Utrecht University
Graduate School of Natural Sciences
Department of Information and Computing Sciences
MSc Artificial Intelligence

# Pixels to Policy: Unravelling Refugee Integration in Urban Istanbul with Mobile Phone Data

A Thesis By

**David Natarajan**
*8034551*

**Project supervisor** Prof. Dr. Albert Ali Salah
**Daily supervisor** Bilgeçağ Aydoğdu
**Second examiner** Dr. Michael Behrisch

# Abstract

In an era where traditional data sources fall short of capturing the intricacies of human mobility and settlement patterns, digital trace data, particularly from mobile phones, emerges as a pivotal resource for policymakers and humanitarians. This study underscores the imperative for policymakers to prioritise the welfare of migrants, especially refugees and asylum seekers, for the sustainable development of cities. By leveraging novel mobile phone extended detail records (xDRs), we delve into the social and economic conditions of Syrian and Afghan refugees in Istanbul, shedding light on their integration and segregation dynamics.

The examination of segregation indices such as dissimilarity and isolation over the day at an unprecedented granular level provides nuanced insights into the spatial distribution of migrant communities, offering policymakers a roadmap for targeted interventions. Additionally, employing behavioural analysis techniques, like that of eigenbehaviour analysis, enriches our understanding of the socio-economic determinants shaping refugee integration.

Our findings unveil the profound impact of COVID-19 lockdown measures on mobility patterns, revealing a significant convergence of residential and workplace locations among the populace. Notably, we identify an ethnic enclave of Afghans in the Zeytinburnu district, emphasising the importance of understanding localised dynamics. Furthermore, our analysis exposes disparities in segregation between Syrians and Afghans, with Syrians exhibiting higher workplace segregation and Afghans experiencing greater residential segregation. Moreover, both groups demonstrate characteristic working-class work habits, underscoring the need for tailored social welfare and housing initiatives to foster inclusive urban environments. Interestingly, our study uncovers a notable trend: Syrians and Afghans tend to reside in areas primarily characterised by informal housing known as 'Gecekondu', reflecting the challenges faced by these communities in accessing formal housing options.

By bridging the gap between data-driven research and policy implementation, this study advocates for holistic approaches that address the multifaceted challenges faced by migrant populations. As cities continue to evolve as hubs of diversity and opportunity, prioritising the well-being and inclusion of refugees and asylum seekers is not only a moral imperative but also a strategic investment in the social and economic resilience of urban landscapes. Through innovative methodologies and

interdisciplinary collaboration, we hope to pave the way for evidence-based policy that uphold the principles of equity, diversity, and sustainable development in the 21st-century urban paradigm.

This study is supported by European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 870661.

# Table of Contents

# 1. Introduction

## 1.1 Problem statement

With the rapid pace of urbanisation worldwide, the demand for sustainable, evidence-driven policy has never been greater. In particular, policy around migration and urban mobility will be instrumental in the success of nations' sustainable development ambitions (International Organisation for Migration (IOM), 2023).

Despite this growing imperative, decision-makers often encounter barriers due to the lack of high-quality and timely data. Traditional data sources like censuses and surveys often fail to capture the dynamic nature of migration and urban mobility, hindering policymakers' ability to respond effectively (International Organisation for Migration (IOM), 2023). Thus, addressing the dearth of reliable data is essential to inform policies that promote sustainable urban development and address the needs of migrant populations.

### 1.1.1 Barriers in policy: traditional data sources

Traditional data sources like censuses and targeted interviews form the bedrock of urban demography and provide essential evidence for designing policy. However, deficiencies in the accuracy, completeness, and timeliness of these sources are well documented and widespread Ahmad-Yar and Bircan (2021).

Censuses are often the largest peacetime data collection operations that countries undergo. The immense costs of this operation mean that censuses are typically conducted only once per decade leaving large gaps in time for which there is no data (Tatem et al., 2023). Additionally, censuses often cannot penetrate remote geographic areas, are blind to unregistered civilians, and struggle to capture short-term mobility patterns.

In conjunction, qualitative data collection strategies like surveys and targeted interviews are also expensive to commission, limited in scope, and are prone to selection and self-reporting bias. In the case of migration data, transnational inconsistencies make data comparison difficult or even impossible. These shortcomings are only exacerbated for poorer nations who do not have the resources to effectively administer these institutions.

Traditional data sources will remain irreplaceable assets in the arsenals of policy

makers, however the severe gaps they leave in our picture of population dynamics are numerous. It is therefore imperative to explore novel data sources intended to supplement and augment traditional ones.

### 1.1.2 Opportunities: novel data sources and digital methods

In this vein, big data sources, in particular digital trace data, have demonstrated a huge potential to fill data gaps left by traditional statistics and, in so doing, open new avenues of investigation into social phenomena. Digital trace data are generated as people interact with technology and leave behind "digital traces" in the form of social media posts, GPS trip summaries, and phone call records to name a few.

Digital trace data, by virtue of their abundance and the useful metadata that they carry like geospatial and demographic markers, are able to reveal behavioural patterns at a societal scale. They create the ability to observe human networking and mobility at an unprecedented resolution, with high regularity, across a wide span of time. Mobile phone data (MPD) are a particularly promising source of digital traces. They have been distinguished as a meaningful and rich data source for policy makers and many works have demonstrated their utility in a variety of real world scenarios.

Mobile phone data possess many characteristics that make them an ideal candidate for supplementing existing migration statistics. Importantly, mobile phones, and especially smart phones, are ubiquitous around the globe and recognised as essential communication tools. With smart phones people can stay in touch over vast distances, send remittances through top-ups, access internet hosted resources, facilitate commercial transactions, and a plethora of other functions that may otherwise require dedicated equipment and services. Smart phones empower communities to act independently and discretely while on the move, making them indispensable tools to economic migrants and refugees.

Because of the widespread usage of mobile phones across geographies, demographics, and cultures, MPD tends to be highly representative of populations. This provides grounds for researchers to use mobile phone data as a proxy for estimating migration and mobility. Furthermore, the spatial resolution of MPD, dictated by the density of cellular base stations, is tremendous and allows for fine measurements of individual mobility. MPD can also be aggregated at different spatial units allowing researchers to limit the risk of identifying customers and such aggregation ensures compatibility with a variety of geospatial data sources. Moreover, the relative simplicity of mobile data types and their standardisation in the telecommunication industry makes it much easier to compare MPD in an international context and reproduce methodologies

across countries (Vanhoof et al., 2020).

### 1.1.3 A growing literature: MPD in practice

Mobile phone call detail records (CDRs) are records maintained by telecommunication service providers that document the details of every call or SMS made or received by a mobile phone (Aydoğdu et al., 2021). These records typically include information such as: 1) call date and time, 2) caller and recipient numbers, 3) call duration, and 4) cell tower information. These records are held by the telecommunication service providers, as they are responsible for managing and maintaining the network infrastructure through which the calls are routed.

CDRs are primarily used for billing customers for the calls they make. The duration and destination of calls are used to calculate charges. Mobile phone call detail records can be processed in various ways for different purposes. Telecommunication companies typically analyse CDRs to understand calling patterns, network usage, and peak hours. This information helps them optimise their network infrastructure and plan capacity upgrades. Other than the mobile phone operator, in research we see a blooming literature on the use of mobile phone CDRs for a variety of applications including epidemiology, demography, public policy, and humanitarian interventions.

MPD has also seen successful deployment in a number of migration and mobility related tasks, often making use of ancillary data to draw conclusions about how migration interact with tertiary phenomena. For example, by combining CDRs and mobile money interactions, Lavelle-Hill et al. (2022) managed to predict migration into poorer neighbourhoods in Dar es Salaam, Tanzania. Two groups, Lu et al. (2016); Acosta et al. (2020), use mobile CDR to observe migration patterns in the wake of tropical storms that struck Bangladesh and Puerto Rico, respectively. Milusheva et al. (2018) find strong correlations between mobile CDR and official migration statistics in Namibia and Senegal. In addition, Bertoli et al. (2021) use mobile phone data to generate a picture of segregation and internal mobility of Syrian refugees in Turkey.

Just as other digital trace data can be used in a variety of contexts, so can MPD, and there is a large body of research that demonstrates the flexibility of this data bordering the realm of migration. Blumenstock et al. (2015) for instance use MPD to reconstruct the wealth distribution of mobile phone users in Rwanda based on their mobility. Mooses et al. (2020) investigated the cross-border mobility of different ethnolinguistic groups in Estonia. They discovered that the less wealthy Russian speaking minority participated in cross-border mobility more frequently and

concluded that international travel is not primarily driven by wealth. Yang et al. (2023) observed mobility data to determine latent activity behaviours of residents in a handful of US cities, learning that the behavioural profiles that they discovered were more predictive of daily visitation habits than demographic variables. Additionally, Salman et al. (2021) used mobile data to plan efficient routes for mobile health services to reach Syrian refugees working on farms in remote areas of Turkey.

### 1.1.4 Detecting Segregation

In a rapidly urbanising world it is imperative that governments adapt to the ebb and flow of migration into and out of their cities. The spatial organisation of culturally, linguistically, and ethnically different groups greatly determines the sustainable growth of a city and the economic outcomes of everyone residing there. Monitoring the distribution of wealth, housing, access to goods and services, etc. therefore takes high priority for policy makers. They may then take steps to support new migrants with housing and financing, as well as design interventions to reduce rates of poverty.

For international migrants, cultural, social, and linguistic integration in the destination country is, in many cases, critical for survival; amplifying employment opportunities and improving one's access to social services like healthcare and education. But the precarious position of economic migrants and refugees often translates to the creation of insular ethnic enclaves due to lack of access to the housing and job markets. While providing a sense of community and security these enclaves may also inhibit integration into the host society and, in its insularity, may escalate into ghettoisation.

This is especially true for the refugees arriving in Turkey as they do not share a common language and are additionally hindered by their legal status under the temporary and international protection regime that limits their access to social services and labour markets. In a report by the Marmara Municipalities Union Erdoğan (2022), it was identified that "the main problems experienced by Syrian temporary protection holders are poverty, being employed as unqualified, cheap labour and housing."

Integration, or conversely, segregation levels are challenging to observe straightforwardly since they are multidimensional phenomena that manifest in a variety of ways. Traditional segregation indices rely on calculations based on population data from census tracts which are often out of date and are not furbished with data on migrants, especially refugees and asylum seekers. Moreover, collecting info on the economic and social well-being of migrants is an endeavour that faces the same

challenges as running censuses and interviews. On the other hand, MPD grants us the opportunity to estimate integration through proxy indicators like the aforementioned segregation indices, the radius of gyration, behavioural indicators and other mobility characteristics.

Using Istanbul and the neighbouring province of Kocaeli as case studies, we try to estimate how segregation evolves in these urban areas over time and at very fine spatial resolutions. More specifically, we calculate the segregation indices of dissimilarity and isolation for the various neighbourhoods of these provinces over 24 hours using MPD aggregated over the whole of 2020. In addition, a behavioural analysis is done using dimensionality reduction techniques to determine if Syrians, Afghans, and Turks are easily distinguishable in their daily habits and whether this is tied to segregation.

### 1.1.5 Increasing temporal resolution with xDRs

CDRs typically contain information about voice calls and text messages, including details such as the anonymised IDs for the caller and recipient, call duration (when available), timestamp, and the location of the cell tower used during the call or text message.

On the other hand, Extended Detail Records (xDR) are more comprehensive and may include additional types of communication activities beyond voice calls and SMS. xDR can encompass a wider range of data, such as multimedia messages (MMS), data usage (e.g., internet browsing, app usage), and location information. This broader scope allows xDR to provide a more comprehensive view of mobile phone usage patterns and user behaviour on the network.

The use of xDR is very new and virtually non-existent in mobility research, leaving a gap in the literature. This research aims to validate the interoperability of these diverging mobile phone data types in the context of social science research. Additionally, we aim to investigate the trade-offs between computational effort and accuracy offered by the increased volume of xDR over CDR.

## 1.2 Research questions

In investigating the properties of xDR as an alternative data source we rely on the following research questions to guide us:

To what extent can mobile phone xDRs provide insights into the spatial distribution

of ethnic enclaves within Istanbul?

What can mobile phone xDRs tell us about the nuance of residential and workplaces segregation as it evolves through the day?

Do mobile phone xDRs capture significant differences in the behavioural patterns of different segments in the population?

## 1.3 Contributions of the thesis

My contributions can be summarised as follows:

- Reproduction of big data methodologies with novel mobile phone xDR data.
- New insights into the nuances of xDRs and how to process them to remove bias.
- High resolution depiction of the distribution and segregation of Syrians and Afghans in the urban environment of Istanbul.
- Behavioural analysis of Syrians, Afghans, and Turks in Istanbul using passive mobile phone data.

## 1.4 Structure of the thesis

Chapter 2 dives into the work that has already been done to make this thesis possible. It gives an overview of research in the digital social sciences, how big data methods continue to evolve in the field, and how this applies to migration and mobility in particular. We discuss mobile phone based research in more depth, including how it can help us answer our research questions, and how an ethics around the use of mobile phone data is important to consider and uphold.

Chapter 3 then proceeds to lay the groundwork for our analyses by describing the scenario that we investigate and describing the data that will drive said analyses. It goes on to explicit how we process the mobile phone data and describes how it is employed thereafter in the methodology section. Here we also illustrate the output of our method and interpret the results.

Finally, chapter 4 makes some concluding remarks about the state of the field, the results we obtain, and future work that we hope this research generates.

Appendix A contains a district map of Istanbul and a close-up choropleth map of its

official population, which may be useful to the reader as they proceed through this thesis.

# 2. Related Work

In the introduction a number of concepts were defined that highlight the motivation and aims of this research. In this chapter the ambition is to position this research within the broader social science literature and set a historical frame for the contributions of the thesis. Firstly, we consider the data-driven trajectory of migration analysis, the methods that utilise traditional statistics, and how these have evolved to leverage big data. Secondly, we provide an overview of mobile phone data, its varieties, and how it came to be a staple in migration research. Thirdly, we look at the different methods that have emerged from the use of mobile phone data and the limitations of using MPD. Finally, arguments around the ethics of big data processing are considered in the context of mobility data for social science research.

## 2.1 Big data based prediction of migration and mobility

Much can be extracted from official data; censuses capture migrant stocks and flows as they ask participants to indicate country of birth and past moves, surveys provide important information on migrant experiences, and administrative data from border checkpoints and immigration offices also collect a large amount of data on a daily basis that is useful for enumerating frequented migration routes and registration of residencies. These data sources have been used to study a variety of social effects related to migration, for instance, several studies have investigated differences between mixed and non-mixed marriages (Smith et al., 2012; Agliari et al., 2018), labour market integration of migrants (Spörlein and Van Tubergen, 2014), language adoption (Van Tubergen and Wierenga, 2011), educational expectations in classrooms (Minello, 2014), and economic prosperity (Alesina et al., 2016).

But, as set out in the introduction, censuses and surveys are often not representative of marginal groups, are few and far between, and have a low spatial resolution. In addition, administrative data collected at border checkpoints and registration offices are rarely centralised, often recorded in analogue with pen and paper, and face standardisation and bias issues. Research that relies on traditional data sources inherit these flaws making it difficult or perhaps impossible, therefore, to build a real-time picture of mobility - the new imperative in migration research and policy (Sîrbu et al., 2021).

The act of predicting migration behaviour in real time, sometimes called nowcasting, can, however, be achieved with big data. They come in many varieties and from many origins but the main sources are (1) social media, (2) satellite imagery, (3) retail data, and (4) mobile phone data. Between the source, the journey, and the destination, there are many aspects that constitute a migration that have different causes and effects (Sîrbu et al., 2021). In this thesis, we focus on specific aspects of migration, particularly the journey and destination phases. We observe the integration and experiences of migrants at their destinations relative to the existing populations. While this research provides useful indicators, it does not encompass the full qualitative depth required for a comprehensive understanding of migration, which is the domain of specialised migration scholars.

These different tasks benefit from different data and approaches, for example, geospatial data is ideal for witnessing migrant trajectories and social media data can be useful for understanding the sociological elements of integration like sentiment and networking. Here we outline some of the main big data types used in migration research, with the exception of mobile data that has been covered in the introduction and that will be expanded upon in the following section.

### 2.1.1 Social media data

Large social media sites like Twitter and Facebook host immense quantities of user data in the form of posts, comments, reactions, and other site interactions that are publicly accessible via APIs or in company curated datasets. Social media data have become more prominent in migration research as they lend themselves to the construction of migrant stocks and flows.

Twitter data for instance contain geolocated and timestamped messages that make localising individuals over time a possibility. Additionally, Twitter features a small, consistent character limit on each post that makes the computational difficulty of natural language processing and semantic analysis much easier to estimate.

Zagheni et al. (2014) for example, use geolocated tweets to analyse migration flows and Moise et al. (2016) compare the language of tweets to their geolocation to approximate the origin of migrants. Moreover, by processing the content of tweets, Coletto et al. (2017) was able to perform sentiment analysis on two polarising migration topics: the "refugee crisis" and "Brexit", demonstrating a means to track locally debated topics and the sentiment polarity towards them over time.

There are limitations here as well in that only a portion of Twitter users enable

geolocation on their tweets and the Twitter user base is generally not representative of population demographics. Nevertheless, Twitter data has been demonstrated to reveal real time migration dynamics that would otherwise be inscrutable from traditional sources.

Similar research has been conducted with data from other platforms. Google Trends[1], a platform that tracks the frequency of search terms in a given country, can be used to predict how many people have the intention to migrate while delivering insights more quickly than official data is released, additionally improving on conventional model estimates (Böhme et al., 2020).

Vieira et al. (2022) have demonstrated how data for advertisers from Facebook and employment, education, and residence data from LinkedIn, as well as language usage data from both sites, can be used to quantify migration, and to study levels of assimilation of migrants based on the interests they express online. Though related, Pötzschke and Weiß (2021) instead conducted an ad campaign on Facebook and Instagram to investigate the possibility of recruiting a nonprobability sample of German emigrants internationally, finding that 98 percent of respondents belonged to the target population.

System drift, self-reporting-, and algorithmic bias were each cited as limitations of these methods, highlighting the need for care in mapping data dependencies and in construct and measurement validity. Documenting data collection and processing steps can also aid in improving the longevity of data sources (Salah et al., 2022).

It is worth discussing the recent developments surrounding Elon Musk's purchasing of Twitter and their impacts on the computational social sciences. Even before the platform was rebranded as "X" Musk had enacted several policy changes including the closure of the research API which has and will continue to have deep repercussions for social sciences research if he decides not to reverse course. Since the opening of it's research API Twitter has become a staple data source in several fields to the point that many third parties have implemented whole workflows to leverage the API that are ubiquitous and popular.

As access to an important data source such as Twitter closes to the public, it will be worthwhile to witness how the community begins to develop and innovate research pipelines for other social media platforms like Instagram that have different principles and mediums of communication, as well as different users and content. For instance,

[1] https://trends.google.com/trends/

with the proliferation of large language models and their various counterparts, the use of image2text processing where stylised text in images are converted to a computer readable format is becoming commonplace after demonstrating itself as a resilient and powerful technology that subverts the medium of images.

### 2.1.2 Satellite data

Big data is also generated by digital sensors and measurement equipment like that of satellites and drones that take high resolution images of the earths surface. This sort of data are unique in their ability to facilitate visual confirmation of environmental conditions, economic factors, and population estimates (Bircan, 2022).

Quinn et al. (2018) highlight the humanitarian applications for remote sensing data with deep learning approaches to estimating the number of structures in refugee settlements across Africa and the Middle East. They note that there is a lot of variability in the characteristics of imagery captured from different sensors and regions and researchers must often work with a relatively small amount of pixel data, and that augmenting human analysis with machine learning approaches is therefore a reasonable strategy to improve efficiency, quality control, and transition away from manual workflows.

In addition to daytime imagery, the distribution of man made light in nighttime images can be very useful in making inferences about mobile populations and settlement characteristics. Niedomysl et al. (2017) for instance use nighttime light data to infer accurate measurements of migration distance by identifying mean population centres, a task that is often problematic with highly aggregate population statistics.

### 2.1.3 Retail data

Supermarkets increasingly collect data about shoppers in order to improve the ergonomics of stores and to tailor customers' experiences with targeted discounts and benefits, but what goes into someone's shopping basket can also reveal a lot about their social proximity to others. The purchasing habits of migrants can, for example, help to quantify the degree of integration over time and whether they converge or diverge from the norms and customs of the host country (Sîrbu et al., 2021).

Employing an unsupervised k-means clustering approach, Guidotti and Gabrielli (2018) are able to distinguish residents, tourists, and occasional shoppers from Italian supermarket data, reporting a high correlation between their predictions and municipal data. Sîrbu et al. (2021) themselves demonstrate the correlation between the cumulative number of new fidelity (rewards) card owners in Albania, France, and Romania with official European immigration statistics. As Albania experienced a

stable rate of fidelity card purchases, Romania a growing rate, and France a declining rate, these trends were mirrored in the immigration data of these countries, giving validity to this method for now-casting immigration.

## 2.2 Mobile phone data

Mobile phone data, another main source of big data in migration literature, is a form of passive digital trace data collected by MNOs and third-party services to facilitate key business operations. To illustrate; airtime top ups, roaming data, CDR, xDR, and CPR are different types of mobile phone data that have appeared in research papers. The word "passive" is used here to express that mobile phone data is a byproduct of phone use, as opposed to "active" data collection where subjects wittingly participate in the data collection. Available to us are two mobile phone datasets, one with CDR and another with xDR. This section discusses CDR and xDR, their attributes, and their applications in migration research.

### 2.2.1 CDRs

CDRs (Call Detail Records) are collected by MNOs (Mobile Network Operators) for billing purposes and are generated when customers make phone calls or send SMS. CDRs log the anonymised user IDs of caller and receiver, the cellular tower of caller and receiver at the time of transmission, the start time and duration of the call (if applicable).

It is possible that roaming data is included in CDR where calls and SMS made from foreign mobile numbers use local infrastructure, but the metadata associated with these records are dependent on the agreements, if any exist, between the foreign and local MNOs. Furthermore, sharing agreements between the foreign and local MNO is also necessary for accessing outbound roaming data since these are collected by the foreign MNO. If roaming data is accessible, CDR can be used to study international migration (Ahas et al., 2018; Mooses et al., 2020).

CDRs enable researchers to observe mobility and social ties at the individual and group level. In terms of mobility, a sequence of base tower interactions serve to reconstruct the geospatial trajectory of customers. These can be used to infer migration events as well as other mobility characteristics like radius of gyration. The inclusion of data about customers' networks make CDRs especially suited to studying social ties and "ego networks" among subscription holders that can deliver rich insights into the community building practices of migrants Coletto et al. (2017).

Although CDR have been demonstrated to possess many desirable characteristics for migration and mobility analysis, limitation exist that make inferences challenging.

Because of the irregularity of phone calls and SMS, CDRs are, as a consequence, also temporally irregular or "bursty". Making calls or sending SMS is highly regulated by peoples' daily habits and schedule, for example spikes in CDRs can be seen after working hours since people avoid conducting personal errands while at work. This burstiness results in relatively sparse or incomplete trajectories for individuals which makes CDRs less dependable for high resolution mobility analysis. Additionally, spatial bias becomes present in the data as users are more likely to engage with their phone at reoccurring locations.

Moreover, a symptom of the way telecommunication infrastructure is operated is that users are not guaranteed to be connected to the nearest base tower. Whether this is to distribute network load, because there are physical obstacles in the way, or because a customer is directly between adjacent towers, inconsistent base tower selection may occur. Further problems occur when we consider the implications of technological drift, market churn, or the accelerating transition to online services.

All of these factors contribute to discontinuity in CDR and care must be taken to account for them during analysis.

### 2.2.2 xDR

Opposed to CDRs, eXtended Detail Records (xDRs) are broader in the types of mobile interactions they capture. In addition to voice calls and SMS, multimedia messages (MMS), mobile data usage (e.g., internet browsing, app usage), and location information can all contribute to the stream of xDRs. xDRs contain the same anonymised user ID, base tower, and time of interaction as CDRs but do not record any networking information like recipients of calls and messages, or the duration of interactions. Roaming data may also appear in xDRs.

As such, xDRs have different strengths that lend them an advantage over CDRs in different scenarios. For instance, the increased bandwidth of xDRs results in a greater regularity and temporal resolution for individuals' trajectories, a benefit for high fidelity mobility analysis. xDRs are still vulnerable to burstiness but because of higher data volumes the effect is less of a hindrance. Conversely, it is not possible to study social ties directly from xDR since, unlike CDR, there is no data about customers' social networks.

Regarding technological drift and market churn, these also affect xDRs but, because the creation of xDRs are driven by mobile data usage, xDRs are in fact enriched by the market shift towards internet based interactions like instant messaging and VoIP. Finally, xDR is also susceptible to incorrect base tower assignment. As customers traverse the boundary between two towers the tower selection may be unpredictable, changing with factors like occlusion by geographic or man-made obstacles, or simply due to the load balancing features of the network infrastructure.

### 2.2.3 Public datasets

In an effort to promote and expedite the deployment of mobile phone data and computational methods as a force for social change, both through policy and activism, several initiatives have sought to bring together stakeholders from government, MNOs, NGOs, international organisations, and academia in the form of open data challenges. The "Data for Development" (D4D) challenges 1 and 2, Blondel et al. (2013) the Telefonica challenge, the Telecom Italia challenge, and the "Data for Refugees" (D4R) Salah et al. (2019a) Challenge, were pioneering in this regard and yielded many new insights, methodologies, and frameworks for utilising mobile data, as well as considerable reflection on the ethics of using mobile phone data for migration research.

The mobile data for the D4D 1 and 2 and D4R challenges were provided by Orange Telecom and Turk Telekom respectively, the former being sourced from Orange Telecom's customers in the Ivory Coast and the latter being sourced from Turk Telekom's domestic customers that had been flagged as "refugees". In both cases, the data was of the CDR variety, generated by phone calls and SMS messages. During the D4D challenge, four separate datasets were prepared covering different aspects of CDR, namely: antenna-to-antenna traffic; 1) hourly totals of calls and call duration between base towers, 2) fine grained trajectories; raw CDR data with full spatial resolution at base tower level but temporally restricted to two weeks, 3) coarse grained trajectories; raw CDR data for a whole year but spatially resolved at the district level, and 4) communication subgraphs to enable the analysis of ego networks. These datasets were curated similarly during the D4R challenge, though the communication subgraph dataset was excluded all together.

To highlight some of the output of these challenges that closely relates to this research, Part II of (Salah et al., 2019a) includes a collection of works on the topic of integration and segregation and Part III touches on the topic of seasonal labour migration. An example of the research on integration is (Rhoads et al., 2019), who evaluate metrics for the social, spacial, and economic dimensions of integration experienced by Syrian refugees and tie in their findings with evidence of voting behaviour shifts due to

refugee presence in the 2015 and 2018 general elections, noting that integration metrics help to better explain the behaviour. In addition, Turper Alışık et al. (2019) provides insights into the potential motivations behind regular and seasonal interprovincial mobility, particularly in relation to accessing services and employment opportunities in both the formal and informal labor markets.

xDR datasets are far less common than CDR datasets like those of D4D and D4R and none have been made public, being mainly curated for isolated case studies. An example is the dataset used by Pappalardo et al. (2021), who obtained written consent from 65 Telefónica Chile employees to use the precise latitude and longitude of their homes in Santiago de Chile.

### 2.2.4 Behaviour descriptors

In elucidating the inherent structure of daily behavioural dynamics, Eagle and Pentland (2009) introduced a novel approach termed "eigenbehaviors," which reduces daily behavioural repertoires into their principle components. These eigenbehaviors are characteristic vectors that encode the fundamental patterns within an individual's behavioural data. By calculating eigenbehaviors at midday, they demonstrate a predictive accuracy of 79% for the remaining day's behaviours, signifying the robustness of this method in capturing underlying behavioural regularities. This finding underscores the notion that human behaviours exhibit discernible patterns amenable to mathematical representation and prediction.

Furthermore, the methodology of Eagle and Pentland (2009) extends beyond individual behaviour analysis, and introduces the concept of a "behavior space" wherein individuals with similar behavioural profiles are grouped together such that a group eigenbehavior can be extracted. Additional individuals' eigenbehaviours can be projected onto this behaviour space and their group affiliation can be thus ascertained. This affords a nuanced understanding of behavioural dynamics, explaining both individual idiosyncrasies and collective behavioural trends.

Expounding upon this seminal work, Farrahi and Gatica-Perez (2010) leveraged the concept of eigenbehaviours in the analysis of mobile phone data. However, in recognising the drawbacks of static eigenbehaviours, they improve the dimensionality reduction technique by incorporating probabilistic topic modelling to capture temporal variations. This augmentation facilitated the discovery of nuanced behavioural patterns that evolve over diurnal cycles, thereby enriching the predictive capacity of the framework.

In a parallel vein, Bouman et al. (2013) endeavoured to extend the application of behavioural analysis to the domain of urban mobility, utilising smart card data from the Dutch public transport systems. Employing analogous dimensionality reduction techniques, they sought to construct an activity model delineating transport network dynamics. Their efforts culminated in the identification of distinct activity patterns deviating from conventional home-to-work commute paradigms.

### 2.2.5 Points of interest

Understanding how people move within cities and interact with their surroundings is crucial for urban planners and policymakers. When we analyse data related to someone's movements, like tracking their location through their mobile phone, it's valuable to make sense of where they go regularly. We call these places "points of interest" or POIs. These can be anything from shops and restaurants to offices and parks. Knowing which types of places someone frequents can tell us a lot about them, such as their income level and lifestyle. On a larger scale, studying these patterns across a city can reveal insights into how neighbourhoods are structured and how accessible amenities are to different parts of the population.

Points of interest are categorised based on their function, like whether they're for retail, leisure, commercial, administrative, or residential purposes. Combining this information with data about how people move around, often collected from sources like mobile phone records, helps us understand urban dynamics better. However, rarely is the data about people's movements detailed enough to precisely pinpoint where they go in relation to specific POIs. In such cases, researchers resort to grouping areas into broader categories, like neighbourhoods or city blocks, to analyse the data effectively.

There are various sources from which we can obtain POI data, including online mapping platforms like OpenStreetMap, official land use statistics provided by governments, and social media platforms like Foursquare, where users voluntarily share information about places they visit. Additionally, advancements in technology, such as computer vision algorithms, enable us to extract information about urban land use directly from satellite imagery (Zhang and Du, 2016).

In recent studies, researchers have used innovative methods to uncover meaningful patterns in urban landscapes. For instance, Zheng et al. (2019) utilise a sophisticated statistical model called a DMR-based topic model along with data on human mobility and POIs to identify distinct functional regions within cities. Similarly, Yuan et al. (2020) emphasised the importance of accurately categorising urban functional areas

(UFAs) for effective urban planning. By combining data on human mobility and POIs, they were able to identify these areas with a high degree of accuracy.

### 2.2.6 Visualisation

Particularly in the field of migration and mobility studies, there is an interest in using visual mediums to convey how people move through space over time, as well as a way to contextualise and ground numerical information. This has the two-fold purpose of making insights more accessible to non-technical audiences, like those in humanitarian or administrative settings, and to take advantage of the dimensionality of the data in ways that cannot be expressed outside of visual means. The most basic visual medium is perhaps the table, its axes demarcating a variety of dimensions of the data, and its rows and columns expressing the enumeration of cases and their differentiating attributes.

The origin-destination (OD) matrix is a good example of how a simple table can be very expressive, in this case for observing the flow of people. The OD matrix is often coloured with a gradient to indicate the flow rate and allows one to see the many-to-many in- and out- flow relationships between places. They play the additional role of being a convenient format for further data processing from the aggregated flow metrics. Figure 1 provides and example OD matrix.

While OD matrices provide valuable insights into the flow of goods, people, or information between different locations, they often fall short in conveying the complexity and nuances of spatial relationships. Additionally, the tabular format can be overwhelming if too many locations are incorporated. Cartographic techniques help to address these limitations by presenting data in a visually intuitive and context-rich manner.

One example is flow charts, which visualise the movement between locations through directional arrows or lines of varying thickness. Flow charts can depict not only the volume of flow but also the directionality and interconnections between different nodes. This visual representation facilitates the identification of patterns, bottlenecks, and dominant pathways within the spatial network.

Bundle charts offer another effective way to visualise spatial flows by aggregating similar movements into bundles or streams. By grouping related flows together, bundle charts reduce visual clutter and highlight major trends within the data. This approach is particularly useful for large-scale transportation networks or migration patterns, where individual flows may be numerous and overlapping.

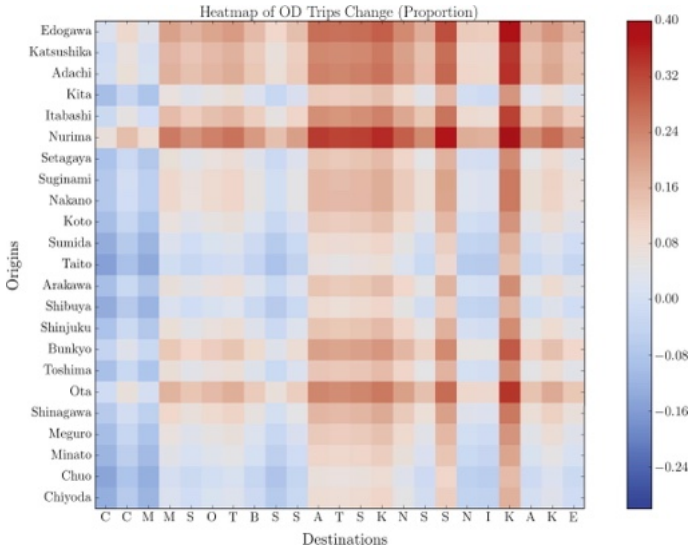


Figure 1. Travel demand in Tokyo. Change in daily trip flows (in proportions) for $23 \times 23$ OD pairs between 2008 and 2012. Figure from (Ge and Fukuda, 2016).

Another example is bubble charts, that represent spatial flows using circles or bubbles whose size and colour correspond to the different variable magnitudes. This method is especially effective for illustrating spatial concentration, dispersion, and disparities in flow volumes across different regions.

In addition, Sankey diagrams are highly versatile tools for visualising flow data. Originally developed in the late $19^{\text{th}}$ century for engineering applications, Sankey diagrams represent the flow of energy, materials, or resources through a system using interconnected pathways of varying width. In the context of spatial analysis, Sankey diagrams excel at illustrating the distribution and redistribution of flows between multiple origins and destinations. By visually encoding flow magnitudes and proportions, Sankey diagrams enable quick interpretation of complex spatial interactions and trade-offs.

In their paper, (Telea and Behrisch, 2022) offer examples of each of these. Figure 2, 3, 4, and 5 depict a flow chart, bundle chart, bubble chart, and Sankey diagram, in that order.

This is just a small sample of the diverse implementations of cartographic methods. Many of these diagrams can be seen in modern infographics in journalism and official statistics as they provide an comprehensive aerial depiction, albeit still static, of the scale of movement. As computer graphics become more advanced and powerful, the use of animated diagrams has additionally allowed the rendering of a temporal dimension. By doing so, we get a much finer sense of a journey. Additionally, instead


Figure 2. An example of a flow chart. Figure from (Boyandin et al., 2011)


Figure 3. An example of a bundle chart. Figure from (Boyandin et al., 2011)

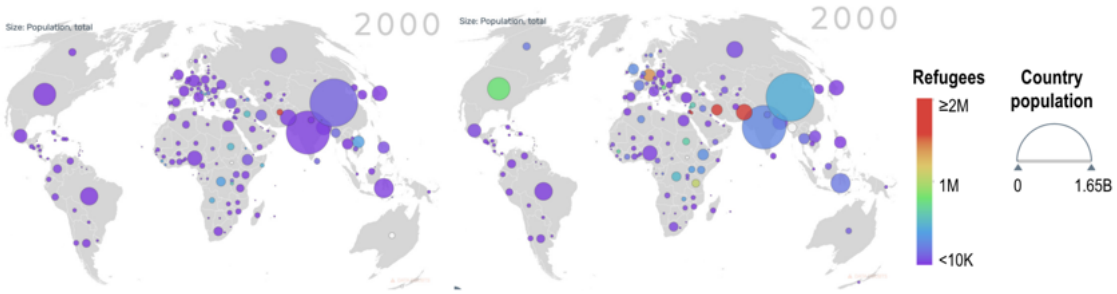


Figure 4. An example of a bubble chart. Figure from (Gapminder Org., 2020)

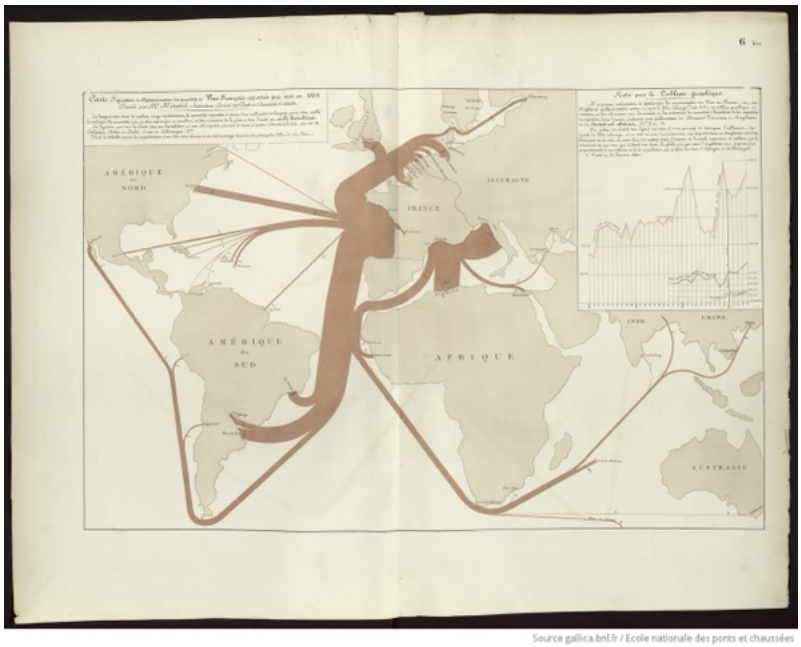


Figure 5. An example of a Sankey diagram. Figure from (Minard, 1844)

of travelling a straight line between a point A and a point B, we may now also see the arteries of road networks and their congestion, and the nuances of human mobility at scale, although appropriate data is also needed to take advantage of these approaches. Mobile data for example is only as spatially fine as the cell-tower or higher administrative levels, which is quite refined in its own right, but makes pinpoint precision impossible, something that GPS data, for example, may handle more precisely.

## 2.3 Detecting Segregation with Big Data

### 2.3.1 What is segregation?

Segregation in urban environments represents the spatial separation of diverse social or ethnic groups within cities or metropolitan areas, influenced by historical, economic, social, and political factors (Massey and Denton, 2003). This segregation manifests across various dimensions, including residential, educational, and economic spheres, resulting in disparities in resource access and opportunities among different groups (Reardon and Firebaugh, 2002).

Measuring segregation is crucial for comprehensively understanding the extent and dynamics of social division within urban landscapes and for informing policy interventions aimed at fostering social cohesion and equity (Logan and Stults, 2011). Quantifying segregation enables researchers and policymakers to identify concentrated pockets of disadvantage or privilege, monitor temporal shifts, and evaluate

the efficacy of interventions, thereby facilitating evidence-based urban planning and governance (Iceland et al., 2002).

Within the migration literature, segregation dynamics garner significant attention due to migration's substantial role in shaping the demographic composition of urban areas (Fischer and Tienda, 2006). The spatial distribution of migrants within cities often mirrors and contributes to segregation patterns, influenced by factors such as ethnicity, socioeconomic status, and cultural affinities (Reardon and Firebaugh, 2002). Ethnic enclaves, for instance, serve as exemplars wherein migrants cluster in specific neighbourhoods due to shared cultural ties, linguistic affinities, or social networks (Fischer and Tienda, 2006).

Understanding these migration-induced segregation dynamics necessitates robust measurement methodologies capable of capturing the intricate spatial patterning and demographic composition of urban locales (Logan and Stults, 2011). Traditional approaches to measuring segregation, such as those reliant on census data, surveys, and administrative records, are indispensable for delineating residential patterns and demographic compositions at the neighbourhood level (Iceland et al., 2002). For example, Massey and Denton (1988) discuss various segregation indices, including the Dissimilarity Index and the Isolation Index, which have been widely used to analyse residential segregation patterns. These indices have been employed in studies to assess the segregation of racial and ethnic groups, revealing disparities in neighbourhood composition and highlighting areas of concentrated disadvantage or privilege. Qualitative methodologies, including interviews and ethnographic research, provide nuanced insights into the social processes underpinning segregation and its ramifications on community dynamics and individual experiences (Massey and Denton, 2003). Consequently, a multidimensional approach to measuring segregation is imperative for comprehensively understanding its complexity and devising targeted interventions aimed at fostering inclusive and equitable urban environments (Reardon and Firebaugh, 2002).

### 2.3.2 The uses of big data

Recent advancements in big data analysis techniques have provided new insights into segregation dynamics, utilising diverse data sources such as mobile phone data, social media, and satellite imagery. These studies employ innovative methodologies to analyse large-scale datasets and uncover patterns of segregation and integration within urban environments.

Bakker et al. (2019) utilise mobile phone data to analyse the integration of Syrian refugees in Turkey, they use mobile phone traffic to estimate segregation levels on the parameters of evenness, exposure, and economic integration to assess integration dynamics. Boy et al. (2019) examine refugee segregation, isolation, homophily, and integration in Turkey using CDR, analysing communication patterns among refugees and the host population. Alfeo et al. (2019) assess the integration of Syrian refugees in Turkey using call data and stigmergic similarity measures. Stigmergy is a communication and organisation strategy used by insects in which a trail of pheromones are deposited while they move. By treating phone activity as pheromone releases, they are able to compare the similarity of stigmergic trails left by phone customers.

Bertoli et al. (2019) integrate data from various sources, including D4R, media events, and housing market data, to gain insights into the integration of Syrian refugees, analysing refugee settlement patterns, socioeconomic integration, and media representations. Hu et al. (2019) quantitatively assess Syrian refugee integration in Turkey using CDR and point of interest data, examining communication patterns and mobility behaviours among refugees to measure integration levels. Bozcaga et al. (2019) perform a individual-level regression analysis against socio-economic, welfare-related, geographic factors to estimate integration levels of Syrian refugees based on mobile CDR.

Sterly et al. (2019) assess the onward mobility of Syrian refugees in Turkey using mobile phone data, examining patterns of mobility and spatial distribution of refugee settlements. Marquez et al. (2019) estimate refugee segregation levels and its effects using sentiment analysis techniques on digital trace data from social networks, providing insights into the spatial and social dynamics of segregation. Rhoads et al. (2019) address behavioural segregation among Syrian refugees in Turkey as an optimisation problem, utilising mobile phone data to model and mitigate segregation dynamics.

Silm et al. (2018) investigate the counter-intuitive phenomenon where segregation is increasing over time. They employ mobile CDR to measure segregation levels at the important locations of individuals and across the activity space and discover that younger individuals of the Russian-speaking minority in Estonia see higher levels of segregation than older individuals. Järv et al. (2015) find, through a similar analysis of activity spaces using CDR, that Russian-speakers and native Estonians are very different in their number of activity locations, the geographical distribution of these spaces, and the overall extent of the spaces. Mooses et al. (2016) encounter evidence

from mobile phone data that the Russian-speaking minority also exhibit segregated behaviour during public and national holidays, featuring stunted mobility out of cities compared to their native counterparts.

Overall, the integration of big data analysis techniques with traditional research methodologies has enriched our understanding of segregation and integration dynamics in urban environments. By harnessing the power of big data, researchers can uncover underlying social processes, provide valuable insights into the integration experiences of migrant populations, and shed light on the factors that shape segregation outcomes within urban environments.

## 2.4 Ethical considerations

There is legitimate and growing concern surrounding the use of trace data for surveillance and the curtailing of civil liberties of the real people who's data is being used. This on top of the already present discussions about the perpetuation of societal biases in the outcomes of black-box algorithms using big data sets such as these. At the same time, there is no argument that the research conducted with these new data sources is filling urgent gaps in many humanitarian spaces and social good initiatives. It is therefore imperative to take these issues seriously and to innovate on frameworks and standardisation that will allow for the safe usage of digital traces.

### 2.4.1 Bias and fairness

In much of migration and mobility research there is a large focus placed on the application of insights in policy making, which can have enormous impacts at a population-wide scale. This raises many of the same ethical issues that can be seen in sectors that are beginning to rely on computational insights for decision making. Mainly this relates to the efficacy of information systems and their culpability, but also to the quality of the data and how they are obtained.

Importantly, any kind of algorithmic decision making is prone to error just as a human arbiter is but, unlike with a human, as algorithms grow in sophistication and size it becomes ever harder to understand how they have arrived at a conclusion. This can obfuscate biases that were inherited from training data and makes it challenging to assign accountability when incorrect inferences lead to people being discriminated against. Moreover, the data itself being the source of the bias in this case is a key part in the fairness of information systems.

The effects of this phenomenon have been seen first hand in several instances and

research has demonstrated how algorithms trained on mobility data exhibit many of the same systematic inequalities that are present in society. Erfani and Frias-Martinez (2023) for instance demonstrate that Covid infection models trained on SafeGraph data favour certain demographics. "Specifically, the models tend to favor large, highly educated, wealthy, young, and urban counties." Similarly, by combining MPD with administrative data Coston et al. (2021) show that allocating public health resources based on the mobile phone data they use could disproportionately harm high-risk elderly and minority groups. Particularly in the case of mobile phone data, that MNOs are not transparent about data collection and transformation practices make it difficult to identify and correct bias (Grantz et al., 2020).

Research into the biases of mobility data from digital traces is still narrow however, and more work needs to be done in establishing protocols for data collection and processing. In the meanwhile, we must heed the advice of scholars of AI ethics and import insights from this field. The themes of transparency, justice, non-maleficence, responsibility and privacy emerge as the key concerns in the debate around ethical AI (Jobin et al., 2019). Personally reflecting on the topic, some concrete actions present themselves. Several layers of audit are necessary to ensure that applications, designed with the intention of improving the lives of people, do not inadvertently end up causing harm. Performing due diligence on the raw data, researchers and policy makers must investigate patterns in the data that reflect social inequalities and account for these when using them during algorithm training. The use of explainable models is also essential if the end purpose is decision making, allowing the sequence of inferences to be traced. Furthermore, incorporating human intervention into the process means that a sensible point of accountability is introduced. The system that is to be modelled should be understood holistically and a decision must be made whether the technological approach is indeed necessary (Selbst et al., 2019).

### 2.4.2 Privacy issues

While mobile phone data is always anonymised, that is to say stripped of any personal identifiers, and sometimes aggregated to eliminate the chance of tracking an individual, it still can contain sensitive personal information including location, communication, and behavioural patterns. In addition, since the data is collected without *informed* consent, it's use becomes ethically dubious especially considering how important the topic of privacy has become as a result of the data economy. There is however precedent in using MPD for social good and the implications of a robust and trustworthy apparatus to continue using it in research is impactful and far-reaching. We consider the privacy pitfalls of using MPD and highlight some outcomes of scholarship in this area.

Many scholars have pointed out that the anonymization of user records does not completely remove the possibility of re-identification by associating the data with the habits of a user. Notably, De Montjoye et al. (2013) discover that four mobile phone records is enough to uniquely identify almost the entire user base of their data set and that the ability to do so does not decay much when coarsening the data in the space and time dimensions. However, others have claimed otherwise, such as the research of Al-Azizy et al. (2016); Cecaj and Mamei (2017); Gambs et al. (2014), produced within the context of the D4D Challenge (Salah et al., 2019a), that claim identifying individuals from data sets prepared in such a fashion with coarse- and fine-grained data aggregated along spatial and temporal dimensions, is not possible. Note that in their research, De Montjoye et al. use a closed data set rendering their results irreproducible, on the other hand, the data made available in the D4D Challenge was an open data set.

The 'data protection by design and default' perspective that drove the implementation of data security for the D4D and D4R challenges focused on the removal of all personally identifiable characteristics of the data during the data collection itself. Once the data leave the servers of the source telecom, it is already impossible to identify individuals in it. To achieve this, the data were aggregated spatially and temporally, removing all links to actual phone numbers and data subjects, and the mapping was discarded (i.e. not stored anywhere). Subsequently, for a base station, the number of people using it at any moment was stored, but this was not linked to the actual records of usage.

When talking about spatio-temporal aggregation, this involves at the most granular spatial level the Voronoi tesselations that demarcate the transmission basin of each cell tower in the network and at the temporal level, hourly mobile phone signals. Both of these dimensions can be coarsened. In the case of Voronoi tesselations, these can either be combined together thereby merging a number of cell towers, or projected onto higher administrative regions like neighbourhoods, districts, and provinces. In terms of time, the data can be aggregated into larger temporal units than single hours. Several hours could be grouped together or units could even be days or weeks. This illustration, Figure 6, from De Montjoye et al. (2013) demonstrates how coarsening is performed.

In another paper, De Montjoye et al. (2018) rightly point out that transforming mobile phone data through aggregation degrades its resolution, which is a hindrance to the majority of applications that use this data. For example if we would like to detect the home and work locations of individuals we would not be able to do so

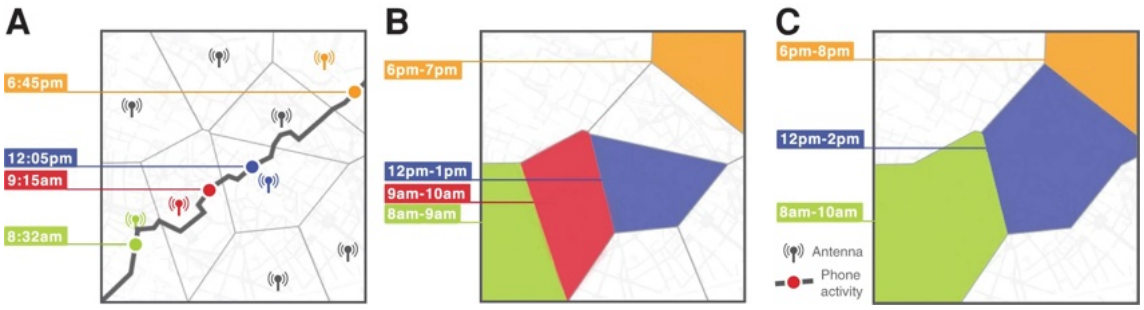


Figure 6. "(A) Trace of an anonymized mobile phone user during a day. The dots represent the times and locations where the user made or received a call. Every time the user has such an interaction, the closest antenna that routes the call is recorded. (B) The same user's trace as recorded in a mobility database. The Voronoi lattice, represented by the grey lines, are an approximation of the antennas reception areas, the most precise location information available to us. The user's interaction times are here recorded with a precision of one hour. (C) The same individual's trace when we lower the resolution of our dataset through spatial and temporal aggregation. Antennas are aggregated in clusters of size two and their associated regions are merged. The user's interaction are recorded with a precision of two hours. Such spatial and temporal aggregation render the 8:32 am and 9:15 am interactions indistinguishable." Figure and caption text from (De Montjoye et al., 2013).

if time steps were aggregated into days. It would be possible however with hourly time steps since we would be able to distinguish where someone is at night or during the day. Similarly, in the spatial dimension, a person would seem stationary if they had a small radius of gyration and our data were aggregated at the provincial level for instance. If instead the spatial dimension remained at the Voronoi tessellation level, we would be able to see their trajectory in great detail, especially in dense urban areas where cell towers are closely arranged. Figure 7 depicts how the spatial granularity of Voronoi cells increases in urban centres and decreases in rural regions. On the other hand, in limiting the precision of traces, aggregation does prohibit us from scrutinising individuals' behaviours and in turn prevents their identification.

Salah et al. (2019b) organised, as a second layer of protection, a project evaluation committee that was entrusted with oversight of the research proposals and how they would employ the mobile phone data. Additionally, the committee decided on rules for the retention, destruction, and archiving of the data. These rules did not permit researchers to keep the data longer than the assigned period without case-by-case permission.

Though the evidence is conflicting, the passage of time will likely bring new techniques that, if wielded inappropriately, will allow the identification of individuals from aggregated data, especially since it is through the disaggregation of data that migration and mobility studies get their most cutting edge insights. The important

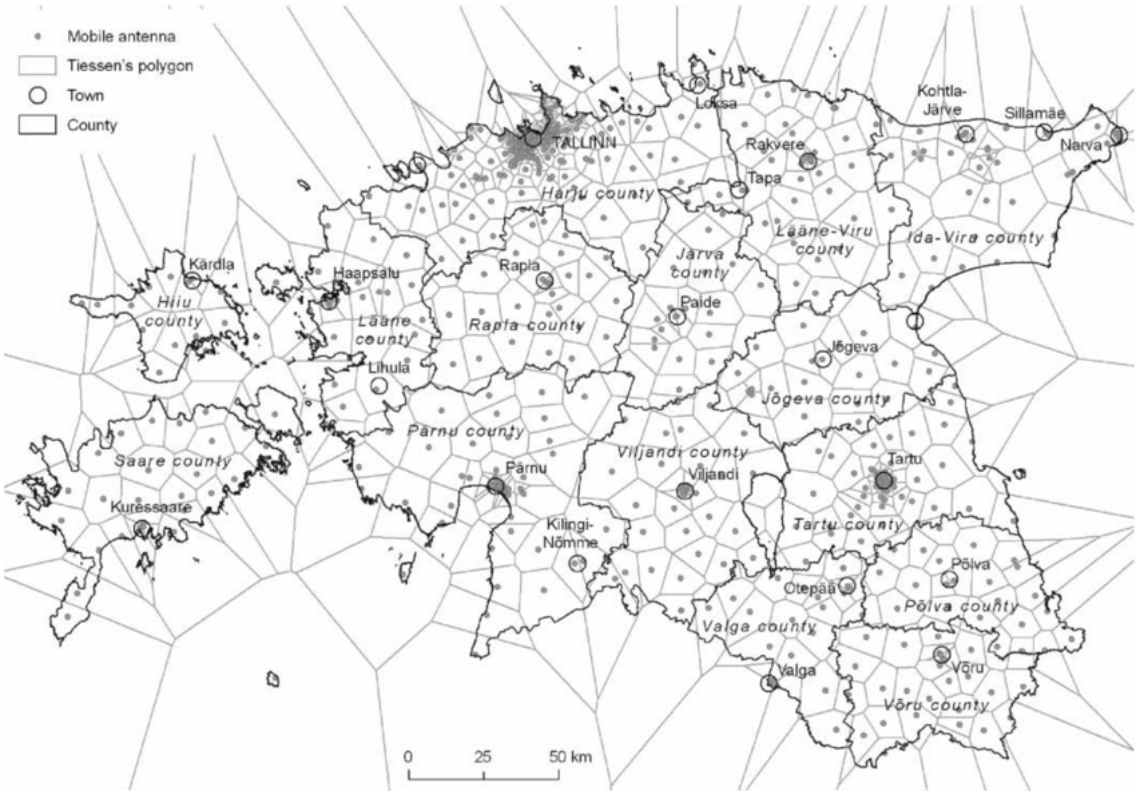


Figure 7. Voronoi tessellation of the mobile phone network in Estonia, demonstrating the variation in spatial granularity between urban and rural areas. Figure from (Ahas et al., 2010).

consideration here is indeed by whom the data is accessed and to what lengths they are willing to go to extract this information. The evidence that the differences in peoples' behaviour uniquely identifies them is, however, clear, and what is also clear is that the process of anonymization does not end at removing identifiers in the data or at the level of spatial aggregation used, these initial processes must be buttressed by rigorous audit, accountability, and transparency at the data sharing level, driven by the stakeholders that wish to repurpose this data.

De Montjoye et al. (2018) import that data sharing practices are the centre point of maintaining privacy. They provide an aspiring framework that attempts to find a balance between preserving the quality of data and preserving the privacy of individuals. Their framework pivots around the idea that data can be hosted and overseen centrally by the MNO or a trusted third party, limiting the need for aggregation. Alternatively, in the same manner as has been seen in the D4D and D4R challenges, data can be partially aggregated and copies of it given to trusted parties under audit.

Crucially, the involvement of many and all stakeholder parties increases the witnesses

to the scholarship produced with this data and by encouraging intersectional perspectives, the knowledge that is produced will similarly drive new understandings of ethical control of trace data processing. (Vinck et al., 2019) is a crucial piece of scholarship that advocates for embracing "ethical complexity and emerging rights," especially in the space of humanitarian work where the high stakes and risks call for advancing the responsible use of data and technologies.

# 3. Case study: Segmenting the cities of Istanbul and Kocaeli

Segmenting a city involves dividing its urban landscape into distinct areas or segments based on various characteristics such as land use, demographic composition, economic activity, or social dynamics. These segments provide a nuanced understanding of the spatial heterogeneity within the city and facilitate targeted interventions to address specific needs and challenges. Common types of segments include residential neighbourhoods, commercial districts, industrial zones, cultural hubs, and transportation corridors. By segmenting a city, planners and policymakers can tailor urban policies and initiatives to the unique characteristics of each segment, thus enhancing the effectiveness and efficiency of urban management.

The utilisation of mobile phone data offers a powerful tool for segmenting cities and understanding urban dynamics in real-time. Mobile phone data provide insights into population movements, social interactions, and spatial behaviour across different segments of the city. For example, a study by Calabrese et al. (2011) analysed mobile phone data to identify functional urban areas within cities, revealing patterns of commuting flows and economic interactions. By segmenting the city based on these mobility patterns, planners can optimise transportation networks, allocate resources, and enhance urban connectivity.

Moreover, mobile phone data can be leveraged to redefine city borders and administrative boundaries based on actual usage and functional relationships between different areas. Researchers have proposed redrawing city borders using mobile phone data to reflect the dynamic nature of urban regions and capture emerging patterns of urbanisation and connectivity. For instance, a study by Zhang et al. (2022) demonstrated how mobile phone data could be used to redefine city boundaries in the Shenzhen-Dongguan-Huizhou area of China. This approach has implications for urban governance, resource allocation, and regional planning, as it enables policymakers to adopt a more flexible and responsive framework for managing urban regions in an era of rapid urbanisation and technological change.

## 3.1 Data

### 3.1.1 xDR

The data we work with comes from a Turkish MNO with a high market share in the mobile networking sector, we do not disclose the company at their discretion (Aydoğdu et al., 2021). The datasets are prepared in the same fashion as the datasets for D4D and D4R, and include separate CDR and xDR data. Here, the idiosyncrasies and validity of the data is explained in detail.

It was decided that these experiments would be limited to the use of xDR since it is under-represented in the mobile phone data literature and the results can serve as a baseline for future experiments. Furthermore, we anticipate a much greater fidelity over CDR in tracking individual mobility which will be advantageous for methodologies that leverage trajectory information.

The xDR dataset covers the entire year of 2020. Notably, 2020 was the year that saw the peak of the Covid 19 pandemic in Turkey, with the highest number of cases and most severe lockdown restrictions occurring during this time. The implementation of curfews and work from home measures had of course shifted several times throughout the year and reports have since suggested that consecutive interventions were not as successful at curbing mobility after the first wave in the beginning of the year (Atahan and Alhelo, 2022; Shakibaei et al., 2021). Nevertheless, we expect to see residual effects on mobility during our analysis.

The xDR dataset can be broken down into the following three parts:

1. **Antenna traffic** The hourly mobile phone traffic volume per antenna in the network for the whole year.
2. **Fine grained mobility** The individual traces of mobile phone usage, the anonymised ID of the customer, as well as the ID of the antenna from which the signal was received. Because of the high spatial granularity, only two weeks of data are made available, specifically from May $16^{th}$ to May $31^{st}$. This measure greatly limits the possibility of profiling users.
3. **Coarse grained mobility** Similar to the fine grained data but spatially aggregated to the province level making it possible to witness interprovincial mobility and migration but impossible to analyse small scale urban mobility patterns. At this spatial resolution there is no possibility of user identification and the data spans the whole year.

**Description: fine-grained xDRs**

The fine-grained data have three primary attributes: the timestamp and location of the event as well as the anonymised customer id. Tuple 3.1 and Table 1 describes the properties of fine-grained xDR in more detail.

```
<timestamp, customer_id, site_id>
```
(3.1)

| Variable | Description |
|---|---|
| `timestamp` | Timestamps are discrete and binned by hour, falling between 12 a.m. 16/05/2020 and 11 p.m. 31/05/2020. |
| `customer_id` | Anonymised user ID of the subscription holder. This number uniquely identifies a customer but is not shared between the different datasets to prevent cross leak. |
| `site_id` | Randomised site ID of the cell tower where the xDR was created. This number uniquely identifies a cell tower and its accompanying Voronoi tessellation across the xDR datasets but does not correspond to any public facing information about cell tower as this is considered sensitive information by MNOs. |

Table 1. Description of the xDR variables

**Description: xDR antenna traffic**

The antenna traffic data are tabulated differently than xDRs in the fine-grained dataset. Aggregated at the cell tower level, this data does not refer to any individuals but rather the number of xDRs that were recorded at each site per hour. The temporal and spatial resolution therefore remain the same, but with the risk of identification removed, this data can span a larger length of time. The antenna traffic data spans the entire year of 2020 as a result. The xDR traffic is, furthermore, aggregated by nationality flags, or what we refer to as segments.

In this work, however, an altered version of the antenna traffic dataset was used. In the altered dataset the traffic counts are aggregated by hour such that each cell tower has a total, mean, and median traffic per hour. Such an altered dataset was used as the original was too large to be processed within the means of this thesis, and the hourly traffic data was sufficient and even convenient for our analysis. Tuple 3.2 and Table 2 describes the properties of the antenna traffic data in more detail.

```
<site_id, hour, [[total, mean, median]_segment]>
```
(3.2)

| Variable | Description |
|---|---|
| `site_id` | Randomised site ID of the cell tower where the xDR was created. This number uniquely identifies a cell tower and its accompanying Voronoi tessellation across the xDR datasets but does not correspond to any public facing information about cell tower as this is considered sensitive information by MNOs. |
| `hour` | Hours 00:00 through 22:00. 23:00 is excluded because there was no way to filter the artefacts as with the fine-grained data. |
| `[[total, mean, median]_segment]` | Separate total, mean, and median columns per segment representing antenna traffic statistics for each cell tower over the entire year of 2020. |

Table 2. Description of the antenna traffic variables

Note that 11 p.m. is excluded from the data. As discussed in the previous section, during 11 p.m. there appear to be artefacts in the data the do not represent actual mobile phone traffic. With the fine-grained data we were able to take the modal location of individuals as a solution to this problem, however with the antenna traffic data this solution is unavailable since we only have aggregated counts. As a result, we simply removed all traffic counts at 11 p.m. as the only course of action to remedy the artefacts.

**Segmentation**

In following the guidelines set out by the D4R challenge, and in the interest of identifying the disproportion of certain migration phenomena for vulnerable groups, the data include a 'refugee flag' where phone lines were registered using a foreign, specifically Syrian, passport. This is possible because the vast majority of Syrians in Turkey are refugees recognised under Turkey's temporary protection scheme, though this excludes refugees of other backgrounds. Furthermore, in order to select a useful sub-sample of the user base for comparative analysis, the accounts were filtered by nationality such that there was an even proportion (per city) of local and foreign nationals.

In total 1,978,373 accounts associated with foreign passports and 2,021,627 associated with Turkish passports were accumulated. These numbers also include customers that went missing from the database over the course of 2020. Some nationalities were

omitted after filtering because there were too few subscribers of those nationalities to be representative in any kind of analysis, especially small countries from Polynesia, with less than 100 customers. After aggregating them, countries in Latin American and the Caribbean were excluded all together as the sum of their subscribers was below 10,000. The remaining countries were classified into the categories seen in Table 3. For continuity, these categories are referred to in the data as 'segments'. These categorisations do not adhere to any standardised system, but we deem this approach appropriate in the sociological context of migration in Turkey which itself does not perfectly observe existing naming conventions.

| Registered nationality | Number of customers |
|---|---|
| Turkey | 2,021,627 |
| Syria | 1,415,588 |
| The Middle East (excl. Syria) | 278,324 |
| South Asia | 79,778 |
| Central Asia | 71,862 |
| The Caucasus | 45,581 |
| North Africa | 60,956 |
| OECD | 49,442 |
| Eastern Europe | 31,357 |
| Sub-Saharan Africa | 29,221 |
| East and Southeast Asia | 14,489 |
| The Balkans | 10,825 |

Table 3. Distribution of nationalities in the database

The classifications are derived from the following logic. Syrians greatly outnumber the citizens of the other Middle Eastern countries combined and thus occupy a separate category. The remaining countries follow the convention of the UN's statistical division with a number of exceptions. Firstly, Western Asia is divided into two groups; the Middle East and the Caucasus. The second exception regards the classification of Balkan countries. If the country is definitively located in the Balkan peninsula or both in the Balkan peninsula and Eastern Europe, such as Romania and Bulgaria, they are considered Balkan. Finally, OECD countries, though geographically diverse, are categorised together and represent a largely middle-class demographic. Transcontinental countries remain grouped according to the UN statistical division conventions.

Figure 8 plots the volume of mobile phone traffic for native Turkish, Syrian, and Afghan segments in the urban areas of Istanbul. It is important to note the legends, which reflect a different maximum volume in each subplot. What can be seen is 1) a greater dispersion amongst Turkish natives compared to the Syrian and Afghan refugees across the entire municipality, 2) both Syrian and Afghan refugees are concentrated on the European side of Istanbul, 3) there is a greater number of Syrian refugees and they are more spread out than Afghan refugees who are clustered by the shoreline.

### 3.1.2 Official neighbourhood population

Granular, official statistics about population and residency distributions are crucial for calibrating and validating the outcomes of our segregation analysis. The Mahallem Istanbul project, lead by the municipality of Istanbul (İstanbul Kalkınma Ajansı, 2017), has produced several datasets enumerating a variety of demographic and socio-economic indicators from the province of Istanbul and neighbouring provinces. Among these is a neighbourhood level population count covering all of the neighbourhoods in Istanbul and Kocaeli. The dataset contains the names of the neighbourhoods and their respective parent districts alongside the population counts.

Temporary protection holders like Syrian and Afghan refugees are not included in these population counts, meaning it cannot be used as a touchstone from which to benchmark our results on these segments. However, we can use these figures as a ground-truth for the native Turkish population, as well as to calibrate a population mapping model that incorporates all segments. Figure 11B visualises the population counts across the various neighbourhoods of Istanbul and Kocaeli.

### 3.1.3 OpenStreetMap administrative boundaries

Geographic information systems (GIS) data, specifically polygon data, for administrative boundaries up to the district level is readily available from Regional IM Working Group - Europe (2021) but smaller administrative units like neighbourhoods have to be extracted by other means. To that end, OpenStreetMap (OSM) has an API through which a range of GIS data like neighbourhood polygons and point of interest data can be downloaded. Geofabrik GmbH (2024) has a series of curated OpenStreetMap data sets conveniently available in their data portal, one of which contains low level administrative boundaries for the provinces of Istanbul and Kocaeli. As Geofabrik regularly updates these datasets, they are usually in concordance with changes to administrative boundaries as they occur.

One of the challenges of working with disparate data sources is a lack of agreement on standards and conventions. Often, manual refactoring is required to align the

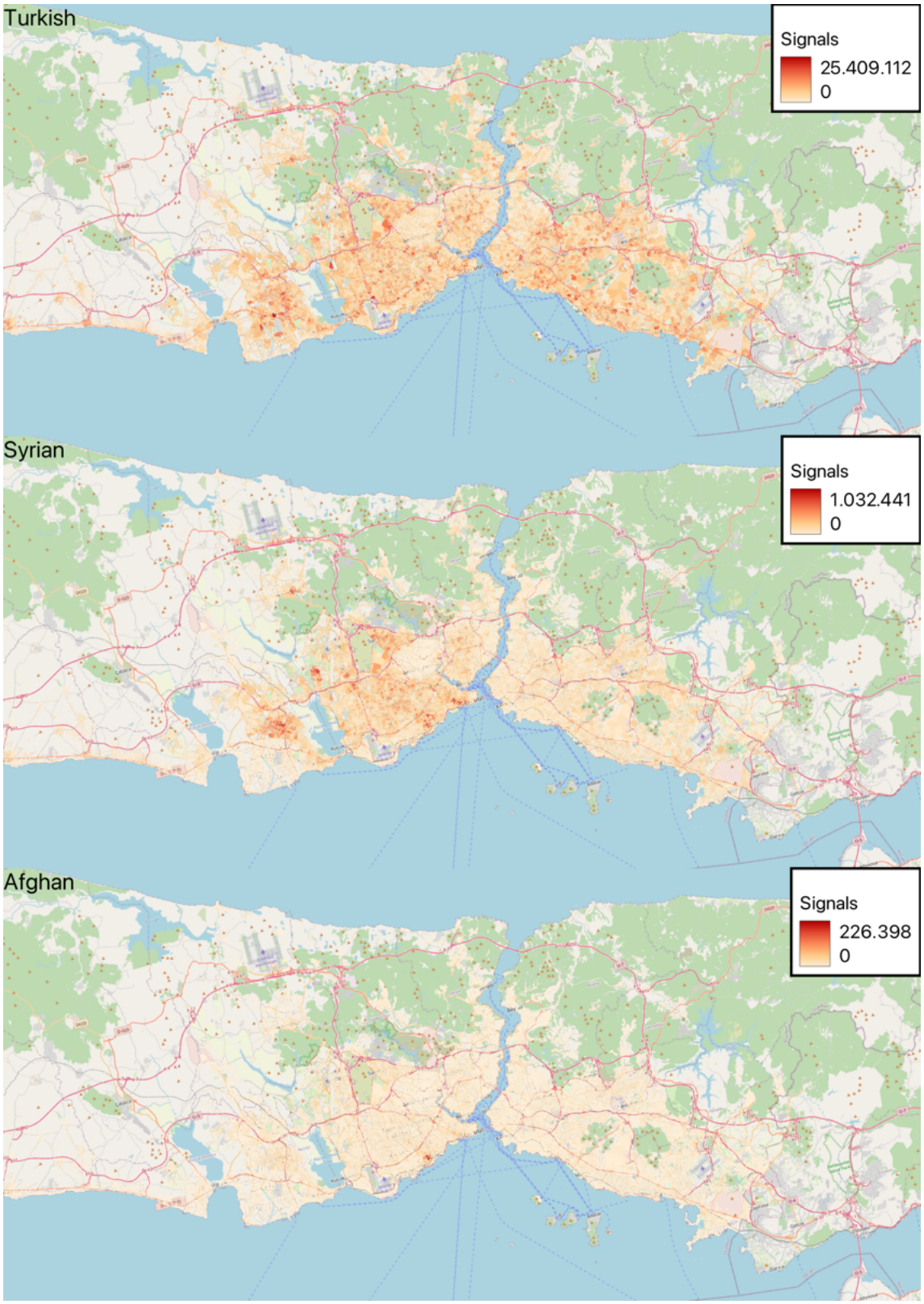


Figure 8. The concentration of xDR signals in the built-up areas of Istanbul for Turks, Syrians, and Afghans. These are the total signal counts per cell tower for one year but distributed according to the area of overlap to smaller spatial units. The smaller spatial units are attained by splitting the Voronoi tesselations according to neighbourhood boundaries.

datasets and render them compatible, though there is no guarantee of compatibility, and data loss is common. A similar situation is encountered here. The OSM data is open source and partly user-generated making it at times contradictory both internally, within the OSM data table, and externally, when compared with official data sources. Through a series of spatial mappings, string matching, and geocoding operations, the OSM data could be reconciled with the HDX and other ancillary datasets.

### 3.1.4 Land use

Urban zoning and land use data is indispensable information in the analysis of segregation within urban environments. By delineating the spatial distribution of various land uses, this dataset offers valuable insights into the allocation of resources, amenities, and opportunities across different neighbourhoods. Through the integration of demographic information with land use classifications, areas characterized by distinct patterns of residential concentration can be identified, thereby unveiling spatial segregation along racial, ethnic, or socioeconomic lines. Furthermore, land use analysis facilitates the identification of segregation hotspots and disparities in access to essential services, such as education, healthcare, and transportation, thus enabling policymakers to target interventions aimed at mitigating segregation and promoting inclusive urban development.

The European Environment Agency and European Environment Agency (2019) land use dataset has 26 distinct classifications. It covers transport infrastructure, industry, agriculture, wilderness, topographic features, and various urban configurations. Although this list is comprehensive, there is no indication of residential or commercial zoning as they all urban sites are considered to have mixed zoning. The classifications and extent of the dataset can be seen in Figure 9.

## 3.2 Processing mobile data

Mobile phone data are generated primarily to enable internal facing processes like billing whose objectives are generally less scrupulous than that of research or policy design. As a consequence, the raw mobile data may contain errors, missing values, and inconsistencies that must be addressed before use in downstream applications. Moreover, several downstream methods require more refined or different input than the tabular xDR data. Feature engineering and extraction allows more sophisticated approaches to be employed.

In this section we inspect the limits of the mobile phone data at our disposal and

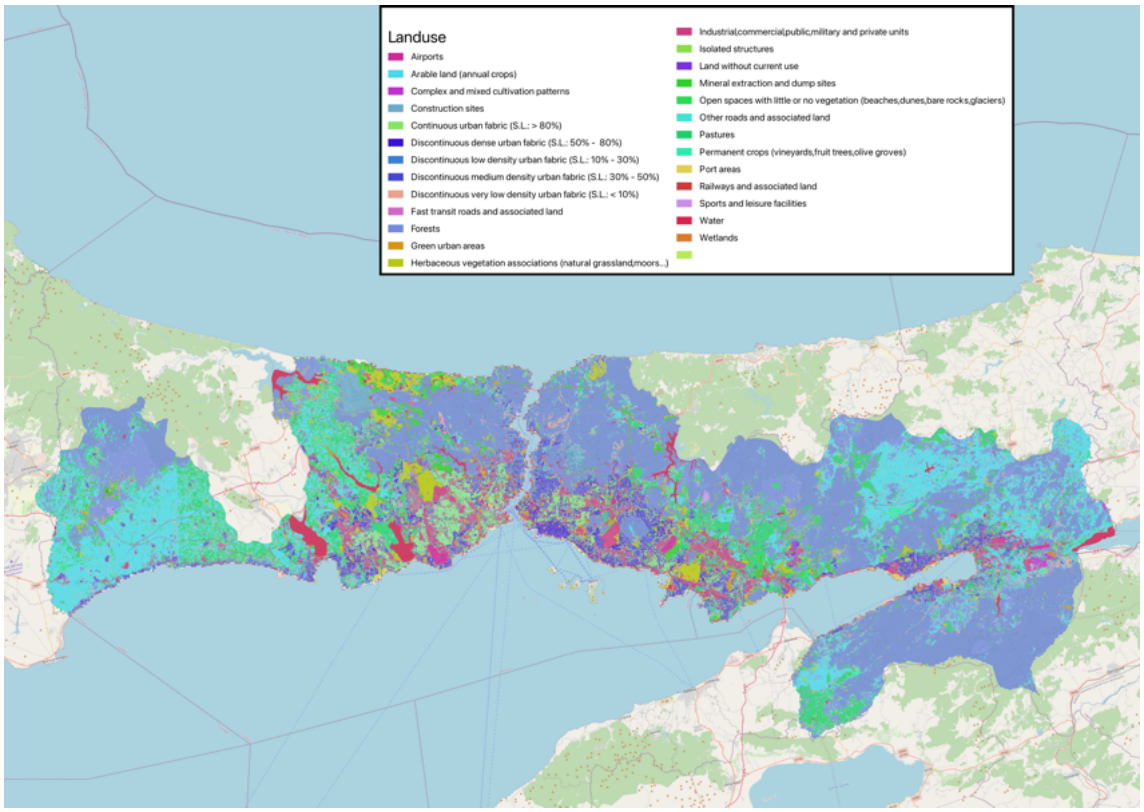


Figure 9. Landuse zones of Istanbul and their geographic extent.

outline the strategies used to clean and process them to within acceptable tolerances. Additionally, we describe some of the ways we extract features from the xDRs for downstream use. Processing mobile phone data is often an ad-hoc, iterative way of working as solutions must be tailored to the dataset at hand, however we ground our processing steps in the literature and provide a discussion on the benefits and drawbacks of alternatives.

There are three separate configurations that comprise the xDR dataset, namely the fine- and coarse-grained configurations, and antenna traffic data. In our analysis, we exclude the coarse-grained data, utilising only the fine-grained data as well as the antenna traffic data at different moments in the methodology.

### 3.2.1 Fine-grained xDRs

#### Volume and temporal resolution

Since xDR are comprised of mobile data and telephony traces as people interact with their phones, the volume of records over time will be a composite of the signals from these various streams of input. These streams will fluctuate with time following the patterns of human behaviour, e.g. awake during the day, asleep during the night, and changes of habit during weekends and holidays.

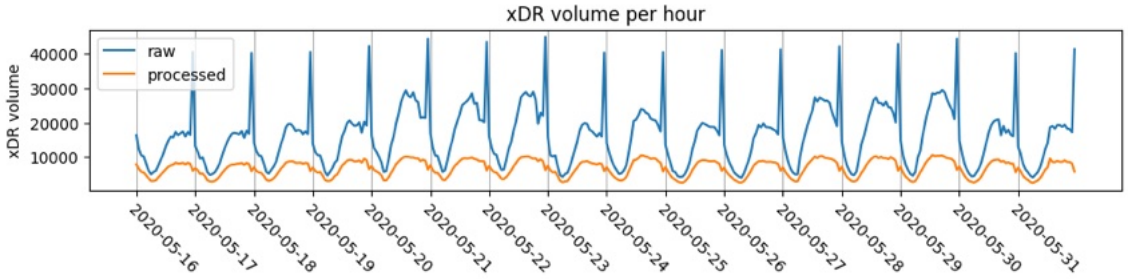


Figure 10. Raw vs. processed xDR volume per hour over the duration of the dataset

Figure 10 depicts the raw volume of xDR present in the fine-grained dataset spanning 16 days in May 2020. In the plot, it can clearly be seen that there is a drop in the volume of events during the the hours of the morning, typically reaching a trough at around 7 a.m.. The signal tends to peak in the afternoon and evening, though the peaks are more sustained than the troughs and less consistent in the hours they occur. Moreover, the peaks change in intensity depending on the day of the week. Here, Wednesday, Thursday, and Friday experience higher xDR volumes than during the rest of the week.

Additionally, we see artificial peaks or artefacts occurring everyday at 11 p.m.. In consultation with the MNO, it was revealed that these peaks were caused by a programmed table update and not an increase in phone usage. Consequently, the false records must be accounted for or filtered out entirely. We resolve this issue in conjunction with the next point.

Importantly, xDRs are aggregated by hour. This has several implications for the temporal resolution of the fine-grained xDRs. Firstly, consider that more than one xDR can occur within one hour. It is useful to localise an individual making short trips in the city as such information can reveal deep insights into the habits of residents and how the city is used. However, since our data is binned by hour, the sequence of such visitations within an hour is not known. This introduces a great amount of uncertainty as to the true location of an individual, especially when they have visited several locations in a short amount of time. They would look as if they occur in all locations simultaneously.

Especially in cases where tracking mobility is important, we must establish a definitive trajectory for each individual, i.e. one location per time-step. In dealing with this issue, two approaches emerge. The first involves selecting a location from the alternatives and the second involves increasing the available time-steps to accommodate all the alternatives.

Because CDRs are less abundant in volume than xDRs, and there is very little chance that individuals have multiple CDRs in an hour, this issue is not extensively discussed in the literature, even more so in the case of the artefacts encountered earlier. However, on greater temporal scales, e.g. over weeks or months, this sort of uncertainty is addressed by taking the modal location of individuals (Marquez et al., 2019; Chi et al., 2020). This approach is akin to our fist option of selecting a location from the alternatives. We apply the model location technique per hour, gaining a more determinate picture of mobility at the expense of greater temporal fidelity.

With regards to the artefacts or artificial peaks, more considerations needed to be made. It would, for example, be possible to simply exclude all xDRs at 11 p.m. but this would also remove data points that are authentic and it is preferable to maintain this contextual information. A large portion of the events during the 11 p.m. peaks reference customer ids that appear nowhere else in the dataset other than the peaks and so can be discarded without consequence. xDRs of any customers who have fewer than two records or more than one modal location during an hour are discarded to avoid situations where only false xDRs appear in the time-step.

Figure 10 demonstrates the result of these data processing steps. The step of taking the modal location per hour acts to reduce the volume of data quite significantly. Furthermore, a result of the processing is that the peaks have been eliminated and the volume of xDRs at 11 p.m. is now more in line with the expectation of lower activity during nighttime.

An alternative to taking the modal location may be to use heuristic methods to infer the order of visitations. Such a method may consider the direction of travel, visitation time, travel time, etc. to probabilistically rank the enumerated visitation sequences. This would be more akin to the second processing option of increasing the available time-steps. We find, however, that taking the modal location provides a sufficient temporal resolution for our analysis and has a much smaller processing overhead than a heuristic method would.

### Cell towers and spatial resolution

In the fine-grained dataset, xDRs are aggregated at the finest spatial resolution, which are the individual cell towers of the network. They are identified by a randomised ID that corresponds to a spatial mapping in the form of a Voronoi tessellation, a series of geographic coordinates that approximate the zone of influence around each cell tower. While the raw xDRs make reference to site ids for cell towers across the whole of Turkey, only Voronoi tessellations of Istanbul and Kocaeli were made available.

The fine-grained data includes anonymised user IDs that uniquely identify each customer in the set, enabling us to do individual level analysis of customers' mobility. There is a caveat that multiple phone numbers can be registered under one main account holder and will all appear under one customer ID in the dataset. This leads to some bias since there may be multiple family members under one umbrella account. Each person will represent independent demographic characteristics and alternative behaviours which together obfuscate each other and introduce noise. Unlike CDRs, there is no networking information in xDRs.

Each account has an associated nationality flag, hereafter referred to as segments. This data is available since the MNO requires customers to provide a passport in order to open an account. These segments are nevertheless a noisy indicator of nationality since many customers, especially refugees who wish to keep some level of anonymity, request native Turks to open accounts on their behalf. The Turkish and Syrian segments are additionally separated by binary gender, though women are more likely to be sub-account holders under some family patriarch and so are underrepresented in the data. An explanation of the segments and their sample size can be found in Table 3.

As previously stated, only Voronoi tesselations of the Istanbul and Kocaeli provinces were made available. Figure 11A shows a map of Turkey and it's various provinces. Istanbul and Kocaeli, highlighted in red, are some of the densest urban environments in the country, with Istanbul alone hosting some 15.4 million residents and Kocaeli hosting about 2 million as of the writing of this article, together capturing roughly 20% of the total population of Turkey. Additionally, according to the Turkish Ministry of Interior (2024), as of 2024 some 530,066 of the roughly 3.2 million Syrian temporary protection holders reside in Istanbul making the city particularly relevant to developing an understanding of urban dynamics between the native and foreign populations.

When spatially filtering the xDR for those that only occur within Istanbul and Kocaeli, some 74% of the xDR are excluded as well as 71% of the customers. On a separate note, despite constituting a mere 1.1% of the area of Turkey, Istanbul and Kocaeli put an enormous load on the network infrastructure and as a consequence many cell towers have been installed within close proximity to ensure the network has enough capacity. The cell towers in these two provinces are very numerous and represent 31% of all cell towers in the network of the MNO across Turkey.

In mapping the xDRs to their respective cell towers, it was discovered that around

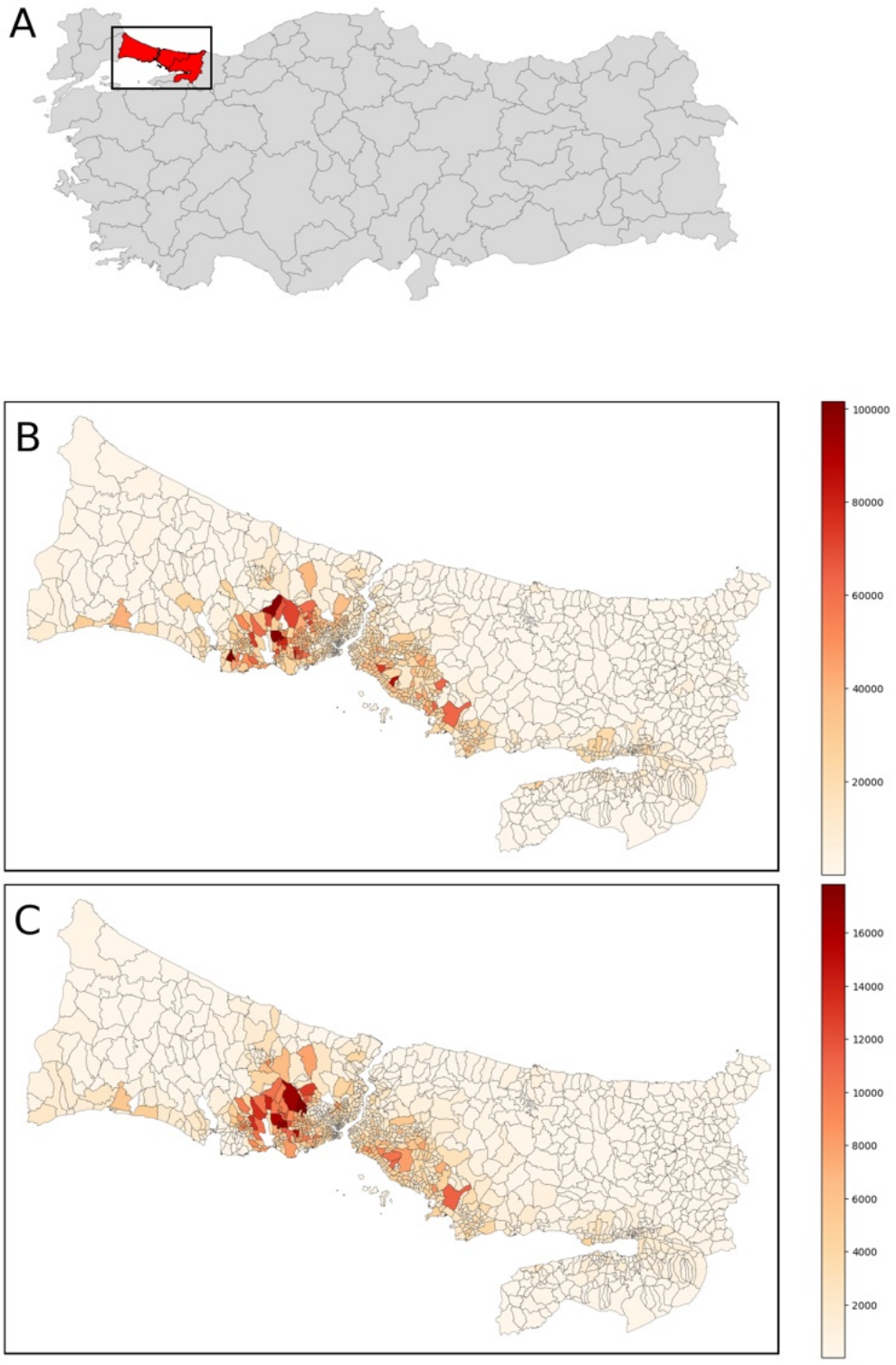


Figure 11. (A) The provinces of Istanbul and Kocaeli highlighted in red, (B) official population counts per neighbourhood, (C) xDR volume per neighbourhood.

29% of the cell towers in Istanbul and Kocaeli experience no activity whatsoever. In these cases, it is assumed that the cell towers at these sites were not in operation for the duration of time the xDRs were sampled from. The area surrounding the offline cell towers would not in reality be devoid of telephone signal since nearby cell towers

would compensate. Thus there is grounds for redrawing the Voronoi tessellations such that the total land surface has coverage with respect to the estimated locations of the online cell towers.

To redraw the Voronoi tessellations, firstly the centre point or centroid of the original Voronoi polygons was identified and used a proxy for the location of the cell tower. Once these centroids were identified, those that were associated with cell towers not in operation were removed, and finally new Voronoi tesselations were constructed from the remaining centroids by employing the Voronoi algorithm available in the Scipy Python package. In keeping with the non-disclosure/competition clauses that were agreed upon with the MNO, it is not possible to publish a map of the original or redrawn Voronoi tessellations as it may reveal the position of cell towers.

In the interest of comparison, the xDR volume is spatially mapped from the Voronoi tesselations onto the administrative neighbourhoods of Istanbul and Kocaeli for which we have official population counts. Figure 11B and C plot the official population counts beside the mapped xDR volume across the neighbourhoods of Istanbul and Kocaeli.

The areas of most activity are highly correlated with residential and commercial built up areas in the two provinces. Conversely, the areas of least activity are overlapping mostly undeveloped areas like mountains, lakes, nature reserves, and agricultural land. When compared to the official population counts, there is a clear similarity between the high and low density areas in the mobile phone data and the official population.

**Trajectory completeness**

Trajectories, briefly mentioned earlier, are a tabulated data format that organises a customer's mobile phone data into a chronological sequence, such that their movements can be plotted on a map. Figure 12 depicts how trajectories can be constructed from trace data like MPD.

Looking at trajectory completeness, defined as the proportion of time steps for which a customer has at least one mobile phone record, can directly indicate the amount of information available in this format.

Figure 13 plots the trajectory completeness for the customers in our dataset after preprocessing. Clearly, the majority of users have a sparse trajectory relative to the total number of time steps in the dataset, though a small number have very

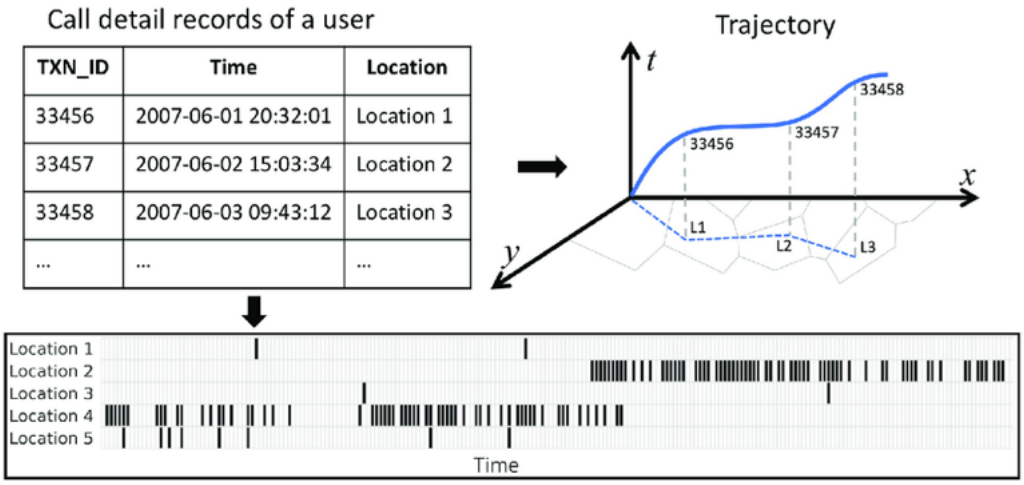


Figure 12. Extracting human trajectories from trace data. Raw data (top left) contains timestamps and geo-coordinates each time each individual is active on the platform (e.g., making a phone call). From these data, the trajectory of the person through space and time can be reconstructed (top right). The bottom figure shows the set of locations (e.g., neighborhoods) in which the individual was observed on each day. Figure from Chi et al. (2020).

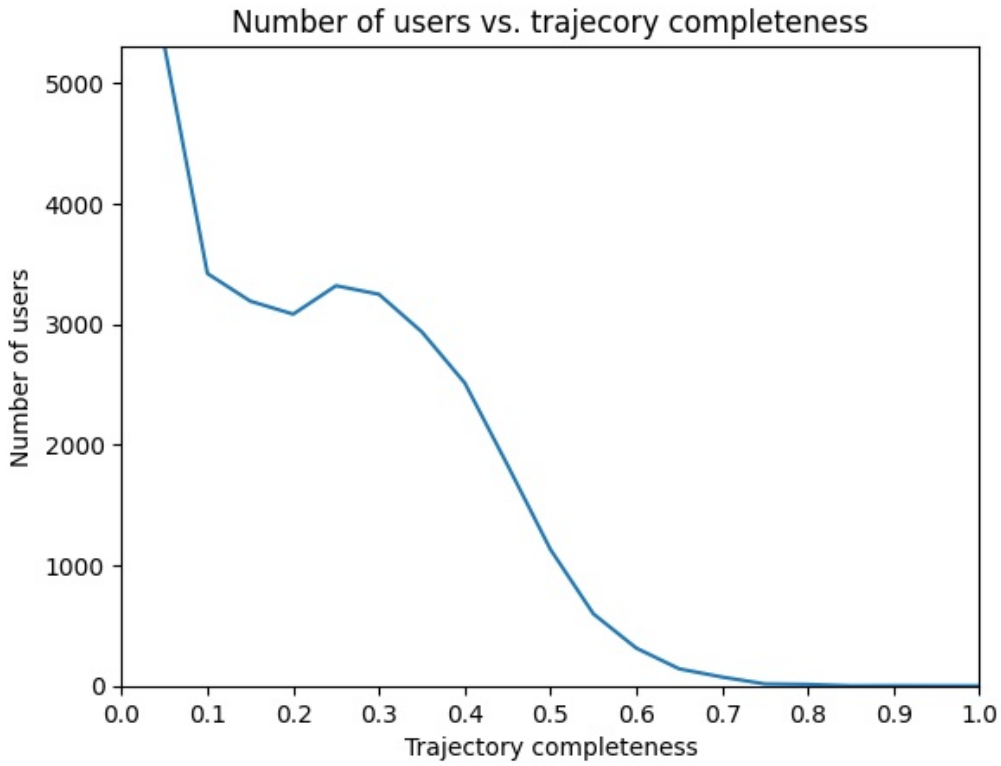


Figure 13. The trajectory completeness for customers in the xDR dataset. Trajectory completeness is the proportion of time steps for which a customer has a mobile phone data record.

complete trajectories. To put this in context, there are 383 hours or time steps in the dataset covering 16 days. A completeness of 10% would be 38.3 hours or 2.4 xDRs on average per day, which can still be enough to identify important locations

and other mobility habits for an individual.

**Home and work location detection**

Residence is an important concept in migration and mobility literature. In virtually all official censuses and surveys, population counts are based on place of residence, making it often the only information available for studying the spatial distribution of humans. The home is considered an anchor point for ones daily activities, being the place that people tend to spend most of their time, and the place from which they depart in order to participate in civil life. It provides context for peoples' behaviours, the extent of their mobility, and, on the longer term, migration events. Where one lives has many implications for access to resources, social capital, jobs, and much more. It is therefore desirable to extract home locations from the mobile phone trace data, not only for it's descriptive properties but also because we are able to compare our findings with official sources.

Work location is often considered the counterpart to home location in mobile phone literature. Knowing a person's work location in complement to their home location can provide a deeper insight into the social boundaries they experience. Working hours, commute times, and zoning codes of the area are some of the properties of the workplace that can help approximate a persons socioeconomic status.

Although they have been talked about here as static and presumed, the home and workplace are increasingly fluid concepts in today's world. Especially in the wake of the Covid-19 pandemic, which saw many industries adapt to lockdowns by virtualising jobs that had previously only been conducted in person. The rise of working remotely, whether from home or from practically anywhere with an internet connection, has fundamentally changed the relationship many people have with home and work. In a sense, there is no categorisation of home and workplace that is universally applicable and this is also true for mobile phone customers.

Choosing, then, a set of criteria for how to extract home and work location from mobile phone data is often done in an ad-hoc manner, using local accounts of work and leisure hours as a baseline for extraction algorithms. Yang et al. (2021) explain that there are two broad categories of approaches to differentiating home and work locations: observing the most visited location during daytime and nighttime and, alternatively, identifying meaningful places in a user's trajectory and assigning home and work locations based on visitation heuristics. Vanhoof et al. (2018) distinguish between continuous and non-continuous trace data, such as GPS and CDR respectively, and how discovering home locations from these different data have required diverging

approaches. In particular they highlight how it had been common, with regards to continuous traces, to cluster spatially the traces into important locations before further processing. They go on to explain that with non-continuous traces this approach of spatial clustering has largely been abandoned. Chi et al. (2020) describe a method for migration detection of which one of the steps of their algorithm can be interpreted as a home location detection step. This method organises an individual's trajectory into a diary where outliers are filtered out revealing a contiguous "segment" in which the individual is considered to be living in a single domicile.

Staying more proximal to the broader literature, we employ the method of most commonly visited location during the night- and day-time to identify home and work location respectively. This approach follows the logic that people tend to return home to rest in the evening and remain there until they resume their activities by leaving the house the following morning. Conversely for work location, the assumption is that people will be at work during typical office hours on business days.

Naturally, the choice of hours that define when people will be home or at work has an impact on results yielded and the interpretation of them. To determine work location, the cell tower with the most activity during local business hours on weekdays is a straightforward and unambiguous criteria. Of course not everyone is employed, so to speak, so the interpretation of "work" is more flexibly understood as the place of occupation, whether that is as a labourer, student, or unemployed. This however does assume that everybody is occupied during the same business hours, in the case of Turkey, 9 a.m. to 5 p.m. are the typical office hours for civil servants and banks, with a Monday to Friday work week.

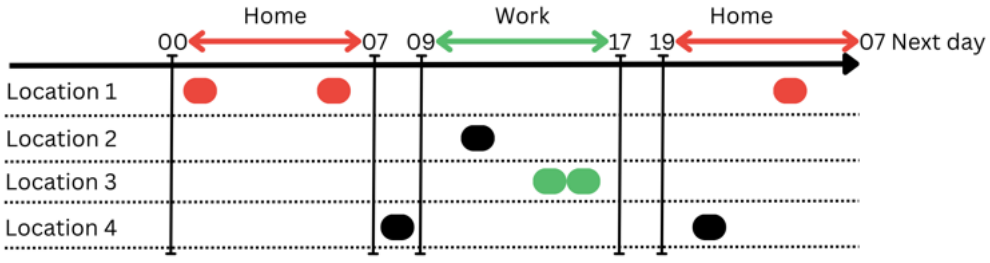


Figure 14. An illustration of how home and work locations are detected from a users' trajectory.

For home location, the choice of hours is less obvious and in practice is largely up to the discretion of the researcher. Results are often given a high-level validation by comparing them to census residence data in the absence of ground truth residence data for individuals. With access to such a ground truth dataset for individuals, Pappalardo et al. (2021) proceed to compare the accuracy of 13 ad-hoc home detection

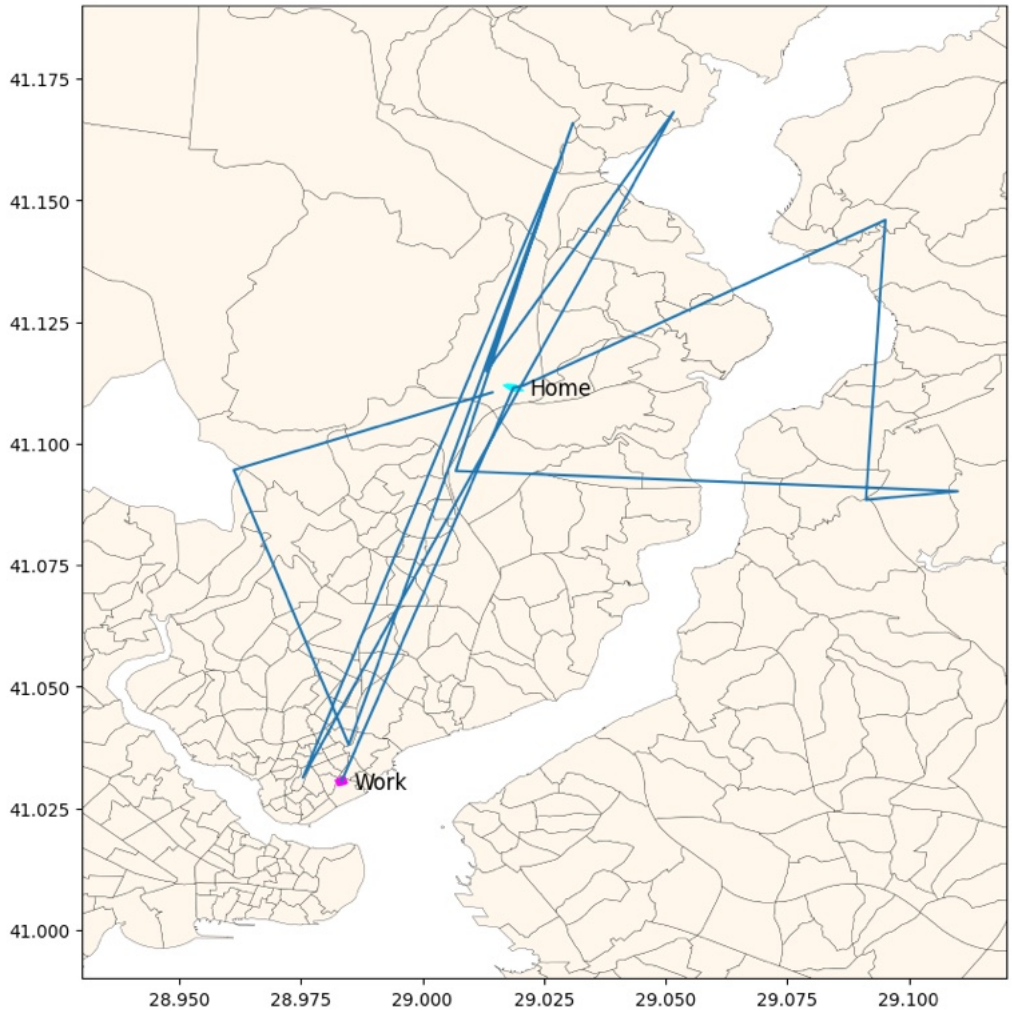


Figure 15. A (hypothetical) trajectory with detected home and work locations labelled in cyan and magenta respectively.

methods commonly applied in the literature on three types of mobile phone data including xDR. They find that the most accurate home detection method for xDR, at 69% accuracy, fixes an individual's home location at the cell tower with the most activity between 7 p.m. and 7 a.m. (nighttime) on weekdays. Since low-level ground truth data is not available to us, the empirical trials in their research help to guide the methodology used here. Following their results, we implement the best candidate home detection method as described above. Figure 14 provides a toy illustration of how the home and work locations are assigned and Figure 15 plots a hypothetical trajectory of a mobile phone customer, labelled with their detected home and work location.

Figure 16 shows the distribution of detected home locations found using this method. Overall, we see a very similar spatial distribution as the raw xDR data, which is sensible considering we assign the home location as the the most active cell tower

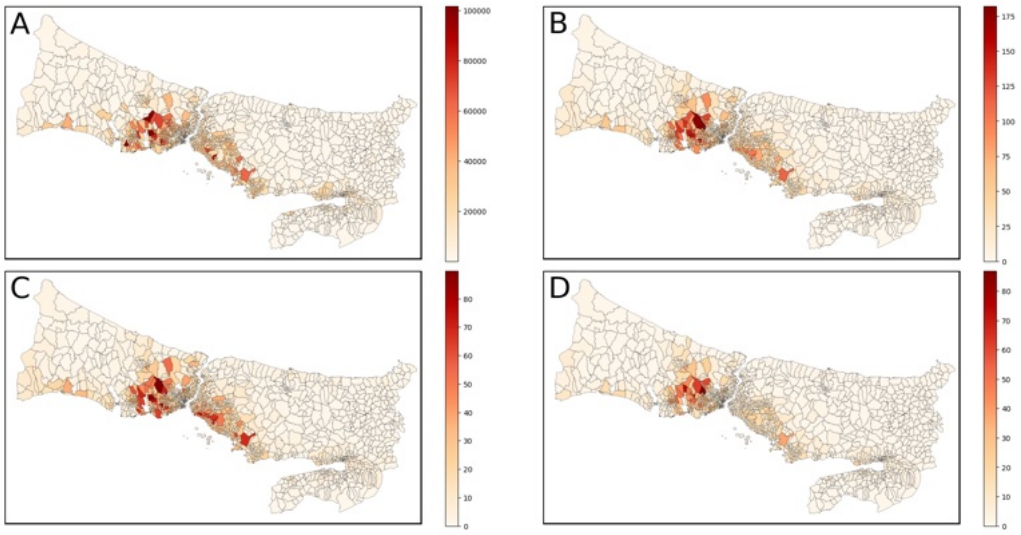


Figure 16. (A) Official population counts, (B) all detected home locations, (C) detected home locations of Turkish natives, (D) detected home locations of Syrians.

during the nighttime. If we compare the distribution of discovered home locations for Turkish natives and Syrians (as dictated by the nationality flags in the xDR dataset), it becomes evident that the native population is well distributed around the neighbourhoods respective to the official population counts and the Syrian population is concentrated in a handful of high density residential neighbourhoods on the Western side of the city as well as near the airport on the Eastern side.

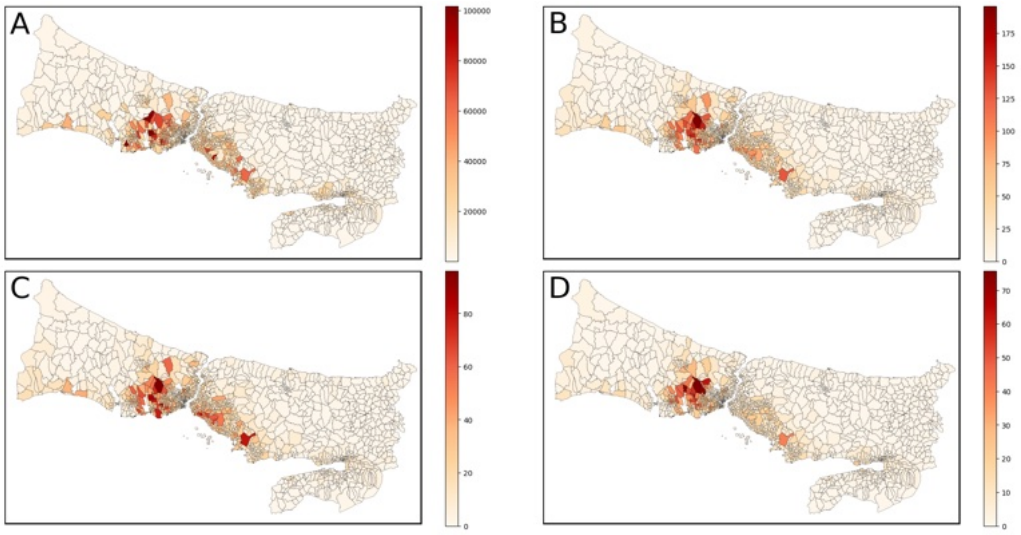


Figure 17. (A) Official population counts, (B) all detected work locations, (C) detected work locations of Turkish natives, (D) detected work locations of Syrians.

Figure 17 displays detected work locations in the same manner. There is very little difference between the home and work location distributions which can suggest that people tend to live close to their place of work or that many people work from home since this dataset captures a moment in time when covid lockdown measures are still

in place (although the measures are more relaxed during this period than during the beginning of the pandemic). To test this, we analyse the amount of customers who were assigned the same location for both home and work and discover that 75% of customers have been assigned identical home and work locations (excluding customers who were not assigned either a home or work location).

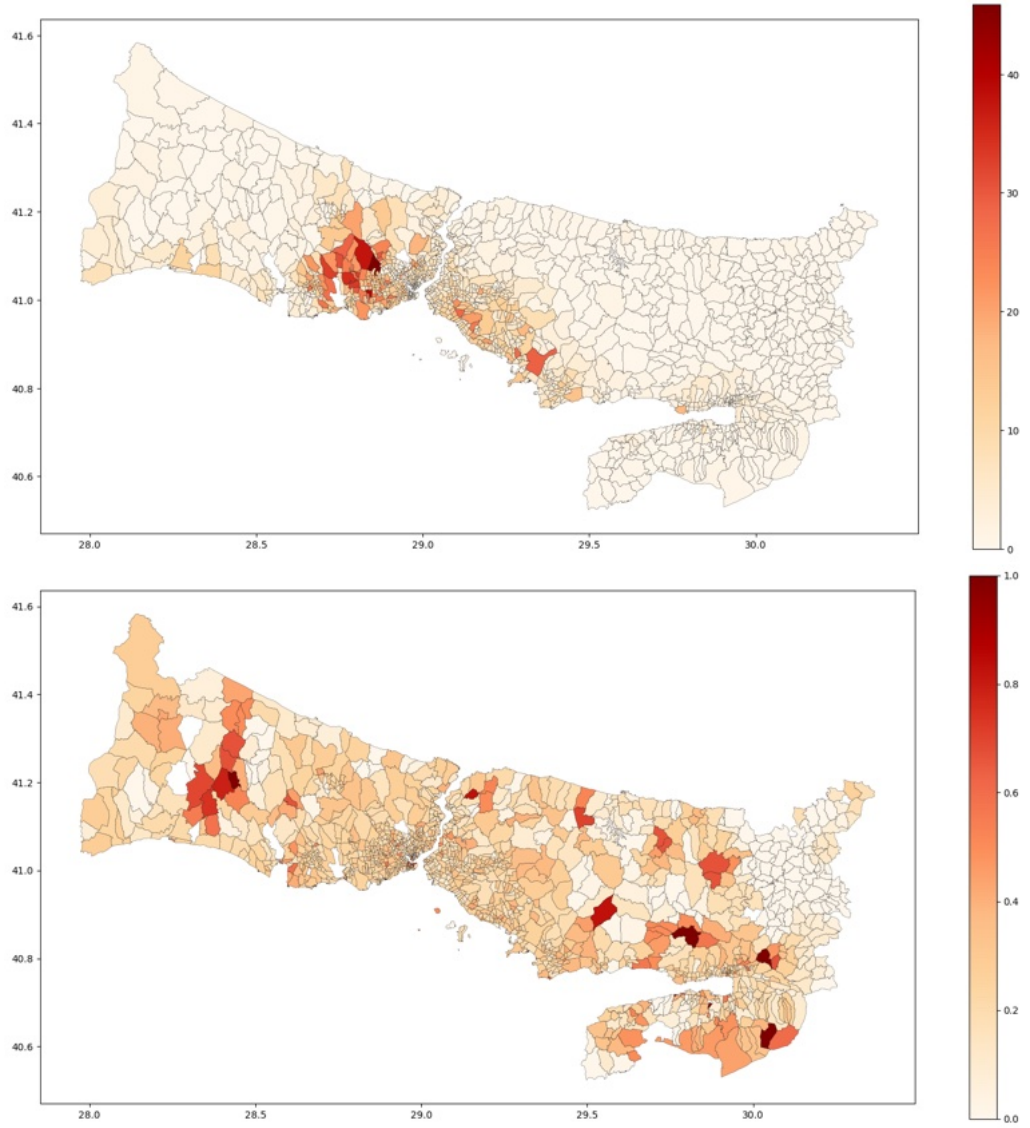


Figure 18. (A) The number of people who have different home and work locations. (B) The proportion per neighbourhood of people who have different home and work locations

Taking this analysis further, we map the distribution of home locations for people who work away from home. The result can be seen in Figure 18. While the overall proportion of Turkish and Syrian people who work away from home is the same as the proportion of Turkish and Syrian people in the population, it can be seen that neighbourhoods which are more populated by Syrians have a higher number of people who commute some distance for work. Additionally, if we consider the proportion of residents in a neighbourhood who go away from home to work, it can

be seen that rural neighbourhoods tend to have a greater proportion of people who go elsewhere for work.

### 3.2.2 xDR antenna traffic

#### Volume

Observing from the antenna traffic the xDR volumes across all cell towers, we expect to see a similar trend across the day as was seen in the fine-grained data. Figure 19 plots the summed xDR traffic per hour for Syrians. We can see that indeed, the same sinusoidal curve emerges with the exception of 11 p.m. being zero since the data for that hour was discarded. All segments demonstrate a virtually identical curve just with different orders of magnitude.

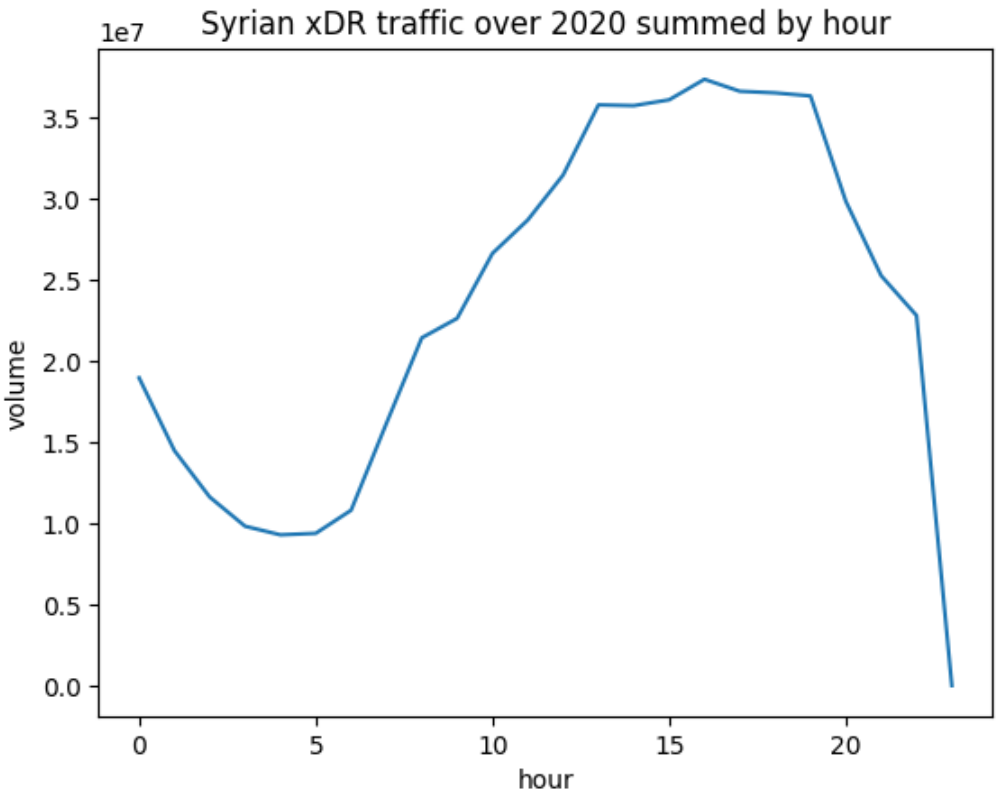


Figure 19. xDR volume for Syrians during 2020, summed by hour.

No further processing was required for this dataset in terms of data cleaning and validation. More however was done in terms of combining it with ancillary data. This is explained in the methodology section.

## 3.3 Methodology

### 3.3.1 Segmentation from segregation indices

Quantifying segregation is challenging due to the interconnected circumstances that bring it about. However, several indices have emerged that have become trusted litmus tests for segregation. Of these, the index of dissimilarity and the index of isolation stand out as the important benchmarks in migration literature.

Traditionally used to measure residential "evenness" and "exposure", respectively. By using xDR traffic as a proxy for the distribution of the different demographic groups across Istanbul, calculating the segregation indices becomes trivial. Additionally, we can move away from static, residence based calculations to reveal how the indices transform over the day. Ideally, this will provide a picture of how the urban environment is utilised by different ethnic groups at an unprecedented resolution.

For the calculating the segregation indices we use the aggregated xDR antenna traffic dataset because the time span of a year that it covers lends itself to greater robustness and stability in simulating population stocks. Moreover, in this situation there is no interest in observing individuals' mobility but rather we are interested in population level dynamics, which this dataset is suitable for.

Here we outline the procedures taken to achieve this, the results of the experiment, and our findings thereafter.

**Intersecting Voronois and Neighbourhoods**

Residential segregation indexes, like the dissimilarity and isolation, almost always depend on census tract data for calculation. That census tracts are intentionally drawn to fit neatly within administrative boundaries is important, as statistics at the census tract level can be imported directly into the calculation. Our situation is less straightforward.

Our statistics are collected at the level of Voronoi tessellations, which importantly do not fit neatly within administrative boundaries, often overlapping several at once. In order for us to calculate the dissimilarity and isolation indices we must first carve up the Voronoi tessellations according to the lowest level administrative boundaries and split the data between those subdivisions. This will result in something akin to census tracts which will serve as the geographical sub-units from which the calculations can be made.

The data is split in a rudimentary fashion using the proportional area of the sub-unit from the original Voronoi tessellation. There is of course the issue that the urban landscape is not uniformly occupied, which is the critical assumption made with this method, and certainly more sophisticated methods could be used to more accurately map the mobile phone data. This is something to be investigated more thoroughly in the future though we tolerate it for this demonstration.

**Land use masking**

The city of Istanbul, despite hosting some 15 million residents, only makes up a fraction of the total land area within the province of the same name. That metrics, like the indexes of dissimilarity and isolation, are typically drawn on a map without consideration for population centres or urban usage contribute to a distorted impression spatial dynamics. By masking the geographic area of the province with land use data that corresponds only to developed urban sites, we intend to improve the interpretability of our findings.

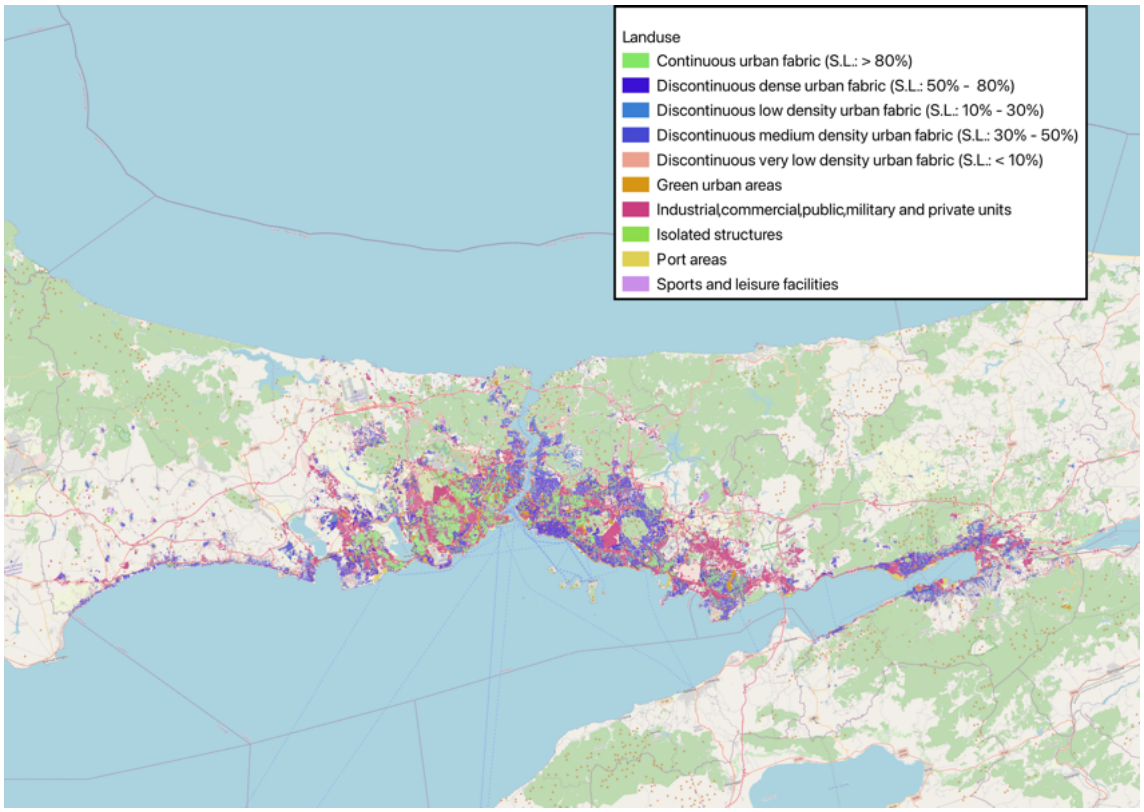


Figure 20. Landuse zones filtered to exclude non-urban areas.

Figure 20 shows the land use mask we employ, along with information about the particular land uses that were left intact by the filter. The presumption in using this mask is that urban areas will account for much more mobile phone traffic than non-urban areas like forests and farmland. By masking the Voronoi tessellations with this filtered landuse profile, we can more accurately localise where mobile phone

traffic is being generated.

**Calculating segregation indices**

We define the index of dissimilarity and index of isolation with respect to Bertoli et al. (2021), who use CDRs in a similar methodology.

Accordingly, the formula for the index of dissimilarity is defined as

$$D = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{a_i}{A} - \frac{b_i}{B} \right|, \tag{3.3}$$

where $a_i$ and $b_i$ are the population of the minority and majority group, respectively, at grid square $i$. $A = \sum_{i=1}^{n} ai$ and $B = \sum_{i=1}^{n} bi$ are, respectively, the total population of the minority and majority groups at the neighbourhood level where $n$ is the number of grid squares in a neighbourhood. The index of dissimilarity $D \in [0, 1]$ is a measure of "evenness" that can be interpreted as the proportion of the minority group that would need to relocate within a neighbourhood to restore an even distribution.

The formula for the index of isolation is defined as

$$I = \sum_{i=1}^{n} \frac{a_i}{t_i} \times \frac{a_i}{A}, \tag{3.4}$$

where $t_i = ai + bi$ is the total population at grid square $i$. The isolation index $I \in [0, 1]$, unlike the dissimilarity index, is sensitive to the total proportion $P = A/T$ of minorities in the neighbourhood population, where $T = \sum_{i=1}^{n} ti$. Massey and Denton (1988) propose an adjusted isolation index $I\_adj$ that is less dependent on the proportion $P$. The formula for the adjusted isolation index is defined as

$$I\_adj = \frac{I - P}{1 - P}. \tag{3.5}$$

The index of isolation can be interpreted as a measure of the probability that a member of the minority group comes in contact with a member of the majority group.

At their extrema, the two indices represent the same result. When both are 1 it means there is complete segregation, conversely there is a uniform distribution of the minority population across every spatial unit when they are both 0.

Both dissimilarity and isolation segregation indices depend on a comparison of population counts between a majority and minority group. In our case the majority

group is the native Turkish population and we compute the segregation indices for two minority groups; Syrians and Afghans.

To ensure the robustness and stability of the population proxy we employ the aggregated xDR antenna traffic data. This dataset aggregates mobile phone traffic counts of an entire year into hourly counts, conveniently separated by segment. Taking the hourly sum of xDR traffic over 365 days effectively eliminates any day-to-day variations in traffic volume, ensuring a stable proxy for population even at very small spatial scale.

For the sake of keeping the amount of plots manageable while still visually demonstrating temporal variation, the traffic data is aggregated again into daytime and nighttime sets. Daytime is chosen as being between 7 a.m. and 7 p.m., nighttime is chosen to be its complement.

### 3.3.2 Segmentation from eigenbehaviours

It can be said of human behaviour that it is highly regular and, at the same time, that individuals' routines are deeply personal and unique. The confluence of these two qualities suggest that much may be learned about the differences between people by observing where their behaviours diverge. In concurrence with our inquiry into patterns of segregation, this part of the methodology explores how behaviour can help delineate differences in socioeconomic status across ethnic lines.

Relevant to our research, mobile phone traces are saturated with behavioural information. Over a period of weeks, it becomes possible to observe frequented locations and phone usage habits that are highly expressive of an individuals behaviour. However, some clever feature engineering is required so that we may directly investigate the behaviours in a computational way.

In their work titled eigenbehaviours, Eagle and Pentland (2009) execute exactly such a method using real-time sensor traces, though they do not discuss the implications for segregation analysis. By reproducing their method with our mobile phone data set, we seek to identify the utility of behavioural analysis for witnessing discrepancies between different demographics on a city wide scale.

#### Data differences

Importantly, our datasets have significantly different characteristics in resolution and scope that affect the outcomes of the method. Foremost, their trajectory data is much denser and more accurate than our mobile phone data but covers a much smaller area and much fewer individuals. They recruit several staff members and students

of their university to participate in the research and, after installing software on their mobile phones, log when these volunteers pass by dedicated beacons scattered throughout campus.

The result of their data collection is trajectories that are very complete, regularly capturing a whole day's worth of location information for each individual with great spatial and temporal accuracy. In comparison, our xDRs, when observed as trajectories, are sparse, containing no more than a handful of locations for each individual per day, additionally only covering a 16 day span as opposed to over 100 days in their case. Furthermore, their data collection method, being contained to the university buildings, make it very easy to determine when individuals are at "work" or at "home", i.e. when they are present at the building or not, as well as having context for the various locations on campus.

Despite the differences in our datasets, as we are able to approximate features similar to the ones used in their paper, it seems plausible that their method can reveal group affiliations and differences in lifestyles between Syrians, Afghans, and Turks. Perhaps we may discover less about ethnic differences and more about class differences, though this will be hard to quantify in the absence of additional data on class status, occupation, and income.

**Vector formatting**

Eagle and Pentland's method requires us to encode trajectory data into behavioural vectors. They do so by first assigning each trajectory time step one of five categories: home, office, elsewhere, nosig [no signal], and off. This is transformed into a binary matrix as seen in Figure 21.

To operationalise this vector format for our own data we must do a similar assignment for the time steps in our trajectory data. Most important are the home and office/work locations. We use the detected home and work locations established earlier in section 3.2.1 to assign this feature in the trajectory of each individual. Additionally, where there is an xDR in the trajectory that is neither home nor work, we assign the "other" tag.

We limit ourselves to three categories: home, work, and other (similar to elsewhere). Since we are unable to know whether a mobile phone is off we can combine this feature with nosig for the case when there is no xDR present at that time step. During our analysis we find that such a feature is not informative so we drop it and stick to the three features mentioned. An example binary behaviour matrix

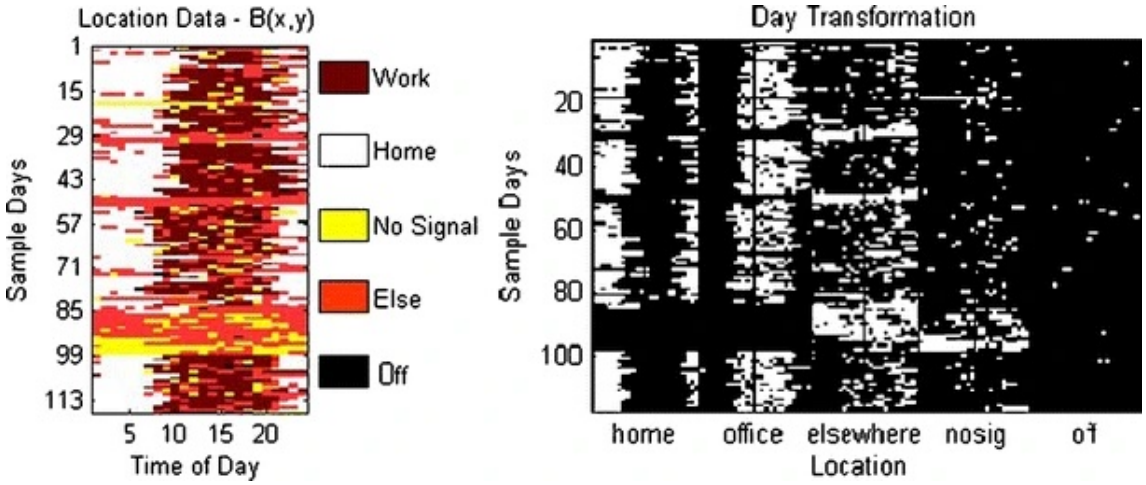


Figure 21. "Transformation from B to B'. The plot on the left corresponds to the subject's behavior over the course of 113 days for five situations. The same data can be represented as a binary matrix of 113 days (D) by 120 (H, which is 24 multiplied by the five possible situations)" Figure and caption from (Eagle and Pentland, 2009)

generated from our data can be seen in Figure 22.

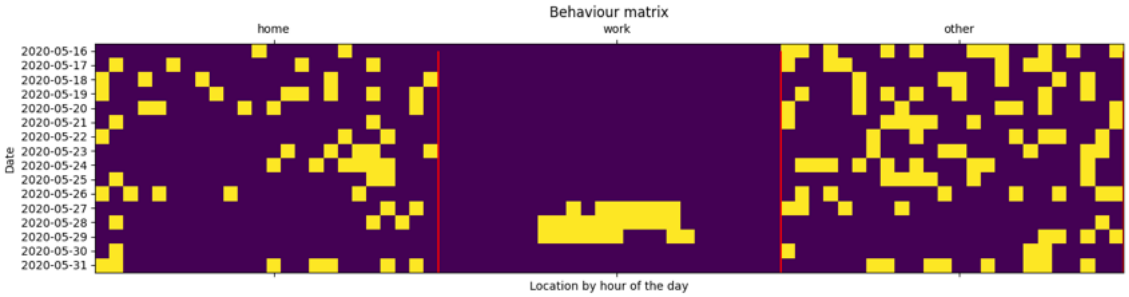


Figure 22. An example binary behavior matrix generated from the xDR data.

Note that Eagle and Pentland use only trajectory data from weekdays. Although some figures show that we include weekends, our final calculations only consider weekdays.

### Matrix decomposition

The next step involves using a matrix decomposition technique that is common in image processing called eigendecomposition. Eigendecomposition decomposes a matrix into three simpler matrices, revealing the underlying structure and properties of the original matrix. Mathematically, given a square matrix $A$ of size $n \times n$ the eigendecomposition of $A$ is given by:

$$A = U\Lambda U^T \tag{3.6}$$

where $U$ is an $n \times n$ orthogonal matrix containing the eigenvectors of $A$ and $\Lambda$ is

an $n \times n$ diagonal matrix containing the eigenvalues of A. We can obtain a square matrix A by calculating the covariance matrix of our behaviour matrix, the set's deviation from the mean. To calculate the covariance matrix the behaviour matrix must first be centred by subtracting the mean of each variable from each observation, such that the data is centred around zero:

$$X_{centred} = X - \bar{X} \tag{3.7}$$

The covariance matrix A can then be calculated by taking the dot product of transposed and centred matrix $X$ with itself and then dividing by the number of observations $m$:

$$A = \frac{X_{centred}^{T} X_{centred}}{m} \tag{3.8}$$

Paraphrasing from Eagle and Pentland (2009), because most people's lives exhibit significant structural similarities, their days do not occur randomly within a vast vector space. Instead, they tend to cluster, defining a relatively lower-dimensional "behaviour space" that characterises individuals. This space is delineated by a subset of vectors that best encapsulate the distribution of behaviours, known as primary eigenbehaviours. Each eigenbehaviour is prioritised based on the variance it explains in the data, reflected by its associated eigenvalue. The eigenbehaviours with the highest eigenvalues are deemed an individual's primary eigenbehaviours. These primary eigenbehaviours serve as a framework onto which all of an individual's days can be mapped with varying degrees of precision. Figure 23 depicts the top three eigenbehaviours of the individual who's behaviour matrix is shown in Figure 22.

From this figure we can see they this individuals first eigenbehviour depicts a situation where they are at home mainly at night but sometimes until midday, they primarily are at work between 7 a.m. and 5 p.m., and they are elsewhere intermittently throughout the day. The other primary behaviours are rather similar. This may be because the assigned locations for this individual come across as dubious and inconsistent making and the interpretation challenging.

To demonstrate how the decomposition is encoding the original behaviour, we can attempt to reconstruct the original behaviour by masking various numbers of eigenvalues while combining the sub-matrices produced in the eigendecomposition.

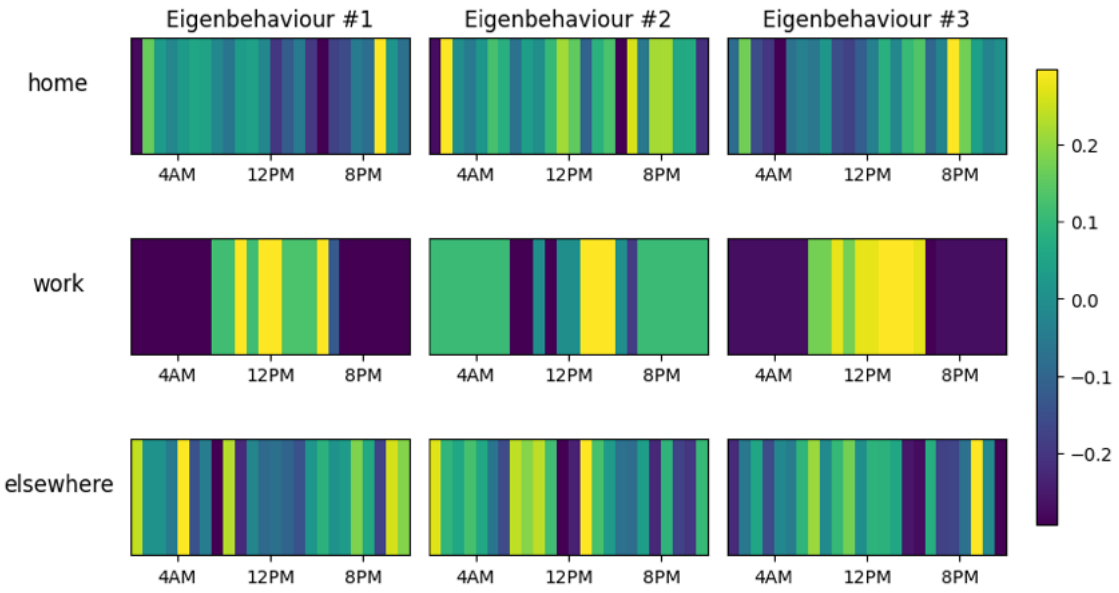


Figure 23. The primary eigenbehaviours for an example individual.

Figure 24 shows the recomposition using one, five, and finally all 16 eigenvalues.

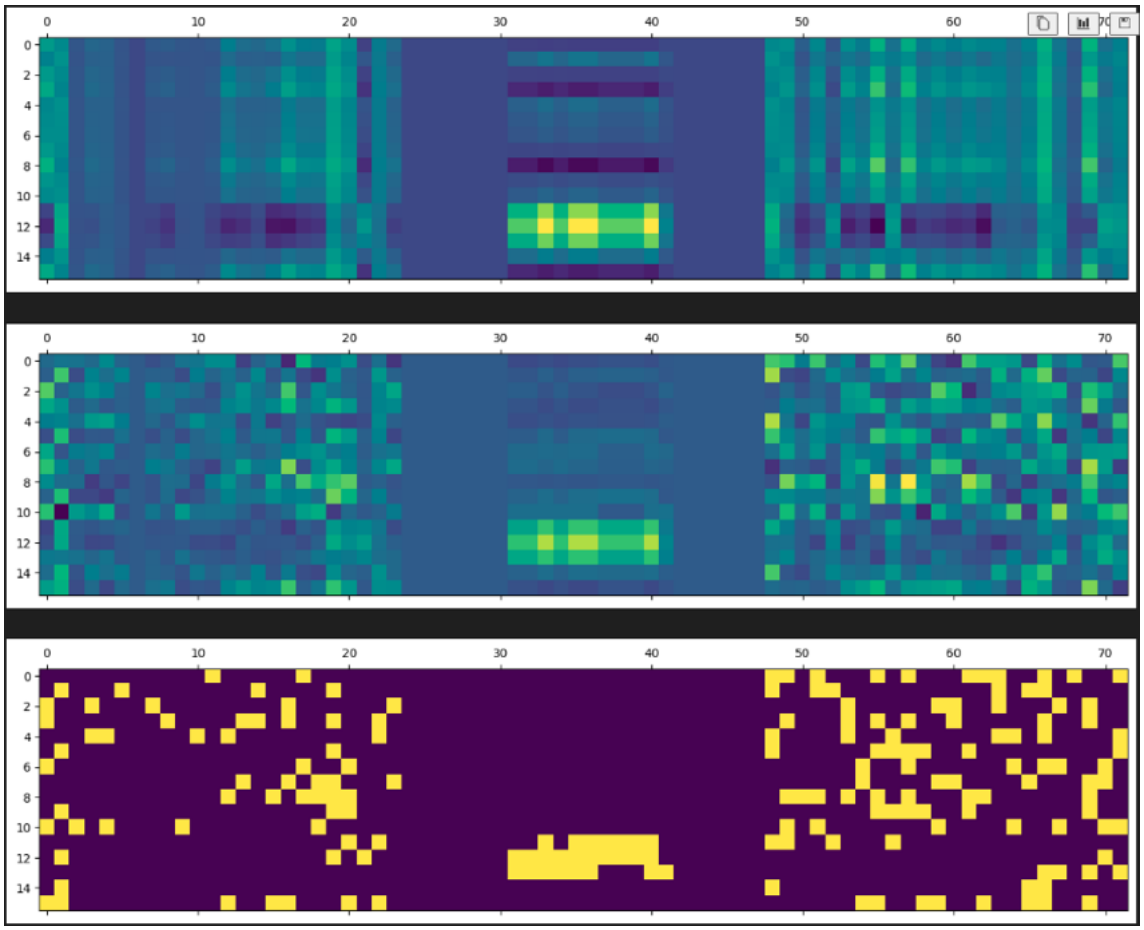

Figure 24. Recomposing the original behaviour matrix using varying numbers of eigenbehaviours.

Furthermore we can use an elbow plot to discern the amount of eigenvalues that explain enough variance to reconstruct the behaviour with accuracy. Figure 25 shows how this looks in practice.

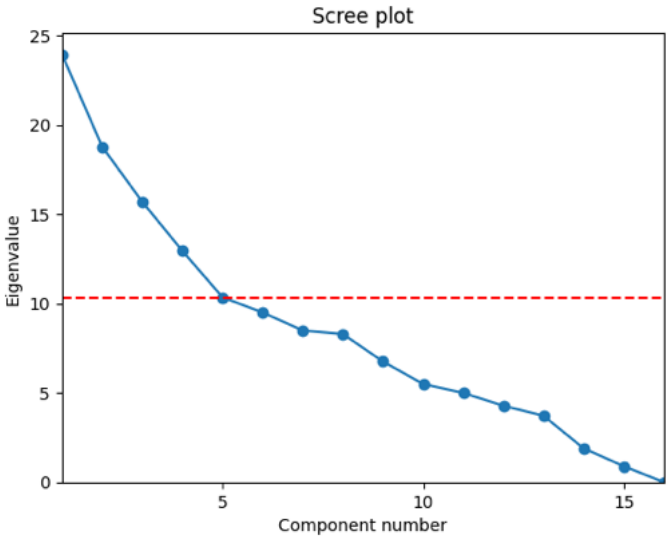


Figure 25. An elbow plot that shows the convex point where including more eigenvalues explains less and less variance.

Although this is a starting point, we are actually interested in the group dynamics of Syrians, Afghans, and Turks and so we move on to generating group eigenbehaviours. This is done in a very similar way, however in this case the behaviour matrix for each group is calculated differently. To generate a behaviour matrix for each group we take the average behaviour of each individual in the group, seen before as $\bar{X}$, and use these as observations or rows in the behaviour matrix. The calculation for the covariance matrix A is then performed as before as well as proceeding eigendecomposition.

## 3.4 Experimental Results

### 3.4.1 Segmentation from segregation indices

#### Neighbourhood level segregation

Firstly, the dissimilarity and isolation scores are calculated per neighbourhood. Therefore the scores should be interpreted as being independent estimations of segregation within each neighbourhood. The segregation indices are not typically calculated at this small scale as the great variety in size and density of neighbourhoods, and districts for that matter, can have a strong effect. Nevertheless, this experiment demonstrates the potential of mobile phone xDRs in generating high resolution insights for policy making.

Figure 26 visualises the dissimilarity index for both Syrian and Afghan refugees at

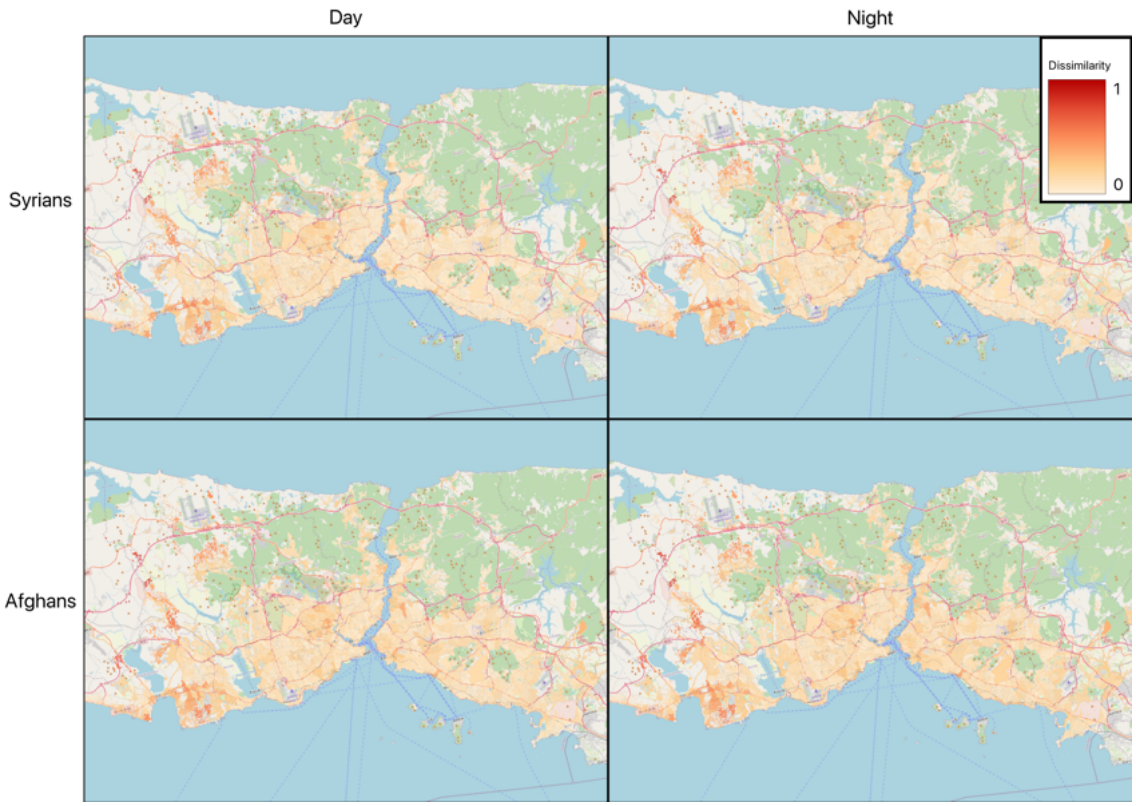


Figure 26. The index of dissimilarity calculated at the neighbourhood level using very fine-grained, Voronoi-neighbourhood intersected polygons as sub units.

the neighbourhood level for night and day. Both groups seem to experience high rates of dissimilarity in the same neighbourhoods especially in the outskirts towards the West of the city but also in select neighbourhoods closer to the Golden Horn and on the Asian side. We see temporal variations in the dissimilarity of neighbourhoods. However, dissimilarity is generally more pronounced at night.

For Syrians on the European side of the city, the dissimilarity is inversely proportional to the size of the minority population, with areas of low minority population experiencing higher dissimilarity. This however does not hold for the Asian side of the city, where both population and dissimilarity remain low.

There are neighbourhoods right along the waterfront called “Yeşiltepe” and “Gökalp”, part of the “Zeytinburnu” district, that have slightly raised dissimilarity for Afghans which match the information from the mobile phone traffic that there is an enclave of Afghans there. Afghans have an overall higher dissimilarity across the city than Syrians.

Figure 27 and 28 visualise the day and night adjusted isolation scores for Syrian and Afghan refugees respectively. Note that the legends do not correspond to the same

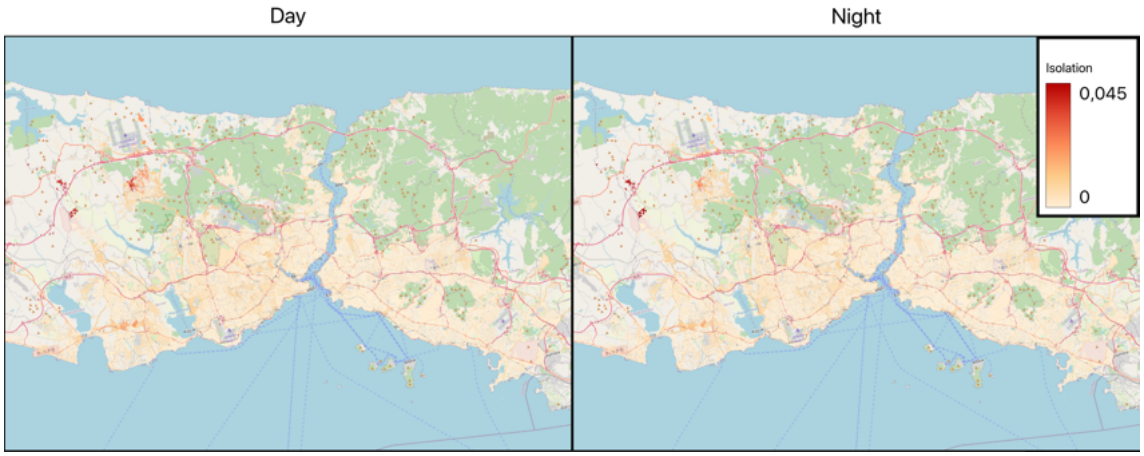


Figure 27. The adjusted index of isolation for Syrians calculated at the neighbourhood level using very fine-grained, Voronoi-neighbourhood intersected polygons as sub units.

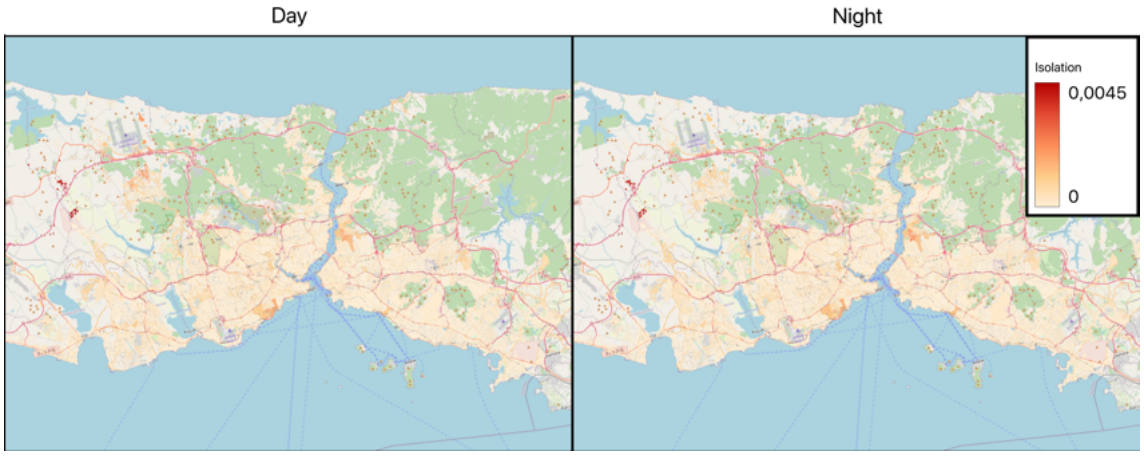


Figure 28. The adjusted index of isolation for Afghans calculated at the neighbourhood level using very fine-grained, Voronoi-neighbourhood intersected polygons as sub units.

scale, Figure 27 has a scale 10× greater than Figure 28. Even so, the isolation scores are very low for both segments, which is to be expected considering that Istanbul is such a populous city and that native Turkish people make up the vast majority of people in the city. It is less clear from the maps whether there is a general trend in increasing or decreasing isolation over the day, we investigate this point later in the section.

A closer look at the maps reveals that Syrian people experience slightly higher isolation scores in the Western outskirts of the city as well as near the international airport in the North West.

Afghans on the other hand experience greater levels of isolation closer to the city centre and where we expect that there is an Afghan enclave along the waterfront of the European side. In addition there is a pocket of isolation near the centre of

the peninsula on the Asian side along the so called “Göksü Creek”, as well as to the south of the Asian peninsula in an area called “Küçükyalı”.

### District level segregation

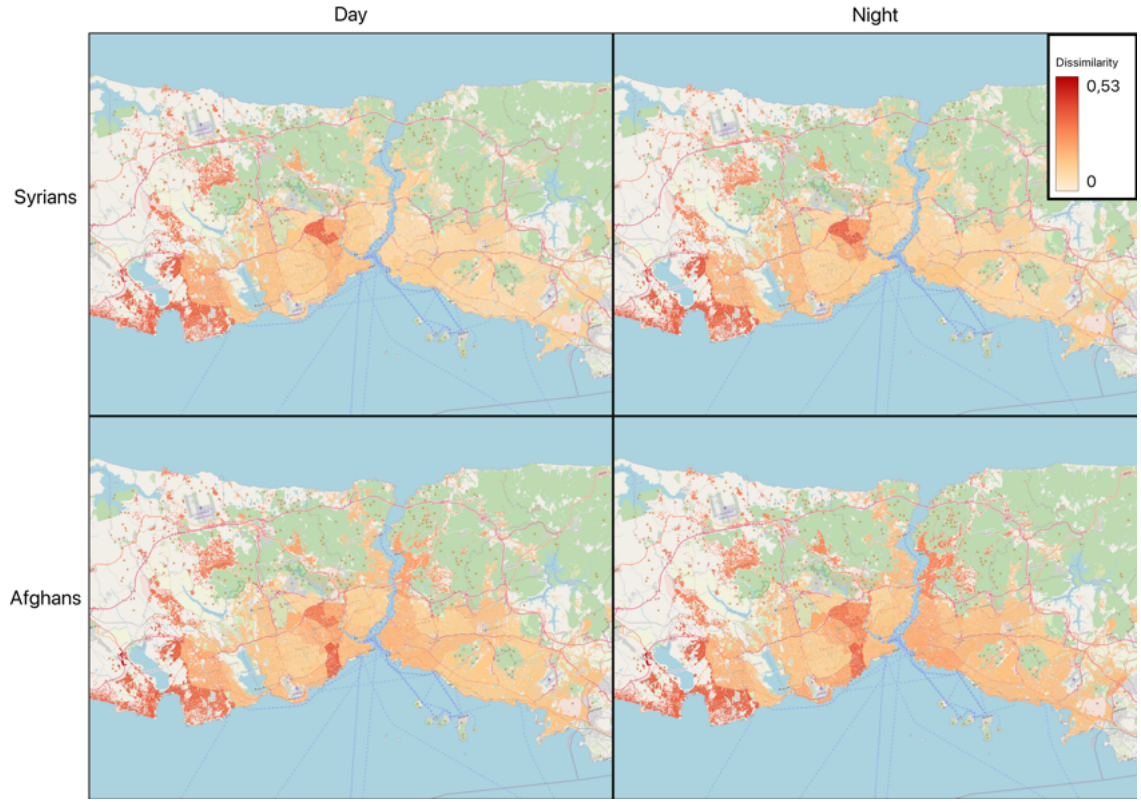


Figure 29. The index of dissimilarity calculated at the district level using very fine-grained, Voronoi-neighbourhood intersected polygons as sub units.

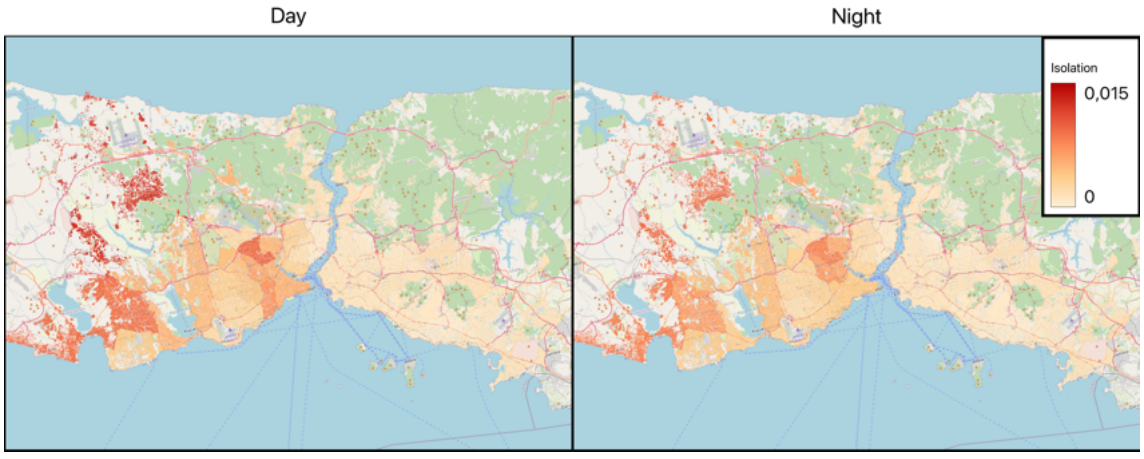


Figure 30. The adjusted index of isolation for Syrians calculated at the district level using very fine-grained, Voronoi-neighbourhood intersected polygons as sub units.

In addition to calculating the segregation indices at the neighbourhood level, we have calculated them also for the district level. In the same manner as before, Figure 29 visualises the dissimilarity index and Figures 30 and 31 visualise the adjusted isolation index for Syrians and Afghans respectively.

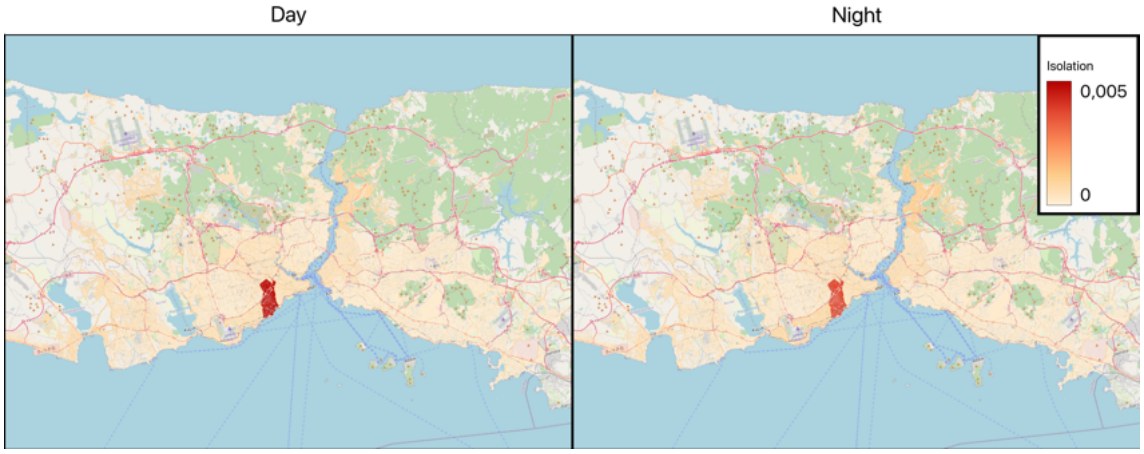


Figure 31. The adjusted index of isolation for Afghans calculated at the district level using very fine-grained, Voronoi-neighbourhood intersected polygons as sub units.

Looking at the dissimilarity plot, there is a clear resemblance with the neighbourhood level calculations. It appears that the areas of concentrated unevenness were projected upwards to both neighbourhood and district level calculations.

In more detail, it can be seen that both Syrian and Afghan refugees are unevenly distributed in the Western outskirts of the city. Again there is a high dissimilarity in district "Zeytinburnu" for Afghans and on the central point of the Asian peninsula. We again see a very slight increase of dissimilarity in the night.

Concerning isolation, we can see that Syrians experience a noticeable shift in the districts where they are isolated, particularly in the inner city locations. Oddly the map appears to depict higher isolation during the day which does not correspond to the other results but this may be due to selection of night and day times.

For Afghans only "Zeytinburnu" stands out. Again we see a higher isolation during the daytime which is unexpected.

Furthermore, we again calculate the district level segregation indices but use neighbourhood aggregated mobile phone traffic as geographic sub units to compare the effect of using differently sized sub units. Notably, the scores for dissimilarity and isolation are lower when using neighbourhood aggregated counts, although, relative to one another, the scores are almost identically distributed as with the sub-Voronoi aggregated counts.

**Evolution of the segregation indices**

While the maps are useful for locating the areas of high and low segregation, it is difficult to accurately tell how the scores change over time. Thus we explore

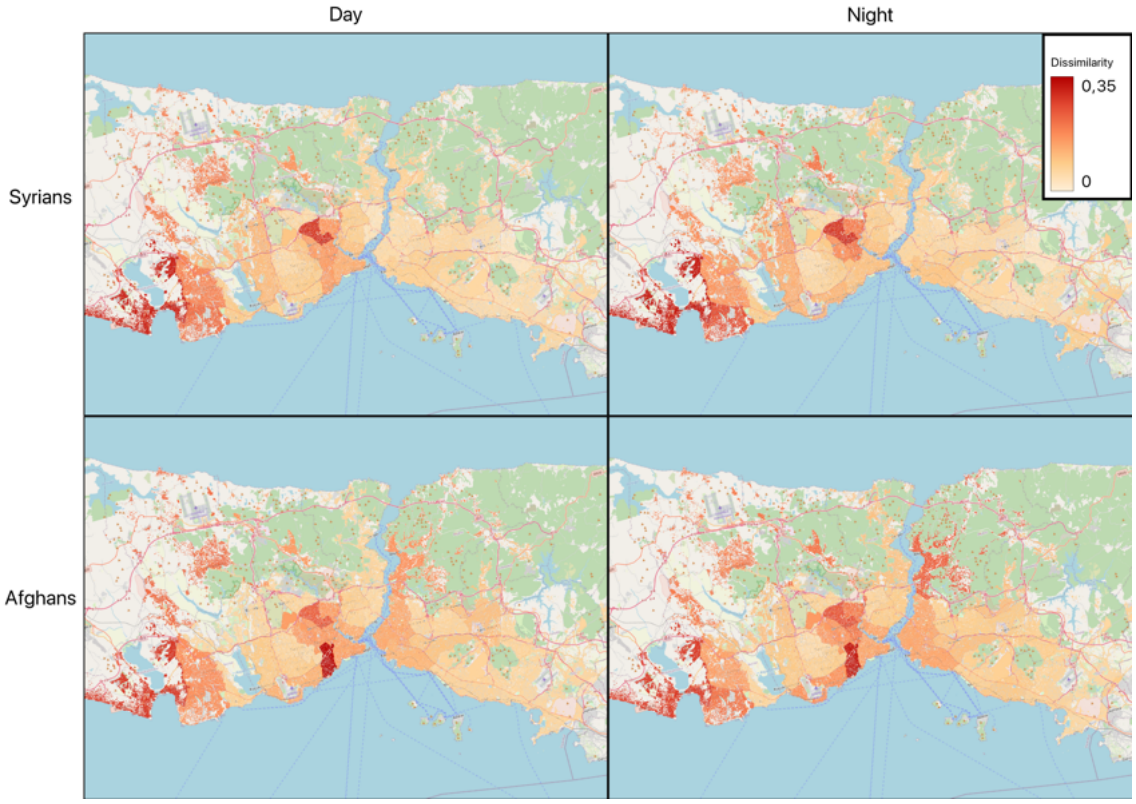


Figure 32. The index of dissimilarity calculated at the district level using neighbourhoods as sub units.

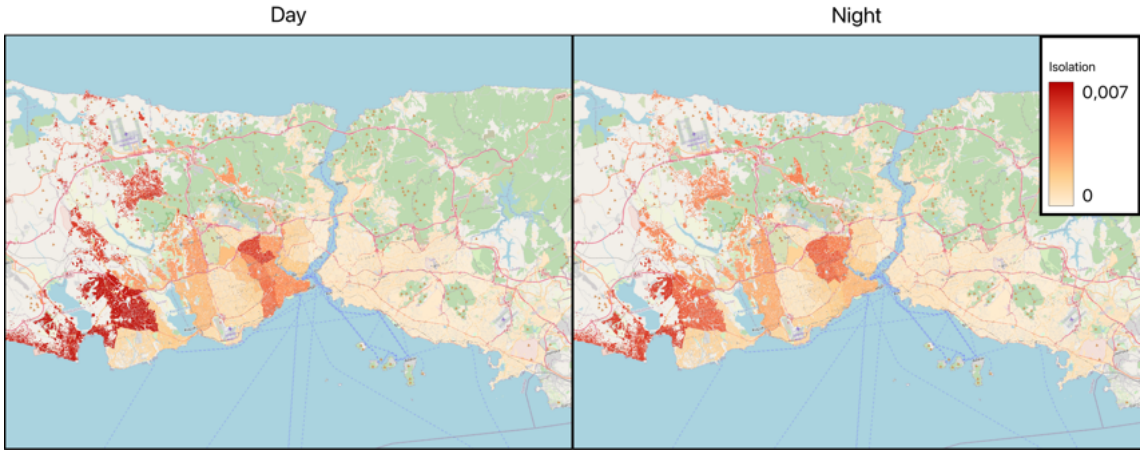


Figure 33. The adjusted index of isolation for Syrians calculated at the district level using neighbourhoods as sub units.

separately the evolution of the dissimilarity and isolation indexes at the district level over the course of 24 hours.

Figures 35 and 36 plot the evolution of the district level segregation indices over the hours of the day for Syrian and Afghan refugees respectively, with dissimilarity plotted in blue and isolation plotted in orange. It should be noted that a few of these districts are part of the Kocaeli province bordering Istanbul. Additionally, note the

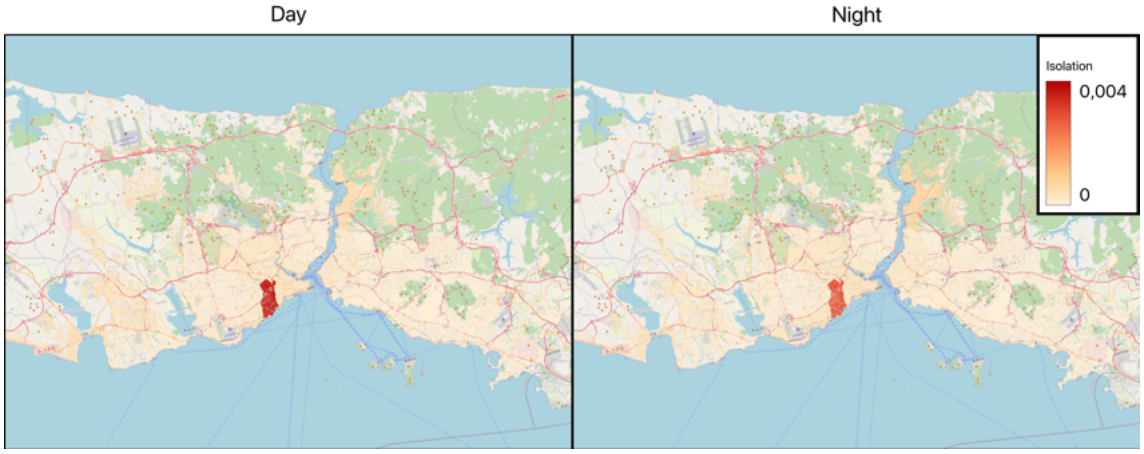


Figure 34. The adjusted index of isolation for Afghans calculated at the district level using neighbourhoods as sub units.

separate dissimilarity and isolation axes on the left and right respectively, with the x axis being the hours of the day.

Beginning with Syrian refugees, it is interesting to see that dissimilarity and isolation scores tend to fluctuate together but isolation scores in almost all districts experience a strong spike at around 5 a.m. that are not tied to dissimilarity. Some districts are very stable in both indices throughout the day like "Maltepe" and "Pendik" but these tend to be places with very little Syrian population on the Asian side of the city. Conversely, there are districts which experience quite intense fluctuation throughout the day in one index or the other like "Bakirkoy" or "Cayirova" and "Gebze" in Kocaeli, each of which experience higher segregation at night. On this point, there does appear to be higher dissimilarity and isolation scores at night in many districts, though not all. In particular, "Esenler", "Esenyurt", "Fatih", and "Zeytinburnu" express higher rates of segregation during the day.

Finally, we look at the evolution of segregation indices for Afghan refugees. Again there is a spike in isolation around 5 a.m. though much less pronounced an effect than with Syrians. The dissimilarity and isolation, like in the case for the Syrians, also mostly follow each other through the day. In most districts the segregation scores are slightly higher during the night. "Catalca" and "Esenyurt" are the only districts where this is contrary. "Zeytinburnu", the district where we expect there is an Afghan enclave, has relatively high but stable dissimilarity and has the highest isolation, peaking at 5 a.m. and dipping during the day.

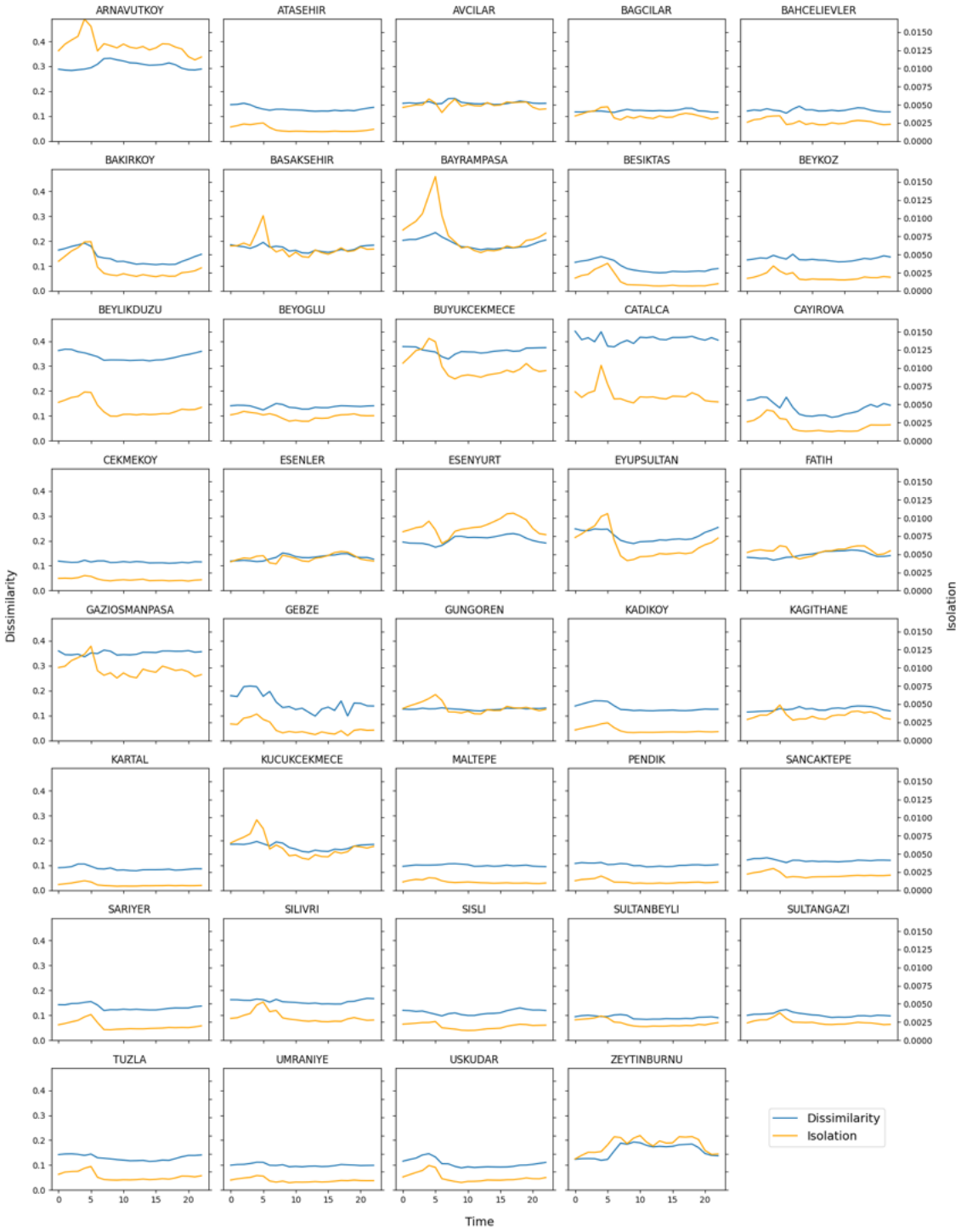


Figure 35. The evolution over 24 hours of the dissimilarity and adjusted isolation index of Syrian refugees for each district of Istanbul. The indices are calculated using mobile phone traffic at the sub Voronoi level (Voronoi cells intersected with neighbourhoods.) Mobile phone traffic was summed over the entire year of 2020 for each hour.

### 3.4.2 Segmentation from eigenbehaviours

Figure 37 shows the average behaviours of all three groups: Turks, Syrians, and Afghans. It is obvious from this plot that each of the groups share a very similar pattern of behaviour. Typically people participate in the workplace most frequently

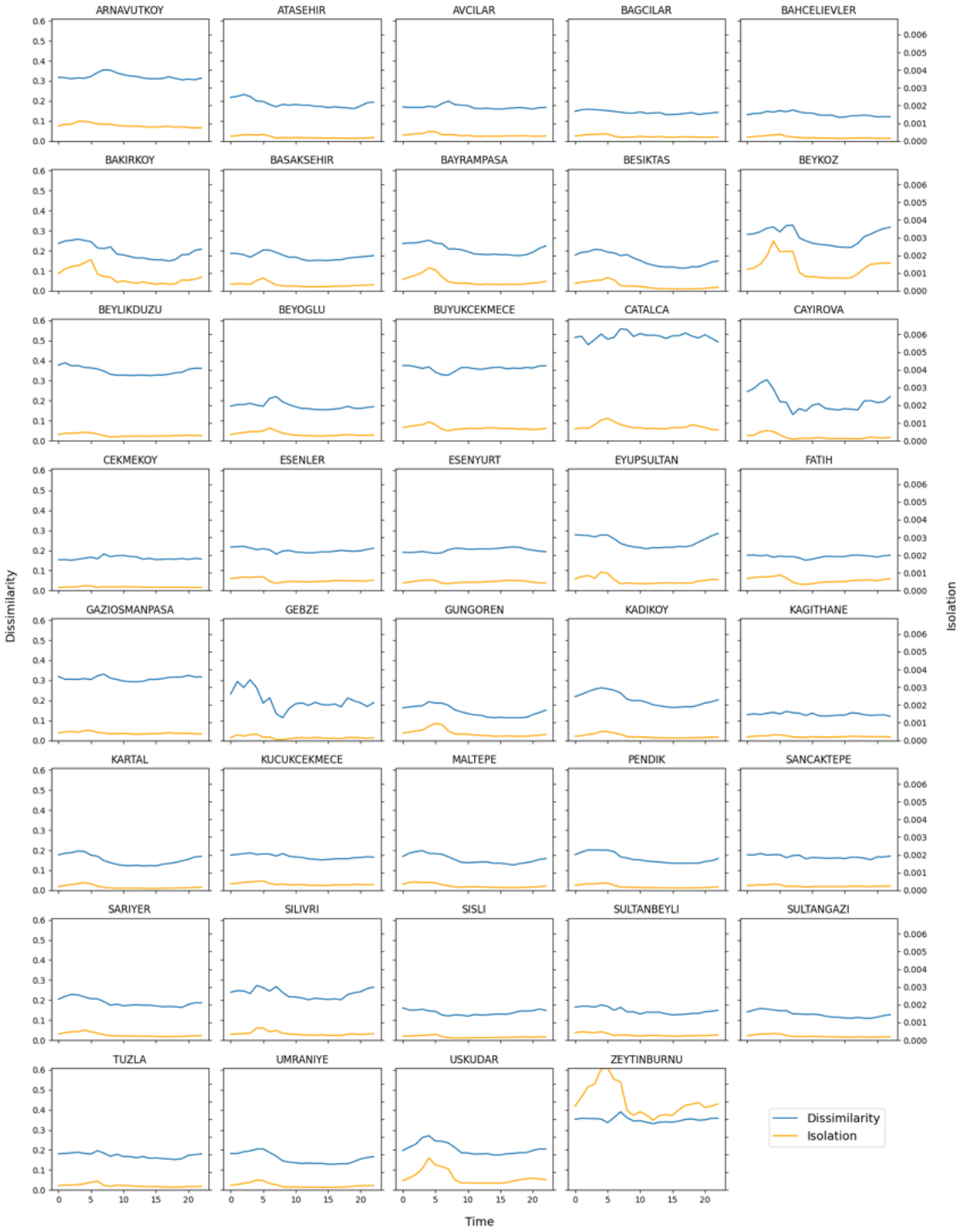


Figure 36. The evolution over 24 hours of the dissimilarity and adjusted isolation index of Afghan refugees for each district of Istanbul. The indices are calculated using mobile phone traffic at the sub Voronoi level (Voronoi cells intersected with neighbourhoods.) Mobile phone traffic was summed over the entire year of 2020 for each hour.

in the early afternoon between 1 p.m. and 5 p.m., although Syrians and Afghans are seen to potentially work longer hours, up to 9 p.m. In terms of time spent at home, we see that all the groups ate most frequently at home around 9 p.m. although Turkish people tend to be home more frequently earlier than the other two groups.

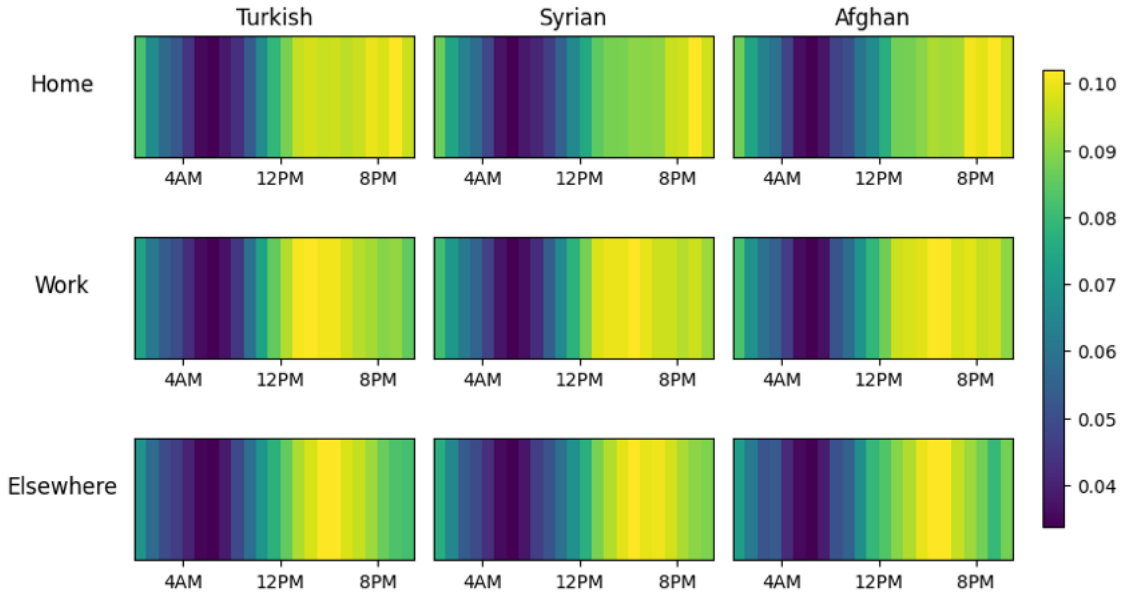


Figure 37. Average behaviour for Turkish, Syrian, and Afghan people.

Additionally, Syrians tend to be spending the least time at home. Other than home and work we see that each group is spending time in the late afternoon at other locations possibly running errands or for leisure.

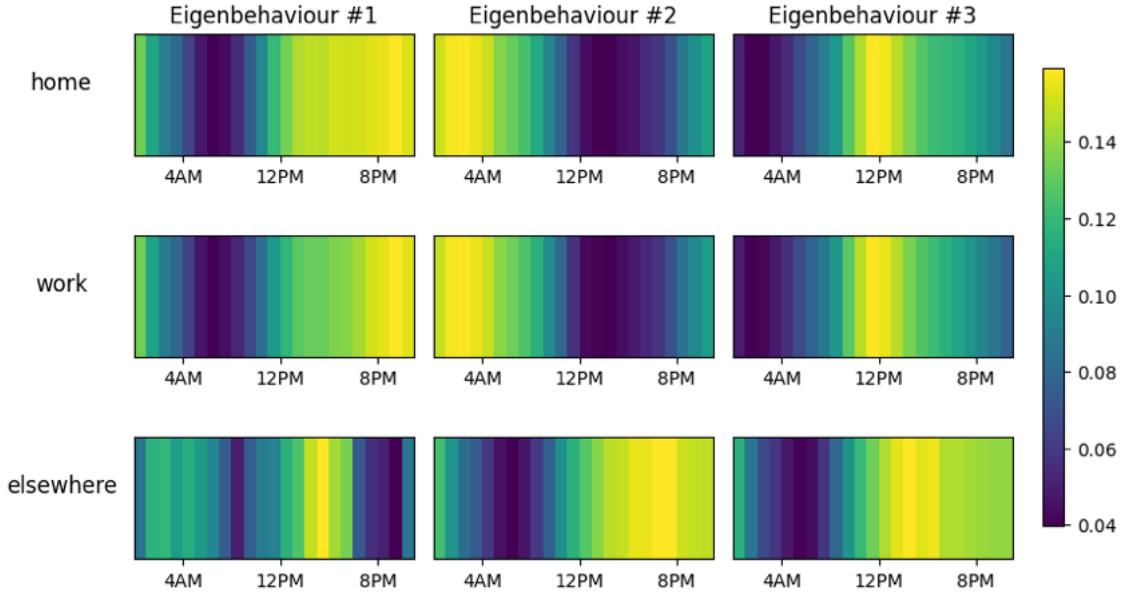


Figure 38. Top three eigenbehaviours of Turkish people.

Figure 38 shows the top three eigenbehaviours of Turks. The primary eigenbehaviour shows that Turkish people are at home and work during very much the same hours, typically the early nighttime, although surprisingly, work hours are more concentrated at 9 p.m. and home hours are more spread throughout the afternoon as well. Other locations that are visited happen most significantly around 3 p.m. For the second eigenbehaviour we see the opposite trend in some ways. Home and work locations

are still very similar but concentrated in the early morning between midnight and 5 a.m., while other locations are frequented in the late afternoon between 5 and 8 p.m. The third eigenbehaviour highlights midday for both home and work, with other locations being visited between 1 and 4 p.m.

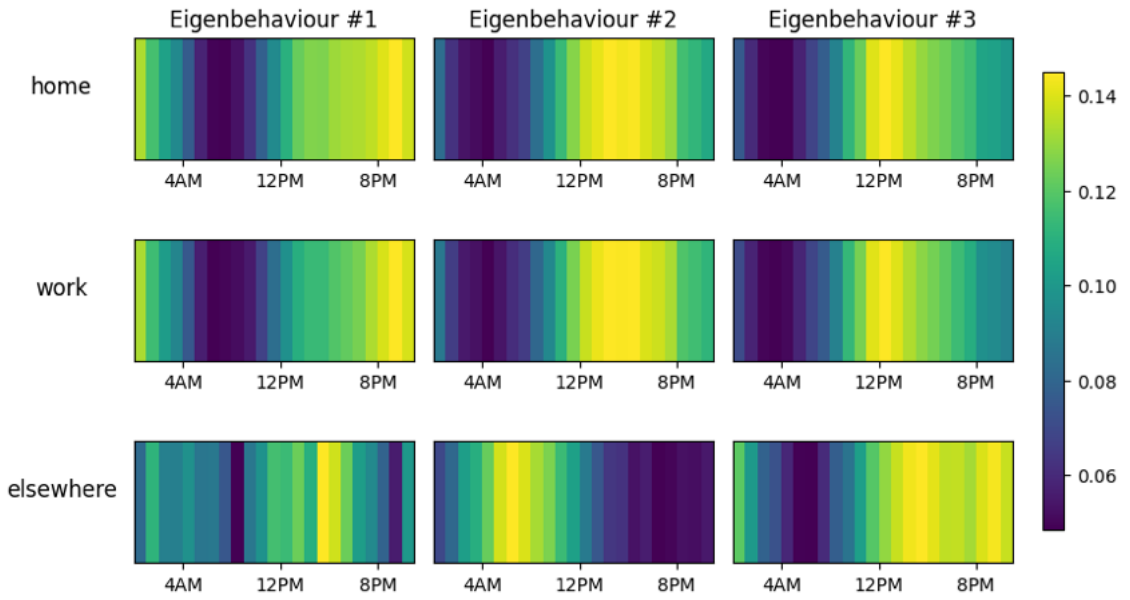


Figure 39. Top three eigenbehaviours of Syrian people.

Figure 39 shows the top three eigenbehaviours of Syrian people. The top eigenbehaviour is almost identical to Turkish people although home and work times are both very concentrated around 9 p.m. The second eigenbehaviour however shows that Syrians are both at home and work locations at more expected work hours, particularly between midday and 8 p.m. While they are at tertiary locations in the morning around 6 a.m. The third eigenbehaviour is again very similar to that of Turkish people with home and work highlighted around midday and other locations highlighted in the afternoon and again at night.

Figure 40 shows the top three eigenbehaviours of Afghans. Here we see a large divergence from the other groups. The primary eigenbehaviour suggest that Afghans are at home and work in the morning, most prominently at 7 a.m. and this slows down towards midday. They appear at other locations scattered throughout the day but more frequently at 8 a.m. The second eigenbehaviour shows that Afghans are at home and work during two separate periods, around 2 p.m. and again at 5 p.m. while they are elsewhere in the morning around 6 a.m. The third eigenbehavior shows afghans to be at home in the early morning hours between midnight and 6 a.m. followed by a brief period in other locations.

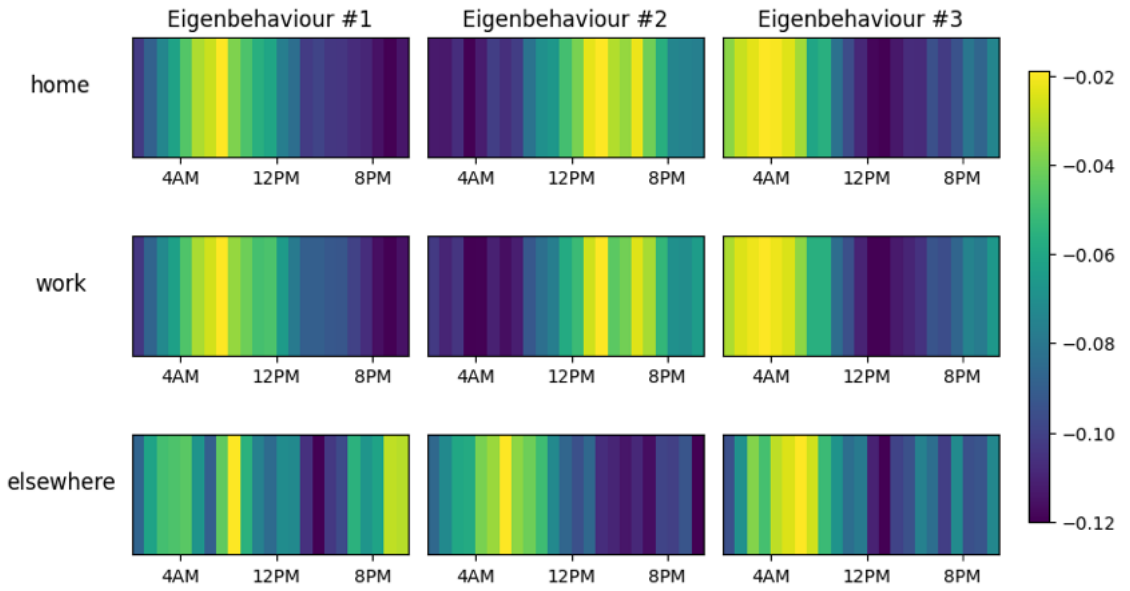


Figure 40. Top three eigenbehaviours of Afghan people.

## 3.5 Discussion

### 3.5.1 Residence

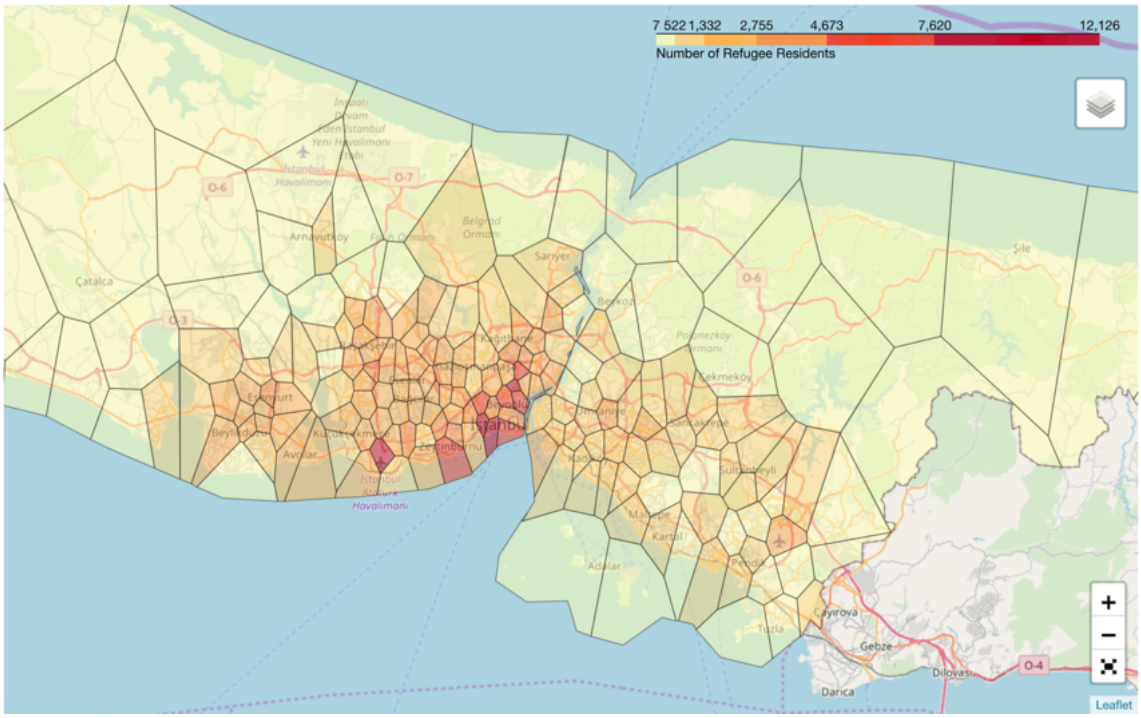


Figure 41. Map shows the number of refugee residents per residential region in Istanbul based on the number of refugee residents metric which we computed using the night time calling activity of refugees in CDR data. Figure from (Altuncu et al., 2019).

If we compare Figures 16, 8, and 41, we can see that previous attempts to map the residence locations of Syrians have localised them to be very present in the inner city locations, while our segregation and home detection analyses have found that

many Syrians in 2020 live also in districts further out of this vicinity, most notably in Esenler.

Furthermore, an interview based study by Balcioglu (2018) claims the following:

> Although each district in Istanbul hosts sizeable numbers, Syrian refugees are mostly clustered in the poorer and more religiously conservative districts of Kucukcekmece, Sultangazi, Bagcilar, and Sultanbeyli. As of March 2016, there were 485,227 Syrian refugees in Istanbul, of whom 20,192 resided in Sultanbeyli, constituting about six percent of the district's total population.

This is very much in line with our findings although it can be hard to notice the neighbourhoods of Sultanbeyli on the choropleth maps. It is expected that Syrian communities regularly reconfigure themselves within the city, adapting to changing conditions in the housing market and in response to affordable social housing initiatives.

According to Bozok and Bozok (2024), "Afghan migrants reside in squatter housing in the deprived neighbourhoods of Beykoz, Zeytinburnu and Fatih districts. Due to their stigmatization as the 'other' in these three districts, Afghan migrants are excluded from solidarity networks." Furthermore, Karadağ (2020) write that "Yenimahalle neighborhood of Küçüksu hosts a large population of Afghans whose number has constantly grown in the last decade. Besides the main districts hosting Afghan population like Zeytinburnu and Esenyurt, other neighborhoods such as Küçüksu and Yenimahalle have become newly emerging settlements predominantly for single Afghan male workers."

In our analysis we find that Afghans are primarily concentrated in Zeytinburnu, aligning somewhat with the statement of Bozok and Bozok (2024) and Karadağ (2020) but we do not see them concentrated in Beykoz, Fatih, and Esenyurt. This concentration in one district leaves the impression that there is an enclave of Afghans in Zeytinburnu. This could be the result of many Afghans having phone numbers that are associated to accounts opened with a Turkish passport, making them invisible in our dataset. On the other hand, it could just be that during 2020 Zeytinburnu hosted the vast majority of Afghans, of which there are comparatively few when looking at native Turks or even Syrians.

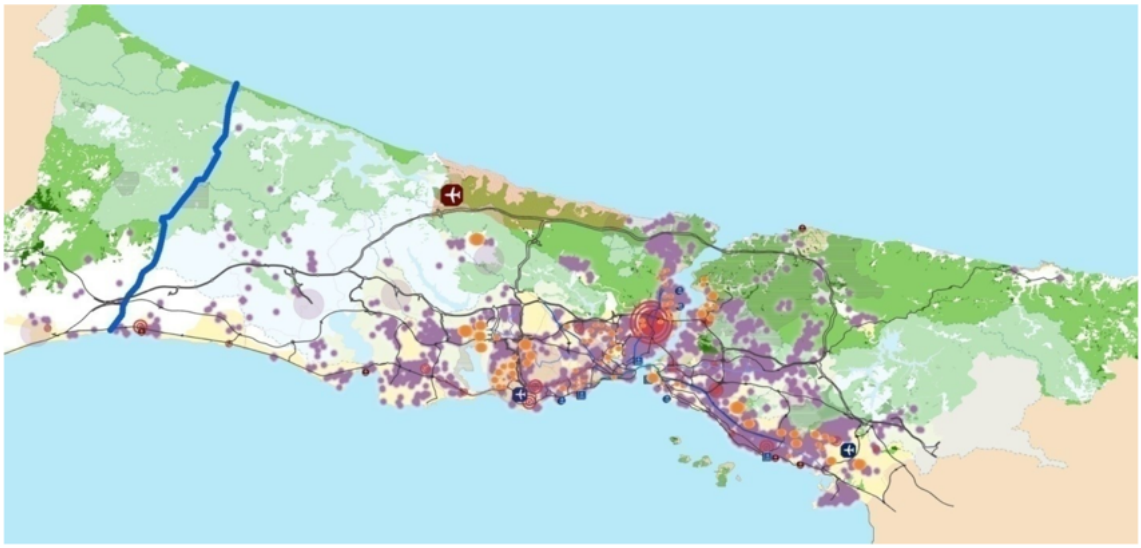

Figure 42. A map of "Gecekondu" in Istanbul ca. 2014. Source: Aksumer et al. (2014); Mansoorian (2018)

What can we learn about the social and economic status of these groups by the places in which they live? Unsurprisingly, the neighbourhoods where Syrians and Afghans live are well known to for their "Gecekondu". Roughly translating to "built overnight", these informal housing developments are characterised by several floors of apartments overhanging rows of shops and commercial spaces on the ground floor. These buildings often lack basic amenities like running water, electricity, and heating, functioning on the basis of communal exchange of services (Gonçalves and Gama, 2020). Figure 42, maps the coverage of Gecekondu across Istanbul, revealing also the strong overlap with refugee occupied neighbourhoods. In reviewing the figure, it becomes evident that informal settlements are widespread throughout the city and even share close quarters with wealthy neighbourhoods and commercial areas.

Thus, there is clear evidence that Syrians and Afghans are disproportionately underprivileged and face spatial segregation into what are effectively ghetto-like neighbourhoods. However, it is also the case that these "Gecekondu" are very diverse, housing native and migrant communities alike. This alludes to the poor condition of the housing market in Istanbul and the lack of social and affordable housing.

### 3.5.2 Residential vs. workplace segregation

Being the larger of the two minority groups, the Syrian refugees are more spread out across Istanbul though they are still concentrated in certain areas. In the inner city district of Fatih, as well as in Zeytinburnu, Bahçelievler, Başakşehir, Esenler, and Esenyurt there are pockets of high Syrian mobile phone activity. This lines up quite well with the home detection analysis done alongside this one despite the expectation that the fine-grained data was severely distorted by pandemic lockdown measures.

Looking at the evolution of the dissimilarity metric, we see that in 5 out of 6 of these districts Syrians experience increased levels of dissimilarity during the daytime. During the workplace detection analysis, it was made clear that many Syrians work in the same place where they live, thought this could be the effect of lockdown measures, meaning the increased segregation is not cause by commuting out of those districts. It is likely then that Syrians experienced higher workplace segregation than residential segregation during 2020. This contradicts the findings of Bertoli et al. (2021), who witness increased residential segregation compared to labour market segregation. Nevertheless, we find that these districts also have an overall low dissimilarity score, each around 0.2, suggesting that, even in the workplace, Syrians are relatively evenly distributed within the districts.

Afghans on the other hand, according to xDR volume, are mainly concentrated in the Zeytinburnu district where we suspect there is an ethnic enclave. Unlike Syrians, the dissimilarity score for Afghans actually decreases in Zeytinburnu during the day, though not by much. Also the overall dissimilarity is quite high in this district, peaking around 0.4, which was not the case for Syrians. This indicates that there is a very high concentration of Afghans in a particular neighbourhood or block of Zeytinburnu and that they live and work within a small radius of this space. This is a very clear example of spatial segregation and highlights the complex relationship this minority group has with the city they live in, as well as the precarious position they occupy within it.

In addition to Zeytinburnu, we know that Afghans occupy other districts in Istanbul. In relation to Bozok and Bozok (2024) and Karadağ (2020) it is worth looking at the following districts in particular: Beykoz, Fatih, Kücükcekmece, and Esenyurt. The European side districts of Fatih, Kücükcekmece, and Esenyurt all have relatively stable and low, around 0.2, dissimilarity over the entire day. Esenyurt does demonstrate a higher dissimilarity during daytime however, indicating a higher workplace segregation in this district. On the other hand, Beykoz, a much more rural district on the Asian side of the city, displays a dramatic shift in dissimilarity over the day. After peaking at 0.32 by 07:00, the dissimilarity falls dramatically to 0.2 over the day only to escalate again after 17:00. It is difficult to say why this rise and fall occurs. It could be due to commuting into the city for work leaving behind a very small population that the dissimilarity metric does not accurately account for at this resolution, or it could be that Afghans travel and disperse within this district for work.

We choose not to speculate on the meaning of the change in the isolation index as

overall it is negligible. Furthermore it is sensible to assume that in a city as populous as Istanbul, it would be difficult to be truly isolated as one will inevitably be living in close proximity to native Turkish residents. This brings us to the dissimilarity index and it's drawbacks. We must acknowledge that the index of dissimilarity is known to have an upward bias, meaning small populations are not well accounted for by the metric. Moreover, the measure is biased by small spatial resolutions like neighbourhood level calculations depending on how many and how uniformly sub-units are distributed within. Another issue with the index is that it does not account for the orientation and configuration of sub-units, meaning it does not reveal the spatial characteristics of segregation.

In countering these drawbacks, we believe that since we used aggregated counts from the entire year, the issue of small population is slightly mitigated since we represent even a small population with some thousands of xDR signals. Finally, we feel that our descision to intersect the Voronoi tesselations and neighbourhoods to produce areal sub-units produced a sufficient amount of sub-units to warrant using the index, even at the neighbourhood level and especially at the district level. Naturally, the greater amount of Syrians in the data set would have accounted for a more accurate picture of their presence in the city, but we find that much of our results corroborate information known about Afghans as well.

### 3.5.3 Eigenbehaviours

For Turks, Syrians, and Afghans we see a reoccurring pattern of shared time at home and work. This makes sense given that our home and workplace detection analysis found that a vast majority of people in our dataset were found to have the same home and work location. There is grounds to believe that this phenomenon occurs because the data is from a period of time in 2020 when there were Covid lockdown measures in place. It could also be an issue with the home and workplace detection strategy that we use.

Furthermore, the average behaviours can be partially accounted for by mobile phone traffic which may bias the results to times of the day when people use their mobile phones the most. This is different from the data of Eagle and Pentland (2009) who collect data actively instead of passively like xDRs. The regularity of their data may make for a less biased analysis of where time is spent.

Nevertheless, we can see that there are slight variations in the behaviours of the groups. For instance, that Syrian and Afghan peoples' second eigenbehaviour is actually during more typical working hours whereas Turkish people are at home

and work during the nighttime for their first two eigenbehaviours. We can also see that Syrians work marginally longer and later hours while Afghans appear to work frequently in the mornings. Turkish people seem to have the most variety in their "elsewhere" visitations, with Syrians coming second and Afghans have the least variety.

It would be worthwhile to trial this analysis with a different dataset featuring a longer duration than 16 days and one that is outside of the time of Covid intervention. Moreover, the analysis of group affiliation could be examined by witnessing how well the eigenbehaviour of each group is able to reconstruct the behaviours of the individuals within them. The analysis could be taken further by employing topic modelling techniques that allow for the tracking of behaviours over time instead of the stationary behaviours we use here (Farrahi and Gatica-Perez, 2010).

# 4. Conclusions

The ongoing civil war in Syria on Turkey's Southern border has resulted in the displacement of millions of people, most of whom have fled to neighbouring countries and which Turkey currently hosts the largest portion at 3.2 million Syrian refugees (Ministry of Interior, 2024). In addition, after the Taliban took control of Afghanistan in 2021, roughly 130,000 Afghans have also claimed asylum in Turkey, making it the country with the most refugees in the world (UNHCR, 2024).

As these migrants flock to cities, it is of utmost importance that the municipalities work in concert with government, humanitarian organisation, and the migrants themselves to alleviate these groups from poverty and restore their self-sufficiency. Orchestrating targeted interventions is a necessary part of this process, but policy makers face the challenge of designing policy based on outdated or redundant information. By using mobile phone traces, we demonstrate how new forms of digital data can supplement existing records from censuses and interviews while providing real-time mobility insights at an unprecedented resolution.

In Istanbul, a city that hosts more than half a million refugees and asylum seekers, the situation for Syrians and Afghans is precarious. Under Turkey's temporary and international protection schemes, many have found themselves unable to participate in the formal job market which requires possession of a valid work permit. The demand for cheep labour and the vulnerable position of refugees mean that informal labour agreements are common and often exploitative. In addition, xenophobic sentiments from the native population, as well as differences in language and culture, remain a challenging obstacle on the path to integration (Uluğ et al., 2023; Bozdağ, 2020). Furthermore, a large deficit in affordable social housing projects promotes a culture of squatting and the construction of informal settlements called "Gecekondu" which are often cramped, lacking in basic amenities or safety measures, and unregistered, making it difficult to access social security.

Segregation in one of the aspects that can be studied to help policy makers gain a clearer picture of the situation for these migrant groups. There is a thorough literature on quantifying segregation that has produced a series of indices that traditionally operate on the basis of census tracts. We demonstrate using mobile phone xDRs how the index of dissimilarity and of isolation can be calculated at very fine spatial and temporal resolutions in the absence of census tract data. Our methodology

results in a highly detailed picture of segregation patterns across Istanbul and the neighbouring province of Kocaeli for different times in the day.

We learn several key insights from the process of refining the mobile phone traces and from the results of the case-study. The mobile phone data from Istanbul in the year 2020 reveal how the Covid pandemic had an impact on mobility in the city, seeing a majority of mobile phone customers keep a close proximity to their place of residence. Similarly we find that most people live and work within the same district, however we also see that, proportional to their local population, more people commute to work when living in rural towns. When looking at the patterns of mobile phone traces through the city, we can see that Syrians and Afghans are concentrated in poorer neighbourhoods more commonly on the European side of the city. In addition we find that there is evidence of an enclave of Afghans in the Zeytinburnu district.

Considering segregation patterns, we see that this group of Afghans display an unusually high rate of segregation suggesting that they are clustered in close quarters and share deep interpersonal connections. For both Syrians and Afghans, segregation is higher in wealthier neighbourhoods and in rural towns. Finally, and surprisingly, it can be said of Syrians that they experience a greater workplace segregation than residential segregation, a discovery that contradicts previous analysis but that could allude to conditions under Covid pandemic measures. Conversely, Afghans for the most part display slightly higher residential segregation though the difference throughout the day is very slight, with one exception. The rural district of Beykoz on the Asian side of Istanbul sees the dissimilarity measure fall dramatically during the day with a high score at night, indicating that Afghans are more evenly distributed during the day. It would be interesting to verify if the Afghan population in this district changes during the day along with the index.

Looking at our eigenbehaviour analysis, some emergent properties appear with regards to working times. The analysis suggests while that Turks, Syrians, and Afghans have and overarching similarity in behaviour, there are nuances in the working times of the minorities that position them outside of the norm. Syrians seem to work longer hours at the night time and Afghans seem to work more during the middle morning. Moreover we see a repeated pattern of shared home and workspace suggesting either a work from home situation and restricted mobility or a situation where many people live very near their workplace.

While we think that this research shows promising results it only touches the surface of what is possible. With better access to data and better official sources with which

to validate results, there is the potential to take the analysis much further. For instance if population counts for temporary and international protection holders at the neighbourhood level were made available alongside existing population counts, it could be possible to accurately model the population composition of neighbourhoods over time. A next step may be to observe outcomes for refugees and native populations alike by comparing segregation insights to housing market and income data.

On final reflection there is a lot to be gained from digital methods in striving for social good, but without the human element and voices of the people whose lived realities we are trying to quantify, we are missing the full picture. It is folly to pursue action before getting all of the stakeholders at the table and taking heed of what is important to the people that will be affected by the decisions made at a high level. Face to face conversations in earnest can help us humanise one another and this is an important starting point if one is to think as a humanitarian.

# A. Appendix

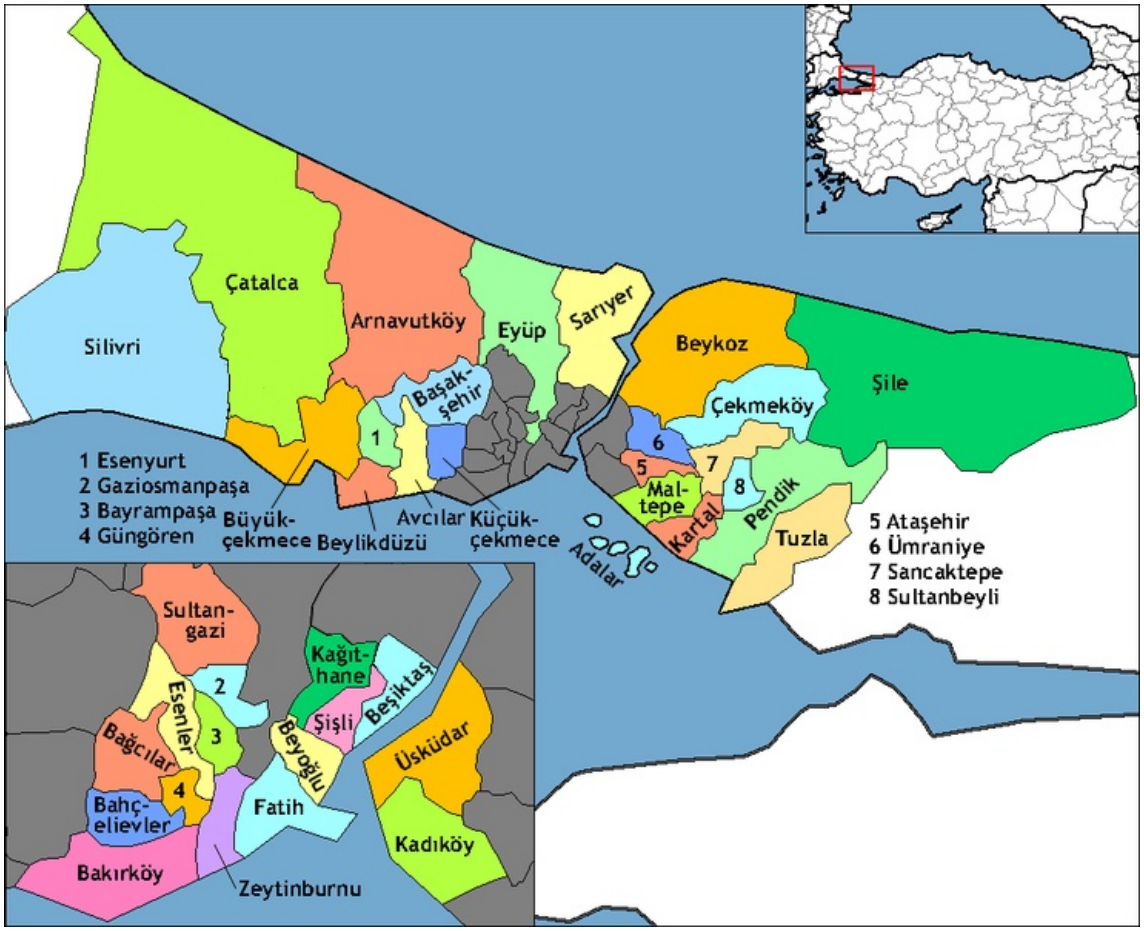


Figure 43. A map with the districts of Istanbul labelled (Rarelibra, 2006).

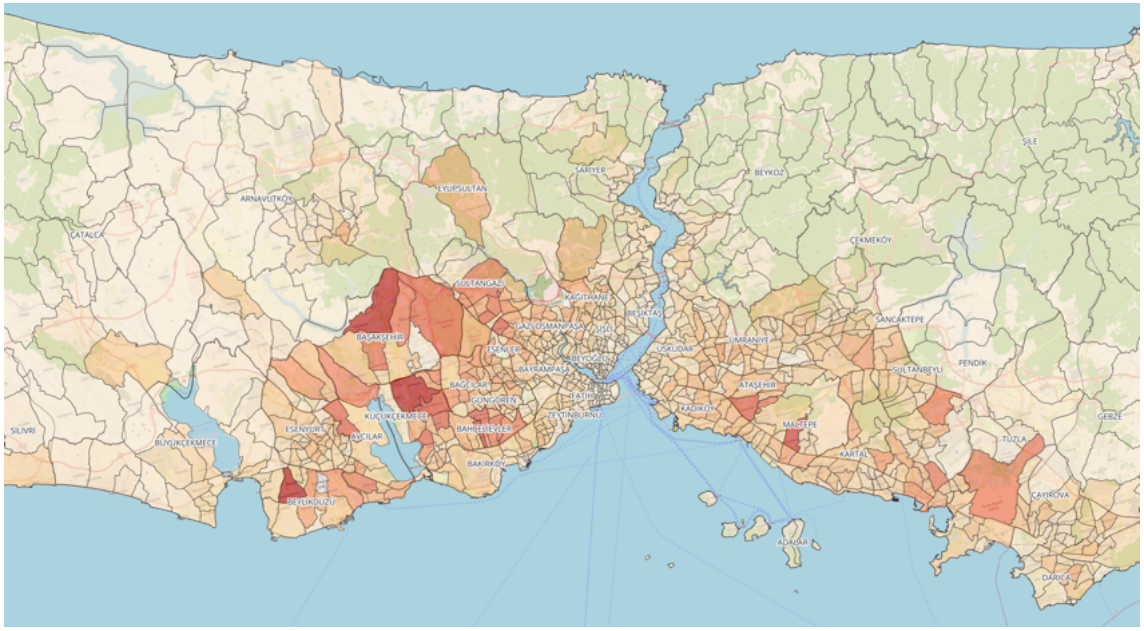


Figure 44. A close-up choropleth of Istanbul with official population data at the neighbourhood level.

# Bibliography


Acosta, R. J., Kishore, N., Irizarry, R. A., and Buckee, C. O. (2020). Quantifying the dynamics of migration after Hurricane Maria in Puerto Rico. *Proceedings of the National Academy of Sciences*, 117(51):32772–32778. Publisher: Proceedings of the National Academy of Sciences.

Agliari, E., Barra, A., Contucci, P., Pizzoferrato, A., and Vernia, C. (2018). Social interaction effects on immigrant integration. *Palgrave Communications*, 4(1):55.

Ahas, R., Silm, S., Järv, O., Saluveer, E., and Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1):3–27.

Ahas, R., Silm, S., and Tiru, M. (2018). Measuring Transnational Migration with Roaming Datasets. In *Adjunct Proceedings of the 14th International Conference on Location Based Services*, pages 105 – 108, Zurich, Switzerland. ETH Zurich. Medium: application/pdf Publisher: ETH Zurich.

Ahmad-Yar, A. W. and Bircan, T. (2021). Anatomy of a Misfit: International Migration Statistics. *Sustainability*, 13(7).

Aksumer, G., Çalişkan, c. O., Yalcintan, M. C., Kap, S. D., Yücel, H., and Çılgın, K. (2014). Sarıyer Gecekondu Mahalleleri Örneğinde Kentsel Dönüşüm Süreçleri ve Bu Süreçlerin Sosyo-Ekonomik ve Fiziki Etkileri. Technical Report 110K404, MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ, Turkey.

Al-Azizy, D., Millard, D., Symeonidis, I., O'Hara, K., and Shadbolt, N. (2016). A Literature Survey and Classifications on Data Deanonymisation. In Lambrinoudakis, C. and Gabillon, A., editors, *Risks and Security of Internet and Systems*, volume 9572, pages 36–51. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Alesina, A., Harnoss, J., and Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, 21(2):101–138.

Alfeo, A. L., Cimino, M. G. C. A., Lepri, B., and Vaglini, G. (2019). Using Call Data and Stigmergic Similarity to Assess the Integration of Syrian Refugees in Turkey. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 165–178. Springer International Publishing, Cham.

Altuncu, M. T., Kaptaner, A. S., and Sevencan, N. (2019). Optimizing the Access to Healthcare Services in Dense Refugee Hosting Urban Areas: A Case for Istanbul. In *Guide to Mobile Data Analytics in Refugee Scenarios: The 'Data for Refugees Challenge' Study*, pages 403–416. Springer International Publishing. arXiv:1903.09614 [cs].

Atahan, A. and Alhelo, L. (2022). The Impact of the COVID-19 Pandemic on Mobility Behavior in Istanbul After One Year of Pandemic. In Akhnoukh, A., Kaloush, K., Elabyad, M., Halleman, B., Erian, N., Enmon Ii, S., and Henry, C., editors, *Advances in Road Infrastructure and Mobility*, pages 933–949. Springer International Publishing, Cham. Series Title: Sustainable Civil Infrastructures.

Aydoğdu, B., Salah, A. A., Ones, O., and Gurbuz, B. (2021). Description of the mobile CDR database. Technical Report Open research data pilot (Deliverable 6.1), Leuven: HumMingBird project 870661 – H2020.

Bakker, M. A., Piracha, D. A., Lu, P. J., Bejgo, K., Bahrami, M., Leng, Y., Balsa-Barreiro, J., Ricard, J., Morales, A. J., Singh, V. K., Bozkaya, B., Balcisoy, S., and Pentland, A. (2019). Measuring Fine-Grained Multidimensional Integration Using Mobile Phone Metadata: The Case of Syrian Refugees in Turkey. In *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 123–140. Springer International Publishing, Cham.

Balcioglu, Z. (2018). Sultanbeyli, Istanbul: A Case Report of Refugees in Towns. Technical report, Refugees in Towns, Turkey.

Bertoli, S., Cintia, P., Giannotti, F., Madinier, E., Ozden, C., Packard, M., Pedreschi, D., Rapoport, H., Sîrbu, A., and Speciale, B. (2019). Integration of Syrian Refugees: Insights from D4R, Media Events and Housing Market Data. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 179–199. Springer International Publishing, Cham.

Bertoli, S., Ozden, C., and Packard, M. (2021). Segregation and internal mobility of Syrian refugees in Turkey: Evidence from mobile phone data. *Journal of Development Economics*, 152:102704.

Bircan, T. (2022). Remote Sensing Data for Migration Research. In *Data Science for Migration and Mobility Studies.*, pages 121–148. Oxford University Press, United Kingdom.

Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot,

F., Smoreda, Z., and Ziemlicki, C. (2013). Data for Development: the D4D Challenge on Mobile Phone Data. arXiv:1210.0137 [physics, stat].

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Böhme, M. H., Gröger, A., and Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142:102347.

Bouman, P., van der Hurk, E., Kroon, L., Li, T., and Vervest, P. (2013). Detecting Activity Patterns from Smart Card Data. In *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence*, Delft. Delft University of Technology.

Boy, J., Pastor-Escuredo, D., Macguire, D., Moreno Jimenez, R., and Luengo-Oroz, M. (2019). Towards an Understanding of Refugee Segregation, Isolation, Homophily and Ultimately Integration in Turkey Using Call Detail Records. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 141–164. Springer International Publishing, Cham.

Boyandin, I., Bertini, E., and Lalanne, D. (2011). Visualizing Migration Flows and their Development in Time: Flow Maps and Beyond.

Bozcaga, T., Christia, F., Harwood, E., Daskalakis, C., and Papademetriou, C. (2019). Syrian Refugee Integration in Turkey: Evidence from Call Detail Records. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 223–249. Springer International Publishing, Cham.

Bozdağ, c. (2020). Bottom-up nationalism and discrimination on social media: An analysis of the citizenship debate about refugees in Turkey. *European Journal of Cultural Studies*, 23(5):712–730.

Bozok, M. and Bozok, N. (2024). Away from home and excluded from local solidarity networks: Undocumented Afghan migrant men in Istanbul. *Population, Space and Place*, page e2775.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., and Ratti, C. (2011). Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151.

Cecaj, A. and Mamei, M. (2017). Data fusion for city life event detection. *Journal of Ambient Intelligence and Humanized Computing*, 8(1):117–131.

Chi, G., Lin, F., Chi, G., and Blumenstock, J. (2020). A general approach to detecting migration events in digital trace data. *PLOS ONE*, 15(10):e0239408.

Coletto, M., Esuli, A., Lucchese, C., Muntean, C. I., Nardini, F. M., Perego, R., and Renso, C. (2017). Perception of social phenomena through the multidimensional analysis of online social networks. *Online Social Networks and Media*, 1:14–32.

Coston, A., Guha, N., Ouyang, D., Lu, L., Chouldechova, A., and Ho, D. E. (2021). Leveraging Administrative Data for Bias Audits: Assessing Disparate Coverage with Mobility Data for COVID-19 Policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, Virtual Event Canada. ACM.

De Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., De Cordes, N., Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., Krings, G., Letouzé, E., Luengo-Oroz, M., Oliver, N., Rocher, L., Rutherford, A., Smoreda, Z., Steele, J., Wetter, E., Pentland, A. S., and Bengtsson, L. (2018). On the privacy-conscientious use of mobile phone data. *Scientific Data*, 5(1):180286.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1):1376.

Eagle, N. and Pentland, A. S. (2009). Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066.

Erdoğan, M. M. (2022). Urban Refugees of Marmara: Process Management of Municipalities. *Marmara Municipalities Union*.

Erfani, A. and Frias-Martinez, V. (2023). A fairness assessment of mobility-based COVID-19 case prediction models. *PLOS ONE*, 18(10):e0292090.

European Environment Agency and European Environment Agency (2019). CORINE Land Cover 2018 (vector), Europe, 6-yearly - version 2020_20u1, May 2020.

Farrahi, K. and Gatica-Perez, D. (2010). Probabilistic Mining of Socio-Geographic Routines From Mobile Phone Data. *IEEE Journal of Selected Topics in Signal Processing*, 4(4):746–755.

Fischer, M. J. and Tienda, M. (2006). Redrawing Spatial Color Lines: Hispanic Metropolitan Dispersal, Segregation, and Economic Opportunity. In *Hispanics and the Future of America*. National Academies Press (US).

Gambs, S., Killijian, M.-O., and Núñez Del Prado Cortez, M. (2014). De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614.

Gapminder Org. (2020). Gapminder Tools.

Ge, Q. and Fukuda, D. (2016). Updating origin–destination matrices with aggregated data of GPS traces. *Transportation Research Part C: Emerging Technologies*, 69:291–312.

Geofabrik GmbH (2024). Geofabrik Download Server.

Gonçalves, J. M. and Gama, J. M. R. F. (2020). A systematisation of policies and programs focused on informal urban settlements: reviewing the cases of São Paulo, Luanda, and Istanbul. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 13(4):466–488.

Grantz, K. H., Meredith, H. R., Cummings, D. A. T., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., Mehta, S., Solomon, S., Labrique, A., Kishore, N., Buckee, C. O., and Wesolowski, A. (2020). The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications*, 11(1):4961.

Guidotti, R. and Gabrielli, L. (2018). Recognizing Residents and Tourists with Retail Data Using Shopping Profiles. In Guidi, B., Ricci, L., Calafate, C., Gaggi, O., and Marquez-Barja, J., editors, *Smart Objects and Technologies for Social Good*, volume 233, pages 353–363. Springer International Publishing, Cham. Series Title: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering.

Hu, W., He, R., Cao, J., Zhang, L., Uzunalioglu, H., Akyamac, A., and Phadke, C. (2019). Quantified Understanding of Syrian Refugee Integration in Turkey. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 201–221. Springer International Publishing, Cham.

Iceland, J., Weinberg, D. H., and Steinmetz, E. (2002). *Racial and ethnic residential segregation in the United States 1980-2000*, volume 8-3. Bureau of Census.

International Organisation for Migration (IOM) (2023). *Harnessing Data Innovation for Migration Policy: A Handbook for Practitioners*. IOM Geneva.

İstanbul Kalkınma Ajansı (2017). Mahallem İstanbul - Kalkınma Kütüphanesi.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Järv, O., Müürisepp, K., Ahas, R., Derudder, B., and Witlox, F. (2015). Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Studies*, 52(14):2680–2698.

Karadağ, S. (2020). Ghosts of Istanbul: Afghans at the Margins of Precarity. Technical report, GAR: the Association for Migration Research, Turkey.

Lavelle-Hill, R., Harvey, J., Smith, G., Mazumder, A., Ellis, M., Mwantimwa, K., and Goulding, J. (2022). Using mobile money data and call detail records to explore the risks of urban migration in Tanzania. *EPJ Data Science*, 11(1):28.

Logan, J. R. and Stults, B. J. (2011). The persistence of segregation in the metropolis: New findings from the 2010 census. *Census brief prepared for Project US2010*, 24. Publisher: Russell Sage Foundation New York, NY.

Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Tatem, A., Canright, G., Engø-Monsen, K., and Bengtsson, L. (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environmental Change*, 38:1–7.

Mansoorian, M. (2018). *Linking conflict and collaboration; bottom-up urban regeneration within top-down structure of urban policy in Istanbul and Tehran*. Doctoral Thesis, Technische Universität Berlin, Berlin. Publisher: [object Object].

Marquez, N., Garimella, K., Toomet, O., Weber, I. G., and Zagheni, E. (2019). Segregation and Sentiment: Estimating Refugee Segregation and Its Effects Using Digital Trace Data. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 265–282. Springer International Publishing, Cham.

Massey, D. S. and Denton, N. A. (1988). The Dimensions of Residential Segregation. *Social Forces*, 67(2):281.

Massey, D. S. and Denton, N. A. (2003). *American apartheid: segregation and the making of the underclass*. Harvard Univ. Press, Cambridge, Mass., 10. print edition.

Milusheva, S., Erbach-Schoenberg, E. z., Bengtsson, L., Wetter, E., and Tatem, A. (2018). Understanding the Relationship between Short and Long Term Mobility. Working Paper 3377c250-d046-4340-947c-129af6d6dc1d, Agence française de développement.

Minard, C.-J. (1844). *Tableaux graphiques et cartes figuratives*. Regnier et Dourdet, Paris.

Minello, A. (2014). The educational expectations of Italian children: the role of social interactions with the children of immigrants. *International Studies in Sociology of Education*, 24(2):127–147.

Ministry of Interior (2024). TEMPORARY PROTECTION.

Moise, I., Gaere, E., Merz, R., Koch, S., and Pournaras, E. (2016). Tracking Language Mobility in the Twitter Landscape. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 663–670, Barcelona, Spain. IEEE.

Mooses, V., Silm, S., and Ahas, R. (2016). ETHNIC SEGREGATION DURING PUBLIC AND NATIONAL HOLIDAYS: A STUDY USING MOBILE PHONE DATA. *Geografiska Annaler: Series B, Human Geography*, 98(3):205–219.

Mooses, V., Silm, S., Tammaru, T., and Saluveer, E. (2020). An ethno-linguistic dimension in transnational activity space measured with mobile phone data. *Humanities and Social Sciences Communications*, 7(1):140.

Niedomysl, T., Hall, O., Archila Bustos, M. F., and Ernstson, U. (2017). Using Satellite Data on Nighttime Lights Intensity to Estimate Contemporary Human Migration Distances. *Annals of the American Association of Geographers*, 107(3):591–605.

Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C., and Bravo, L. (2021). Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ Data Science*, 10(1):29.

Pötzschke, S. and Weiß, B. (2021). Realizing a Global Survey of Emigrants through Facebook and Instagram. preprint, Open Science Framework.

Quinn, J. A., Nyhan, M. M., Navarro, C., Coluccia, D., Bromley, L., and Luengo-Oroz, M. (2018). Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170363.

Rarelibra (2006). File:Istanbul districts.png - Wikipedia.

Reardon, S. F. and Firebaugh, G. (2002). Measures of Multigroup Segregation. *Sociological Methodology*, 32(1):33–67.

Regional IM Working Group - Europe (2021). Türkiye - Subnational Administrative Boundaries - Humanitarian Data Exchange.

Rhoads, D., Borge-Holthoefer, J., and Solé-Ribalta, A. (2019). Measuring and Mitigating Behavioural Segregation as an Optimisation Problem: The Case of Syrian Refugees in Turkey. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 283–301. Springer International Publishing, Cham.

Salah, A. A., Bircan, T., and Korkmaz, E. E. (2022). New data sources and computational approaches on migration and human mobility. In *Data Science for Migration and Mobility*. Oxford University Press.

Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors (2019a). *Guide to Mobile Data Analytics in Refugee Scenarios: The 'Data for Refugees Challenge' Study*. Springer International Publishing, Cham.

Salah, A. A., Pentland, A., Lepri, B., Letouzé, E., De Montjoye, Y.-A., Dong, X., Dağdelen, O., and Vinck, P. (2019b). Introduction to the Data for Refugees Challenge on Mobility of Syrian Refugees in Turkey. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 3–27. Springer International Publishing, Cham.

Salman, F. S., Yücel, E., Kayı, I., Turper-Alışık, S., and Coşkun, A. (2021). Modeling mobile health service delivery to Syrian migrant farm workers using call record data. *Socio-Economic Planning Sciences*, 77:101005.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, Atlanta GA USA. ACM.

Shakibaei, S., De Jong, G. C., Alpkökin, P., and Rashidi, T. H. (2021). Impact of the COVID-19 pandemic on travel behavior in Istanbul: A panel data analysis. *Sustainable Cities and Society*, 65:102619.

Silm, S., Ahas, R., and Mooses, V. (2018). Are younger age groups less segregated? Measuring ethnic segregation in activity spaces using mobile phone data. *Journal of Ethnic and Migration Studies*, 44(11):1797–1817.

Smith, S., Maas, I., and Van Tubergen, F. (2012). Irreconcilable differences? Ethnic intermarriage and divorce in the Netherlands, 1995–2008. *Social Science Research*, 41(5):1126–1137.

Spörlein, C. and Van Tubergen, F. (2014). The occupational status of immigrants in Western and non-Western societies. *International Journal of Comparative Sociology*, 55(2):119–143.

Sterly, H., Etzold, B., Wirkus, L., Sakdapolrak, P., Schewe, J., Schleussner, C.-F., and Hennig, B. (2019). Assessing Refugees' Onward Mobility with Mobile Phone Data—A Case Study of (Syrian) Refugees in Turkey. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 251–263. Springer International Publishing, Cham.

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., Guidotti, R., Bertoli, S., Kim, J., Muntean, C. I., Pappalardo, L., Passarella, A., Pedreschi, D., Pollacci, L., Pratesi, F., and Sharma, R. (2021). Human migration: the big data perspective. *International Journal of Data Science and Analytics*, 11(4):341–360.

Tatem, A. J., Dooley, C. A., Lai, S., Woods, D., Cunningham, A., and Sorichetta, A. (2023). Geospatial Data Integration to Capture Small-area Population Dynamics. In *Harnessing Data Innovation for Migration Policy: A Handbook for Practitioners*, pages 10–24. IOM Geneva.

Telea, A. C. and Behrisch, M. (2022). Visual Exploration of Large Multidimensional Trajectory Data. In *Data Science for Migration and Mobility Studies*. Oxford University Press.

Turper Alışık, S., Bayraktar Aksel, D., Yantaç, A. E., Kayi, \., Salman, S., \.Içduygu, A., Çay, D., Baruh, L., and Bensason, I. (2019). Seasonal Labor Migration Among Syrian Refugees and Urban Deep Map for Integration in Turkey. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios: The 'Data for Refugees Challenge' Study*, pages 305–328. Springer International Publishing, Cham.

Uluğ, O. M., Kanık, B., Tekin, S., Uyanık, G. D., and Solak, N. (2023). Attitudes towards Afghan refugees and immigrants in Turkey: A Twitter analysis. *Current Research in Ecological and Social Psychology*, 5:100145.

UNHCR (2024). TÜRKİYE FACT SHEET. Technical report, UNHCR, Turkey.

Van Tubergen, F. and Wierenga, M. (2011). The Language Acquisition of Male Immigrants in a Multilingual Destination: Turks and Moroccans in Belgium. *Journal of Ethnic and Migration Studies*, 37(7):1039–1057.

Vanhoof, M., Lee, C., and Smoreda, Z. (2020). Performance and Sensitivities of Home Detection on Mobile Phone Data. In *Big Data Meets Survey Science*, pages 245–271. Wiley, 1 edition. arXiv:1809.09911 [cs].

Vanhoof, M., Reis, F., Ploetz, T., and Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, 34(4):935–960. arXiv:1809.07567 [cs].

Vieira, C. C., Fatehkia, M., Garimella, K., Weber, I., and Zagheni, E. (2022). Using Facebook and LinkedIn Data to Study International Mobility.

Vinck, P., Pham, P. N., and Salah, A. A. (2019). "Do No Harm" in the Age of Big Data: Data, Ethics, and the Refugees. In Salah, A. A., Pentland, A., Lepri, B., and Letouzé, E., editors, *Guide to Mobile Data Analytics in Refugee Scenarios*, pages 87–99. Springer International Publishing, Cham.

Yang, Y., Pentland, A., and Moro, E. (2023). Identifying latent activity behaviors and lifestyles using mobility data to describe urban dynamics. *EPJ Data Science*, 12(1):15.

Yang, Y., Xiong, C., Zhuo, J., and Cai, M. (2021). Detecting Home and Work Locations from Mobile Phone Cellular Signaling Data. *Mobile Information Systems*, 2021:1–13.

Yuan, G., Chen, Y., Sun, L., Lai, J., Li, T., and Liu, Z. (2020). Recognition of Functional Areas Based on Call Detail Records and Point of Interest Data. *Journal of Advanced Transportation*, 2020:1–16.

Zagheni, E., Garimella, V. R. K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from Twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444, Seoul Korea. ACM.

Zhang, B., Zhong, C., Gao, Q., Shabrina, Z., and Tu, W. (2022). Delineating urban functional zones using mobile phone data: A case study of cross-boundary integration in Shenzhen-Dongguan-Huizhou area. *Computers, Environment and Urban Systems*, 98:101872.

Zhang, X. and Du, S. (2016). Learning selfhood scales for urban land cover mapping with very-high-resolution satellite images. *Remote Sensing of Environment*, 178:172–190.

Zheng, S., Xie, S., and Chen, X. (2019). Discovering Urban Functional Regions with Call Detail Records and Points of Interest: A Case Study of Guangzhou City.

In *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6, Xi'an, China. IEEE.