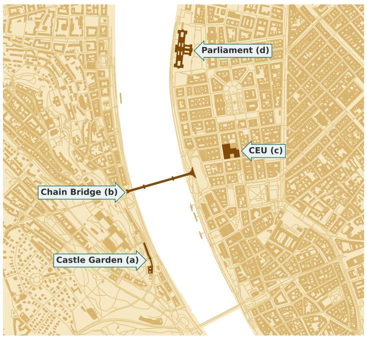


**Figure 5.8.:** The main locations of the demonstration, that started at the Castle Garden and ended at the Parliament.

On April 4, 2017, the Hungarian Parliament passed amendments to the Higher Education Law of Hungary that restricts the operation of international universities, such as CEU. Many people considered it a threat to academic freedom, so on April 9, 2017, people took to the streets to protest for the Central European University and against the modification of the Hungarian higher education law. This case study focuses on the movement of the people who participated in a demonstration against the modification of the Hungarian higher education law on April 9, 2017. It was reported [141] that tens of thousands of people attended the event, which started at the Castle Garden (Figure 5.8a) at 17:00. Then, the demonstrators headed north along the riverbank and went through the Széchenyi Chain Bridge (Figure 5.8b), past the campus of the Central European University (Figure 5.8c) to the Hungarian Parliament Building (Figure 5.8d). At 19:00, the demonstration officially ended in the Kossuth Lajos square (at the Hungarian Parliament Building), but it took a while for the crowd to disperse.

The cells covering the path of the demonstration are selected (see Figure 4.7b), and the devices that were present in these cells at the time of the demonstration are considered participants. The query resulted

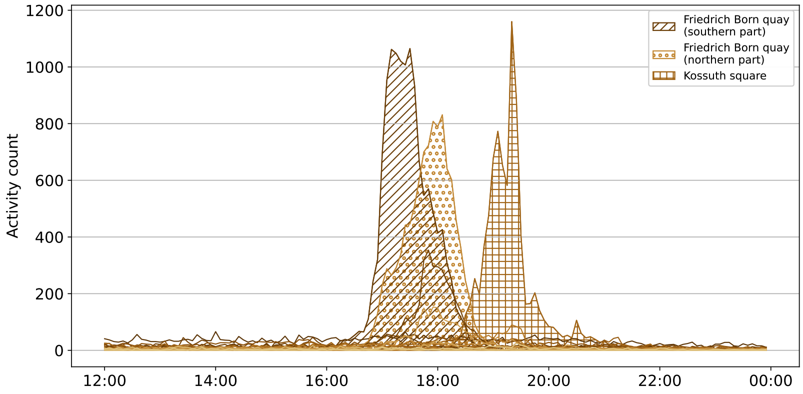


**Figure 5.9.:** Cell activities in the path of the CEU demonstration. Friedrich Born quay is located between the Elisabeth Bridge and the Chain Bridge. Kossuth Square is at the parliament.

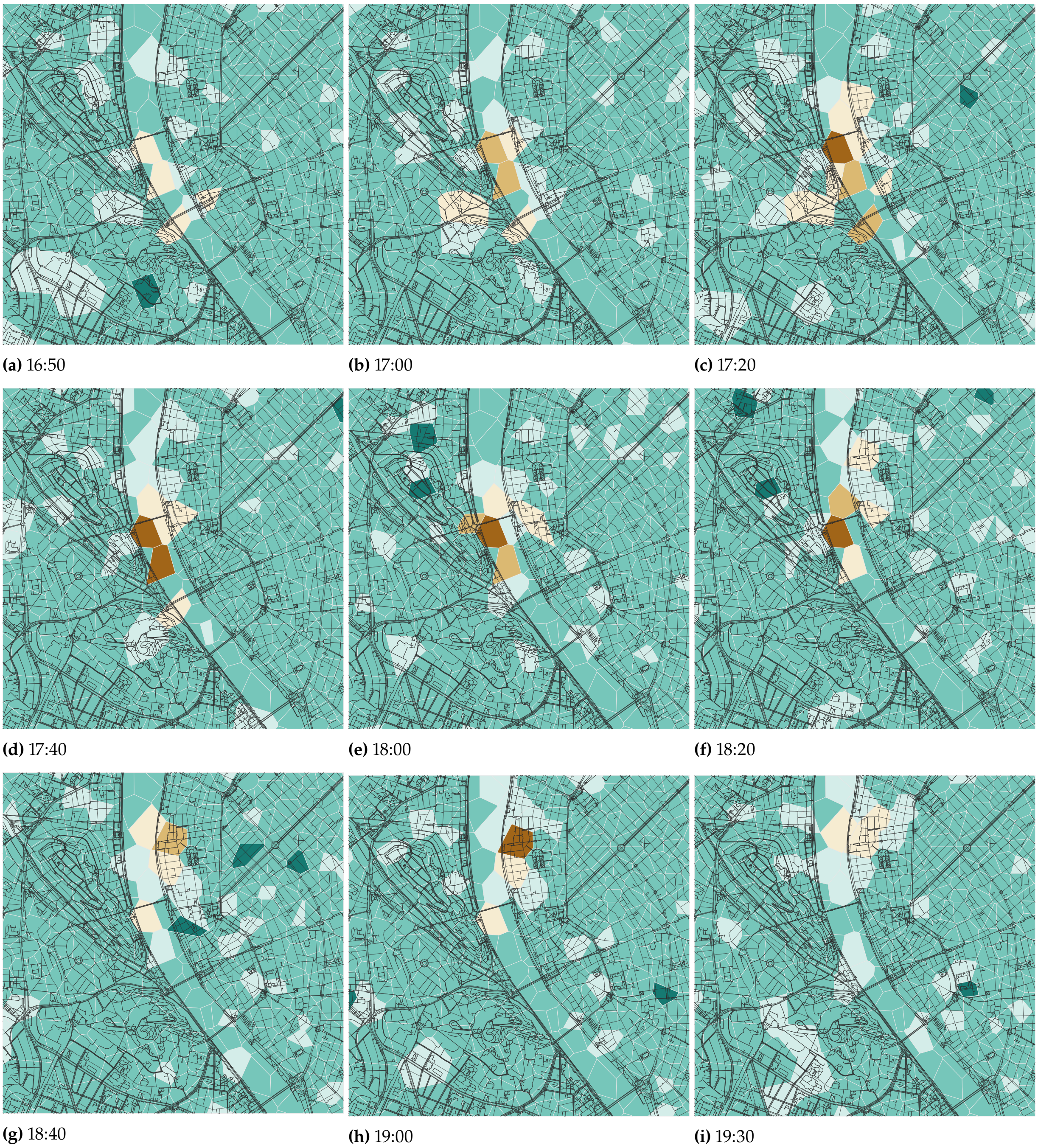
in more than 10,000 unique devices. Note that using the phone from some of these cells does not mean that the user is a protester. Also, it is not guaranteed that every protester used their phones during the demonstration. Supposing that the choice of cell phone operator among the demonstrators corresponds with the nationwide trends, then the total number of the demonstrators could be more than fourfold this number, as the operator has about 25.5% market share in 2017 Q2 [109].

Figure 5.9 shows the activity counts – aggregated by 5 minutes – of the cells that cover the path of the demonstration. Figure 5.10 displays the activity in the Voronoi zones during the demonstration. The activity along the Friedrich Born quay and at the Castle Garden was above average even before the official start of the demonstration. Then, the high activity areas highlight the path of the event. Note that the people only passed by the CEU building, so many people did not assemble there.

According to the reports [142], the first protesters were already at the parliament when the end of the march was still at Buda, and it took one and a half hours for everyone to arrive at the Kossuth Square. The mobile network data can confirm that.

### 5.2.1. In Social Media

The motto “I stand with CEU” was used, also in social media to express solidarity and support with the CEU. Figure 5.11 displays the distribution of the tweets, with the hashtag “#IstandwithCEU”, using different colors for genuine tweets and retweets. There was a notable peak during the protest, while the retweets showed some delay. On the other hand, the retweets still circulated the following day, while not too many new tweets were posted.



**(a)** 16:50

**(b)** 17:00

**(c)** 17:20

**(d)** 17:40

**(e)** 18:00

**(f)** 18:20

**(g)** 18:40

**(h)** 19:00

**(i)** 19:30

**Figure 5.10.:** Mobile network activity, during the mass protest.

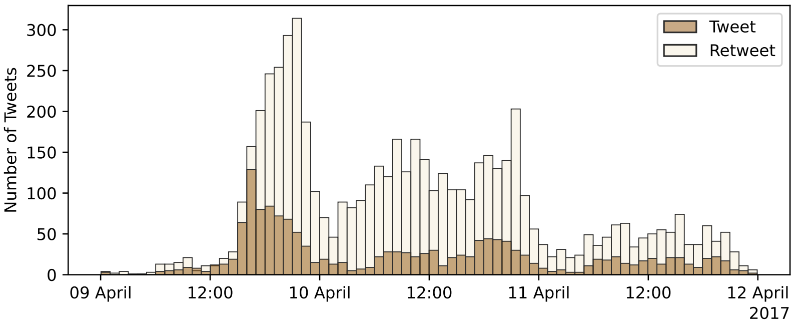


**Figure 5.11.:** The distribution of tweets with the hashtag "IstandwithCEU", on the day of the protest and the following two days.

# Commuting | 6.

Hungary is a typical capital-oriented country. Budapest is the political, economic, logistical, and cultural center of the country, where almost 18% of the population lived [119]. In 2017, the population of Budapest was 1,749,734, and the population of Pest county was 1,247,372. In the agglomeration of Budapest, 837,532 people lived according to KSH [143]. The river Danube divides the city into the Buda (Boo-da) and the Pest (Pesh-t) side. The former is the supposed more fashionable area, and the property values are remarkably higher (see Figure 3.14 and Figure 8.1a).

Due to its central role, Budapest attracts a workforce from a relatively large area. This process makes contact between the capital and the surrounding settlements, called commuting. According to Kiss and Matyusz [13], commuting is the relation between two locations. The inhabitants of one source location travel to work to another location; this is called "out-commuting". The target location receives a workforce that is called "in-commuting". The loss of the source location is: (i) the out-commuter does not use the local resources, (ii) does not create value, (iii) loosens their relation, as it is partly relocated to the target location, and (iv) although it brings back income, that is partly spent elsewhere. On the other hand, the target location (i) gains human resources, (ii) can create more value in place, (iii) the local relations and society become stronger, and (v) the local consumption increases.

Pálóczi analyzed the country-wide commuting in Hungary, using the methods of complex network analysis, based on the data of census 2011 [144]. He considered settlements as nodes, and commuters as a directed edge between the nodes, then applied the disparity ($Y_2$) parameter [145], which was developed to measure the heterogeneity of weighted relations. If $Y_2$ is close to one, it means that one destination dominates the commuting.

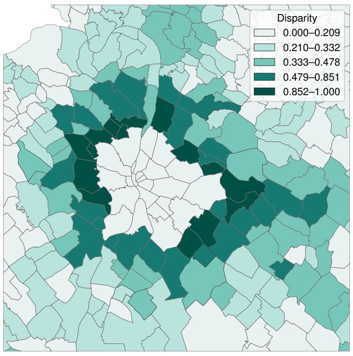


**Figure 6.1.:** Disparity of out-commuting, based on census 2011.
Partial replot of [144, Figure 3].

Figure 6.1 displays a replotted part of [144, Figure 3] to illustrate Pálóczi's results in respect of the area discussed in this chapter. As Figure 6.1 shows, settlements next to Budapest are darker, meaning that the connections of these settlements are one-sided. Pálóczi demonstrated that the out-commuting dependency is not the greatest around Budapest, but around Győr and Székesfehérvár [144], but also high at the centers of the employment regions (e.g., chief towns of the counties).

Kiss and Matyusz state that [1], although commuting is an important and common phenomenon, its measurement is occasional and inadequate [13]. Commuting is predominantly analyzed by the census data, but that is performed only once in a decade and thus cannot follow sudden but permanent changes. They also stress that commuting should be examined continuously, and its methodology should be established.

As questioning the population is a slow, tedious and expensive task, it would be obvious to automate the process with the available info-communication technologies (ICT). In this chapter, the application of CDR processing is proposed to examine commuting, mainly to Budapest,

[1] In respect of Hungary, at least.

and the findings are validated by the results of studies that analyzed commuting using census data. In many cases, the findings are presented in a form as close to those results as possible to aid the comparison.

## 6.1. Home and Work Locations

To analyze commuting, the subscribers' home and work locations have to be determined. As the CDRs are anonymized, do not contain information like a residential address. The workplace is even not mandatory for a subscription, the operators cannot have that information. The process of home and work location estimation is described in Section 4.6. After these locations have been determined, they should be validated. Although the applied approach is practically equivalent to what can be found in the literature, the validity of the results is hard to confirm. Pappalardo managed to validate the home location in the case of sixty-five subscribers [16], but this is not possible in this case. So, the settlement and — in the case of Budapest — district based population data [143] is applied from the KSH.

Figure 6.2, shows a comparison of the population between the KSH and the CDR data. When comparing the two maps, there are a few differences: Some parts of Budapest are not dark enough on the CDR map. Districts 16 and 17 seem not as populated as in the KSH data. Nevertheless, the divergent districts are on the Buda side: Districts 2 and 3. The difference may have ensued from the inaccuracy of the home detection in that area or simply from the different preferences in mobile operators of the inhabitants of the Buda Hills.

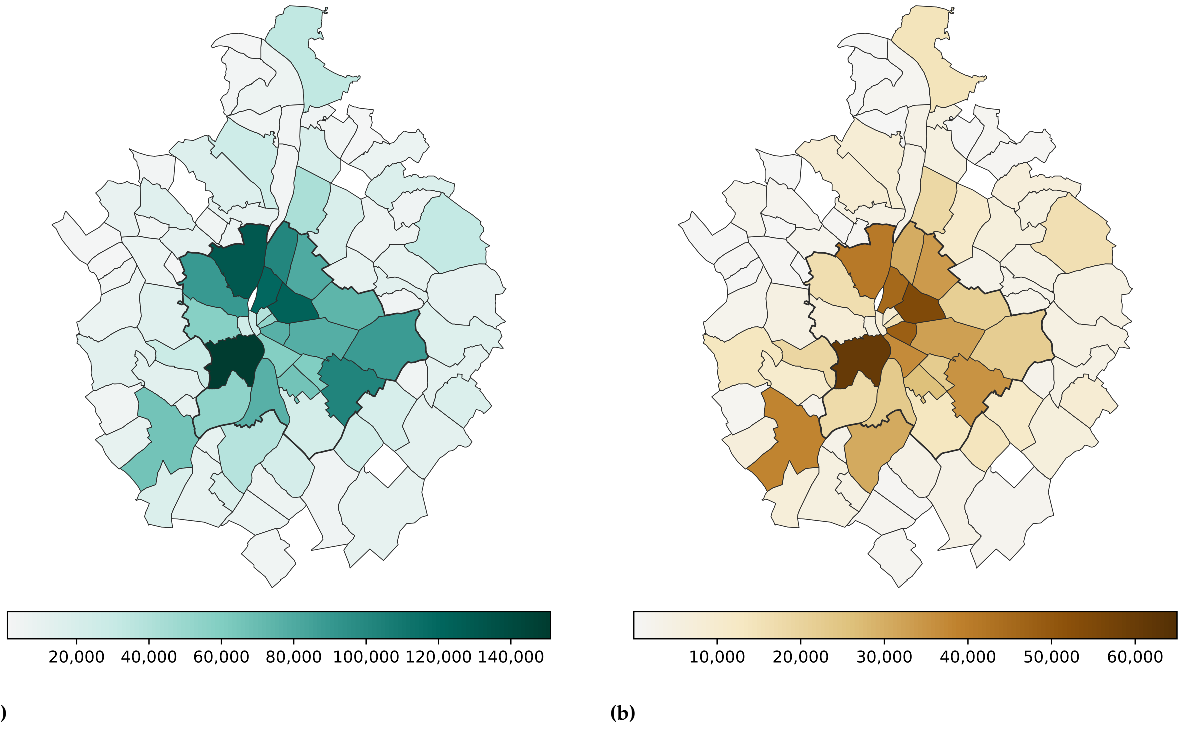
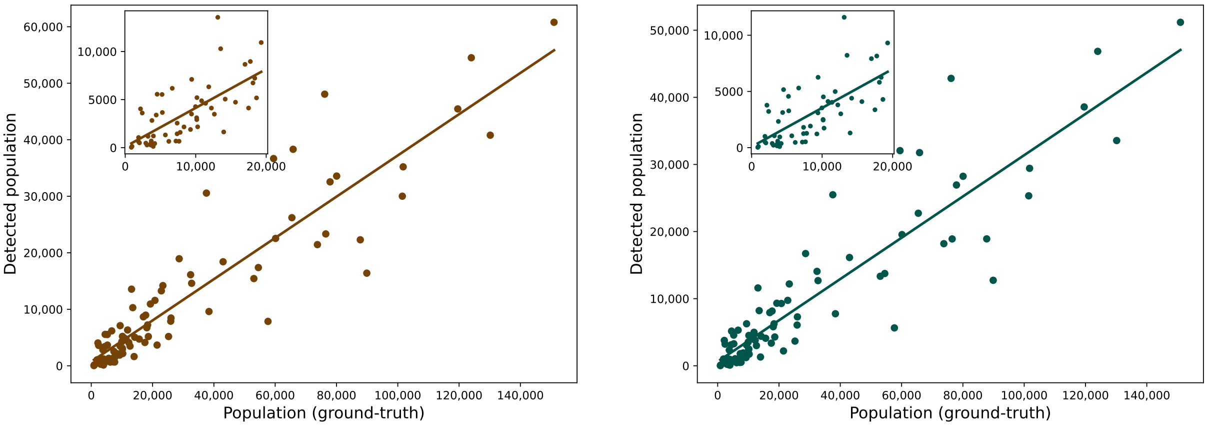


**Figure 6.2.:** Comparing the ground truth [143] (**a**) and the estimated population based on mobile network data (**b**).

Apart from this, the findings based on CDRs correlate with the statistical data; the Pearson's R is 0.9213 ($R^2$ is 0.8488), counting every SIM cards. Figure 6.3a, shows this as a regression plot. As described in Section 3.3, the obtained Call Detail Records contain information (TAC) about the devices that use the mobile network. Based on this information, more than 300,000 SIM cards have been identified that do operate other types of devices like cellphones (e.g., 3G modem), indicating that they do not represent people. Figure 6.3b, illustrates the correlation without these SIM cards. Although the population values are decreased, the correlation does not seem significantly affected: Pearson's R is 0.9125.



**(a)**

**(b)**

**Figure 6.3.:** Correlation between the population of the agglomeration and the districts of Budapest based on KSH and mobile network data. In left figure (**a**), all the SIM card were used, in the right figure (**v**), SIM card that certainly operate in non-phone devices are excluded.

Numerically, the CDR data often shows significant mismatches, but they are not easy to objectively compare. The available mobile network data originated only from one operator, which had about 25% market share in the observation period [109]. This market share is about the subscriptions, not the number of unique people. Furthermore, it also has to be noted that this ratio represents a nationwide value. As spatially more detailed market share is not available, it has to be supposed that Vodafone Hungary had the same market share in every subregion to make this comparison. Although this is unlikely, one-fourth of the population values can be used as a rule of thumb.

### 6.1.1. Work Locations

Along with the home locations, the workplaces are the most important element of the mobility and commuting analysis. During the COVID-19 pandemic, this has changed. As part of the social distancing directive, to slow down the spread of the disease, working from home has come to the fore. Presumably, the prevalence of home-office will be higher than before the pandemic, as both the employers and the employees got used to this situation, but many scopes of activities will still require a work location, so the importance of this topic will remain the same. However, as the data sources used in this work predates the pandemic, this question could only be answered in another work.

The workplaces have been determined, defined as the most frequent place where a subscriber appears during work hours. See Section 4.6 for the details. For example, querying the work locations of the inhabitants of an area, a settlement can be the initial step of the commuting analysis. Figure 6.5 shows the typical workplaces of three selected settlements and one district of Budapest, using Gaussian kernel density plots, in two different versions: with (left column) and without (right column) those subscribers who work in their home settlement. When the local workers are included, the darkest areas are the selected area itself, as many people work in the vicinity of their homes.

### 6.1.2. Connectivity

As the origin and the destination of the commuting are determined, it is possible to build a network, for example, considering the districts of Budapest nodes that are connected by the commuters. Figure 6.4 shows the connections between the districts of Budapest. The Buda districts are placed to the left, whereas Pest districts are to the right, and the colors of the nodes match with the district groups in Figure 3.15a. The edges represent commuters between districts, removing self-links, and the weight of the edges denotes the number of commuters. The weight is expressed by colors, using darker colors for the stronger edges. The weakest links ($w < 250$) are omitted to improve visibility.

Extending this topic to the level of the agglomeration, or the country, could be another research direction: for example, to analyze the in-commuting and out-commuting. Pálóczi's work [144] could serve as a census-based reference.

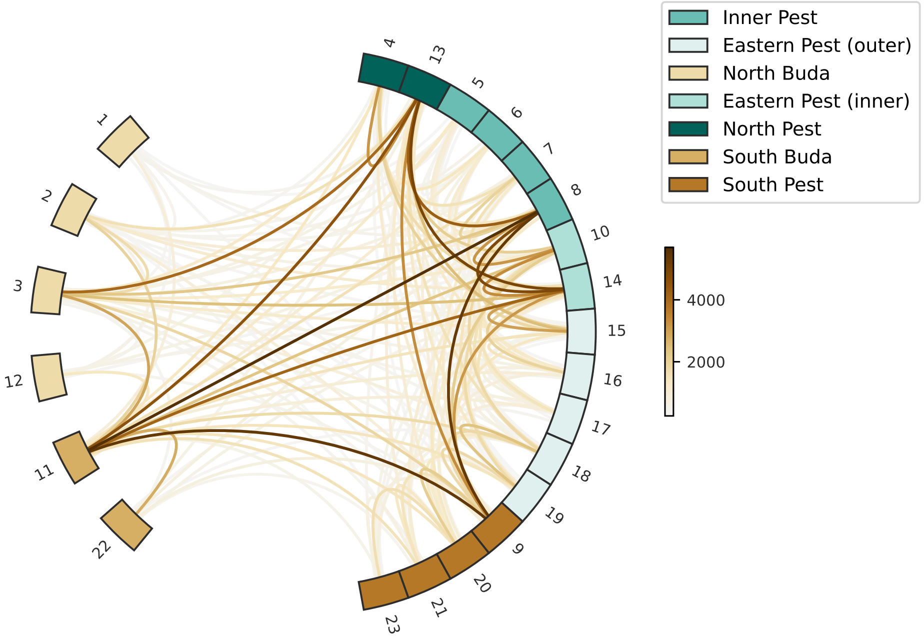
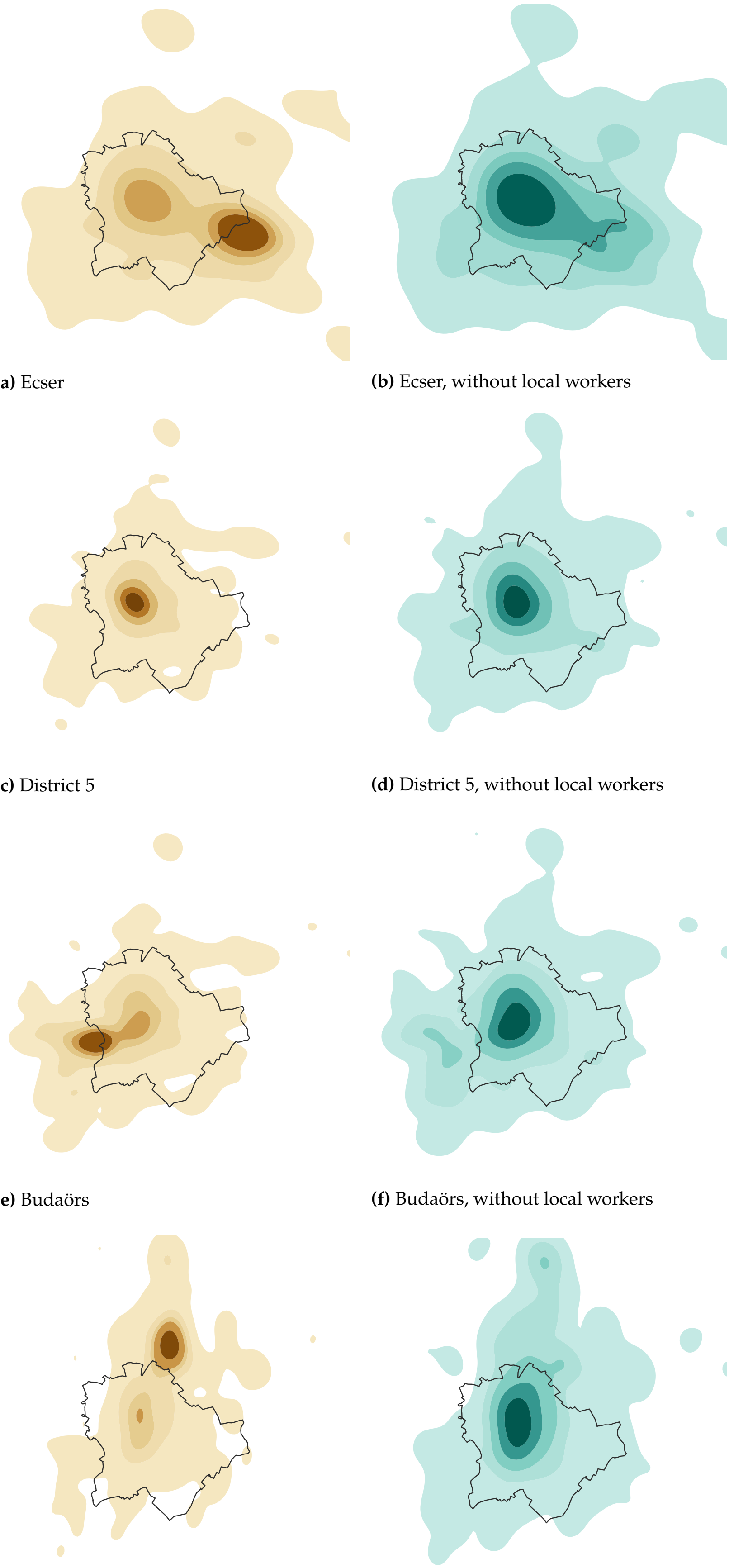


**Figure 6.4.:** Connectivity between Budapest districts, using the home and work locations. A link originates from the home district to the work district, excluding the ones who where lives. The weak links are omitted to improve visibility. The district nodes are colored by the district groups.



**(a)** Ecser

**(b)** Ecser, without local workers

**(c)** District 5

**(d)** District 5, without local workers

**(e)** Budaörs

**(f)** Budaörs, without local workers

**(g)** Dunakeszi

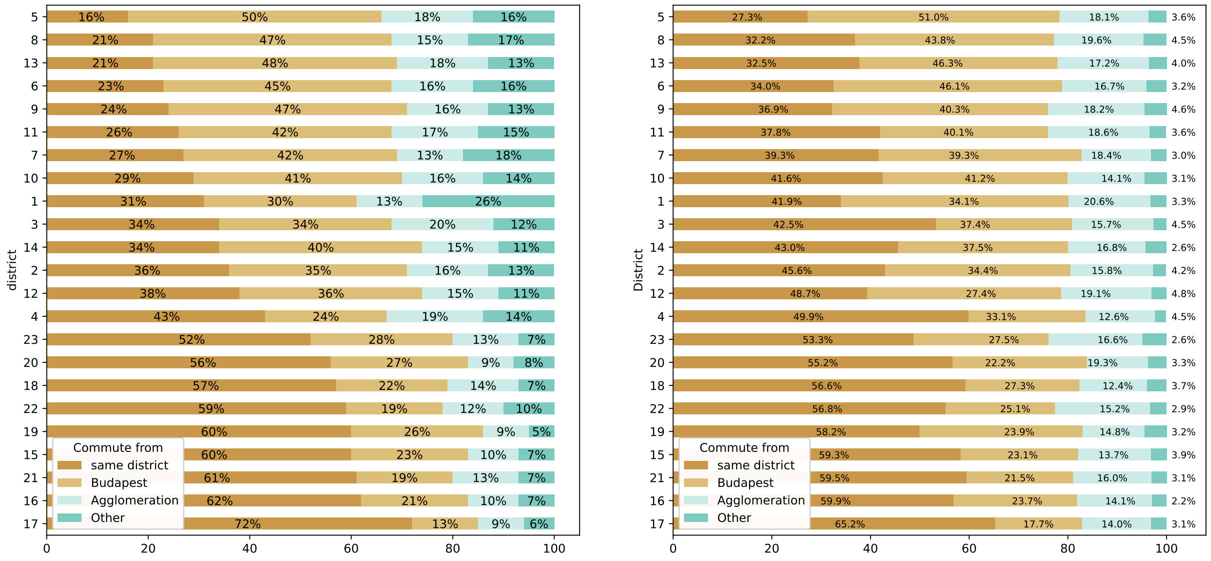
**(h)** Dunakeszi, without local workers

**Figure 6.5.:** Using kernel density plots (with Gaussian kernel) to display the typical working locations for three selected settlements and a district of Budapest, with (**a**, **c**, **e**, **g**) and without (**b**, **d**, **f**, **h**) local workers.

## 6.2. Validation by Census

In order to verify the reliability and accuracy of the method proposed for the home and work location estimation, a comparative study is performed on the mobile network data and the information processed from the census. In Hungary, a census is obtained every ten years and a micro-census with a 10% corpus at halftime. The last census was performed in 2011, while the last micro-census was in 2016. [2] Based on these surveys, commuting to Budapest (and generally in Hungary) is analyzed in studies like [13, 144, 146, 147]. These studies are used as the reference for comparing the results.

Figure 6.6 shows the comparison between the CDR and the census-based [147, Figure 1] traveling ratios of the commuters by the districts of Budapest and the home location category. People who work in Budapest are represented, and the home location can be (i) the same district where one works, (ii) another district of Budapest, (iii) the agglomeration, and (iv) other settlements outside the agglomeration.



**(a)** **(b)**

**Figure 6.6.:** Comparison between the census based (**a**) [147, Figure 1] and the CDR (**b**) commuting ratios to the districts of Budapest, from the same district, other parts of Budapest, the agglomeration or out of the agglomeration.

Good agreement mostly within Budapest has been found on the proportions of the commuters. The most significant difference can be seen with the "outside agglomeration" category. This deviation, however, originated from the content of the data source, as the mobile network data used in this study covers mainly the area of Budapest and its close vicinity. It also contains phone activities from the surrounding county, but by moving away from Budapest, the available data decreases.

The fraction of workers who have their homes in the same district is very close to the census data in the outer districts (15–23) but generally overestimated in the core districts (1, 5–9) and the inner districts (2–4, 10–14). The workers from other district groups show the best match to the census data (where the CDR should have the best quality), while the agglomeration is somewhat overestimated in many districts.

[2] The next census should have been performed in 2021, but it was postponed due to the COVID-19 pandemic.

## 6.3. The Agglomeration

In [146], there is a more detailed analysis regarding the commuting from the agglomeration that is divided into six sectors, and the commuting was examined by origin (home sector, occasionally by towns) and destination (district group of Budapest).

Figure 6.7 shows the commuters' distribution in the districts of Budapest, from the six sectors of the agglomeration, based on the CDR evaluation. In representation, Figures 6.7a–6.7f are analogous to [146, Figures 2–9], and show to which districts the inhabitants commute from the given sector of the agglomeration.

Lakatos and Kapitány analyzed the commuting tendencies of some settlements to the districts of Budapest between censuses: 1990, 2001, and 2011 [146]. The same analysis has been made using CDR processing, and six settlements of the 13 thoroughly analyzed of [146] are presented in this study. The results are summarized in Figure 6.8, compared with the censuses. It contains a settlement from every sector of the agglomeration, so it also serves as a more focused analysis of Figure 6.7. The location of the settlements, in relation to Budapest, is also displayed on small maps to give context to the findings. For example, from towns west to the capital, the most common commuting targets are the Buda-side and the inner districts, for example. Moreover, in many cases, the mobile network based findings, which are six years older than the last census, indicate a clear continuation of the previous tendency.

In the case of Budaörs (Figure 6.8a), albeit North Pest, outer Eastern Pest, and South Pest are not significant commuting destinations, census data show an increasing tendency, which is confirmed by the mobile network data. The CDR based results of South Buda and Inner Pest also fit the trend, but in an opposite tendency. The most considerable discrepancy lies in the cases of North Buda and the inner Eastern Pest district groups. The Pearson correlation coefficient, regarding all the six district groups, between the census 2011 and the mobile network data is 0.8976.

Dunakeszi (Figure 6.8b) is east of the River Danube and north of Budapest, which implies the dominance of North Pest as the commuting destination, although its importance has been decreasing over the last few decades, as well as Inner Pest. While South Buda, South Pest, and the outer Eastern Pest have an increasing tendency, the inner Eastern Pest and North Buda do not show such clear tendencies. The correlation coefficient (Pearson's R) between the census (2011) and the CDR based results is 0.9416.

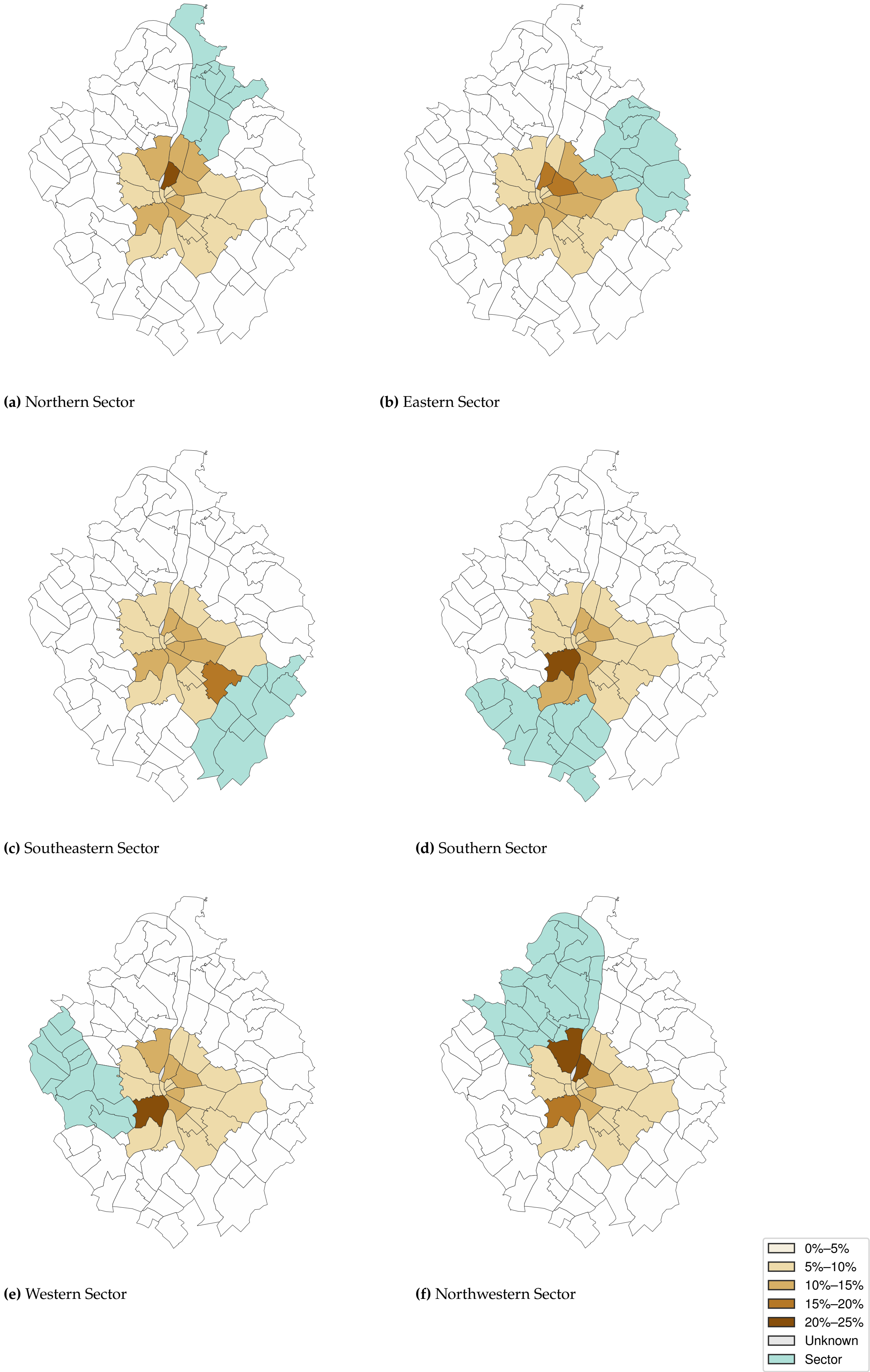
Vecsés is in the southeastern sector of the agglomeration, from where the majority of the commuters work in the inner and outer Eastern Pest, Inner Pest, and South Pest regions. North Pest and Buda was not a notable destination for the commuters, but the results show increasing trends (Figure 6.8c). The correlation coefficient, in the case of Vecsés, is 0.924.

Dunaharaszti is in the Southern sector of the agglomeration and east of the Danube. Consequently, the main destination of the commuters was South Pest. Besides that, Inner Pest received considerable in-commuters, but its importance seems to be decreasing. The rest of the district groups had roughly the same trends (Figure 6.8d). Dunaharaszti has the strongest correlation out of the examined settlements: Pearson's R is 0.971.

Érd has the largest population (65,857 in 2017 [143]) in the agglomeration and also in the Southern sector. The detected commuting ratios fit into the trends of the last three censuses, although Eastern and South Pest seem overestimated, and North Buda underestimated by the mobile network data based approach (Figure 6.8e). The correlation with the ground-truth is 0.8488 (Pearson's R).

In the case of Szentendre (Figure 6.8f), the mobile network based results might show the most significant discrepancy. Still, the correlation coefficient (Pearson's R) is 0.9127. As located in the Northwestern sector, west of the Danube, the most obvious destination for commuting is North Buda. According to census data, it had the most in-commuters, even with a slightly increasing tendency. However, the CDR based results underestimate it, whereas Eastern and South Pest seem overestimated. The result of Inner Pest lags behind the census the latest census data, but that fits into the trend.

These detailed results demonstrate the applicability of the CDR processing for commuting analysis. It would be interesting to compare these results with the next census. That would reveal how precisely these findings fit into the trend of the changing commuting customs of the population of the agglomeration.



**(a)** Northern Sector

**(b)** Eastern Sector

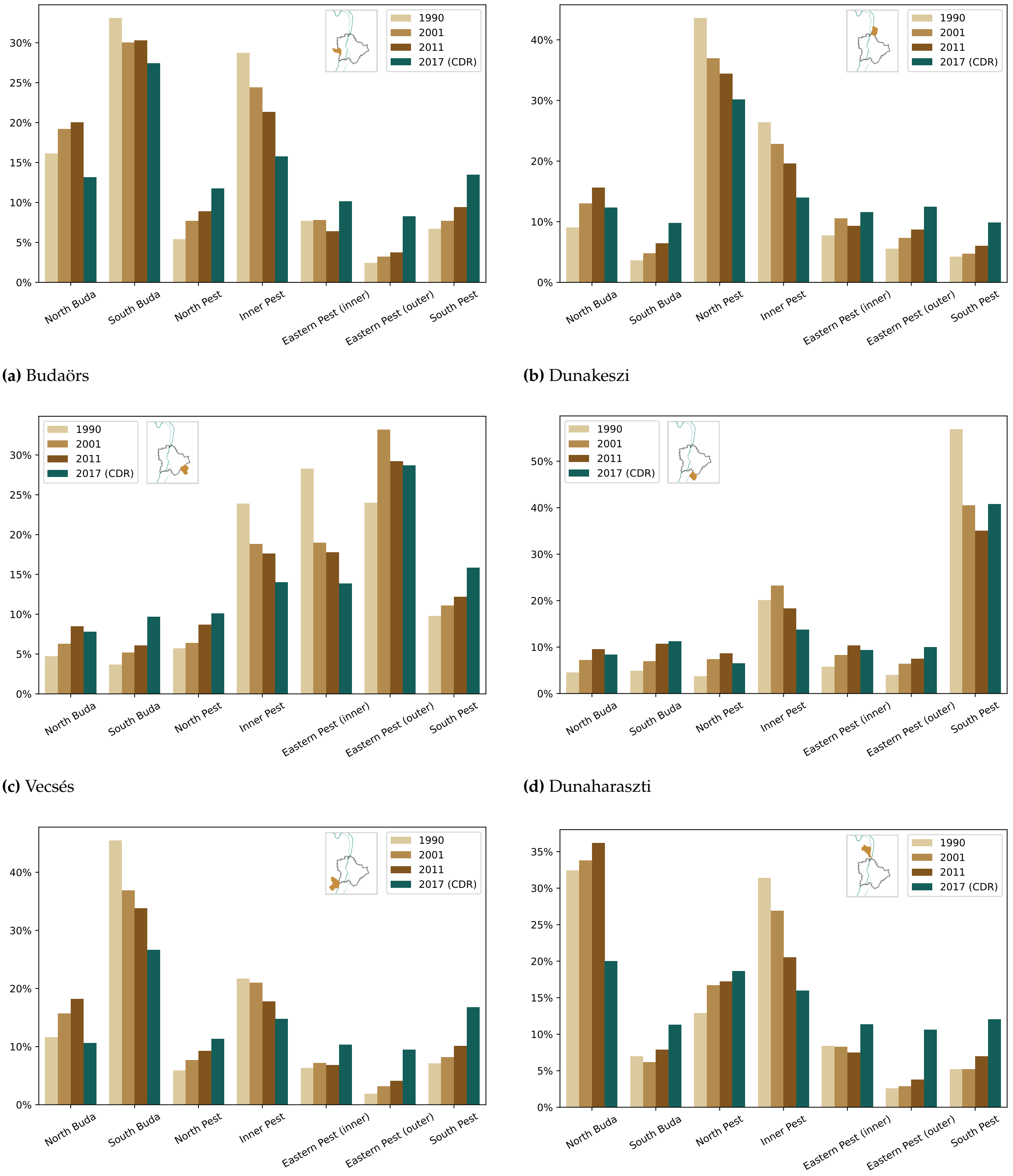
**(c)** Southeastern Sector

**(d)** Southern Sector

**(e)** Western Sector

**(f)** Northwestern Sector

**Figure 6.7.:** Commuting from the six sectors of the agglomeration, based on CDR evaluation.



**(a)** Budaörs

**(b)** Dunakeszi

**(c)** Vecsés

**(d)** Dunaharaszti

**(e)** Érd

**(f)** Szentendre

**Figure 6.8.:** Commuting to the seven districts groups of Budapest from selected settlements of the agglomeration, comparing census (1990, 2001 and 2011) and mobile network data. Next to the legends, the location of the settlements in question is displayed in a map.

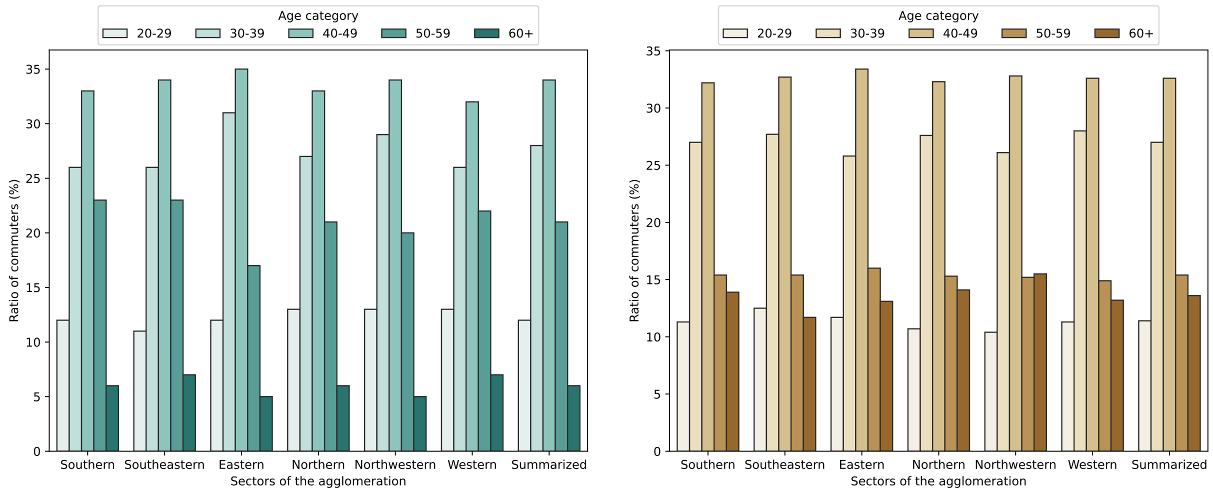
## 6.4. Demography

As the available mobile network data contains information about the age and the gender of the subscribers — in the case of the 66.17% and 70.76% of the subscriptions, respectively — the commuting trends can be studied by age-groups.

Koltai and Varró provide reference data for this analysis [147, Table 1]. Figure 6.9a shows the distribution of the commuters by age categories and the sector as the home location. Only those commuters were examined who work in Budapest.

It is not clear from the paper what is the upper limit of the "60+" age category. The people who usually go to work are assumed to be younger than 65 years old (the current retirement age in Hungary), although people can work over 65. In the CDR based figure (Figure 6.9b), the 60+ means over 60 and less than 100. However, there are not many subscribers over 70, only 2.48% of SIM card owned by people older than 70 years. Furthermore, it has to be noted that the age information belongs to the owner of the subscription, not necessarily to the actual user of the phone.

Comparing data obtained by the micro-census and the cellular information, good agreements (Pearson's R is 0.8977) have been found on the trends and measures of the distribution of the commuters by age categories. The most significant difference between the census and the CDR based data are within the "60+" and the "50–59" categories. The number of people in their fifties seems underrepresented by the CDR data, while the "60+" category is overrepresented, that might be caused by the different interpretation of the upper limit. On the other hand, the values are very similar in the other categories. Based on the similarity of the results (Figure 6.9), it is confirmed that mobile network data can be a reliable method for commuting analysis even regarding the demographic features.



**(a)**

**(b)**

**Figure 6.9.:** Distribution of the commuters by age categories and the sectors of the agglomeration (%). Comparison between micro-census data [147, Table 1] (**a**), and mobile network data (**b**).

## 6.5. Describing Mobility

Up to this section, the mobility is described by only the home and work locations, and the commuting between these two locations was in focus. However, mobility is a much broader concept. It should contain all the subscribers' activities, not only the work-related ones. There are some widely used indicators in the literature (Section 2.2) to characterize the mobility.

These indicators — namely Activity Location Number, Radius of Gyration, and Entropy — have been calculated (Section 4.8) for every subscriber in the data sets. It is important to note that the home-work (beeline) distance can also be a numeric indicator of commuting, but as the Radius of Gyration takes into consideration every location where a subscriber appeared, it can give more insight about the mobility.

The distribution of the distance between the home and work locations shows exponential characteristics Figure 6.12c. The majority of the population investigated have chosen workplaces closer than 5 km, and a relatively small number of people need to travel more than 10 km to work.

The radius of gyration (Figure 6.12a) shows normal distribution. However, it is significantly skewed to the right (seems to be similar to the observations performed in Boston [10]). The majority of the inhabitants have a relatively small Radius of Gyration; on the other hand, some subscribers need to travel significant distances [87].

The Entropy also shows the normal distribution. The majority of Budapest's population visits plenty of locations (Entropy between 0.2 and 0.6), and besides a solid fraction of people who remain in their home's neighborhood (Entropy less than 0.2), there are only a few who visit many places in the city. This is in agreement with the findings of other studies [5, 10, 87].

Figure 6.10 and 6.11 show the average Radius of Gyration of the individuals whose home location is estimated to be in the given area. The former shows the Budapest districts and the settlements of the agglomeration, and the latter focuses on Budapest at a cell level, using Voronoi polygons as representation. The broad tendencies are the same: The farther one lives from the center, the more one travels, but the cell level map reveals some local city centers that make the image more nuanced. The impact of the local city centers on mobility trends could be a separate topic of investigation.

As the activity time series shows (Figure 3.3a), the mobile network data has a daily periodicity. Besides, there are also fewer activities on holidays. The quantity is one thing, but the quality is another. People do not just use less the mobile network on holidays, but their behavior is quite different. As for one, most of them do not need to go to work, so the usual commuting-related mobility patterns are not applied. This appears in the daily aggregated mobility indicators (Figure 6.13). People are active at a fewer number of locations (Figure 6.13c), which might be caused that they stay at home to rest, which is also supported by the entropy results (Figure 6.13b).

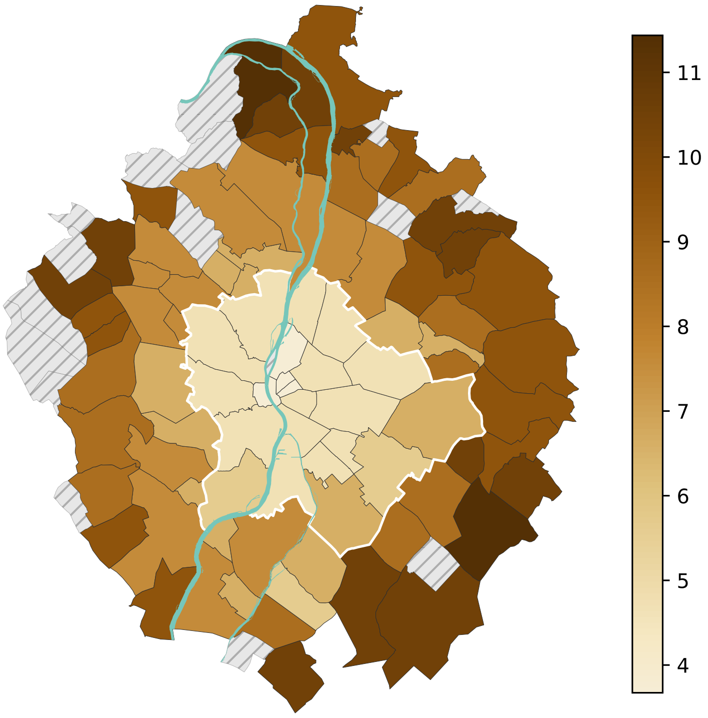


**Figure 6.10.:** Budapest districts and the settlements of the agglomeration colored by the mean Radius of Gyration (km) for workdays. The tendency is that the farther someone lives from the city center, the more one travels. The white border denotes the administrative border of Budapest, and the Danube is displayed in green. Settlement without sufficient data is represented by gray.

The activities of the people during working days are significantly higher than on the weekends. In addition to, higher values of Gyration Radius, Entropy, and the number of active locations have been recorded on Fridays, which implies that the last working day of the week has a sort of privileged role (Figure 6.13). On the weekends, more activities and more vibrant travels can be observed on Saturdays, and many of the dwellers rest on Sundays.

## 6.6. Limitations

The evaluation and the validation have been performed based on the results of other studies that analyzed commuting based on census data. With direct access to the statistical data from KSH and other sources, more and finer aspects of the validation could be performed.

## 6.7. Conclusion

In this chapter, the evaluation of the subscribers' home and work locations are presented, and the results were compared to the ground truth. Though the detected population numerically differs from the actual population, the distribution across the settlements shows a strong correlation. It can be explained by the fact that the CDRs were obtained from only one mobile network operator.

Based on the home and workplace detection, it was demonstrated that mobile network data could be an effective solution for commuting analysis.

The findings are presented in a form as close to the results of other studies that examined commuting in the agglomeration of Budapest as possible to aid the comparison.

It was examined to which district people commute from the sectors of the agglomeration. In the case of some selected settlements, the destination districts of the commuters are also presented in contrast to the last three censuses. It was found that mobile network based results fit into the three-decade tendencies. The commuters were also analyzed by age groups, which also agree with the census-based studies.

These results confirm that mobile network data is capable of commuting analysis. Using activity records from all the operators of a country, a more precise and representative analysis could be performed. Given the fact that mobile networks are available in the most populated areas, mobile network data should be standardized for statistical and sociological usage while respecting privacy and personal data.

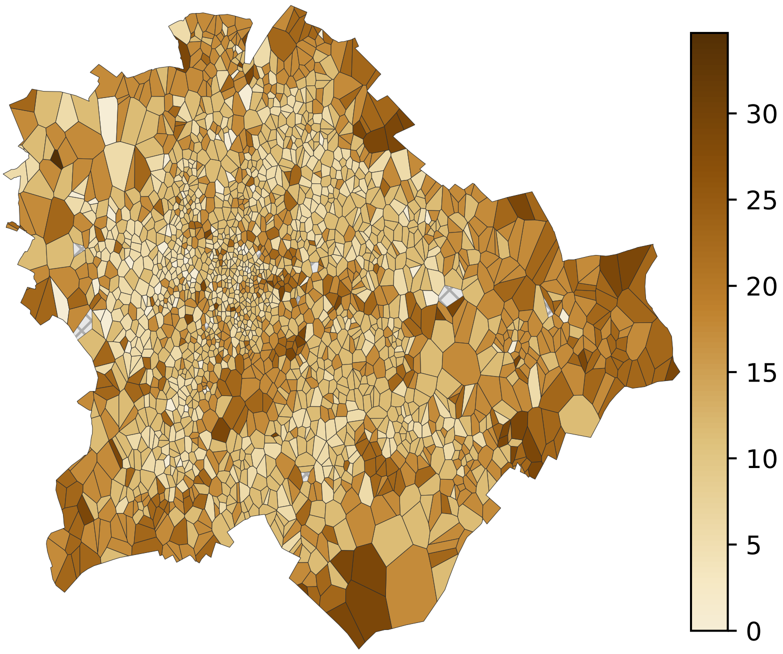
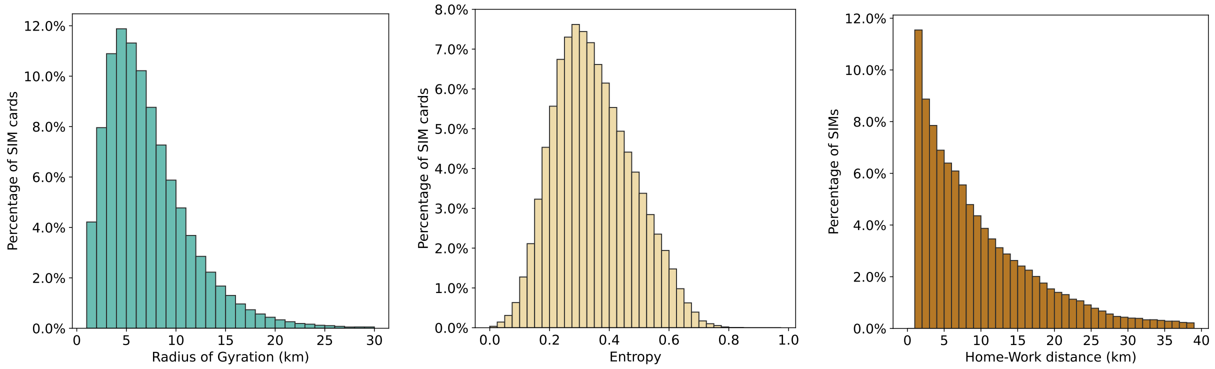
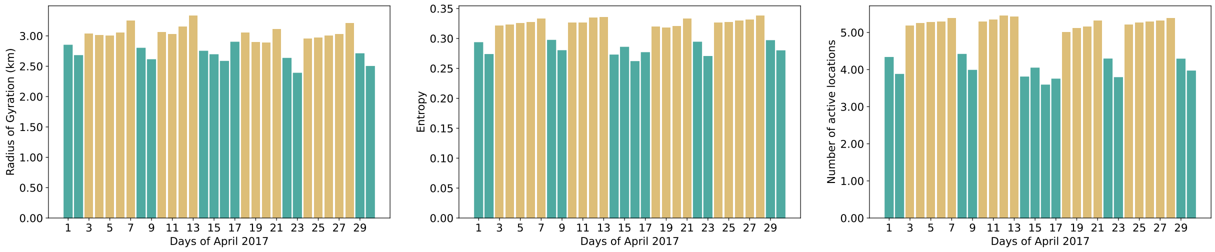


**Figure 6.11.:** Cell voronoi polygons colored by Radius of Gyration (km), dark green polygons represent cell without enough data.



**(a)** Radius of Gyration distribution **(b)** Entropy distribution **(c)** Home-work distance distribution

**Figure 6.12.:** Histogram of the Radius of Gyration (a), Entropy (b) and the Home-Work location distance (c)



**(a)** Radius of Gyration **(b)** Entropy **(c)** Number of Activity Locations

**Figure 6.13.:** The differences between workdays (light brown) and holidays (green) are clear. Orange columns represent the holidays, April 14, 2017, was Good Friday and April, 17 2017, was Easter Monday that are holidays in Hungary.

# Awakening City 7.

A mobile phone activity precisely follows the subscribers' daily routines. It is enough to see the daily periodicity in Figure 3.3a. As the daytime activity is so clearly distinguished from the nighttime activity, the question has arisen: "When does the day exactly begin?". And a set of follow-up questions along with it: When do the inhabitants awaken? Are there any differences between the given type of subscribers? When does the city awaken? Are there any differences between the given areas of the city? Furthermore, is it possible to cluster city areas by the time when the activity of the inhabitants, the workers, or the passers-by starts their activity in the morning or halts in the evening? Do city parts have "chronotypes"? Can neighborhoods or districts be described by the terms "morningness" or "eveningness"?

This chapter aims to answer these questions.

## 7.1. Wake-up Time

The Call Detail Records were aggregated for every cell and 10-minute time intervals. Then, the moving average was applied with the window of 12. To determine the wake-up time, the positive edge of the curve needs to be detected when the number of mobile phone activity increases drastically in a short period of time. The wake-up time is considered when the activity value reaches the arithmetic mean of the minimum and the maximum value on the positive edge of the activity curve. The minimum of the aggregated records is usually in the middle of the night, and the maximum is in the afternoon. Figure 7.1, illustrates the concept.

The content of this chapter has been published in [116*], and the early concept appeared in [148*].

This process was repeated in the case of every day, then the median of the daily wake-up times was determined. Analogously to the wake-up time, the bedtime can be calculated to describe when the mobile phone activity decreases significantly by selecting the mean value on the negative edge. Note that these values are naturally not the actual times when people wake up or fall asleep. Those moments cannot be determined using only the mobile phone network. Using the screen-on events of the phone

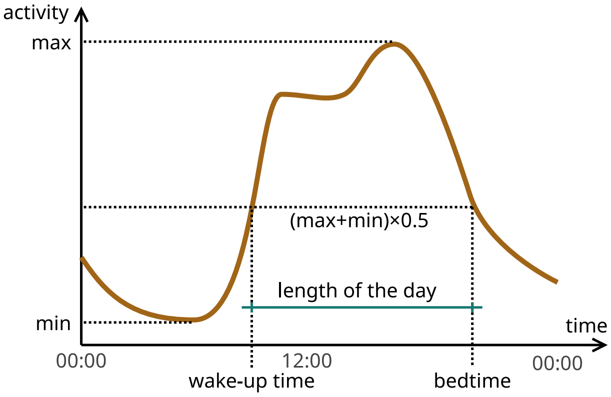


**Figure 7.1.:** Calculation of the wake-up and the bedtimes

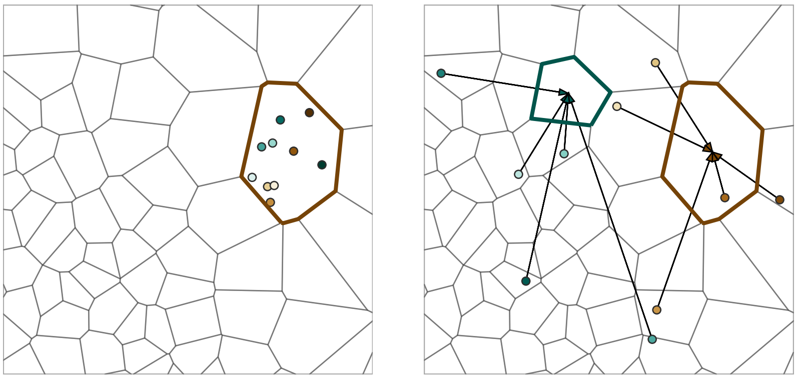
[90] can be much closer to the actual values. Especially in the case of the wake-up time, if the phone is used as an alarm clock. In spite of this, it is supposed that this approximating method can reveal the rough tendencies of the daily routine. Still, the terms "wake-up time" and "bedtime" are used to refer to the time of the positive/rising and negative/falling edges of the daily activity curve, respectively.

## 7.2. Aggregation of the Subscribers

As the type of phone activity is unknown, the wake-up time of the individual is hard to estimate. Since making a call requires to be awake (whereas message could be received and data could be transferred autonomously), knowing which activities represent phone calls would provide more accurate information about when people are certainly awake. Instead of this, the significant activity increase is used to identify the wake-up time.

The activity of a single device is also sporadic and does not provide enough data to identify an activity increase for all the days. Because of the sporadic nature, the devices were grouped, which could be performed in two ways: calculating the wake-up time of an area (a cell or a cell group) or the inhabitants of an area.

In the first approach, the activity records were aggregated that took place in a given cell, regardless of which SIM cards produce them. In the case of the second version, those activity records were used, which were produced by the inhabitants of the given cell, regardless of where the activity took place. The first approach could be called cell-based grouping, and the latter inhabitant-based grouping. The two approaches are illustrated in Figure 7.2. Cells can be grouped further to examine a larger area (e.g., residential, suburb, district).

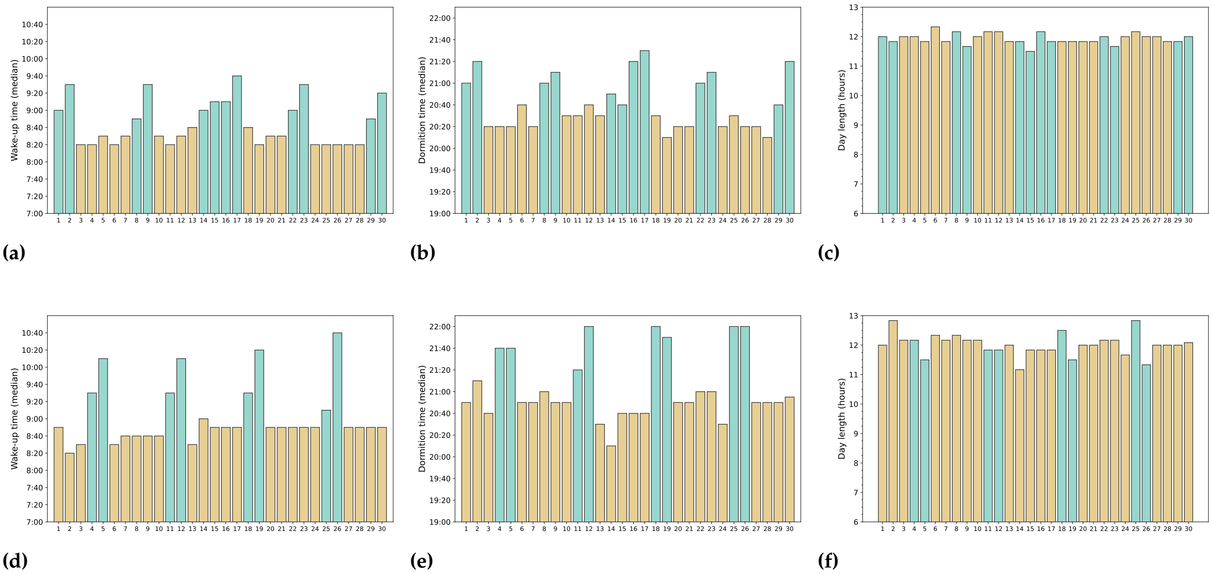


**(a)** **(b)**

**Figure 7.2.:** Visualizing the difference between the cell-based (**a**) and the inhabitant-based (**b**) approaches. The former considers subscribers present in a given cell wherever they live. The latter aggregates activity of the inhabitants to the home cell, regardless where the activity occurred.

## 7.3. Inhabitant-based Approach

Based on the results of the "April 2017" data set (Figure 7.3a and 7.3b), the wake-up time almost always around 7:10 on workdays. On Sundays (or Easter Monday in the case of the long holiday), it is later by about



(a) (b) (c)

(d) (e) (f)

**Figure 7.3.:** Daily wake-up times (**a**,**d**), bedtimes (**b**,**e**), and the day lengths (**c**,**e**), for the "April 2017" (**a**–**c**) and the "June 2016" (**e**–**e**) dataset. Holidays are represented by a different color.

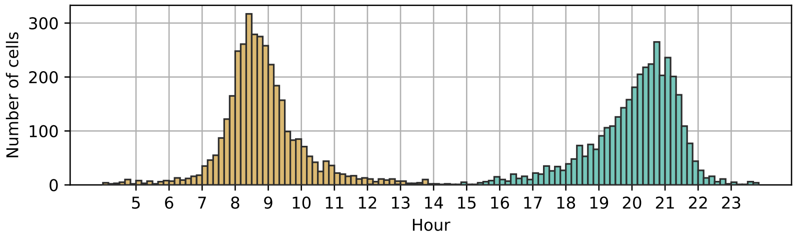


**Figure 7.4.:** Inhabitant-based wake-up (brown) and bedtime (green) values of the cells, April 2017.

one hour. On Saturdays (and the other days of the holiday), it is later by about 30–40 minutes compared to an average workday.

The bedtimes are not so even, but also have a nice trend. The activity decreases between 19:40 and 20:10 on workdays, but the holiday values shifted by 30–50 minutes, similarly to the wake-up times. Interestingly, as bedtimes follow the wake-up times on the weekends, it results that the average day length remains approximately the same. Figure 7.3c shows the day lengths, calculated as a difference between the bedtime and the wake-up time, in minutes. The day lengths are between 12 hours 30 minutes and 13 hours 10 minutes. On average, the day length is 12 hours 45 minutes (the standard deviation is 10 minutes).

The same evaluation was performed on the "June 2016" data set. Figure 7.3d and 7.3e, show the daily wake-up times and bedtimes, respectively. Basically, the same tendencies can be observed: wake-up times and bedtimes are shifted on holidays. However, there are some irregularities, especially within the bedtimes. On June 22, the bedtime is almost on the same level as a weekend. The mobile phone activity decrease occurred 30 minutes later than on the other days of that week.

Figure 7.4 displays the inhabitant-based wake-up and bedtimes of the cells. This result is comparable to [90, Figure 9], however in that case, the wake-up times ($t_{wake}$) were earlier and bedtimes were later ($t_{sleep}$). Note that the difference ensues from the nature of the approach. In [90], the screen-on events of the phones are considered, while in Figure 7.4, the active usage of the mobile phone network is aggregated for a larger area. The “SensibleSleep” application [90] observes offline (from a mobile phone network perspective) smartphone activity that adjusts better to the actual SWC of an individual.

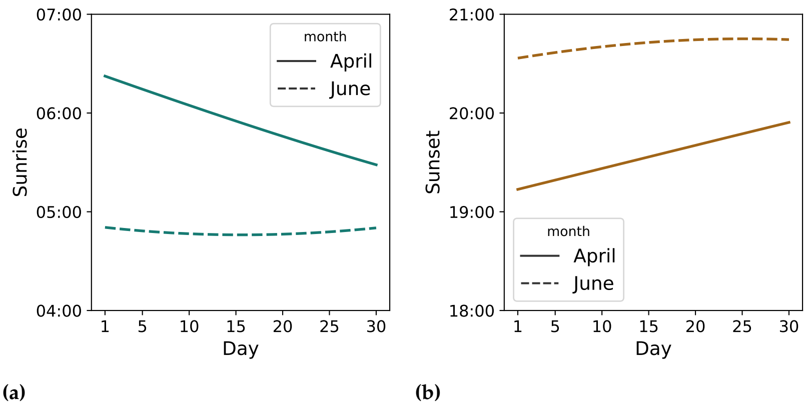
## 7.4. The Length of the Day

Figure 7.3a and 7.3b show that people start and end their days later on the holidays. This makes perfect sense since they do not need to go to work, they do not need to spend time traveling to get somewhere in time, and they probably like to rest more. But, the bedtime is shifted as well, which results that the length of the day remaining practically the same, see Figure 7.3c.

Although the two data sets (Section 3.2 and 3.1) were not recorded in the same year, they still cover two different months of the year (April and June). Considering that a mid-spring month is well represented by the “April 2017” dataset and an early summer month by the “June 2016” dataset (for Budapest), the differences between the wake-up time, bedtime, and the day length can be compared between the two seasons.

Figure 7.3d shows a small increase in the wake-up times during the second half of the month. Although that 10-minute increase in the averages cannot be considered significant, it may reflect the end of the school term (June 15). During the intersession, the time schedule of the public transport services is adjusted. For many lines, the headway is increased on workdays, while for some lines, it is decreased, especially on weekends. In this way, the intersession affects the whole transportation system of Budapest and even its agglomeration to some extent.

Until the summer solstice (June 20 in 2016 [149]), the days are getting longer, but can the longer daylight have an effect on mobile phone network activity? As a reference, astronomical information (sunrise and sunset) has obtained from `www.visualcrossing.com` [121] for Budapest. Figure 7.5 shows the difference between the sunrises and the sunsets during the two observation periods of the data sets, projected to the same figure. The dashed lines display the values of June, and the solid lines display the values of April. As the summer solstice is in June, the differences



(a)

(b)

**Figure 7.5.:** Sunrise (**a**) and sunset (**b**) times of June 2016 and April 2017 projected to the same figure to highlight the differences between the seasons, using data from [121].

during the month are negligible, but in April, the differences between the beginning and the end of the month are much more significant.

So, in June, the Sun rises earlier and sets later than in April. Consequently, lighter periods of days are longer. June 15 is more than 2-hour longer than April 15, using the astronomical definitions of the sunrise and sunset. According to the calculated day lengths, the average workday length is 12 hours in the "June 2016" data set and 10 hours and 40 minutes in the "April 2017" data set, resulting in a difference of 80 minutes.

Although this is less than the astronomical difference, the Sun rises very early in the morning, when people are still sleeping. It may be more practical to compare the results with the sunset differences as the people do not wake up earlier to organize an activity before work but may do it after work if it is still bright and the weather is good. The Sun went under at 19:33 on April 15 and at 20:42 on June 15, which is a 69-minute difference that is more comparable to the calculated values. The average workday bedtime values are 19:43 and 20:47, respectively. This finding — the day lengths are longer in June — is in good agreement with [97], where the seasonal influence of the daylight was examined via the length resting period.

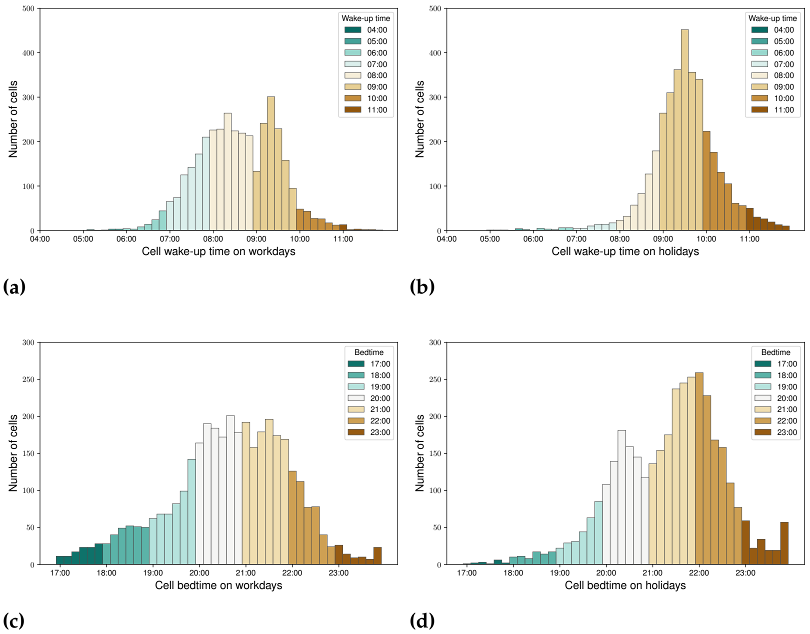
The wake-up times are 8:57 and 8:44 on average (workdays), which shows a slight decrease in the summer. Figure 7.3d shows slightly later wake-up times in the second half of the month. The average workday wake-up time in the first half of June 2016 is 8:39, and 8:50 in the second half. As mentioned in Section 7.3, it might be caused by the end of the school term.

## 7.5. Area-based Approach

Based on the cell/area activity, the wake-up, and the bedtime has been determined for every cell. Figure 7.6a and 7.6c, show the distribution for workdays, and Figure 7.6b and 7.6d, show distribution for holidays, respectively. The results clearly show that the usage of the mobile phone network intensifies and reduces later on holidays, indicating that people wake up and go to sleep later when they do not work.

Figure 7.7 shows the spatial distribution of the cell-based wake-up times. The Voronoi polygons, representing the mobile phone cells, are colored by the calculated "wake-up" times. The map clearly shows some extrema, where the times are around 10:00. Extrema 1-8 are all malls (WestEnd City Center (1), Arena Mall (2), Árkád (3), KÖKI Terminál (4), Lurdy Ház (5), Allee (6), MOM Park (7) and Mammut (8).), that uniformly open at 10:00, though 5 and 8 partly serve as office buildings.

Figure 7.8 shows the bedtime values of the sites on workdays. As the Figure 7.6c illustrates in respect of cells, the bedtime is usually between 20:00 and 22:00. There are, however, some sites with a later bedtime, and a few of them are denoted in the figure. Marker 1 at the party district (Appendix A). The site at marker 2 does not have any distinctive object that could explain this result. However, East of that site, there is a beer factory that might have notable activity in the evening — compared to the neighborhood — and the distortion of the Voronoi tessellation could



**(a)**

**(b)**

**(c)**

**(d)**

**Figure 7.6.:** Cell based wake-up (**a**, **b**) and bedtime (**c**, **d**) distribution for workdays and holidays, respectively, in the "April 2017" dataset.

have resulted in this late bedtime. At marker 3, there is a student hostel and, in the neighborhood, there are sport and concert venues.

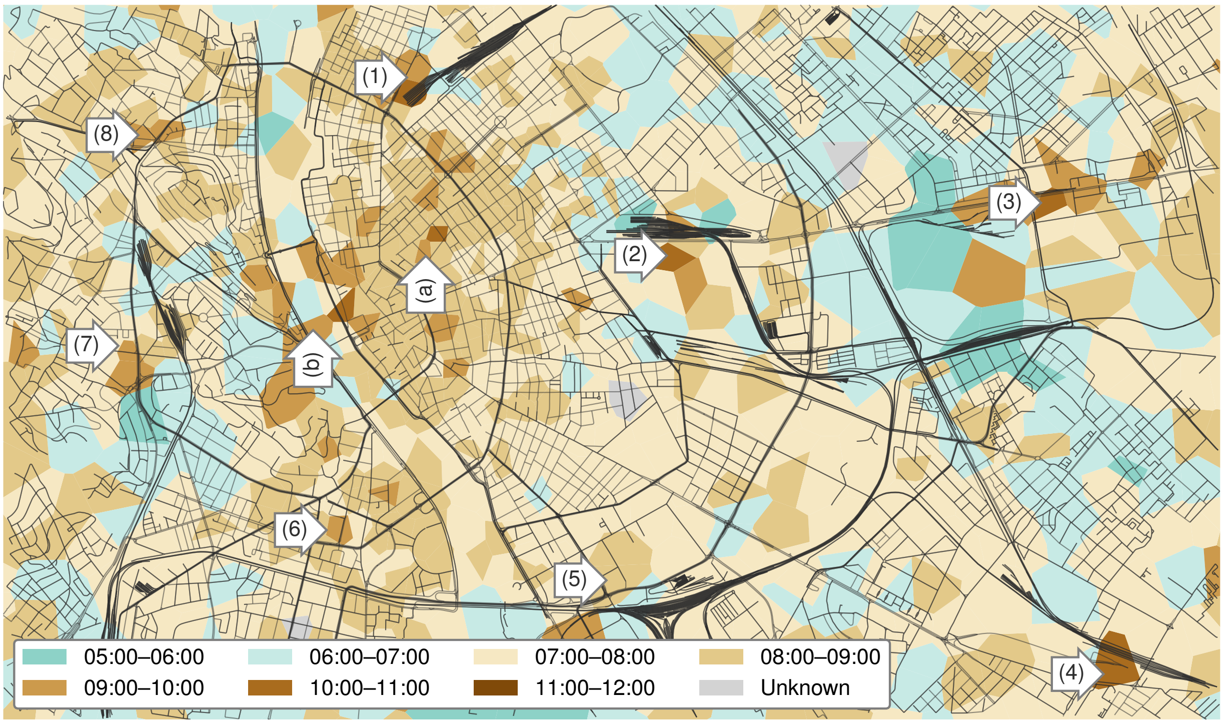


**Figure 7.7.:** Spatial distribution of the cell-based wake-up times, malls (1-8) opens at 10:00.

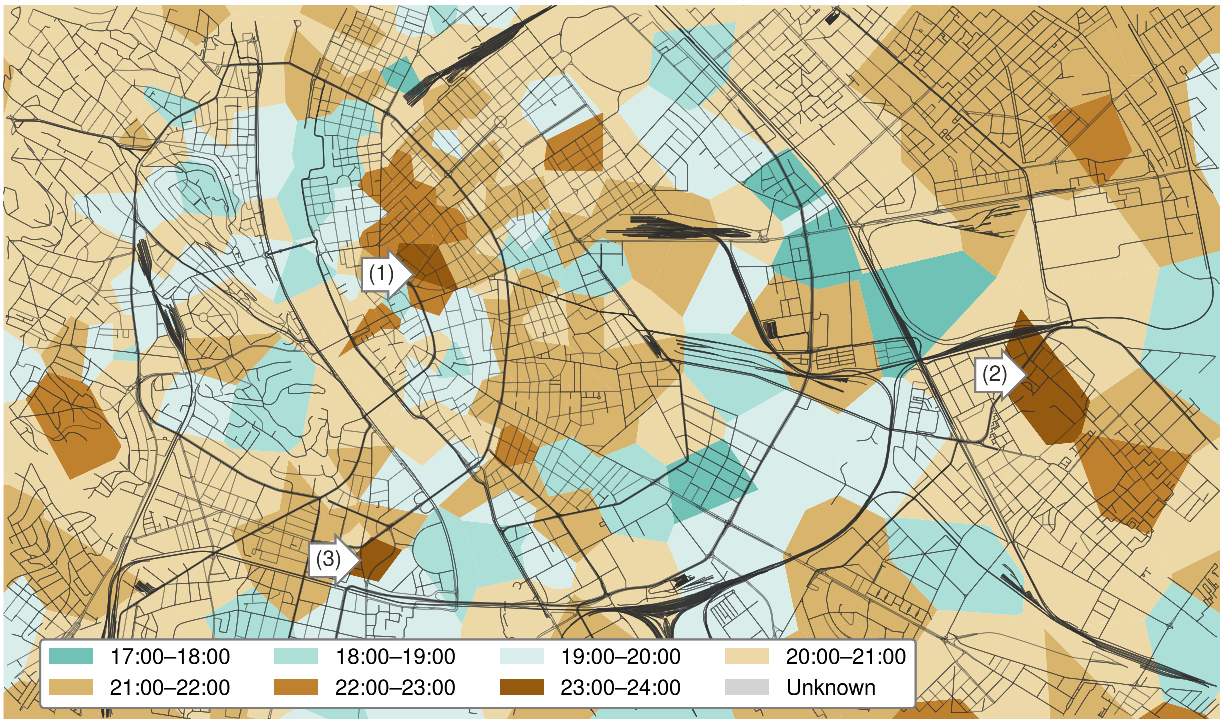


**Figure 7.8.:** Spatial distribution of the cell-based bedtimes, aggregated to sites.

## 7.6. Working Hours

Figure 7.9, shows the activity distribution by days of week and hours, based on the "April 2017" data set, separating the workplace (7.9a) and the home (7.9b) activity, but using the same scale. At the workplace, most of the activity was recorded during the working hours on workdays, as expected. The activity increases fast in the mornings but decreases more slowly in the afternoon. The home activity was mostly clustered on the weekends, but it had a notable activity peak after the work hours in the evening. The home activity before the working hours does not seem so significant.

This procedure can be applied for every cell (or group of cells, like sites), then the activity of the workers and the inhabitants will represent the workplace and home activity tendencies. Note that a cell can have both workers and inhabitants, so every cell has two aspects. Figure 7.10a and 7.10b, illustrates the activity of the subscribers who work and live a selected site, respectively. Figure 7.10a, also demonstrates the concept of Figure 7.1, using actual data.

As expected, Figure 7.10a and 7.10b is in accord with Figure 7.9a and Figure 7.9b. Workers' activity increases in the morning and decreases late in the afternoon, whereas the inhabitants' activity decreases in the morning and increases late in the afternoon and reaches its peak in the evening. The two aspects of the same site have the opposite tendency.

Applying the same inhabitant-based approach as in the case of the wake-up time (Figure 7.1) to the workers' activity, it is possible to detect the positive and the negative edge of the activity curve. The positive edge could indicate the start of the working hours, and the negative edge could indicate the end of the working hours in a given cell (or site). Moreover, the difference between the two times can determine the length of the working hours.

In most of the sites, the working hours are about 8 hours (Figure 7.11), just as expected. In the rest of the sites, — especially where the working hour is less than seven hours or over nine and a half hours — the mobile phone activity proved to be so low during the working hours that the results cannot be considered appropriate. Nevertheless, the mobile phone network reflects the length of the working hours.

The work hours do not necessarily start and end at the same time in every workplace. Are there any differences in this regard from a mobile phone perspective?

Figure 7.11 shows the average workday length of the sites. The workday length in most of them is around eight hours, which agrees with the

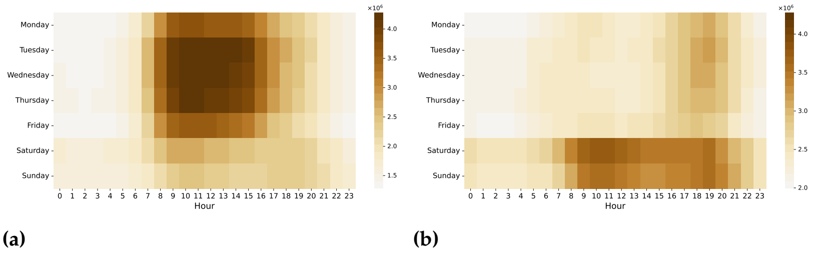


**Figure 7.9.:** The mobile phone activity distribution by days of week and hours based on the "April 2017" data set, for the work places (**a**) and the homes (**b**).

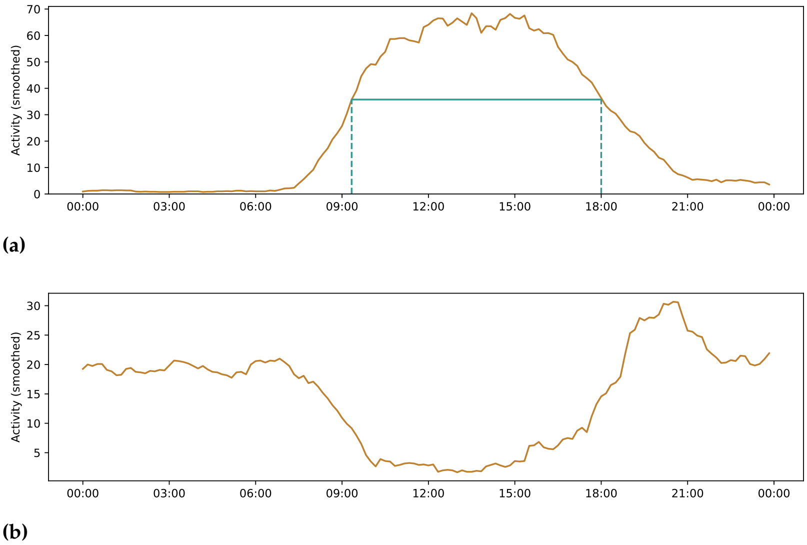


**Figure 7.10.:** Activity of the workers (**a**) and the inhabitants (**b**) of a selected site, on a selected day.

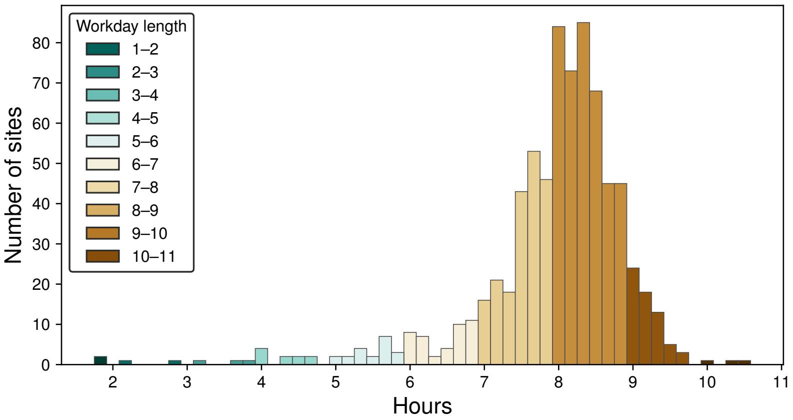
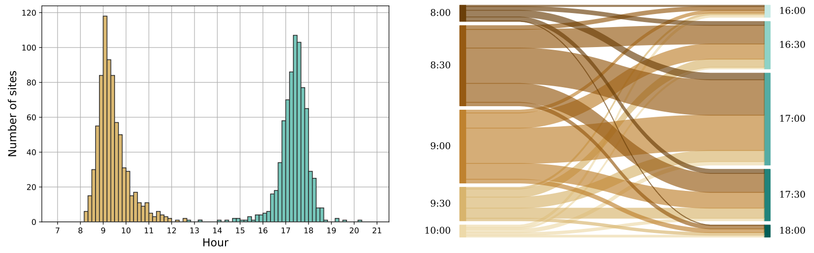


**Figure 7.11.:** Distribution of workday length in sites.

practice in Hungary. Moreover, the distribution of the working hour lengths also verifies the workplace detection approach, summarized in Section 4.6. Although there are some sites with very low (less than seven hours) or high (more than nine and a half hours), these sites have very few activities during the observation period: less than 2% of the total activity occurs in these sites.

The length of the working hours is usually around eight hours in most of the sites, but there might be differences at the beginning and the end of the labor-time. Figure 7.12b shows the five most frequent beginning and ending labor-time (defined as the positive and negative edge of the workplace activity curve of the site) and the connection between them, using a type of Sankey diagram. It is clear that in most of the sites, the mobile phone network activity increases between 8:30 and 9:30. The most notable group That is not surprising, considering the lunch break. Note that the observed time values may have a delay compared to the actual start of the work since an employee may not actively use the mobile phone network as soon as they arrive at their workplace (or when they leave it).



**(a)**

**(b)**

**Figure 7.12.:** The distribution of the starting (brown) and the ending (green) of the working hours, in sites (**a**); and the connection between the starting and the ending times (**b**).

## 7.7. In Respect of Mobility

In Chapter 6, I showed that the mobility has a weekly seasonality (Figure 6.13). As Figure 7.3 shows, the wake-up time also have a seasonality. To compare it with the mobility indicators described in Section 2.2, the daily wake-up times and the mobility metrics were normalized using min-max feature scaling. Figure 7.13a and Figure 7.13b display the normalized Entropy and the normalized Radius of Gyration — without the non-phone devices — in contrast to the wake-up times, respectively. As the mobility indicators are determined per subscriber, the inhabitant-based version of the wake-up time is used.

During workdays, both the Entropy and the Radius of Gyration were high, but the wake-up time was low. On holidays, it is on the contrary. The wake-up times were higher, and the mobility indicators were lower. This is not surprising since people tend to wake up earlier on workdays to go to work. Thus people usually need to travel to their workplace. On holidays, it is common to spend more time at home, which also reduces the mobility values. Figure 7.13 visualizes this clearly, especially in the case of Entropy. Numerically (Pearson's R), the correlation between the wake-up times and the mobility indicators (Entropy, Radius of Gyration) are $-0.9019$ and $-0.6869$, respectively.

Bedtimes show a similar trend (Figure 7.13c and Figure 7.13d) in contrast of the two mobility metrics (Entropy and Radius of Gyration), however, the correlations numerically are not that strong: $-0.874$ and $-0.6479$, respectively. This might have been caused by leisure activities after work. While people may go straight to work in the morning, they do not necessarily go home right after work.

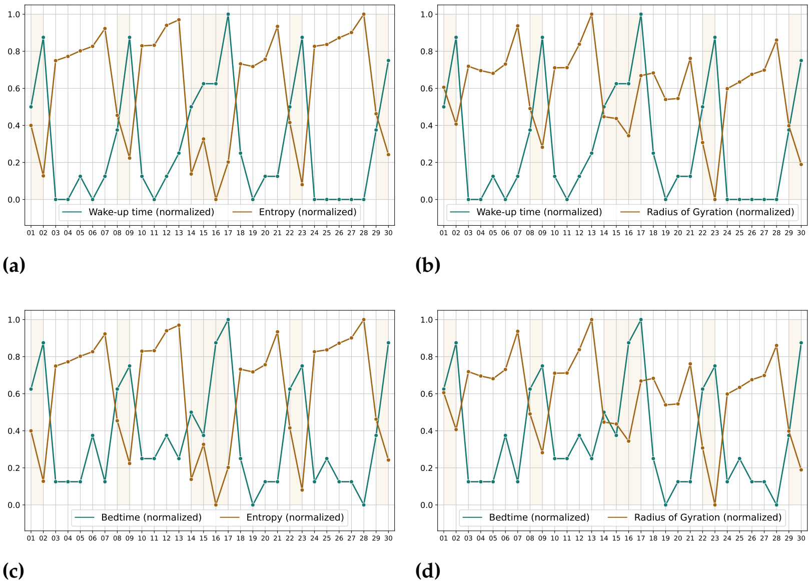


**Figure 7.13.:** Normalized, inhabitant-based wake-up times in contrast of the normalized daily Entropy (**a**) and Radius of Gyration (**b**). Pearson's Rs are −0.9019 and −0.6869, respectively. Second row: Normalized, inhabitant-based bedtimes in contrast of the normalized daily Entropy (**c**) and Radius of Gyration (**d**). Pearson's Rs are −0.874 and −0.6479, respectively.

## 7.8. Limitations

As the activity curves do not reach their maxima immediately, the selected time does not accurately represent neither the wake-up time nor the time when the working hours actually start. These results can only be used relatively to other cells or areas of the city. The reason for this is that the CDRs in this study only represent the active (in other words, billed) usage of the mobile phone network. However, the employees usually do not actively use the network the moment they arrive at their workplace or home. There is a certain gap between that moment and the first activity. If the available data also contained the passive (cell-switching) communication, the time when a SIM card enters the home or work cell could be used. In this case, the terms "leaving home" and "arriving at the workplace" would be more accurate. Furthermore, the difference between the two could serve as a basis for a precise travel time estimation.

As the active mobile phone network activity is sporadic, most subscribers do not have enough activity records to trace back their daily movement accurately enough to determine when they leave their homes or arrive at their workplaces. Commonly, the subscribers do not have activity at home before work. It would be necessary to have multiple activities in the morning to select the first and last home activity at home to conclude when the subscribers tend to wake up and leave their homes. As this is not assured with this kind of data, this approach is unsuitable for individual-level analysis.

The phone price database does not contain launch prices (see Section 3.3.3), and the applied depreciation method is unknown. The results, utilizing the cell phone price as a socioeconomic indicator, should be interpreted by keeping that in mind.

## 7.9. Possible Enhancements

With the home and work locations, the distance can be determined. On the other hand, the travel time is hard to estimate. The subscriber could use different transportation modes that have different time demands. Furthermore, the time when the subscriber left the home cell and arrived at the work cell cannot be exactly determined without cell-switching information. Using the worker or inhabitant filtered activity curves — described in Section 7.6 —, the morning fall (and the evening rise) of the home location activity and the rise (and the fall) of the work location activity can be calculated. Considering that the work location activity rises when the workers usually arrive, and the home activity of the inhabitants drops when they usually leave to work, the difference between these values could be applied to estimate the travel time of a group of subscribers. It would, naturally, require a larger number of subscribers in every home–work cell pair, or the locations should be aggregated by base stations or even larger areas of the city.

Although the "edge detection" method used in this paper gives reasonable results to determine the positive and negative edge of the activity curve, it would be worth comparing it with other approaches.

## 7.10. Conclusions

In this chapter, I introduced "wake-up time" as an indicator to describe the behavior of a group of subscribers. Due to the sporadic nature of the mobile network data, the subscribers were aggregated for this analysis by home and work locations. With wake-up time and its counterpart, called "bedtime", the beginning, the ending, and the length of the working hours were also estimated. Tendencies between the starting and the ending time of the working hours at the work locations were also presented. This indicator was used to classify the groups of subscribers (inhabitant-based approach) and the parts of Budapest (area-based approach); thus, city parts can also be characterized by chronotypes. It was demonstrated by real-life examples such as the opening hours of the malls in Budapest or the late activity fall of the party district.

Wake-up time as a proposed indicator was compared to common indicators, such as Radius of Gyration or Entropy, and a clear negative correlation was found. On workdays, both Radius of Gyration and Entropy values were higher, while the wake-up times were lower. On holidays, it is quite the contrary. The correlation between bedtime, the counterpart of the wake-up time, and the mobility metrics were not that strong but still considerable.

The day length, calculated as a difference between bedtime and wake-up time, was constant between the workdays and holidays: the start and the end of the day were also shifted. On the other hand, the day length reflected the seasonal differences between the two data sets: it was found that the days are longer from the perspective of a mobile phone network when there is more daylight. The longer days are the reason for the delayed activity fall, as the wake-up times were marginally affected by the earlier sunrise.

# Socioeconomic Status 8.

In this chapter, the relations between mobility indicators and measures of socioeconomic status are explored. I have used two different approaches to characterize socioeconomic status: a more direct and a more indirect one. The latter is the housing price. As I demonstrated in Chapter 6, the home location estimation is reliable. When the home location of a subscriber is known, the typical housing price of that area can be associated with the subscriber, considering that if someone lives in a more expensive area, their socioeconomic status is higher.

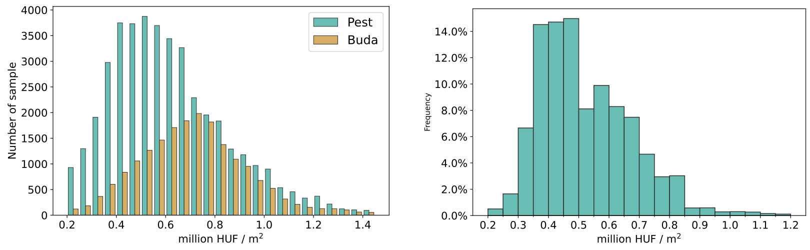
The other approach uses the price and the age of the cellphone that a subscriber was using. As it is described in Section 3.3, the device in which the SIM card operates is known, so the price and the release date of the cellphone are also available.

## 8.1. Housing Prices

Applying housing prices as a socioeconomic indicator has precedent in the literature (Section 2.8). My contribution, in this regard, is to consider the housing prices at both the home and the work locations. The data source was the *ingatlan.com* estate selling portal (Section 3.4).

Based on my mobile network data processing, Amir Mosavi presented a hybrid machine learning approach to model real estate prices based on the subscribers' mobility metrics [150*].

Figure 3.14 shows the spatial distribution of the estate price, which clearly illustrates that the Buda side of Budapest is much more expensive. Figure 8.1a displays the normalized housing prices by the real estate advertisements, distinguished by the Buda and Pest side of the city, while Figure 8.1b shows the average housing price distribution per cell. The cells with a very light color in the middle of the city are Margaret Island, a recreational area with large parks, sport establishments, and hotels without any residential zone. So, there was no estate for sale in the data.

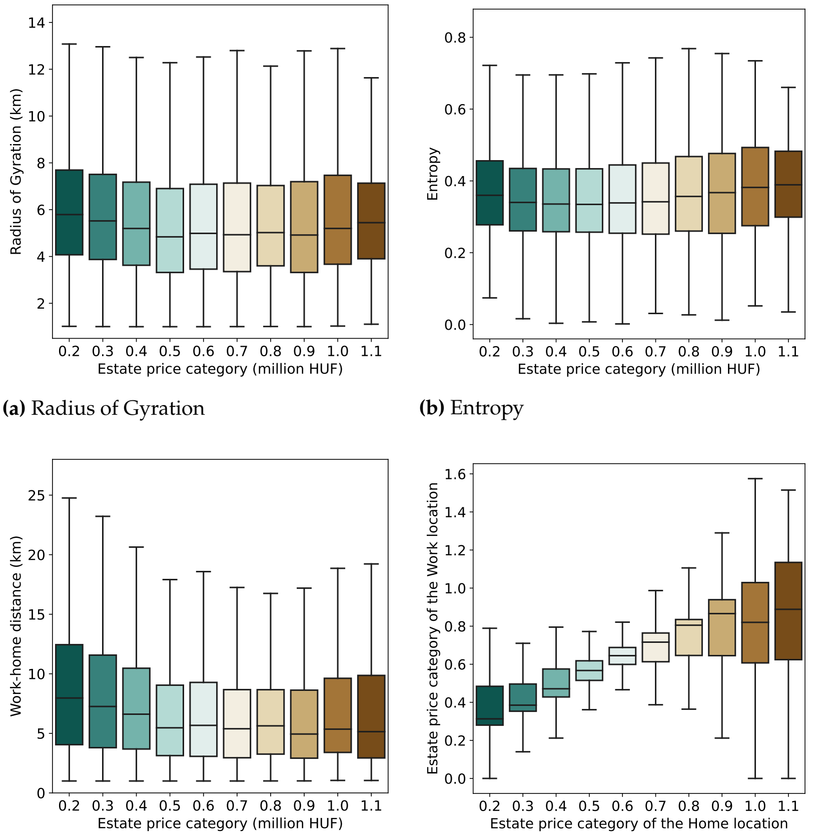


**(a)** Distribution of the estate price samples by Buda and Pest

**(b)** Normalized price distribution of the home locations (cells)

**Figure 8.1.:** Distributions of the estate price: (a) shows the distribution of the estate price samples distinguished by the Buda and Pest of the city, and (b) shows the distribution of the home location prices of the individuals.

I also analyzed the correlation between property prices at home and work areas. It was found that the economic status of the home and work locations for each mobile phone user has a directly proportional relationship (Figure 8.2d). Higher mean values of housing prices, where the people spend their working hours, belong to the higher home price classes. Notably, the mean value of property prices in work locations is higher than the home classes for residents whose housing price is



**(a)** Radius of Gyration

**(b)** Entropy

**(c)** Home-Work distance

**(d)** Estate prices of the work locations in contrast to the workers' home location estate prices

**Figure 8.2.:** Indicators by financial category.

between 0.2–0.6 and 0.9–1.1 million HUF. The total range (max–min) and the Q1–Q3 range of the property prices at workplaces is inversely proportional with home housing prices, in the 0.2–0.6 and directly proportional in the 0.7–1.1 million HUF range. This observation suggests that a few workplaces are in the cheaper areas of Budapest, and therefore, most of the dwellers spend their working hours in the more expensive regions. The smallest scatter of the property values at workplaces is given for the residents living in the 0.6 million HUF price regions. Nevertheless, the results in Figure 8.2d suggest that job and home locations have a notable relationship.

## 8.2. Principal Component Analysis

The individuals were classified into ten categories, in the price range of 0.2 to 1.2 million HUF / m$^2$, based on the mean estate price of their home location (cell). Then the mean value of the indicators was determined within the classes. Figure 8.2, shows the Radius of Gyration, the Entropy, and the Home-Work distance of the financial classes, where only the Home-Work distance (Figure 8.2c) shows a slight tendency between classes.

The Radius of Gyration defines a circle in which an individual lives their life, and the Entropy measures the diversity of an individual's visited locations. However, both of them are highly affected by the data quality. The data used for this work contains only those visited locations where billed activity (phone calls, short messages, or data transfer) happened. Everything beyond that is invisible. Even with only the active communication, the mobility customs of an individual can be described, but this seems to be insufficient to distinguish using financial status criteria.

Figure 8.2d shows the typical Estate Prices of the work locations in contrast to Estate Prices of the home locations using a box plot without the outliers. For every home location estate price category, the box plot (Figure 8.2d) splits the SIM cards into three groups: (i) those between the first and the third quartile (Q1–Q3), (ii) those below the first quartile (between the minimum and the Q1: min–Q1), and (iii) those above the third quartile (between Q3 and the maximum, Q3–max). The idea was to treat the SIM cards within the interquartile range (IQR) separately from those below and above the IQR.

The SIM cards were aggregated by the home and workplace estate price categories (0.2–1.2 million HUF) and the twenty Radius of Gyration and Entropy categories (using 0.5 km distance ranges between 0.5 and 20 km for the Radius of Gyration and Entropy values with 0.05 steps between 0.05 and 1.00). The data structure used for the Principal Component Analysis is defined as follows. Every row consists of 40 columns representing 40 gyration radius bins between 0.5 and 20 km and 20 columns representing 20 Entropy bins between 0.05 and 1.00. The bins contain the number of SIM cards that have been normalized by metrics to compare them. Although the workdays and the holidays were treated separately during the whole study, the data were not explicitly labeled by them. The same table was constructed using holiday metrics, and its rows were appended after the workday ones. The home and workplace estate price columns were not provided to the PCA algorithm.

### 8.2.1. Evaluation

The relationship between the mobility metrics (during weekdays and weekends) and the Social Economic Status (SES) was investigated using the result of Principal Component Analysis (PCA). By applying the PCA, the 60-dimension vector was reduced to two dimensions, and those variables were plotted to be analyzed. The results of this section were published in [151*].

Figure 8.3, shows the Pareto Histogram for the 60 components. The first two components (about 50% of the explained variance) are sufficient to represent the variance of the variables. Figure 8.4 shows the first two components of an unsupervised PCA analysis applied. The marker styles, colors, and sizes each represent one label of the data. First, the brownish colors represent the workdays, the greenish ones the holidays, and they clearly form two separate clusters, though the workday and the weekend rows were not distinguished. This means that the first principal components are expressive enough to distinguish between workday and holiday mobility habits.

Marker size represents the estate price at the home location. There is a noteworthy tendency along the PC2 axis that the markers increase.

Although the trend-line is more like a curve bending to the right, the results seem to distinguish the wealthier subscribers. The low-cost region's mobility can be found at the bottom part of the chart, and the higher the property prices, the higher the vertical positions of the mobility metrics. The second principal component seems to be a proper variable for representing the influence of housing prices on gyration and entropy. It has to be noted that the smallest markers look out of place. Those represent extremely inexpensive areas that are not residential but more like industrial zones, and only a minority of the subscribers are classified into those categories. The mobility habits of dwellers living in the wealthiest regions are visibly separated from the others, especially on the working and rest days.

The mentioned quartile groups are represented with different markers and hues. The symbols representing the mobility metrics of the min–Q1 group are to the left, and the markers representing the mobility metrics of the Q3-max group are to the right of the Q1–Q3 group in most of the house pricing groups. This means that the first principal component looks to be a determinative parameter to distinguish the workday and weekend mobility patterns and the quartile range of the housing price categories at the home location.

The range of PC1 between 0.0 and 2.5 includes markers of the first quartile for the most part, while a mixture of representatives belonging to Q1–Q3 and Q3–max groups located in the 2.5 – 10.0 range. The weekday mobility customs of people belonging to the min-Q1 group significantly differ from the other inhabitants' groups. This suggests that even though these residents are living in a similarly valuable neighborhood to the ones of the groups Q1–Q3, they are managing their lives differently. Notably, the deviation is smaller in the lowest and highest housing price categories while bigger in the middle-level property price range. It seems that people having homes in the middle price level area in the city have more diverse habits than those living in the poorer or wealthier districts.

The mobility characteristics related to the dwellers associated with the Q1–Q3 and Q3–max groups are quite similar. In the range of PC1 between 2.5 and 5.0, several markers from different groups and different housing price categories overlap each other. The markers on the chart (i.e., in the range of PC2 between 0.0 and -0.2) probably show that residents living in the cheapest and the middle price level categories have very similar traveling patterns independently if they belong to the Q1–Q3 and Q3–max groups. It is also remarkable that during the weekdays and the weekends, the mobility customs of the wealthiest districts' population are completely separated from others. This observation suggests that the daily traveling routines of the residents having homes in the most expensive area of Budapest are significantly different from the others.

Figure 8.2a represents the relationship between the radius of gyration and the normalized housing prices. I ranked the subscribers into ten groups (q = 10). I found that subscribers living in the middle price range (0.5–0.7 million HUF) areas have smaller Radius of Gyration on average than the cheaper (poorer) regions. Dwellers in poorer zones travel around 5 km on average, while the Radius of Gyration in the inhabitants from the middle class and more expensive regions is lower. These findings can be explained by the fact that mean of people living in middle price houses

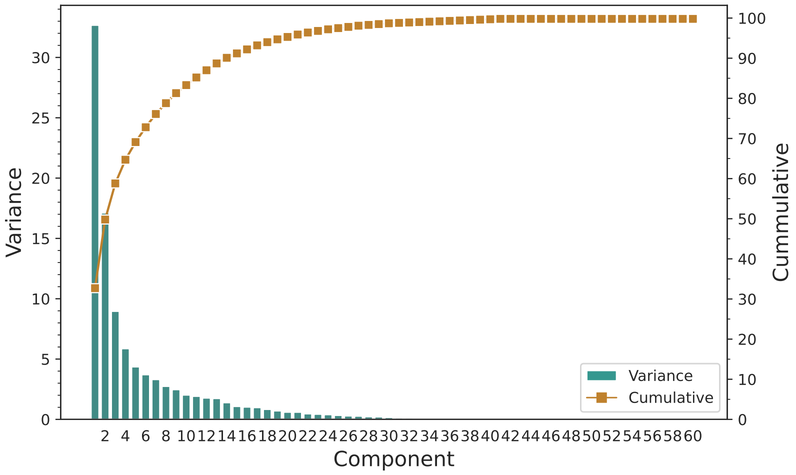


**Figure 8.3.:** The Pareto histogram for the 60 components of the Principal Component Analysis.

(Figure 8.2a) have a better chance to have different job and recreation opportunities, and therefore they do not need to travel long distances. The average and the interquartile range of subscribers living in the most expensive area (>1.0 million HUF) is larger than in the less costly area. This also follows that fewer people have homes in more wealthy regions, and therefore, the mean and the scatter of Gyration Radius are higher.

Figure 8.2b illustrates the relationships between SES and activity entropy, which indicator of mobility is dedicated to describing the regularity of subscribers' daily travel and activity patterns. The results show a low variability across SES for activity entropy. The mean values of each house pricing class remained in the range of 3.5–4.5. One potential reason is that for most people, their daily activities mainly concentrate in a few locations (e.g., home and work locations). However, a slightly increasing tendency can be observed, which suggests that the occupants living in more expensive sectors were visiting more places. The regularity of activities at the wealthier locations has a slight impact on the activity diversity of the subscribers. This finding differs from the results of [10]. Xu et al. showed that the wealth level of people, at least in Singapore and Boston, is not a limiting factor that affects how they travel around in the city. It seems that the residents living in more expensive areas in Budapest — are more likely or forced to — visit more places than the ones having homes in the less expensive regions.

We examined the relationship between the work-home distances and SES. Figure 8.2c illustrates the mileage between the home and job locations as a function of home location prices. The descending trend mean values of distance related to home prices from the lower to the higher property prices suggest that SES has a significant role on how much the dwellers have to travel to their workplaces. The cheapest homes are the farthest from the job location (the mean is 8.4 km and the maximum is 27.5 km), while the shortest daily work trips belong to the homes located in the 0.9 million HUF category (the mean is 5.2 km and the maximum is 21.8 km). In other words, the people living in the poorer area have to travel more to their work locations. More time spent on daily commuting by the people living in cheaper places might affect their opportunities to visit more attractions in the city. This capability may also explain the relationship between activity entropy and SES (Figure 8.2b).

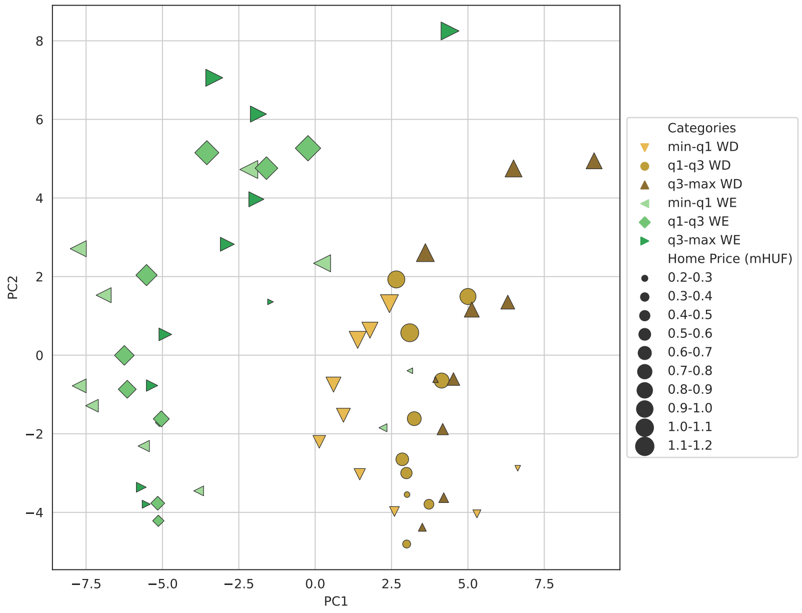


**Figure 8.4.:** Scatter plot of the 2-component PCA. Marker sizes indicates the home price category, the color/type work price category and also the day type (Weekday/Weekend).

## 8.3. Cellphone Price and Age

Note that the results of this section are based on the "June 2016" dataset (Section 3.2). The contents of this section were published in [152*].

Earlier in this chapter, the real estate price of the subscribers' home locations were used to describe their socioeconomic status. In this section, the CDRs were enriched by phone prices, and the phone price is assumed to apply as a socioeconomic indicator. To demonstrate the applicability of the mobile phone price as a socioeconomic indicator, it was examined in respect of the mobility indicators, applying Principal Component Analysis (PCA), as described before.

The SIM cards were aggregated by the subscriber age categories (5-year steps between 20 and 80) and the phone price categories (100 EUR steps to 700 EUR), the Radius of Gyration, and Entropy categories. For the Radius of Gyration, 0.5 km distance ranges were used between 0.5 and 20 km, and the Entropy values were divided into twelve bins between 0.05 and 1.00. The data structure used for the Principal Component Analysis is defined as follows.

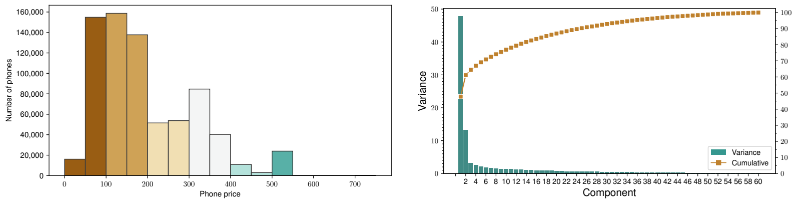
A table has been generated where every row consists of 40 columns representing 40 gyration radius bins between 0.5 and 20 km and 20 columns representing 20 Entropy bins between 0.05 and 1.00. The subscribers belonging to each bin are counted, and the cardinality has been normalized by metrics to be able to compare them. The categories were not explicitly labeled by them, so the subscriber age and the phone price descriptor columns were not provided to the PCA algorithm. The same table was constructed using weekend/holiday metrics, and its rows were appended after the weekday ones.

When the PCA was applied, the 60-dimension vector was reduced to two dimensions based on the mobility customs, where the bins were weighted by the number of subscribers. The cumulative variance of the two best components is about 61% (see Figure 8.5b). The bins representing the two new dimensions (PC1 and PC2) were plotted (see Figure 8.6), where the markers were colored by the phone price, and marker sizes indicate the subscriber age category, using larger markers for younger subscribers.

As Figure 8.6 shows, the markers are clustered by color — in other words, the phone price — that is proportional to PC1 but inversely proportional to PC2. Within each phone price group, the younger subscribers (larger markers) are closer to the origin, indicating that the mobility custom of the younger subscribers differs from the elders, although this difference is smaller within the higher price categories. This finding coincides with [153], where Fernando et al. found a correlation between subscribers' age and mobility metrics.

To give context to Figure 8.6, Figure 8.5a, shows the phone price distribution: most of the phones are within the 50–200 EUR range. Note that there are only a few phones over 550 EUR, but the owners of those have significantly different mobility patterns.

Figure 8.6 does not only show that the phone price forms clusters but also reveals the effect of the subscription type on the mobility. Within the phone price categories, except for the highest with only a very few subscribers, the postpaid groups are usually closer to the origin. Prepaid subscriptions are usually for those, who do not use their mobile phone extensively, and it seems that people with a prepaid subscription have similar mobility customs as people with less expensive phones but postpaid subscription. That is most notable at (-6, 2) and (-5, -1).



**(a)** Phone price distribution.

**(b)** The Pareto histogram for the PCA.

**Figure 8.5.:** Phone price distribution and the Pareto histogram for the 60 components of the Principal Component Analysis.

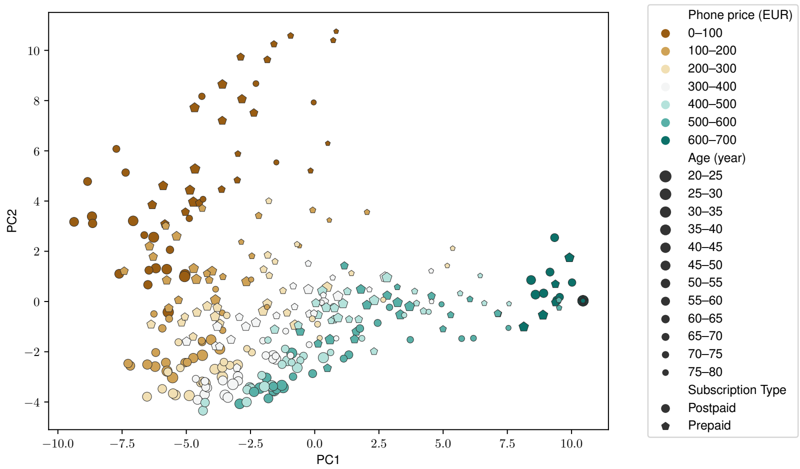


**Figure 8.6.:** Scatter plot of the 2-component PCA. Marker sizes indicate subscriber age category, the color represents the phone price category and the subscription type (Prepaid/Postpaid) is distinguished by the marker type.

Sultan et al. identified areas in Jhelum, Pakistan, where more expensive phones appear more often [12]. Using the same method, Budapest and its agglomeration were evaluated: the average phone prices from the activity records are determined for every site. The ground truth is that the real estate prices are higher on the Buda side (West of river Danube) of Budapest and downtown, and this tendency can be clearly seen in Figure 8.7. The airport area has a significantly higher average than its surroundings, which is not surprising. The spatial tendencies of the mobile phone price, along with the result of the PCA (Figure 8.6), clearly demonstrates the expressiveness of the phone price as a socioeconomic indicator.

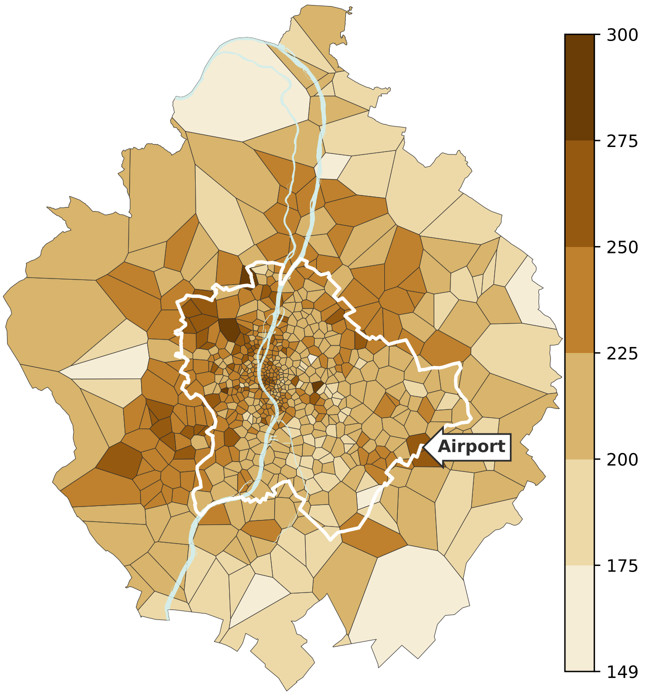


**Figure 8.7.:** Average price (in EUR) of the mobile phones, that generated the activity records in each site, during the whole observation period (June 2016).

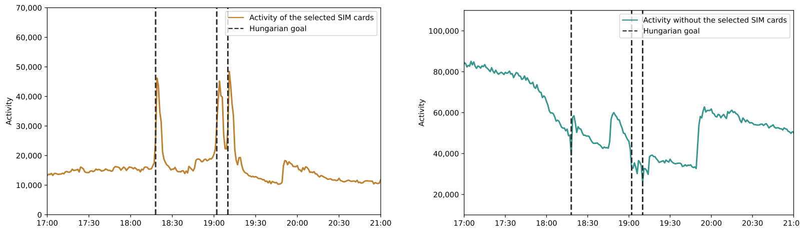
## Socioeconomic Status of the Football Fans

During Hungary vs. Portugal match, there were three Hungarian goals (Figure 5.3), so there were three peaks, starting at 18:18, 19:02, and 19:18. All of them had about 5-minute fall times. Who was responsible for these peaks? To answer this question, the SIM cards that were active during any two of the peaks were selected. Selecting SIM cards that were active during any of the peaks would also include many subscribers that cannot be considered football fans. The participation of all the three peaks, on the other hand, would be too restrictive.

Figure 8.8a, presents the activity of the selected 44,646 SIM cards and the owner of these cards, which may belong to the football fans. Removing these SIM cards from the data set should result in an activity curve without peaks and at the same time similar, in tendency, to the average activity. However, as Figure 8.8b shows, the activity still drops during the match. Therefore, the 'football fan' category should be divided into 'active' and 'passive' fans from the mobile phone network perspective. Active fans are assumed to express their joy using the mobile phone network (presumably to access social media) and cause the peaks. It seems that the passive fans ceased the other activities and watched the game, which caused some lack of activity compared to the average. By removing the active fans from the observed set of SIM cards, the activity level decreased in general (Figure 8.8b). However, this is not surprising, as these people reacted to the goals, they probably often use the mobile phone network. There are also some negative peaks, indicating that the selection is not perfect.

Is there any difference between the active fans regarding the phone age and price compared to the other subscribers? Figure 8.9a shows the relative age of the phones in respect of the subscribers' behavior after the goals. No significant difference has been realized between the active fans and other subscribers, the median of the relative phone age is about two years, and there are some much older (nearly ten years old) phones in use. It should be noted that older devices are used by elderly people. The phone prices show an opposite tendency: the younger subscribers own more expensive phones (Figure 8.9b).

Naturally, not all of these 169,089 SIM cards (without the ones operating non-phone devices) generated activity after all the goals. 83,352 devices were active after the first goal, 70,603 after the second and 68,882 after the third. After at least two goals 44,646, and after all the three goals, only 9102 devices had activity within 5 minutes.

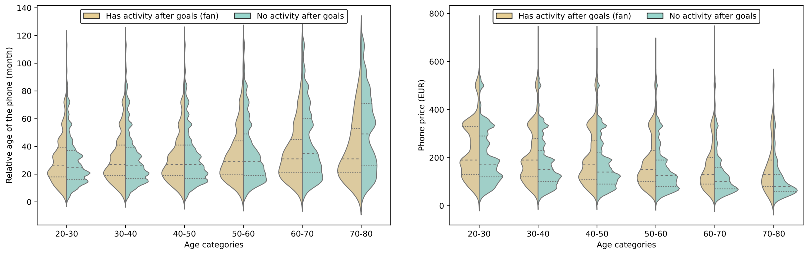


**(a)** Activity of fans.

**(b)** Activity without the fans.

**Figure 8.8.:** Mobile phone network activity of the SIM cards (fans), that had activity right after any two of the Hungarian goals, and the mobile phone activity of the other SIM cards.

Why would they use the mobile phone network to access social media? If they were at home, they would have used the wired connection via Wi-Fi for mobile devices. In Hungary, the 79.2% of the households had wired internet connection, according to the KSH[154], and it could be even higher in Budapest. However, if they were at fan zones, for example, in Szabadság Square, using the mobile network is more obvious.



**(a)** Relative age of the phones.

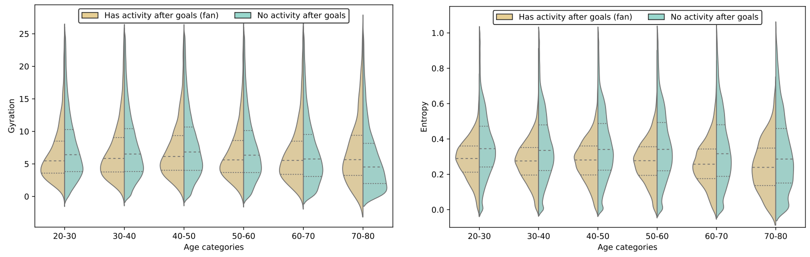
**(b)** Price of the Phones.

**Figure 8.9.:** Mobile phone relative age and the price distributions in different age categories, comparing the fans, who had activity right after any two of the Hungarian goals, and the rest of the SIM cards.

As Figure 8.9 shows, there is no significant difference in the phone age between the active football fans and the rest of the subscribers. The medians are almost the same within the young adult and the middle-age categories, but elders tend to use older devices, especially those who did not react to the goals. The active football fans' median phone price is 180 EUR, in contrast to the 160 EUR median of the rest of the subscribers.

However, older subscribers tend to use less expensive phones. This tendency is also present within the football fans but stronger within the other group.

Figure 8.10 illustrates the mobility metrics in different age categories, also comparing the football fans and the rest of the subscribers. The Radius of Gyration median is almost the same in all the age categories and groups. The Entropy medians have a notable difference between the two groups but do not really change between the age categories. This means that the mobility customs of the football fans, who use the mobile phone network more actively, are similar, regardless of the subscribers' age.



**(a)** Radius of Gyration of the subscribers.

**(b)** Entropy of the subscribers.

**Figure 8.10.:** Radius of Gyration and Entropy distributions in different age categories, comparing the fans, who had activity right after any two of the Hungarian goals, and the rest of the SIM cards.

## 8.4. Wake-up Time and Socioeconomic Status

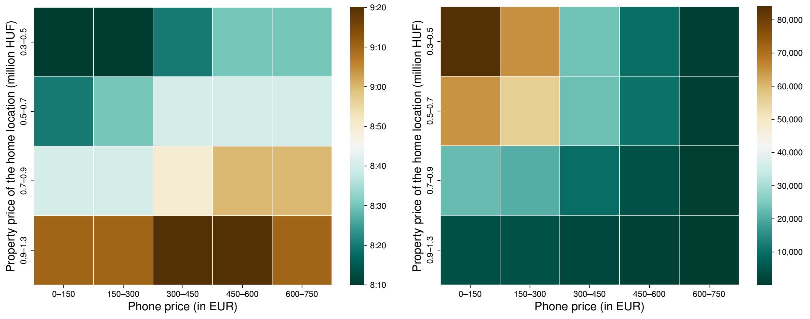
Earlier in this chapter, the correlation has been demonstrated between the mobility customs, the distance between the home and work locations, and the socioeconomic status. We found that people who live in less expensive parts of Budapest tend to travel more to their workplace. The larger distance should indicate longer travel times and earlier wake-up times, as the work mostly starts at the same time in the morning. Lotero et al. previously found that "rich people do not rise early" [96], analyzing two cities of Colombia.

As for Budapest, there is a positive correlation between socioeconomic status and wake-up time. The inhabitant-based approach is used for this analysis, as the socioeconomic indicators are applied to subscribers. Besides the real estate prices (Section 8.1), two properties of the mobile phones (Section 8.3) were considered: the price and relative age. Figure 8.11a shows the wake-up times in contrast to the real property price and the mobile phone price categories. To give context to these categories, Figure 8.11b illustrates the number of subscribers in each category. Figure 8.12 has the same structure but applied to the age of the cell phones.

Four property price categories and five phone price categories were formed. As Figure 3.13 shows, most of the price of one square meter in most of the real estate advertisements are under 0.6 million HUF. The property price categories are: (i) 0.3–0.5, (ii) 0.5–0.7, (iii) 0.7–0.9 and (iv) 0.9–1.3 million HUF. Mobile phone price categories are: (i) 0–150, (ii) 150–300, (iii) 300-450, (iv) 450–600 and (v) 600–750 EUR. As for the mobile phone age, five categories were formed from 0 to 5 years. The older phones were omitted. As Figure 3.10a indicates, most of the subscribers lived in less expensive homes and used less expensive cell phones that were 1 to 3 years old.

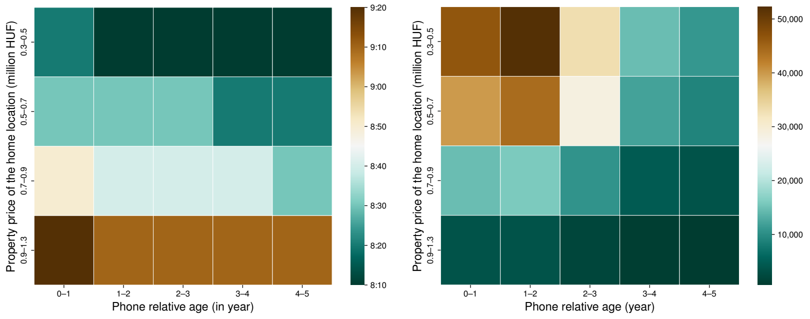
In Figure 8.11a, the wake-up times are increasing in both dimensions. The lowest wake-up value is in the top-left corner, where the owners of the least expensive home locations and cell phones are located. Towards the bottom-right corner, the wake-up times are increasing. This observation clearly indicates that the richer people start their days somewhat later than the less wealthy. The differences are numerically not large between the categories, but the observation area is also relatively small. Budapest is 525 square kilometers, and its diameter is about 30 km. The farthest part of the agglomeration is about 40 km from the city center.

Figure 8.12a reveals a negative correlation between phone age and wake-up time. This result implies that wealthier people do not only use more expensive phones but tend to use the latest models. Presumably, they replace their phones more often.



**(a)** **(b)**

**Figure 8.11.:** The inhabitant-based wake-up times in the socioeconomic categories, based on the property price of the home location and mobile phone price (**a**), and the number of the subscribers in each category (**b**).



**(a)** **(b)**

**Figure 8.12.:** The inhabitant-based wake-up times in the socioeconomic categories, based on the property price of the home location and the relative age of the mobile phone (**a**), and the number of the subscribers in each category (**b**).

## 8.5. Limitations

As for the housing-price-based approach, the inhabitants were associated with the mean property value at their home cell location (i.e., Voronoi polygons). In some regions of the city, where the density of the cellphone towers is low, this approach could lead to inaccuracy on the SES of residents (the housing prices vary in relatively bigger zones). Only one variable (housing price) was used, and other sociodemographic factors were not considered. Family and household size and education level can also be essential dimensions of SES.

As for the cellphone price base approach, subscribers' SES was associated with the release price of their cell phones. However, it is not necessary for them to buy their phones at that price. Many people buy their phone on sale or discount via the operator in exchange for signing an x-year contract.

Also, the subscribers can change their phone devices at any time, so only those subscribers have been taken into consideration, who had used only one device during the observation period or had a dominant device that generated most of the activity records of the given subscriber. It also has to be noted that people can even have more than one cellphone (e.g., a private and a work-related), but the presented model cannot take this into account.

I have fused three data sets to exclude the non-phone SIM cards, but the identified devices are not complete. There remained devices whose models are unknown, and there are phones whose release date and price are unknown. It is not possible to determine SES of these subscribers with the proposed solution. Furthermore, as Section 3.3.3 demonstrated, the phone prices seem to be depreciated, which might have distorted the results, but using actual release prices should fix this.

## 8.6. Possible Enhancements

One possible research enhancement would be to use more precise data sources for the financial classification, for example, the type and price of mobile phones. The combination of the housing prices and the price of the mobile phone device might provide a better result. Moreover, the relative age of the cell phone might be used as a weight for the phone price when applied as SES indicator to distinguish between the phone price categories as an expensive but old phone is not worth as much as a newer one with the same price.

Using a data set containing information about both participants of a call, not only the social network could be built, but that network could be weighted with the financial status of the subscribers.

It may be possible to enhance this work by examining the visited places of the wealthiest classes, whether they tend to appear in areas, for example, with elegant restaurants or theaters outside of normal work hours.

Although the current solution to select the football fans' SIM cards that caused the peaks gives a reasonable result but could be improved by analyzing the activity during the whole observation period. For example, applying a machine learning technique.

Extending the list of the non-phone TACs could also help to refine the results, and combining the mobile phone prices with the real estate prices of the home location would most certainly enhance the socioeconomic characterization.

## 8.7. Conclusion

In this chapter, I presented a method based on Principal Component Analysis to evaluate socioeconomic status (SES) depending on the indicators of human mobility. I have utilized two features to characterize SES, a direct and an indirect one. The indirect is the housing prices because the level of a neighborhood was used to infer the SES. When the home location is known, the typical housing price of that area can be associated with the subscriber.

I found that the mean distance between home and work locations is significantly correlated to housing prices. The people living in cheaper regions have to take longer routes to their work sites. Analyzing the residents' real estate values at home and work locations, I have found a strong positive relationship. Living in a high-priced neighborhood in Budapest requires working in more expensive regions and vice versa.

A more direct feature for the SES is the price of the subscriber's cellphone. Although the actual purchase price of a mobile phone is unknown, the recommended retail price was available along with the release date. The mobile phone price was also proved to be an expressive socioeconomic indicator that can be applied to distinguish the subscribers by mobility customs.

Using socioeconomic classes derived from housing prices at the home location, mobile phone prices, and the age of the cell phone, a correlation between the wake-up time and the socioeconomic status was also identified. The subscribers living in less expensive apartments get up earlier, and this tendency holds true in respect of the mobile phone prices: subscribers who own more expensive cell phones tend to get up later.

# Conclusions 9.

After describing the utilized data sources of my research (Chapter 3) and briefly introducing my data processing framework (Chapter 4), I demonstrated its social sensing capabilities with two case studies in Chapter 5. I showed that mobile phone network activity precisely shadows the behavior of football fans or attendees of a mass protest, so capable of social sensing. The first presents the football fan's behavior during and after the Hungarian matches of the Euro 2016 championship. The joy — felt after the Hungarian goals — is manifested in the data as sudden activity peaks.

The spontaneous festival after the "Hungary vs. Portugal" match and the welcoming event at the Heroes' Square is comparable to mass protests from a data perspective. Just like the second case study, which presents the protest for the Central European University and against the modification of the Hungarian higher education law. During these events, the mobile network activity was significantly higher than the average in affected areas, clearly representing the people who participated.

In Chapter 6, the commuting customs of the population have been extensively investigated using Call Detail Records. First, I presented the results of my home and workplace estimation solution, then the commuting trends between these locations. To aid the comparison, I presented these results in a form as close as possible to other studies' results, focusing on commuting in Hungary and the Budapest agglomeration. The examinations show that commuting analysis can be automatized using mobile network data. It could provide a fast, cost-effective, and relatively punctual method of commuting analysis and a powerful tool for the sociologists interested in commuting.

In Chapter 7, I introduced "wake-up time" as an indicator to describe the behavior of a group of subscribers. This new indicator can describe when a group of subscribers starts to use the mobile network in the morning. Due to the sporadic nature of the mobile network data, the subscribers were aggregated for this analysis by home and work locations.

With wake-up time and its counterpart, called "bedtime", the beginning, the ending, and the length of the working hours were also estimated. Tendencies between the starting and the ending time of the working hours at the work locations were also presented. This indicator was used to classify the groups of subscribers (inhabitant-based approach) and the parts of Budapest (area-based approach); thus, city parts can also be characterized by chronotypes. It was demonstrated by real-life examples such as the opening hours of the malls in Budapest or the late activity fall of the party district.

Wake-up time as a proposed indicator was compared to common indicators, such as Radius of Gyration or Entropy, and a clear negative correlation was found. On workdays, both Radius of Gyration and Entropy values were higher, while the wake-up times were lower. On holidays, it is quite the contrary. The correlation between bedtime, the counterpart of the wake-up time, and the mobility metrics were not that strong but still considerable.

The day length, calculated as a difference between bedtime and wake-up time, was constant between the workdays and holidays: the start and the end of the day were also shifted. On the other hand, the day length reflected the seasonal differences between the two data sets: it was found that the days are longer from the perspective of a mobile phone network when there is more daylight. The longer days are the reason for the delayed activity fall, as the wake-up times were marginally affected by the earlier sunrise.

Using socioeconomic classes (Chapter 8) derived from housing prices at the home location, mobile phone prices, and the age of the cell phone, the correlation between the wake-up time and the socioeconomic status was also identified. The subscribers living in less expensive apartments get up earlier, and this tendency holds true in respect of the mobile phone prices: subscribers who own more expensive cell phones tend to get up later.

These results may help to analyze further the city structures by identifying "early bird" or "night owl" areas and possible connections between them. City parts with early morning or late night activities may require different public transport services, for example, and can aid the transportation infrastructure planning. Business development could also benefit from the detailed insight of the neighborhood chronotypes, especially with the associated information on the home locations and the socioeconomic status of the subscribers.

The socioeconomic aspect of these findings can also contribute to a better understanding of the social structure of urban environments. In this regard, further studies need to be made, possibly with detailed census data, to evaluate the commuting and working habits of the different socioeconomic classes.

In Chapter 8, I presented the utilization of housing prices and phone prices as socioeconomic indicators, and I introduced an analytical method to express the correlation between the Social Economic Status (SES) and the mobility patterns of the subscribers.

The Radius of Gyration, Entropy, and Euclidean distance between home and work locations has been derived from the mobile phone network data and used as mobility indicators to characterize the travel patterns of subscribers. The property price of homes has been applied to represent their socioeconomic status. I performed a data fusion method to quantify the SES of mobile phone users at an aggregate level.

By investigating the relationships between SES and three mobility indicators, I found that the wealth level in Budapest has a certain influence on travel customs. The real estate price of home sites does not have a remarkable effect on the number of visited places (Entropy) and Gyration, which suggests that the travel diversity of the inhabitants is not strongly dependent on their socioeconomic status.

On the other hand, the mean distance between home and work locations was significantly correlated to property values. The people living in cheaper regions have to take longer routes to their work sites. This relationship can be explained by the less expensive districts located at the perimeter of the city, while most of the work locations are in the wealthier regions. Analyzing the residents' real estate values at home and work locations, I found a strong positive relationship. Living in a high-priced neighborhood in Budapest requires working in more expensive regions and vice versa.

The Principal Component Analysis revealed similarities and differences in movement behavior of residents living in different price level categories. The correlations of mobility customs and SES found could be used efficiently for supporting the activities of public transportation, real estate developments, and retail.

I also evaluated the connection of movement behavior and SES by performing Principal Component Analysis (PCA). The results showed that the first two components could be sufficient to distinguish the workday and weekend mobility patterns, and the home location's estate price categories. The mobile phone network data can also be used to analyze commuting, similar to census data, even though its spatial resolution is not as accurate. The commuting trends between the given sectors of the agglomeration and the district groups of Budapest match enough to say that Call Detail Records can be a cheaper and temporally more accurate solution for commuting analysis and possibly other kinds of sociological studies.

This research focused on identifying hidden or unclear links between human mobility and socioeconomic status. To that end, this work has been successful. The approach and the framework developed herein can be applied to various locations. The findings can be used to understand more precisely and efficiently the conditions and circumstances, of day-to-day life, for people living in our modern urban environments.

I also demonstrated that mobile phone price is an expressive socioeconomic indicator. It can cluster the areas of a city and distinguish the subscribers by mobility customs. On the other hand, it does not seem to affect the interest in football.

The mobile network data and the cellphone specification databases have been fused. The data fusion also allowed me to remove a considerable number of SIM cards from the examination that certainly operates in other devices than mobile phones. Although there are some still unidentified Type Allocation Codes in the data set, this way, the activity records involved in this study have a significantly higher possibility of being used by an actual person during the events.

## 9.1. Practical Applicability

Sociological studies are usually performed based on censuses or direct questionnaires. An application of this research would be to support these studies using the anonymized mobile network data as a frequent, countrywide, and cost-effective alternative. Naturally, this application would require legal regulation, as a cooperation with the mobile operators and the Central Statistical Office.

These results may help to analyze further the city structures by identifying "early bird" or "night owl" areas and possible connections between them. City parts with early morning or late night activities may require different public transport services, for example, and can aid the transportation infrastructure planning. Business development could also benefit from the detailed insight of the neighborhood chronotypes, especially with the associated information on the home locations and the socioeconomic status of the subscribers. The socioeconomic findings of this research can also contribute to a better understanding of the social structure of urban and rural environments if the analyses are performed at the country level.

## 9.2. New Scientific Results

### Thesis 1: New Method for Evaluation of the Commuting based on Mobile Network Data

*I have designed a method to describe the commuting patterns of the population quantitatively. The method is based on the statistical detection of the home and work locations using anonymized mobile network data. I have validated the results by comparing them to census-based commuting analyses and found good agreement between the determined mobility patterns and census-based data.*

I have compared the detected population of the districts of Budapest and the settlements of the agglomeration with the population data of KSH [143] and the correlation coefficient (Pearson's R) was 0.9213. The commuting from the sectors of the agglomeration to the districts of Budapest was also evaluated, based on [146]. Lakatos and Kapitány analyzed commuting directions in the case of 14 settlements. Using the census data, they determined to which district group the inhabitants of these settlements commute. I have replicated this analysis using mobile network data, and the correlation was between 0.8488 (Érd) and 0.971 (Dunaharaszti). Figure 9.1, summarizes the correlation coefficients of this analysis. Figure 9.2, shows the detailed results of Érd and Dunaharaszti along with the results of [146]. As can be seen, the results fit into the trends of the last three censuses.

Koltai and Varró analyzed the commuters' composition based on the age from the sectors of the agglomeration to the district groups of Budapest [147]. The results were compared to theirs, and a strong correlation was found in this regard as well: Pearson's R was 0.8977.

My publication pertaining to this thesis: [151*].

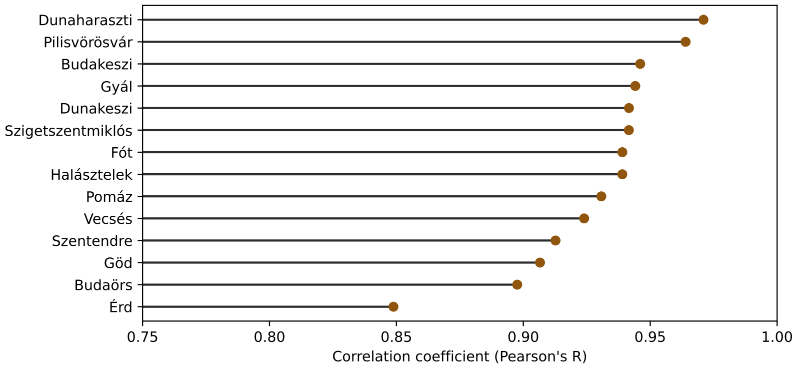
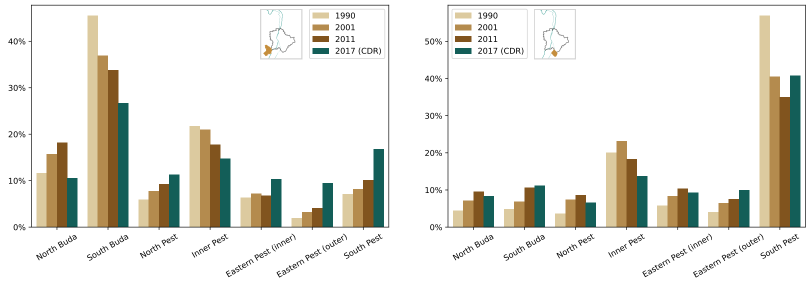


**Figure 9.1.:** Correlation coefficients (Pearson's R) of commuting distribution with [146].



**(a)** **(b)**

**Figure 9.2.:** Commuting to the seven districts groups of Budapest from Érd (**a**) and Dunaharaszti (**b**), comparing census (1990, 2001 and 2011) and mobile network data.

## Thesis 2: Correlation between Home and Workplace Price-levels

*Using anonymous mobile network data, I have demonstrated that people living in a less expensive neighborhood usually work in a less expensive area, based on housing prices of the home and the work locations. It has also been presented that people, who live in a more expensive neighborhood, tend to work in a more expensive area.*

Figure 9.3a shows a moderate correlation between the home-work distance and the price level of the home location. Figure 9.3b illustrates the correlation between the price level of the home and the work location, supporting the thesis.

My publications pertaining to this thesis: [151*].

## Thesis 3: New Indicators for Characterizing Mobility Customs

*I have introduced new indicators for quantitative evaluation of wake-up time and bedtime in an urban environment. The wake-up and bedtime conditions were determined by the rate of mobile network activity in the morning and evening hours. Two subscriber aggregation methods (area and inhabitant-based) have been developed to determine the wake-up characteristics of a geographical area or a group of subscribers.*

In the area-based approach, the activity records are aggregated that take place in a given cell, regardless of which SIM cards produce them. In the inhabitant-based approach, activity records produced by the

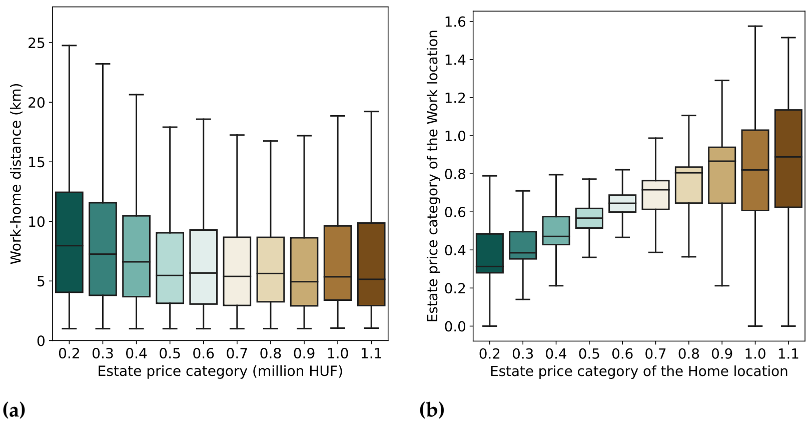


**Figure 9.3.:** Home-work distances (**a**) and the workplace housing prices in contrast to the housing prices of the workers' home locations (**b**).

inhabitants of the given cell were considered, regardless of where the activity occurred.

Bedtime, the counterpart of this indicator, describes when a group of people cease to use the mobile network in the evenings. The day length can be estimated using the wake-up and bedtime, as this indicator correlates with the astronomical day lengths. The Sun went under at 19:33 on the 15th of April 2017, and at 20:42 on the 15th of June 2016, which is a 69-minute difference. The average workday bedtime values are 19:43 and 20:47, respectively. Moreover, the length of the working hour can also be estimated using the inhabitant-based approach.

My publications pertaining to this thesis: [116*, 148*].

## Thesis 4: New method to Analyze the Correlation of Mobility and Socioeconomic Status

*I have designed a method using Principal Component Analysis to evaluate socioeconomic status depending on the indicators of human mobility. Housing prices have been used to characterize the socioeconomic status of the population. I have found differences in the mobility customs within the different socioeconomic classes, so that the socioeconomic status can be inferred from the mobility.*

Figure 9.4, shows the first two components of an unsupervised PCA analysis. The brown colors represent the workdays, the green ones the holidays, and they form two separate clusters. Marker size represents the housing price at the home location. While there is a marked tendency along the PC2 axis as the markers increase, the PC1 axis separates the estate price of the workplaces.

My publication pertaining to this thesis: [151*]

## Thesis 5: Introduction of Cellphone Price as Socioeconomic Status Indicator

*I have fused cellphone prices and release dates with the mobile network data to analyze the mobility customs in contrast to the price and the age of the subscribers' cellphone. I found that the cellphone price and age are eligible to characterize a subscriber's socioeconomic status.*

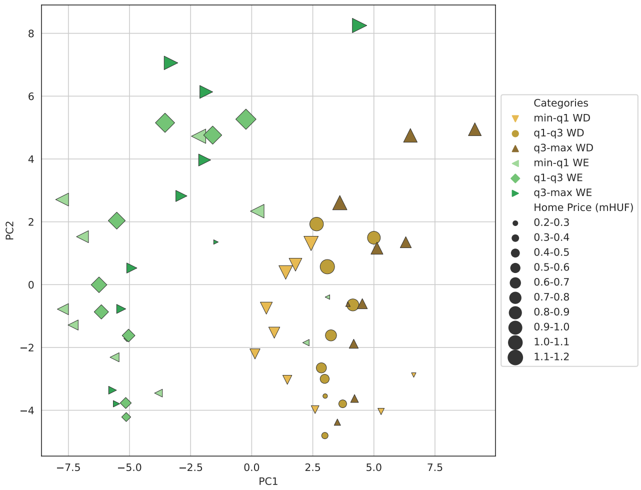


**Figure 9.4.:** Scatter plot of the 2-component PCA. Marker size indicates the home price category, the type denotes work price category and the color refers to Weekdays or Weekends.

Figure 9.5 does not only show that the phone price forms clusters but also reveals the differences between workdays and holidays in the mobility, which is the most notable in the case of the subscribers with the least expensive cellphones.

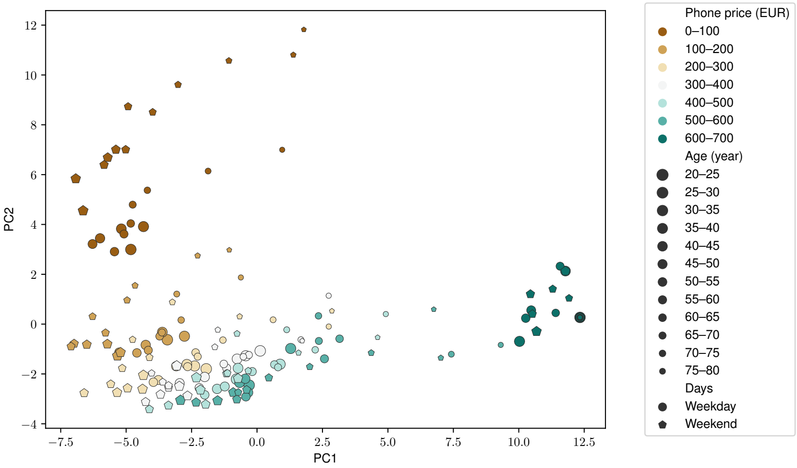


**Figure 9.5.:** Scatter plot of the 2-component PCA. Marker size indicates subscriber age category, the color represents the phone price category and the workdays and holidays are distinguished by the marker type.

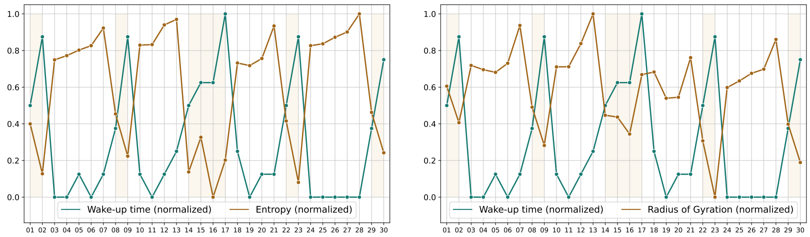
My publication pertaining to this thesis: [152*].

## Thesis 6: The Relation of Wake-up Time and Socioeconomic Status

*I demonstrated a relationship between the wake-up time and the mobility customs, as well as the socioeconomic status. The subscribers living in less expensive apartments get up earlier than those who live in pricier neighborhoods. The same tendency holds regarding mobile phone prices: subscribers who own more expensive cellphones tend to get up later.*

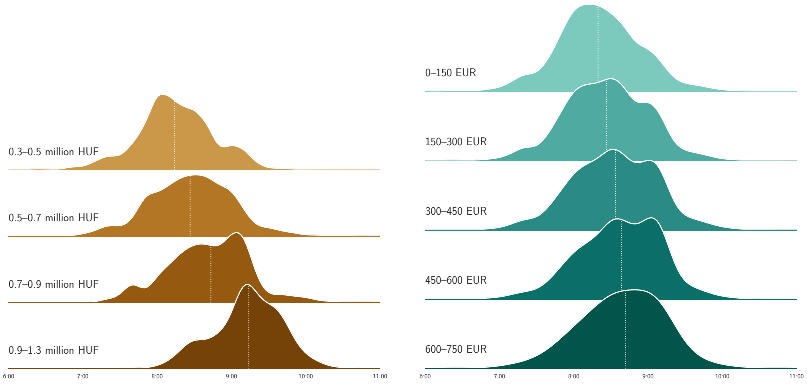
Figure 9.6 shows the daily wake-up times in contrast to daily mobility indicators. A strong negative correlation was found, especially in the case of Entropy.

Figure 9.7a and 9.7b, illustrate the wake-up time distribution by housing price and phone price categories, respectively. Although there is an increasing tendency in both cases, the connection between the housing price and the wake-up time is stronger.



**(a)** **(b)**

**Figure 9.6.:** Normalized, inhabitant-based wake-up times in contrast of the normalized daily Entropy (**a**) and Radius of Gyration (**b**). Pearson's Rs are −0.9019 and −0.6869, respectively.



**(a)** **(b)**

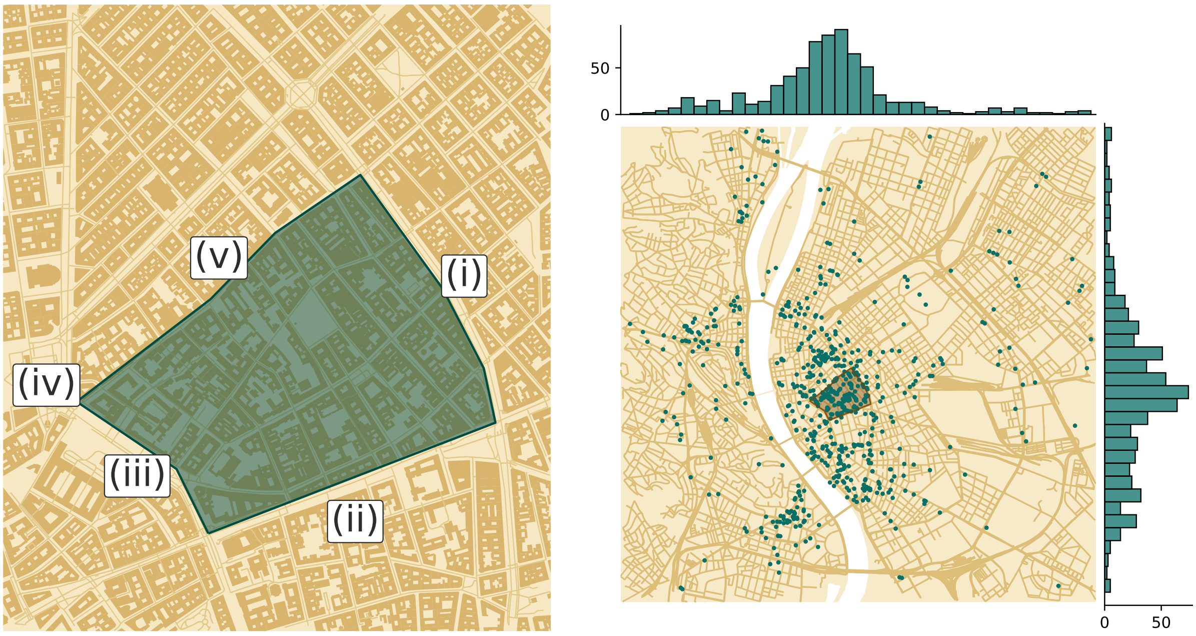
**Figure 9.7.:** Wake-up time distribution by housing price (**a**) and phone price (**b**) categories, with medians denoted by dotted vertical lines.

My publication pertaining to this thesis: [116*].

# Party District A.

The so-called party district is not an official city area with definite borders. Usually, the area bounded by the Károly Boulevard (iii), Király Street (v), Erzsébet Boulevard (i), and Rákóczi Street (ii) is referred to as "party district" (see Figure A.1a). This area is famous for ruin-pubs and nightlife, which also involves not being the most active area early in the morning. The vivid nightlife of downtown is not confined only to this area. The Deák Ferenc Square (iv), the northern side of the Király Street (v), and the eastern side of the Erzsébet Boulevard (i) can also be considered part of the party district.

Actually, how big is the party district? To answer this question, the points of interest (POI) have been downloaded from the OpenStreetMap (OSM) in the categories of "bar", "pub", "biergarten" and "nightclub", using the OSM terminology, and plotted to a map (see Figure A.1b). There are 192 bars, 472 pubs, 12 beer gardens, and 31 nightclubs in the displayed, 9 km by 9 km part of Budapest, and the area from Figure A.1a is also highlighted. Note that the POIs are queried in 2021 and cannot show the actual state of 2017. However, only the tendencies are important, which did not change fundamentally in the last decade. The concentration of bars and pubs is the highest within the highlighted "party district" area, but still very high on the whole Pest-side of the downtown, and there are also three smaller groups on the Buda-side of the city.



**(a)**

**(b)**

**Figure A.1.:** The borders of the party district (**a**), within District 7; and bars, pubs, beer gardens and nightclubs in Budapest downtown (**b**), based on OpenStreetMap data.

# Bibliography

[1] Michael Batty et al. “Smart cities of the future”. In: *The European Physical Journal Special Topics* 214.1 (2012), pp. 481–518.

[2] Tamás Tettamanti and István Varga. “Mobile phone location area based traffic flow estimation in urban road traffic”. In: *Columbia International Publishing, Advances in Civil and Environmental Engineering* 1.1 (2014), pp. 1–15.

[3] Tamás Egedy and Bence Ságvári. “Urban geographical patterns of the relationship between mobile communication, social networks and economic development–the case of Hungary”. In: *Hungarian Geographical Bulletin* 70.2 (2021), pp. 129–148.

[4] Miklos Szocska et al. “Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis”. In: *Scientific Reports* 11.1 (2021), pp. 1–9.

[5] Luca Pappalardo et al. “Returners and explorers dichotomy in human mobility”. In: *Nature communications* 6 (2015), p. 8166.

[6] Alket Cecaj, Marco Mamei, and Nicola Bicocchi. “Re-identification of anonymized CDR datasets using social network data”. In: *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE. 2014, pp. 237–242.

[7] Zhiren Huang et al. “Modeling real-time human mobility based on mobile phone and transportation data fusion”. In: *Transportation research part C: emerging technologies* 96 (2018), pp. 251–269.

[8] Angelo Furno et al. “Fusing GPS probe and mobile phone data for enhanced land-use detection”. In: *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE. 2017, pp. 693–698.

[9] Luca Pappalardo et al. “Using big data to study the link between human mobility and socio-economic development”. In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE. 2015, pp. 871–878.

[10] 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.

[11] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. “Predicting poverty and wealth from mobile phone metadata”. In: *Science* 350.6264 (2015), pp. 1073–1076.

[12] 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.

[13] Ambrus Kiss and Zsolt Matyusz. “Az ingázás, mint forgalomkeltő tényező”. In: *Munkaügyi szemle* 59.5 (2015), pp. 20–34.

[14] Andrew McAfee et al. “Big data: the management revolution”. In: *Harvard business review* 90.10 (2012), pp. 60–68.

[15] Matt Williams. *Meso-computing and meso-data: the forgotten middle*. (Online; accessed on 19 May 2021). Aug. 2020. URL: `https://milliams.com/posts/2020/mesocomputing/`.

[16] Luca Pappalardo et al. “Evaluation of home detection algorithms on mobile phone data using individual-level ground truth”. In: *EPJ data science* 10.1 (2021), p. 29.

[17] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. “Understanding individual human mobility patterns”. In: *nature* 453.7196 (2008), p. 779.

[18] Luca Pappalardo et al. “An analytical framework to nowcast well-being using mobile phone data”. In: *International Journal of Data Science and Analytics* 2.1-2 (2016), pp. 75–92.

[19] Clémentine Cottineau and Maarten Vanhoof. “Mobile phone indicators and their relation to the socioeconomic organisation of cities”. In: *ISPRS International Journal of Geo-Information* 8.1 (2019), p. 19.

[20] Vincent A Traag et al. "Social event detection in massive mobile phone data using probabilistic location inference". In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE. 2011, pp. 625–628.

[21] Faber Henrique Zacarias Xavier et al. "Analyzing the workload dynamics of a mobile phone network in large scale events". In: *Proceedings of the first workshop on urban networking*. 2012, pp. 37–42.

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

[23] Barbara Furletti et al. "Discovering and understanding city events with big data: the case of rome". In: *Information* 8.3 (2017), p. 74.

[24] Dániel Kondor et al. "Prediction limits of mobile phone activity modelling". In: *Royal Society open science* 4.2 (2017), p. 160900.

[25] Humberto T Marques-Neto et al. "Understanding human mobility and workload dynamics due to different large-scale events using mobile phone data". In: *Journal of Network and Systems Management* 26.4 (2018), pp. 1079–1100.

[26] 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*. Ed. by Hocine Cherifi et al. Cham: Springer International Publishing, 2020, pp. 415–426.

[27] Assaf Rotman and Michael Shalev. "Using Location Data from Mobile Phones to Study Participation in Mass Protests". In: *Sociological Methods & Research* (2020), p. 0049124120914926.

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

[29] 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), e0156794.

[30] Yang Xu et al. "Towards a multidimensional view of tourist mobility patterns in cities: A mobile phone data perspective". In: *Computers, Environment and urban systems* 86 (2021), p. 101593.

[31] Yang Xu et al. "Understanding the movement predictability of international travelers using a nationwide mobile phone dataset collected in South Korea". In: *Computers, Environment and Urban Systems* 92 (2022), p. 101753.

[32] Chen Qian et al. "Using mobile phone data to determine spatial correlations between tourism facilities". In: *Journal of Transport Geography* 92 (2021), p. 103018.

[33] Sanja Brdar et al. "Unveiling spatial epidemiology of HIV with mobile phone data". In: *Scientific reports* 6.1 (2016), pp. 1–13.

[34] Marcel Salathé. "Digital epidemiology: what is it, and where is it going?" In: *Life sciences, society and policy* 14.1 (2018), pp. 1–5.

[35] Hyeoun-Ae Park et al. "Digital epidemiology: use of digital data collected for non-epidemiological purposes in epidemiological studies". In: *Healthcare informatics research* 24.4 (2018), p. 253.

[36] Elias Willberg et al. "Escaping from cities during the covid-19 crisis: Using mobile phone data to trace mobility in finland". In: *ISPRS International Journal of Geo-Information* 10.2 (2021), p. 103.

[37] Gustavo Romanillos et al. "The city turned off: Urban dynamics during the COVID-19 pandemic based on mobile phone data". In: *Applied Geography* 134 (2021), p. 102524.

[38] Won Do Lee, Matthias Qian, and Tim Schwanen. "The association between socioeconomic status and mobility reductions in the early stage of England's COVID-19 epidemic". In: *Health & Place* (2021), p. 102563.

[39] Matan Yechezkel et al. "Human mobility and poverty as key drivers of COVID-19 transmission and control". In: *BMC public health* 21.1 (2021), pp. 1–13.

[40] Hamid Khataee et al. "Effects of social distancing on the spreading of COVID-19 inferred from mobile phone data". In: *Scientific Reports* 11.1 (2021), pp. 1–9.

[41] Kristi Bushman, Konstantinos Pelechrinis, and Alexandros Labrinidis. "Effectiveness and compliance to social distancing during COVID-19". In: *arXiv preprint arXiv:2006.12720* (2020).

[42] Song Gao et al. "Association of mobile phone location data indications of travel and stay-at-home mandates with covid-19 infection rates in the us". In: *JAMA network open* 3.9 (2020), e2020485–e2020485.

[43] Songhua Hu et al. "A big-data driven approach to analyzing and modeling human mobility trend under non-pharmaceutical interventions during COVID-19 pandemic". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102955.

[44] Ahmad Ilderim Tokey. "Spatial association of mobility and COVID-19 infection rate in the USA: A county-level study using mobile phone location data". In: *Journal of Transport & Health* 22 (2021), p. 101135.

[45] Lorenzo Lucchini et al. "Living in a pandemic: changes in mobility routines, social activity and adherence to COVID-19 protective measures". In: *Scientific Reports* 11.1 (2021), pp. 1–12.

[46] Takahiro Yabe et al. "Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic". In: *Scientific reports* 10.1 (2020), pp. 1–9.

[47] Adam Sadowski et al. "Big data insight on global mobility during the Covid-19 pandemic lockdown". In: *Journal of big Data* 8.1 (2021), pp. 1–33.

[48] Nuria Oliver et al. *Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle*. 2020.

[49] Balázs Cs Csáji et al. "Exploring the mobility of mobile phone users". In: *Physica A: statistical mechanics and its applications* 392.6 (2013), pp. 1459–1473.

[50] Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore". In: *IEEE Transactions on Big Data* 3.2 (2017), pp. 208–219.

[51] Pierdomenico Fiadino et al. "Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the" Always Connected Era"". In: *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*. 2017, pp. 43–48.

[52] Guilherme Augusto Zagatti et al. "A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR". In: *Development Engineering* 3 (2018), pp. 133–165.

[53] Marco Mamei et al. "Evaluating origin–destination matrices obtained from CDR data". In: *Sensors* 19.20 (2019), p. 4470.

[54] Hugo Barbosa et al. "Uncovering the socioeconomic facets of human mobility". In: *Scientific reports* 11.1 (2021), pp. 1–13.

[55] Qi Wang et al. "Urban mobility and neighborhood isolation in America's 50 largest cities". In: *Proceedings of the National Academy of Sciences* 115.30 (2018), pp. 7735–7740.

[56] Eszter Bokányi et al. "Universal patterns of long-distance commuting and social assortativity in cities". In: *Scientific reports* 11.1 (2021), pp. 1–10.

[57] Mi Diao et al. "Inferring individual daily activities from mobile phone traces: A Boston example". In: *Environment and Planning B: Planning and Design* 43.5 (2016), pp. 920–940.

[58] Yang Xu et al. "Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach". In: *Transportation* 42.4 (2015), pp. 625–646.

[59] Maarten Vanhoof et al. "Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics". In: *Journal of official statistics* 34.4 (2018), pp. 935–960.

[60] Ling Yin, Nan Lin, and Zhiyuan Zhao. "Mining daily activity chains from large-scale mobile phone location data". In: *Cities* 109 (2021), p. 103013.

[61] Teodoro Dannemann, Boris Sotomayor-Gómez, and Horacio Samaniego. "The time geography of segregation during working hours". In: *Royal Society open science* 5.10 (2018), p. 180749.

[62] Roberto Trasarti et al. "Discovering urban and country dynamics from mobile phone data with spatial correlation patterns". In: *Telecommunications Policy* 39.3-4 (2015), pp. 347–362.

[63] Kwang-Sub Lee et al. "Urban spatiotemporal analysis using mobile phone data: Case study of medium-and large-sized Korean cities". In: *Habitat International* 73 (2018), pp. 6–15.

[64] Mohammadhossein Ghahramani, MengChu Zhou, and Chi Tin Hon. "Mobile phone data analysis: A spatial exploration toward hotspot detection". In: *IEEE Transactions on Automation Science and Engineering* 16.1 (2018), pp. 351–362.

[65] Mohammadhossein Ghahramani, MengChu Zhou, and Chi Tin Hon. "Extracting significant mobile phone interaction patterns based on community structures". In: *IEEE Transactions on Intelligent Transportation Systems* 20.3 (2018), pp. 1031–1041.

[66] Zide Fan et al. "Estimation of urban crowd flux based on mobile phone location data: A case study of Beijing, China". In: *Computers, Environment and Urban Systems* 69 (2018), pp. 114–123.

[67] Linglin Ni, Xiaokun Cara Wang, and Xiqun Michael Chen. "A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data". In: *Transportation research part C: emerging technologies* 86 (2018), pp. 510–526.

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

[69] Rahul Goel, Rajesh Sharma, and Anto Aasa. "Understanding gender segregation through Call Data Records: An Estonian case study". In: *Plos one* 16.3 (2021), e0248212.

[70] 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.

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

[72] Chandreyee Roy et al. "Morningness–eveningness assessment from mobile phone communication analysis". In: *Scientific Reports* 11.1 (2021), pp. 1–13.

[73] Susan Hanson and Perry Hanson. "The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics". In: *Economic geography* 57.4 (1981), pp. 332–347.

[74] Mei-Po Kwan. "Gender, the home-work link, and space-time patterns of nonemployment activities". In: *Economic geography* 75.4 (1999), pp. 370–394.

[75] Thiri AUNG, Ko Ko LWIN, Yoshihide SEKIMOTO, et al. "Identification and Classification of Land Use Types in Yangon City by Using Mobile Call Detail Records (CDRs) Data". In: *Journal of the Eastern Asia Society for Transportation Studies* 13 (2019), pp. 1114–1133.

[76] Neeti Pokhriyal and Damien Christophe Jacques. "Combining disparate data sources for improved poverty prediction and mapping". In: *Proceedings of the National Academy of Sciences* 114.46 (2017), E9783–E9792.

[77] Sanja Šćepanović et al. "Mobile phone call data as a regional socio-economic proxy indicator". In: *PloS one* 10.4 (2015).

[78] 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.

[79] Galo Castillo et al. "The silence of the cantons: Estimating villages socioeconomic status through mobile phones data". In: *2018 International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE. 2018, pp. 172–178.

[80] Iñaki Ucar et al. "News or social media? Socio-economic divide of mobile service consumption". In: *Journal of the Royal Society Interface* 18.185 (2021), p. 20210350.

[81] Salvatore Vilella et al. "News and the city: understanding online press consumption patterns through mobile data". In: *EPJ Data Science* 9.1 (2020), pp. 1–18.

[82] Mariano G Beiró et al. "Shopping mall attraction and social mixing at a city scale". In: *EPJ Data Science* 7 (2018), pp. 1–21.

[83] 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.

[84] Marco De Nadai et al. "Socio-economic, built environment, and mobility conditions associated with crime: a study of multiple cities". In: *Scientific reports* 10.1 (2020), pp. 1–12.

[85] Yannick Leo et al. "Socioeconomic correlations and stratification in social-communication networks". In: *Journal of The Royal Society Interface* 13.125 (2016), p. 20160598.

[86] Kevin S Kung et al. "Exploring universal patterns in human home-work commuting from mobile phone data". In: *PloS one* 9.6 (2014), e96180.

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

[88] Albert-Laszlo Barabasi. "The origin of bursts and heavy tails in human dynamics". In: *Nature* 435.7039 (2005), pp. 207–211.

[89] Hang-Hyun Jo et al. "Circadian pattern and burstiness in mobile phone communication". In: *New Journal of Physics* 14.1 (2012), p. 013055.

[90] Andrea Cuttone et al. "Sensiblesleep: A bayesian model for learning sleep patterns from smartphone events". In: *PloS one* 12.1 (2017), e0169901.

[91] Talayeh Aledavood, Sune Lehmann, and Jari Saramäki. "Social network differences of chronotypes identified from mobile phone data". In: *EPJ Data Science* 7.1 (2018), p. 46.

[92] Taha Yasseri, Robert Sumi, and János Kertész. "Circadian patterns of wikipedia editorial activity: A demographic analysis". In: *PloS one* 7.1 (2012), e30091.

[93] Taha Yasseri, Giovanni Quattrone, and Afra Mashhadi. "Temporal analysis of activity patterns of editors in collaborative mapping project of OpenStreetMap". In: *Proceedings of the 9th International Symposium on Open Collaboration*. 2013, pp. 1–4.

[94] Fabon Dzogang, Stafford Lightman, and Nello Cristianini. "Circadian mood variations in Twitter content". In: *Brain and neuroscience advances* 1 (2017), p. 2398212817744501.

[95] Rein Ahas et al. "Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data". In: *Transportation Research Part C: Emerging Technologies* 18.1 (2010), pp. 45–54.

[96] Laura Lotero et al. "Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes". In: *Royal Society open science* 3.10 (2016), p. 150654.

[97] Daniel Monsivais et al. "Seasonal and geographical impact on human resting periods". In: *Scientific reports* 7.1 (2017), pp. 1–10.

[98] Daniel Monsivais et al. "Tracking urban human activity from mobile phone calling patterns". In: *PLoS computational biology* 13.11 (2017), e1005824.

[99] T Alakörkkö and J Saramäki. "Circadian rhythms in temporal-network connectivity". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.9 (2020), p. 093115.

[100] Aleksey Ogulenko et al. "Probabilistic positioning in mobile phone network and its consequences for the privacy of mobility data". In: *Computers, Environment and Urban Systems* 85 (2021), p. 101550.

[101] Fabio Ricciato et al. "Towards a methodological framework for estimating present population density from mobile network operator data". In: *Pervasive and Mobile Computing* 68 (2020), p. 101263.

[102] Wei Tu et al. "Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns". In: *International Journal of Geographical Information Science* 31.12 (2017), pp. 2331–2358.

[103] Alessandro Galeazzi et al. "Human mobility in response to COVID-19 in France, Italy and UK". In: *Scientific reports* 11.1 (2021), pp. 1–10.

[104] Harry ER Shepherd et al. “Domestic and international mobility trends in the United Kingdom during the COVID-19 pandemic: an analysis of facebook data”. In: *International journal of health geographics* 20.1 (2021), pp. 1–13.

[105] Johannes Scholz and Janja Jeznik. “Evaluating Geo-Tagged Twitter Data to Analyze Tourist Flows in Styria, Austria”. In: *ISPRS International Journal of Geo-Information* 9.11 (2020), p. 681.

[106] Bartosz Hawelka et al. “Geo-located Twitter as proxy for global mobility patterns”. In: *Cartography and Geographic Information Science* 41.3 (2014), pp. 260–271.

[107] Raja Jurdak et al. “Understanding human mobility from Twitter”. In: *PloS one* 10.7 (2015), e0131469.

[108] Dong Li and Jiming Liu. “Uncovering the relationship between point-of-interests-related human mobility and socioeconomic status”. In: *Telematics and Informatics* 39 (2019), pp. 49–63.

[109] National Media and Infocommunications Authority, Hungary. *A Nemzeti Média- és Hírközlési Hatóság mobilpiaci jelentése 2015. IV. – 2019. II. negyedév*. Tech. rep. 23-25. Ostrom u., Budapest 1015, Hungary: National Media and Infocommunications Authority, Dec. 2019.

[110] MM Al-Akaidi and H Ali. “Performance analysis of antenna sectorisation in cell breathing”. In: (2003).

[111] Központi Statisztikai Hivatal. *22.1.1.3. Népesség korév és nem szerint, január 1.* (Online; accessed on 20 May 2021). URL: `https://www.ksh.hu/stadat_files/nep/hu/nep0003.html`.

[112] Keszthelyi Christian. *Sziget Festival sees record attendance of 441,000.* (Online; Accessed on 5 July 2021). URL: `https://bbj.hu/budapest/culture/awards/sziget-festival-sees-record-attendance-of-441-000`.

[113] Mohit Sainani. *GSMArena Mobile Phone Devices*. (Online; Accessed on 28 June 2020). URL: `https://www.kaggle.com/msainani/gsmarena-mobile-devices`.

[114] Vijay Janapa Reddi, Hongil Yoon, and Allan Knies. “Two billion devices and counting”. In: *IEEE Micro* 38.1 (2018), pp. 6–21.

[115] Aidin Zehtab-Salmasi et al. “Multimodal price prediction”. In: *Annals of Data Science* (2021), pp. 1–17.

[117] Emil Protalinski. *iPhone prices from the original to iPhone X*. (Online; accessed on 14 February 2022). Sept. 2017. URL: `https://venturebeat.com/2017/09/12/iphone-prices-from-the-original-to-iphone-x/`.

[118] Ravindra Dissanayake and Thustan Amarasuriya. “Role of brand identity in developing global brands: A literature based review on case comparison between Apple iPhone vs Samsung smartphone brands”. In: *Research journal of business and management* 2.3 (2015), pp. 430–440.

[119] Központi Statisztikai Hivatal. *Budapest – Gazdaság és társadalom*. ISBN: 978-963-235-541-2. Keleti Károly utca 5–7., 1024 Budapest, Hungary: Központi Statisztikai Hivatal, 2018.

[120] Hungarian Central Statistical Office. *Summary Tables (STADAT)*. (Online; accessed on 26 March 2022). URL: `https://www.ksh.hu/stadat_eng`.

[121] Visual Crossing Corporation. *Visual Crossing Weather (2016-2017)*. (Online; accessed on 30 November 2021). 2021. URL: `https://www.visualcrossing.com/`.

[122] Twitter. *Twitter API for Academic Research*. (Online; accessed on 26 March 2022). URL: `https://developer.twitter.com/en/products/twitter-api/academic-research`.

[123] Ed Summers et al. *DocNow/twarc: v2.9.5*. Version v2.9.5. Mar. 2022. DOI: `10.5281/zenodo.6327291`.

[124] Maarten Vanhoof et al. “Comparing regional patterns of individual movement using corrected mobility entropy”. In: *Journal of Urban Technology* 25.2 (2018), pp. 27–61.

[125] 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.

[126] Olivera Novović et al. “Uncovering the Relationship between Human Connectivity Dynamics and Land Use”. In: *ISPRS International Journal of Geo-Information* 9.3 (2020), p. 140.

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

[128] Lino Galiana, Benjamin Sakarovitch, and Zbigniew Smoreda. "Understanding socio-spatial segregation in French cities with mobile phone data". In: *Unpublished manuscript* (2018).

[129] Rein Ahas et al. "Using mobile positioning data to model locations meaningful to users of mobile phones". In: *Journal of urban technology* 17.1 (2010), pp. 3–27.

[130] Iva Bojic et al. "Choosing the right home location definition method for the given dataset". In: *International Conference on Social Informatics*. Springer. 2015, pp. 194–208.

[131] Eurostat. *Employed persons working at nights as a percentage of the total employment, by sex, age and professional status*. (Online; accessed on 31 March 2020). 2020. URL: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfsa_ewpnig&lang=en.

[132] Luca Pappalardo et al. *scikit-mobility*. Version 1.1.0. If you use this code, please cite this software. July 2019. DOI: 10.5281/zenodo.5979518.

[133] Md Tahmid Rashid and Dong Wang. "CovidSens: a vision on reliable social sensing for COVID-19". In: *Artificial intelligence review* 54.1 (2021), pp. 1–25.

[135] John Hopewell and Ed Meza. *Soccer's Euro 2016: Big Ratings, Strategic Losses*. (Online; Accessed on 15 July 2021). URL: https://variety.com/2016/tv/global/soccers-euro-2016-huge-ratings-nets-losses-1201812565/.

[136] Központi Statisztikai Hivatal. *22.1.1.1. A népesség, népmozgalom főbb mutatói*. (Online; accessed on 15 June 2021). URL: https://www.ksh.hu/stadat_files/nep/hu/nep0001.html.

[137] hirado.hu/MTI. *Csoportgyőztes az M4 Sport is*. (Online; Accessed on 5 July 2021). URL: https://hirado.hu/2016/06/24/csoportgyoztes-az-m4-sport/.

[138] hirado.hu/MTI. *Tízezrek ünnepelték a magyar válogatottat a Hősök terén*. (Online; Accessed on 5 October 2021). URL: https://hirado.hu/2016/06/27/hatalmas-fiesztara-keszul-este-budapest/.

[141] TheGuardian. *Thousands protest in Hungary over threat to Soros university*. (Online; Accessed on 1 August 2018). URL: https://www.theguardian.com/world/2017/apr/09/thousands-protest-in-hungary-over-bill-threat-to-soros-university.

[142] Dull, Szabolcs. *Ekkora tömeg régen vonult utcára Orbán ellen*. (Online; accessed on 4 March 2022). Apr. 2017. URL: https://index.hu/belfold/2017/04/10/ekkora_tomeg_regen_vonult_utcara_orban_ellen/.

[143] Központi Statisztikai Hivatal. *Calculated population data by settlement - Resident population in Hungary (2017 - 2020)*. (Online; Accessed on 22 June 2021). URL: http://statinfo.ksh.hu/Statinfo/QueryServlet?ha=NT6B01.

[144] Gábor Pálóczi. "Researching commuting to work using the methods of complex network analysis". In: *Regional Statistics* 6.1 (2016), pp. 3–22.

[145] Marc Barthelemy et al. "Characterization and modeling of weighted networks". In: *Physica a: Statistical mechanics and its applications* 346.1-2 (2005), pp. 34–43.

[146] Miklós Lakatos and Gabriella Kapitány. "Daily Mobility of Labour Force (Commuting) and Travel in Budapest and in the Metropolitan Agglomeration Based on Data of the Population Census. Part II". In: *Területi Statisztika* 56.2 (2016), pp. 209–239. DOI: 10.15196/TS560206.

[147] Luca Koltai and András Varró. "Ingázás a budapesti agglomerációban". In: *Új munkaügyi szemle* 1.3 (2020), pp. 26–37.

[149] Fred Espenak. *Solstices and Equinoxes: 2001 to 2050*. (Online; accessed on 16 December 2021). Feb. 2018. URL: http://astropixels.com/ephemeris/soleq2001.html.

[153] Lasantha Fernando et al. "Predicting population-level socio-economic characteristics using Call Detail Records (CDRs) in Sri Lanka". In: *Proceedings of the Fourth International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets*. 2018, pp. 1–12.

[154] Központi Statisztikai Hivatal. *12.8.1.9. A háztartások internetkapcsolat típusainak aránya*. (Online; Accessed on 15 September 2021). URL: https://www.ksh.hu/stadat_files/ikt/hu/ikt0047.html.

# Publications

[116*] 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. DOI: 10.3390/info13030114.

[134*] 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. DOI: 10.1109/SACI.2018.8440943.

[139*] Gergő Pintér, László Nádai, and Imre Felde. “Analysis of Mobility Patterns during a Large Social Event”. In: *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE. 2018, pp. 339–344. DOI: 10.1109/SISY.2018.8524674.

[140*] Gergő Pintér et al. “Activity Pattern Analysis of the Mobile Phone Network During a Large Social Event”. In: *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE. 2019, pp. 1–5. DOI: 10.1109/RIVF.2019.8713741.

[148*] Gergő Pintér and Imre Felde. “Evaluation of urban daily routines by using Mobile Phone Indicators”. In: *2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE. 2019, pp. 314–319. DOI: 10.1109/SACI46893.2019.9111495.

[150*] Gergő Pintér, Amir Mosavi, and Imre Felde. “Artificial intelligence for modeling real estate price using call detail records and hybrid machine learning approach”. In: *Entropy* 22.12 (2020), p. 1421. DOI: 10.3390/e22121421.

[151*] 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. DOI: 10.3390/ijgi10050328.

[152*] 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. DOI: 10.3390/info12110468.

If you immediately know the candlelight is fire, the meal was cooked a long time ago.

– Oma Desala in “Meridian”, Stargate SG-1