LEVERAGING INNOVATIVE DATA SOURCES FOR ANALYSIS OF MIGRATION PATTERNS

by
MERT GÜRKAN

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of the requirements
for the degree of Master of Science

Sabancı University
June 2021

LEVERAGING INNOVATIVE DATA SOURCES FOR ANALYSIS OF MIGRATION PATTERNS

Approved by:

Date of Approval: July 2, 2021

ABSTRACT

LEVERAGING INNOVATIVE DATA SOURCES FOR ANALYSIS OF MIGRATION PATTERNS

MERT GÜRKAN

COMPUTER SCIENCE AND ENGINEERING M.Sc. THESIS, JULY 2021

Thesis Supervisor: Prof. Selim Balcısoy

Keywords: Migration Patterns, Innovative Data Sources, Exploratory Visual Analysis, Behavioral Analysis

The globalized world of the 21st century hosts various types of migration movements as a common phenomenon. Understanding these movements and the conditions that cause these movements is crucial as the effects of migration range from economic outcomes to the social integration of different communities. Because of this, it is common to approach this phenomenon with data-driven studies. Many studies utilize statistical and administrative data sources to study migration patterns and their demographic and socio-economic drivers. However, factors such as varying definitions of migration in available data sources and gaps between data collection periods limit the success of data-driven studies. To address these, recent studies utilize innovative big data sources. In this thesis, we propose two different studies with innovative data sources for various definitions of migration. The first study adopts a transactional data source of credit card expenditures of customers of a private bank. These geo-located transactions are employed to infer possible internal migration movements in Turkey. The latter study utilizes data obtained from Facebook to contribute to a better understanding of global migration patterns. Obtained dataset from Facebook is combined with migration stock datasets from international organizations in a visual exploratory tool. This way, the visual tool creates a medium where the innovative data source utilized can be validated. The tool is also used for visualizing the results of the case study with the transactional data source. The advantages and shortcomings of utilized innovative data sources are thoroughly discussed.

ÖZET

GÖÇ HAREKETLERI İÇIN YENILIKÇI VERI KAYNAKLARINDAN YARARLANMA

MERT GÜRKAN

BİLGİSAYAR BİLİMİ VE MÜHENDİSLİĞİ YÜKSEK LİSANS TEZİ, TEMMUZ 2021

Tez Danışmanı: Prof. Dr. Selim Balcısoy

Anahtar Kelimeler: Göç Hareketleri, Yenilikçi Veri Kaynakları, Görsel Keşif Analizi, Davranışsal Analiz

21. yüzyılın küreselleşen dünyası, birçok çeşitli göç hareketlerine ev sahipliği yapmaktadır. Göçün etkileri ekonomik sonuçlardan farklı toplulukların sosyal entegrasyonuna kadar çeşitlilik gösterdiğinden, bu hareketleri ve bu hareketlere neden olan koşulları anlamak çok önemlidir. Bu nedenle, bu olguya veriye dayalı çalışmalarla yaklaşmak yaygındır. Birçok çalışma, göç hareketlerini ve bunların demografik ve sosyo-ekonomik etkenlerini incelemek için istatistiksel ve idari veri kaynaklarını kullanır. Ancak, mevcut veri kaynaklarında farklı göç tanımları ve veri toplama dönemleri arasındaki boşluklar gibi faktörler, veriye dayalı çalışmaların başarısını sınırlandırmaktadır. Bu sorunları ele almak için, son çalışmalar yenilikçi büyük veri kaynaklarını kullanmaktadır. Bu tezde, çeşitli göç tanımları için yenilikçi veri kaynaklarıyla iki farklı çalışma öneriyoruz. İlk çalışma, özel bir bankanın müşterilerinin kredi kartı harcama veri seti üzerinde oluşturulmuştur. Bu coğrafi konumlu işlemler, Türkiye'deki olası iç göç hareketlerini anlamak için kullanılmaktadır. İkinci çalışma, küresel göç modellerinin daha iyi anlaşılmasına katkıda bulunmak için Facebook'tan elde edilen verileri kullanır. Facebook'tan elde edilen veri seti, uluslararası kuruluşlardan gelen göç stok veri setleri ile görsel bir keşif aracında birleştirilmiştir. Bu şekilde görsel araç, kullanılan yenilikçi veri kaynağının doğrulanabileceği bir ortam yaratmaktadır. Araç ayrıca, işlemsel veri kaynağıyla vaka çalışmasının sonuçlarını görselleştirmek için de kullanılmıştır. Kullanılan yenilikçi veri kaynaklarının avantajları ve eksiklikleri kapsamlı bir şekilde tartışılmıştır.

ACKNOWLEDGEMENTS

Firstly, I would like to express my gratitude to my thesis supervisor Prof. Selim Balcısoy. Thanks to his guidance, understanding, and patience, this thesis is completed.

Additionally, I thank my thesis jury members Prof. Burçin Bozkaya and Dr. Günet Eroğlu. Prof. Burçin Bozkaya assisted me many times during my masters' study, and Dr. Günet Eroğlu motivated me for this journey. I also thank Alfredo Morales-Guzman for his valuable feedback for this study and the inspirational study they produced with his colleagues.

I thank all members of BAVLAB for their companionship. Many thanks to Yasin Fındık and Eray Alpan for the good times together. And I sincerely thank Hasan Alp Boz for his support and friendship.

I would like to thank my family for their support and understanding. Apart from the past months where the conditions imposed by the pandemic, we have usually been distant. But, I have never felt the lack of their support and understanding.

Finally, I would love to thank the feline fellows Sarman and Sultan with whom I share my living space. Their presence introduced a lot of joy to my life during these past months.

Sometimes life is like this dark tunnel.
You can't always see the light at the end of the tunnel,
but if you just keep moving you will come to a better place.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATONS

Important note regarding references in this thesis:

Many forms of media ranging from online available data sources to news and reports published by international organizations are utilized in this thesis. To better distinguish academic publications from online sources, mentioned online sources are referenced through footnotes rather than referencing in a bibliography. This way possible confusion for the reader is also averted as the referencing style of the thesis is not fit to reflect some of the online sources utilized.

1. INTRODUCTION

Migration is a complex social phenomenon as it is connected to many demographic, economic, social, and political factors leading to it and it produces many social dynamics. Due to this complexity, many of the studies aiming to generate insights on migration dynamics rely on data and data-driven approaches. This complexity is also reflected in the data-driven approaches as the increase in the type and number of datasets utilized.

Global migration dynamics is one of these problems with many different data sources with various characteristics. With the recent increase in the number available from various domains, analysis of migration and related phenomena can be possible by addressing challenges with efficient origin-destination data visualizations. Studies aiming to analyze migration dynamics should reflect these demographic and socio-economic conditions of the source and destination regions of these flows.

The more connected and globalized world of the 21st century created dynamics and metrics that traditional data sources have difficulty keeping up and conveying. Because of this, the utilization of innovative data sources is becoming a common practice in migration studies. Mostly, innovative data sources include datasets obtained from social media and their services. High spatial resolution and public availability can be considered as the primary advantages of innovative data sources. These datasets can support traditional data sources for migration to address their shortcomings.

To facilitate the analysis of global migration flows, this study provides an exploratory visual tool. The tool creates an environment where traditional data sources for migration can be compared with recent innovative data sources to address the challenges stated above. Additionally, the exploratory tool allows users to transform the aggregated migration flow estimates into estimates with higher resolutions. To this end, the visual tool is developed with demographic and socio-economic metrics that contribute to a better understanding of migration patterns.

Internal migration is one of the common ways of how people choose to migrate. This

type of migration takes place within the borders of the original states of migrant individuals. As this type of movement is easier to achieve for most individuals, much larger estimates for internal migration are pointed out. The study of United Nations Development Programme published their estimates[1] as 740 million internal migrants worldwide, which is more than two times larger than the estimated international migrants in 2017. Understanding socio-economic and demographic roots and the results of these movements are also crucial for addressing these movements with effective policy-making. Because of this, it is also common for internal migration problems to be addressed by data-driven studies. To this end, in addition to the efforts for international migration dynamics, a case study is also designed to discover possible internal migration movements with transactional data.

Along with these, the primary contributions of this thesis can be summarized as;

- Presenting an exploratory visual tool designed for displaying migration patterns of countries. The developed tool is also a medium for comparing different datasets for migration data with its various components. Additionally, as will be discussed further in Chapter 6, the visual tool also facilitates migration data transformations into different spatial resolutions. This way, higher resolution data not available for more specific regions can be estimated and more insights can be gathered about migration in different scales.

- Employing a transactional dataset of credit card expenditures for analysis of internal migration of a country. As presented thoroughly in Chapter 4, a sample dataset that contains credit card transactions of customers in one year is analyzed to infer possible internal migration movements of individuals in the sample. Following the methodology defined in the literature, the case study presented in Chapter 4 investigates various scenarios to study the movement of individuals. The chapter also discusses metrics derived from transactional data for a better understanding of the behavior of customers.

Additionally, some of the secondary contribution can be listed as;

- Providing country-level emigration estimations of 90 countries available in the data obtained from Facebook.

- Presenting a detailed literature review and discussion of data-driven migration studies. The discussion also covers recent studies with innovative data sources. As utilized in this study, datasets obtained through Facebook are discussed in all its aspects.

[1] Human Development Report of United Nations Development Programme, (2009) Accessed May 2021 from the following URL: www.hdr.undp.org/sites/ default/files/reports/269/hdr_2009_en_complete.pdf.

- Presenting the visual capability of the developed exploratory tool with the findings from Chapter 4. The results of the chapter are displayed with the visual tool as the second use case.

For the remainder of the thesis, the structure can be summarized as the following. Chapter 2 presents the literature review. A thorough review of both data-driven studies regarding migration and data visualization studies aimed at migration flows and patterns can be found in the chapter. Chapter 3 describes the utilized datasets in this study. Details such as data collection mechanisms and some of the statistics of these datasets are also present in the chapter. Chapter 4 includes the case study to infer internal migration patterns of Turkey conducted with transactional data of credit card records. Chapter 5 presents the visual exploration tool developed for the exploration of available data sources for migration patterns. Lastly, Chapter 6 provides a discussion of findings and refers the future work.

2. RELATED WORK

In this chapter, the literature related with studies conducted for this work will be elaborately covered. The chapter is divided into two section named as Data-Driven Migration Studies and Visual Analytics. While the former section discusses the significant studies that concentrate on migration patterns and related phenomena, the latter is more focused on demonstrating the use cases of data visualization and visual analytics to study concepts such as movement, migration patterns, and migration flows. The relevance of sections here to future chapters should not be considered as a one-to-one mapping to future chapters. Both sections can provide important background for upcoming chapters.

2.1 Data-Driven Migration Studies

Traditional data-driven migration studies, mainly utilize statistical and administrative data sources for studies concerning migration patterns. As defined by the International Organization for Migration (IOM)[1], statistical datasets refer to migration estimates obtained from census and survey results. Administrative datasets refer to estimates derived from regulatory processes and procedures such as visa, residence, and work permits, or border data collection systems. However, these data sources are stated to be not adequate for migration studies due to their shortcomings. Fatehkia, Coles, Ofli & Weber (2020); Sîrbu, Andrienko, Andrienko, Boldrini, Conti, Giannotti, Guidotti, Bertoli, Kim, Muntean & others (2020) state these shortcomings in inconsistencies between available data sources, different definitions used for migration, and lack of dataset at higher spatial resolutions.

As the dimensions for possible analysis for migration movements and related phe-

[1]Resources on Data Sources of Migration Data Portal, (2021) Accessed June 2021 from the following URL: https://migrationdataportal.org/resources/data-sources

nomena increase, the shortcomings of traditional data sources are becoming more evident. Because of this, studies that aim to analyze migration with data seem to be utilizing innovative data sources more commonly. Innovative data, in this case, maybe considered as the data which is originally not designed for migration studies, however, due to available high dimensionality and metrics these data sources can be utilized when studying migration movements.

The recent study of Sîrbu et al. (2020) introduces a framework where the migration studies with big data sources are divided into three categories called the journey, the stay, and the return. While the journey includes studies aiming to infer migration flows in different scales, studies under the stay category are aimed to provide a better understanding of the demographic and social results of migration movements. An example study can be given as the study by Carmon (1996) on policies regarding migrant integration.

This section first discusses the studies that aim to predict migration movements by utilizing available data sources. Hence, these studies can be considered in *the journey* category. Then, the discussion of related work is followed by recent studies with innovative data sources. As some of the studies also include analysis of other social phenomena and additional socio-economic and demographic metrics, these studies can be considered as *the stay* studies.

2.1.1 Studies For Modeling and Now-casting of Migration Patterns

Data-driven migration studies can generate significant insights into migration patterns on different scales. Usually, these studies utilize the available statistical and administrative datasets published by international organizations. Additionally, these datasets are often complemented with publicly available demographic and socio-economic metrics datasets of these organizations. However, existing gaps in data collection for these datasets constitute the primary limitation of studies with these data sources. Because of these long periods, it may be difficult or not possible to gather current migration stock estimations. Due to this, the ability to nowcast these migration patterns can be valuable. Nowcasting in this context refers to predicting current migration flows between countries. To this end, gravity models modified to forecast migration movements can be utilized.

Although gravity models originated with economic flows and trade relations, recent studies demonstrate gravity models to infer migration movements. When employed

for this context, researchers define push and pull factors among the indicators related to migration. These indicators tend to be in the demographic and socio-economic domains. Similar to gravity models for trade, utilizing the distance between countries as a push factor is also a common usage while forecasting migration stocks. As an example, Lewer & Van den Berg (2008) utilize distance between countries among their independent variables for the gravity model. Their study includes the total population of countries, differences between income per capita between countries, the existing stock of immigrants, sharing common languages and borders, sharing colonial histories, and some of the metrics also utilized in prior studies in their model.

The utilization of gravity models for migration dynamics seems to be first coined in the work by Borjas (1989). Karemera, Oguledo & Davis (2000) expand on the prior definitions to obtain a gravity model that can represent factors affecting migration and model the international migration patterns in a more successful way. The gravity model presented by Ramos & Suriñach (2016) aims to infer migration movements between European and European neighboring countries. Push and pull factors range are in various domains ranging from demographic to political factors. Their findings suggest that distance, contiguity, and GDP differences between countries are the most significant indicators to predict migration forecasts between neighboring countries.

The International Migration Drivers study by Migali, Natale, Tintori, Kalantaryan, Grubanov-Boskovic, Scipioni, Farinosi, Cattaneo, Benandi, Follador & others (2018) also utilizes gravity models to infer the effects of chosen indicators on migration patterns between countries. With gravity models, their work aims to discover the coefficients for socio-economic indicators such as GDP per capita, expenditure in education, networks between states, etc. The study presents the results of four different gravity models designed and developed for varying definitions and settings of types of migration. One of the different approaches taken in the study is the categorization of countries. The study divides countries into three groups based on their income level for the models. This way, relationships between migration estimates and reported coefficients of socio-economic indicators can be analyzed also in terms of the income level of the country.

2.1.2 Studies With Innovative Data Sources

Even though many statistical and administrative data sources are publicly available for the analysis of global migration dynamics, limitations discussed earlier still tend to persist. Those limitations on the quality of available data are often reflected in the practices with these data sources. It is often a challenging task to utilize data sources from various statistical and administrative sources and blend them for migration analysis. As stated earlier by Spyratos, Vespe, Natale, Weber, Zagheni & Rango (2018), mismatches in data types, data collection dates, and missing statistics are the main obstacles to analyses in which these datasets from official international organizations are combined. On top of these complications, the presented granularity of the available data also produces another layer of complexity to working with these datasets. As these datasets are generally aggregated at the country level, studies aiming to analyze migration patterns of areas with higher spatial resolution cannot utilize these datasets directly. For such cases where sub-national migration data is needed, predictive models or other estimation methods seem to be necessary.

To address the shortcomings of the traditional data sources listed above, recent studies introduce innovative data sources for studying the mobility and migration patterns of masses. As described by IOM[2], big data sources create the primary portion of these types of data sources. Although these datasets are not targeted for studies in migration or related subjects, indicators or present attributes in these datasets can allow researchers to include them in their inquiries. The main contributing factor of this availability arises from this type of data being reflective of the behavior or movement of individuals. In this context, examples such as social media usage, call details records, spending behaviors, mobility patterns of individuals inferred from big data sources can allow researchers to gain insights about migrant communities. On top of that, this type of data can help researchers to model global migration flows.

The utilization of innovative data sources for the analysis of migration dynamics is slowly becoming to be a common practice. As the data obtainable from these sources are not often limited only to emigration and immigration estimates, studies can also focus on other social phenomena related to migration flows too. From the metrics and indicators gathered from these data sources, studies try to gather insights on subjects such as migration intent and migrant integration. To exemplify the utilization of these non-traditional data sources; Böhme, Gröger & Stöhr (2020)

[2]Migration data sources of Migration Data Portal, (2020) Accessed May 2021 from the following URL: https://migrationdataportal.org/themes/migration-data-sources

employs Google Trend Index data for measuring the migration intentions in the origin countries, Blumenstock, Chi & Tan (2019) utilized call detail records of migrants to create networks between migrants for analysis of immigrant communities, Fiorio et al. (2017) and Zagheni, Garimella, Weber & State (2014) uses data from Twitter for inferring migration patterns of individuals, Spyratos, Vespe, Natale, Weber, Zagheni & Rango (2019) show that and similar data obtained from Facebook can be utilized for the better understanding of mobility patterns related with migration.

Alexander, Polimis & Zagheni (2020) demonstrates the forecasting or the nowcasting ability of innovative data sources with studies on migration mobility. As argued by Abel & Cohen (2019), these studies hold the potential to be alternatives to statistical models aiming to now-cast global migration patterns with statistical and administrative data. As noted long gaps in data collection can decrease the predictive performance of models built with traditional data sources. A concrete example of this is the availability of country-to-country migration statistics data of the UN being collected at intervals of five years. On the other hand, many of the available innovative data sources are accessible on-demand and can be collected for shorter intervals. This way, models relying on the analysis of time-series of migration patterns can be built with much more instances. Besides, due to being generated from digital traces of behaviors and actions of their users, innovative data sources can also provide indicators and metrics that cannot be matched by traditional data sources. Herdağdelen, State, Adamic & Mason (2016) and Dubois, Zagheni, Garimella & Weber (2018) demonstrate the capability of innovative data sources in reflecting demographic and socio-economic metrics of migrants and produce significant inferences with such metrics.

The predictive capability of Facebook data for nowcasting the emigration patterns is demonstrated in the study by Zagheni, Polimis, Alexander, Weber & Billari (2018) as well. Their study combines data obtained from Facebook with the statistical survey data from an official organization to build a Bayesian Hierarchical model. They showcase their model to nowcast - predicting the current trends - of migration from Mexico to California.

The utilization of data from Facebook for similar studies is not only limited to emigration estimates. In recent studies, researchers also utilize this innovative data source to gather socio-economic well-being insights on communities in different regions. The common approach for such studies are described in Giurgola, Piaggesi, Karsai, Mejova, Panisson & Tizzoni (2021) and Fatehkia, Tingzon, Orden, Sy, Sekara, Garcia-Herranz & Weber (2020). To elaborate, the approach is usually based on analyzing ownership percentages of different mobile devices and analyzing

connection or satellite technologies individuals use for browsing Facebook. Fatehkia et al. (2020) employ this approach to study to demonstrate ownership of different device types and using different connection technologies can be successful estimators for well-being indicators developed prior by international organizations. Similarly, Giurgola et al. (2021) use these available metrics for comparative poverty estimates for four different cities in the world.

Large-scale population displacements can also occur due to incidents and natural disasters. In such circumstances, referring to statistical and administrative data to study migration patterns can be challenging. The study by Acosta, Kishore, Irizarry & Buckee (2020) makes use of Facebook data for a comparative study together with data generated by mobile phone usage. Their work relies on the data shared by Facebook's Disaster Maps[3] repository.

Call Detail Record (CDR) datasets are another source that can be utilized for understanding individual and group mobility. The mobility aspect of these datasets is not only limited to migration studies, as the work of Lu, Bengtsson & Holme (2012) exemplifies another utilization of these datasets for investigating mobility patterns of individuals after incidents. However, when used for understanding migration dynamics, they are usually used for creating networks that aim to uncover the social networks existing between individuals. As these networks can hint at the communication networks and behaviors of migrants in destination countries, as performed in Blumenstock et al. (2019), they are usually linked with studies concerning metrics and indicators on immigrant integration.

As owners of these big data sources, telecommunication companies can host their big data of CDRs for studying social phenomena such as migration flows and migrant integration. The study of Salah, Pentland, Lepri, Letouzé, Montjoye, Dong, Dağdelen & Vinck (2019) is an example with such a data source. Similar studies involving network analysis of migrants can also be performed with data obtained through online communication services and network platforms. Kikas, Dumas & Saabas (2015) perform a similar study with Skype networks and Rodriguez, Helbing, Zagheni & others (2014) focuses LinkedIn networks of individuals.

As one of the innovative data sources, Twitter data reports geolocation data of users from their tweets, focusing on the changes of these locations or the trend of these changes is one of the main strategies of studies performed with the Twitter data. The combination of the availability of geolocated data of individuals and the languages of tweets shared by users can also lead to a better understanding of

[3]Disaster Maps of Facebook, (2021) Accessed June 2021 from the following URL: https://dataforgood.fb.com/tools/disaster-maps/

international mobility patterns. Referring to Twitter data to analyze the temporal and spatial distributions of language compositions is a combined method to uncover these patterns. Such analysis is shown to be beneficial by Moise, Gaere, Merz, Koch & Pournaras (2016) to estimate the number of individuals who tweet with another language than the language spoken in the current location of the person.

One of the challenges of using Twitter data seems to be deriving the definition of migrant from the tweets of users. Although this complexity does not originate from this data source, Twitter data is one of the innovative data sources where a definition of a migrant can be produced within data collection and filtering procedures. As claimed by Hannigan, O'Donnell, O'Keeffe & MacFarlane (2016), the origins of this problem however are rooted in the variances of migrant definitions of different countries. Fiorio et al. (2017) also states the complexity to be reflected in Twitter data as the importance of a well-defined 'migrant' terminology that categorizes individuals as migrants from the geographical and temporal analysis of their tweets. Since the analysis of Twitter data can allow researchers to infer mobility patterns of individuals over time, a definition of migration or migrant based on a selected duration of the inferred mobility can be crucial for successful estimations for migration patterns. Nevertheless, the challenge of distinguishing migration from other types of mobility with Twitter data is stated by Armstrong, Poorthuis, Zook, Ruths & Soehl (2021) to be persisting as many obtained samples have a large number of instances with missing geolocations.

As also is the case in the other innovative data sources, it is stated that Twitter data should also not be considered as the true representative of the underlying population. It is due to sampling performed for studies not being able to represent the whole Twitter user base and Twitter users not being a representative sample of the underlying populations. On the other hand, Zagheni et al. (2014) points out that similar to other innovative data sources described in this section, careful analysis of this data can reveal its ability to infer international and national migration patterns.

2.2 Visual Analytics

In Foundation for a Science of Data Visualization, Ware (2004) lists the advantages of data visualizations. Providing mediums for individuals to comprehend large amounts of data, providing insights about the collection methods, and allowing properties otherwise hidden to be observed are some of the advantages listed in the study. Data visualizations are stated to be more informative as compared to other ways of conveying information as humans can acquire more information through vision compared to other senses. Claims of Stevens (2017) also support this statement through the Stevens' Power Law.

As one of the fields of visualization, Card (1999) defines information visualization as the inquiry of representing underlying data with computer-supported and interactive systems.Shneiderman (2003) summarizes the seven primary abstracted tasks for visual systems with the Visual Information Seeking Mantra. The Information Seeking Mantra is one of the well-known paradigms that contribute to the taxonomy of information visualization. These tasks are displaying an overview of the data, zooming, filtering and displaying details on demand, displaying relationships, supporting iterative usage, and lastly allowing users to extract filtered data. The study also describes how these abstracted tasks can be utilized for different underlying data types. Supporting tasks and principles are also pointed out by Tufte (1986) and Tufte, Goeler & Benson (1990).

The study by Keim (2002) describes the tasks and guidelines for visual exploration systems for different data types. lists the potential benefits of visual data exploration. Similar to the advantages listed earlier, the benefits of this type of information visualizations are pointed out to be easing the exploration of unstructured data, facilitating better understanding, and conducting data exploration in a faster way.

Visualization of movement is one of the areas that is often addressed with information visualization techniques. Andrienko & Andrienko (2013) categorizes the most significant approaches to be taken for movement visualization. Their work is aimed to distinguish different methodologies for visualizing movement data. As argued, methods can involve movement analysis through various approaches that involve efficient visualizations of trajectories and flows. Many examples of the methodologies listed in their study can be found in the work of Slingsby & van Loon (2016) for movement in the ecology domain.

2.2.1 Visualization of Migration Flows & Patterns

Since the phenomena itself denotes a movement from an origin location to the destination location, many of the successful applications of data visualization and visual analytics to study migration flow concentrate on algorithmic solutions to origin-destination visualizations. In the data visualization and visual analytics literature, approaches to address origin-destination data firstly utilize origin-destination matrices. In this context, an OD matrix is a data structure that conveys available flow statistics for origin and destination pairs in the data. Studies by Robillard (1975) and Willumsen (1978) are example studies that create or estimate OD (origin-destination) matrix from the underlying data sources.

These studies are followed by studies with OD maps and OD flow maps. As defined by Wood, Dykes & Slingsby (2010) OD maps encode the movement from the origin to destination by the accurate encoding of the geographic locations. This way a grid structure that reflects the geography and location of the regions visualized can be achieved. Then, cells in these grid structures can be filled with visual encoding to denote the density of the movement. As noted by Andrienko & Andrienko (2008), aggregation of estimates by the origin and destination is a very common approach to obtain OD maps.

OD flow maps also add encoding of the movement from the origin to destination location on map visualizations. The main advantage of this approach is that it enables the user to observe the movement or flow on the original geography. The study of Tobler (1987) can be considered as one of the earlier studies for OD flow maps. However, the increasing number of movement patterns to be visualized this approach is likely to introduce visual clutter on maps. This may cause these approaches to be difficult to interpret by the users. Many of the recent works concentrate on better representations of flows and optimum ways of visualizing them such as filtering and bundling.

Figure 2.1 displays examples of OD map and OD flow map from the study by Yang et al. (2016). The figure on the left displays flows on the map. The OD map in the middle encodes origin-destination data in the smaller maps for encoded locations. The right-most figure is the combined approach of the authors called MapTrix.

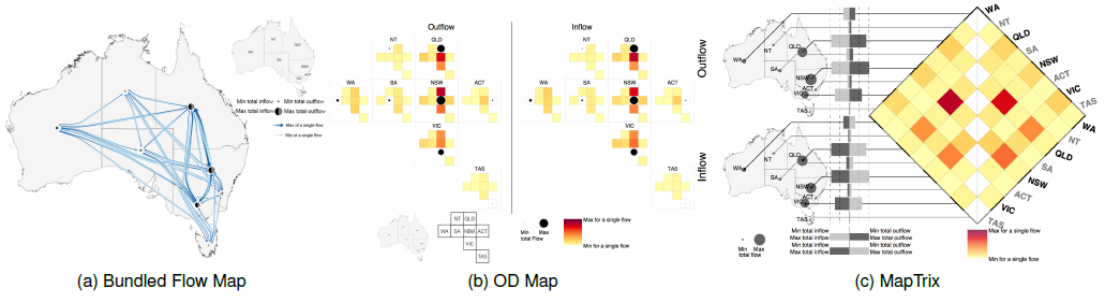

Figure 2.1 The MapTrix visualization method depicted in the study of Yang et al. (2016)

As described above, flow map layouts are one of the existing solutions to display origin-destination typed data. As noted in Buchin, Speckmann & Verbeek (2011), recent and earlier implementations of flow maps tend to differ. Initial adoptions of the method did not seem to introduce any methods to decrease visual clutter. On the other hand, many of the recent studies are aimed to introduce novel algorithms to increase the visual quality of the presented flow. As examples to these; Buchin et al. (2011) employs edge-bundling to simplify visualized flows and Guo & Zhu (2014) utilizes a flow smoothing method based on the spatial neighboring of flow representations. A resulting flow map of their work is presented in Figure 2.2.

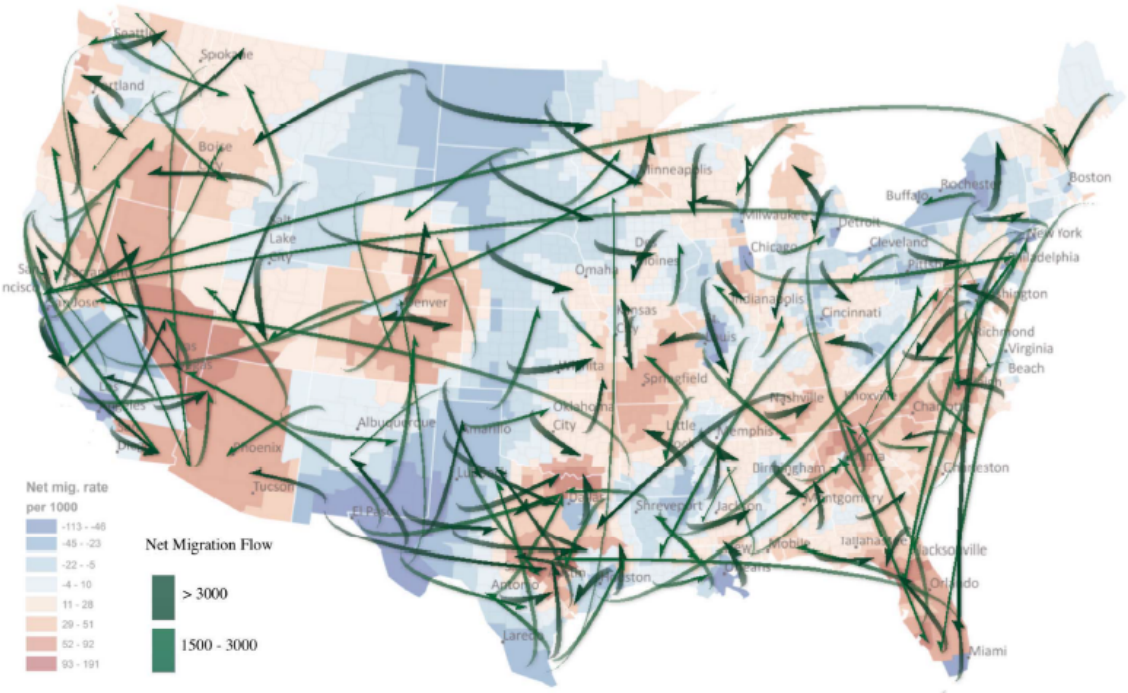

Figure 2.2 The study by Guo & Zhu (2014) combines a flow visualization with the underlying choropleth map of migration estimates

To visualize origin-destination data, visual abstraction is one of the well-utilized

methods in the literature. What is often aimed by applying these abstraction techniques for origin-destination data to achieve visualization that is easier to conceive. It is usually obtained by representing more complex origin-destination flows with transformed representations that are more simple. The methods listed in the study by Zhou, Meng, Tang, Zhao, Guo, Hu & Chen (2018) demonstrates some of the possible ways to achieve the desired abstraction. Similar efforts are also present for automated and optimized representations of flows in the maps. The work of Stephen & Jenny (2017) concentrates on the optimal placements of nodes and flows in flow maps. Their implementation is then displayed with internal migration patterns of the United States. Figure 2.3 displays an exemplary visual from the discussed study.

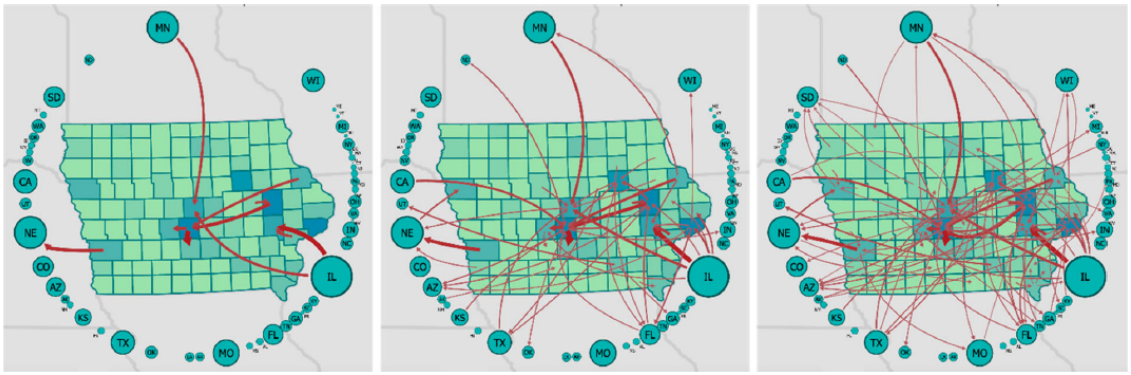

Figure 2.3 Visualization of migration flows in the United States in the study of Stephen & Jenny (2017)

As the presence of origin-destination data is prevalent in most of the data-driven studies regarding migration, studies that are not primarily focused on data visualization often utilize effective ways of conveying their findings in visual components. The study by Abel & Cohen (2019) can be stated as an example of such inquiries. The authors of the study utilize chord diagrams to demonstrate the country-to-country migration patterns. These chord diagrams also encode the region and position of the country in the region information. Figure 2.4 share examples of these chord diagram visualizations of the authors.

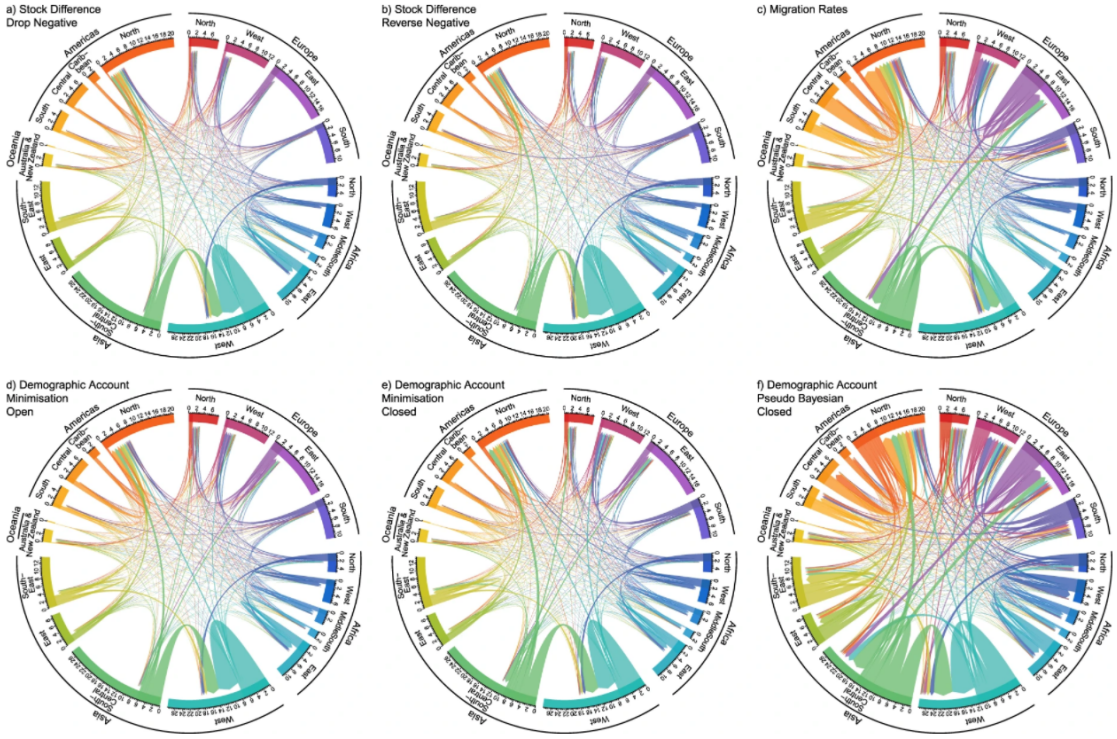

Figure 2.4 Estimated migration flows are visualized with chord diagrams in the study by Abel & Cohen (2019)

International organizations for migration also utilize data visualizations. It is also common for these organizations to host their visual findings on online platforms. As one of these platforms, Migration Data Portal[4] shares the visualizations of migration patterns and other related indicators of migration. The portal also contains articles on migration-related phenomena and data sources for migration.

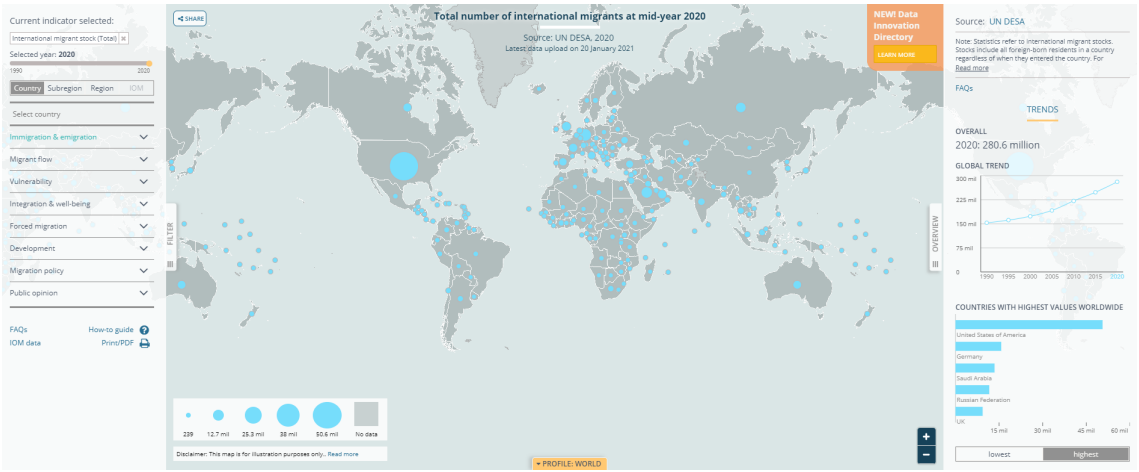

Figure 2.5 A screenshot from the main page of the Migration Data Portal

[4]Migration Data Portal of IOM, (2017) Accessed April 2021 from the following URL: https://migrationdataportal.org/

3. UTILIZED DATASETS

In this chapter, datasets utilized for the proposed visual tool and the case study are covered in detail. Each section of this chapter summarizes the data source and characteristics, also the data collection methodologies if applicable. Section 3.1 and Section 3.2 discuss statistical datasets obtained by the United Nations and Eurostat respectively. Section 3.3 will explain data obtained by Facebook Marketing API and how it can be utilized for migration studies. Section 3.4 will clarify the transactional data employed for the case study for internal migration patterns of Turkey. Lastly, Section 3.5 illustrates the Internal Migration Estimates dataset of TURKSTAT.

As datasets utilized in this work can display migration patterns in different scales, each section of the chapter also covers important statistics of these datasets. These statistics are presented with figures when necessary.

3.1 United Nations Migrant Stock

As the primary data source of statistical data, United Nations Migrant Stock data is utilized [1]. Datasets shared by UNDP include multiple datasets of country-to-country migration estimates aggregated by various demographic indicators. The main dataset shared by UNDP includes total international migrant stock, migrant stocks aggregated by age groups and sex, and migrant stock estimations by origin and destination. The utilization of this data source is discussed below.

The total international migrant stock dataset is employed for the Main Map component of the visual tool. The Main Map component uses this dataset to render one of the choropleth map options. Then, this choropleth creates a selection mechanism

[1] International migrant stock 2019 of UNDP, (2019) Accessed May 2021 from the following URL: https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp

of the visual tool.

The migrant stock estimations are utilized by multiple components of the visual tool. The Comparison Panel of the tool uses this dataset to display the origin countries for emigration to the selected country, and destination countries of emigration. The same dataset is also used in the Comparison Panel and the Ladder Plot components of the tool. Comparison Panel component takes this dataset to distribute it to NUTS centers based on the selected metric. The latter takes this dataset to generate a ladder plot which enables comparison of metrics from origin and destination.

Below, some of the important statistics of the United Nations Migrant Stock dataset are visualized.

The United Nations Migrant Stock data is shared with periods of five years starting from 1990. The only exception is that the last version of the dataset is from 2019. Having five-year periods allows the comparison of primary migration patterns with time. Table 3.1 below displays the total migrant stock in the world for all available years.

Year	Estimated Migrant Stock
1990	153,011,473
1995	161,316,895
2000	173,588,441
2005	191,615,574
2010	220,781,909
2015	248,861,296
2019	271,642,105

Table 3.1 Total migrant stock estimates of United Nations for available years

It can be seen that the total migrant stock of the world is in an increasing trend for all available time intervals. As also reported by other sources, the 2019 estimations for migrant stocks refer approximately to 3,5% of the whole population of the world[2].

The United Nations Migrant Stock data also categorizes countries and regions based the on development metric of UN. Figure 3.1 display the total emigration estimates of these regions for all available years. It can be seen that the primary destination of the emigration patterns present in the world is countries categorized as developed.

[2]Population, total of The World Bank, (2021) Accessed May 2021 from the following URL: https://data.worldbank.org/indicator/SP.POP.TOTL

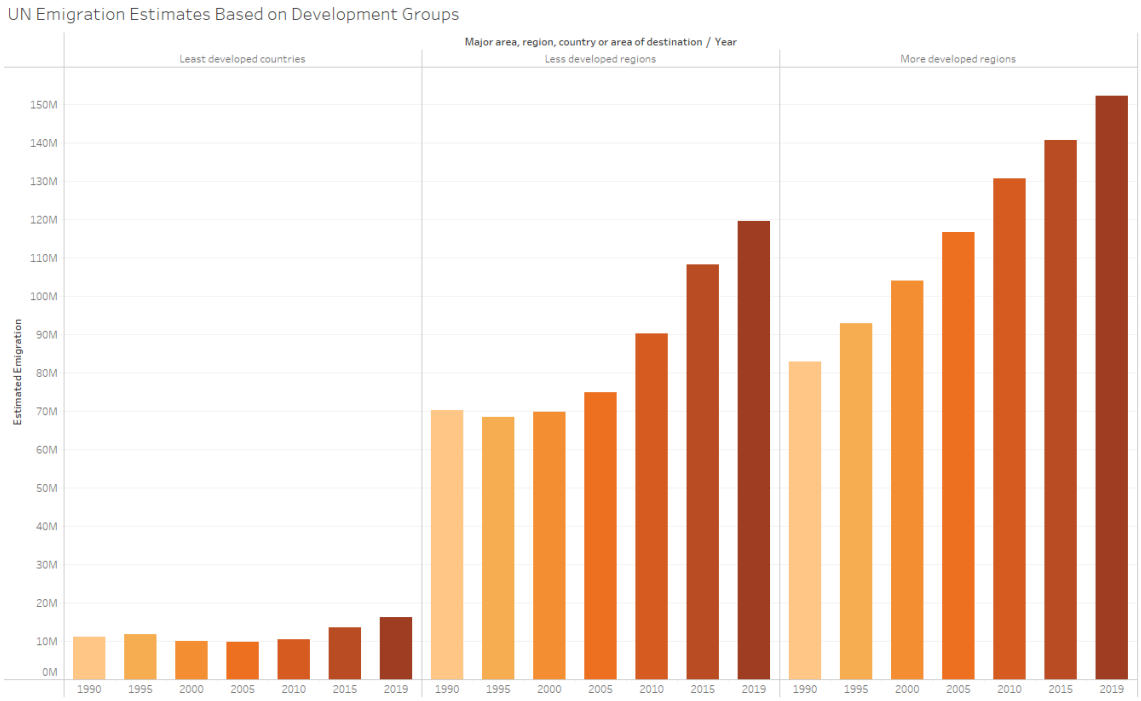

Figure 3.1 United Nations emigration estimates based on development groups

Below, The 10 countries with the most emigration estimates are visualized for 1990, 2005, and 2019. The countries on the bar chart convey the destination countries of migration patterns.

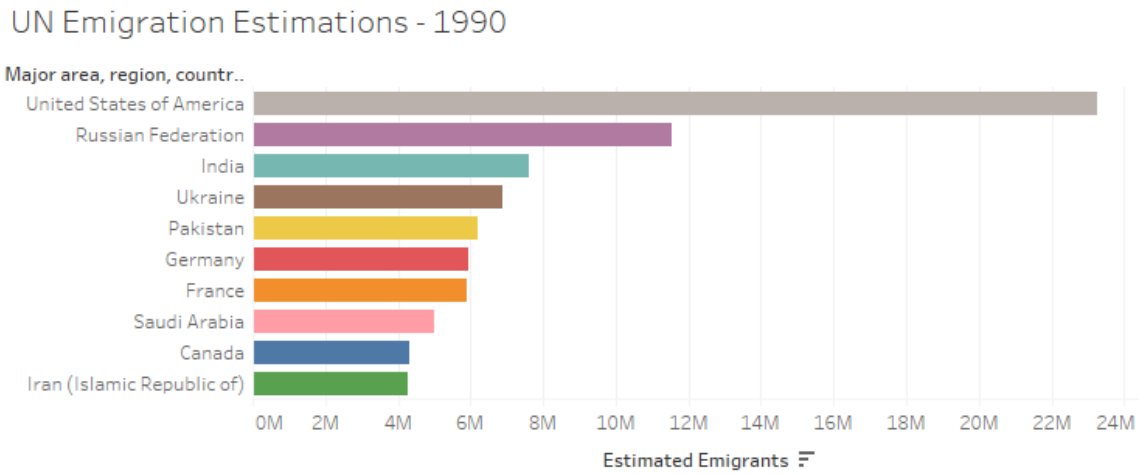

Figure 3.2 United Nations emigration estimates rankings - 1990

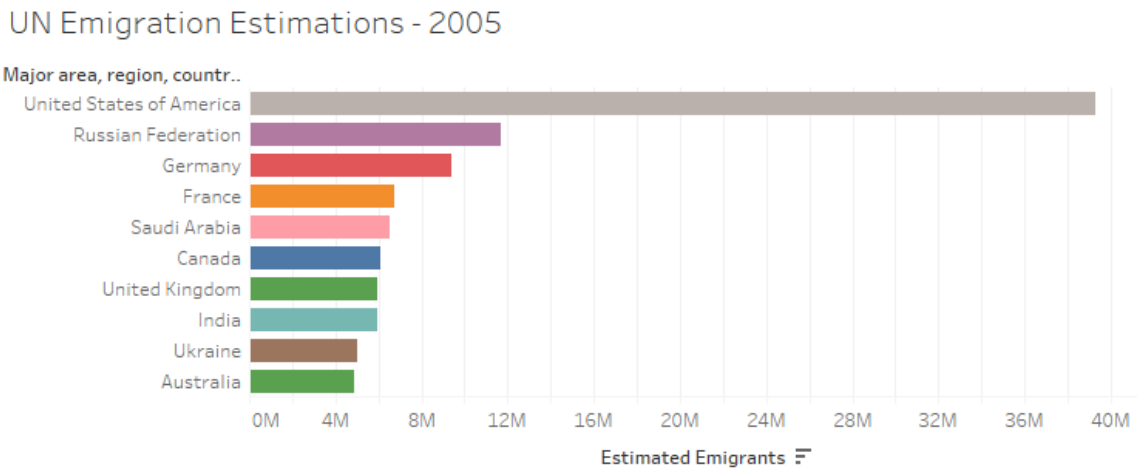

Figure 3.3 United Nations emigration estimates rankings - 2005

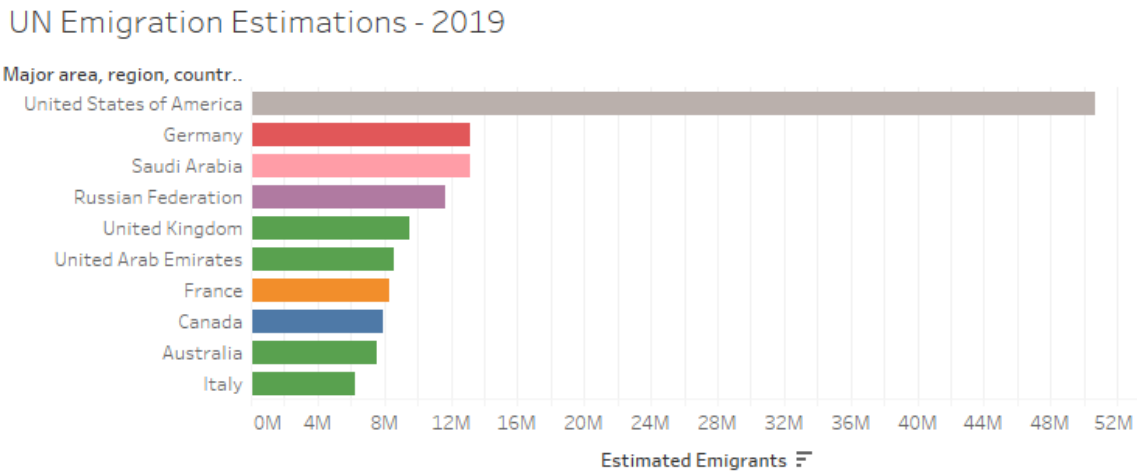

Figure 3.4 United Nations emigration estimates rankings - 2019

It can be seen that the United States of America is the primary destination of the migration patterns in the world. This trend did not seem to change for almost thirty years covered by the data source. Additionally, countries such as Germany, Russian Federation, Saudi Arabia, India, and France also keep on being highly preferred destinations for migration movements. As can be observed, the scale of the emigration is increasing for countries displayed.

The figures below visualize the immigration estimates for 1990, 2005, and 2019. Country names on the vertical axis of these bar charts convey the origin country information.

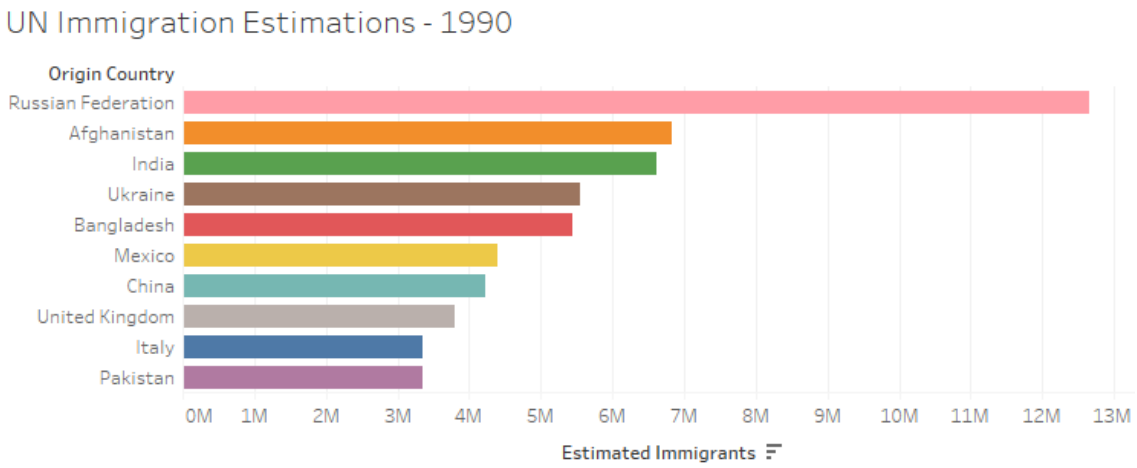

Figure 3.5 United Nations immigration estimates rankings - 1990

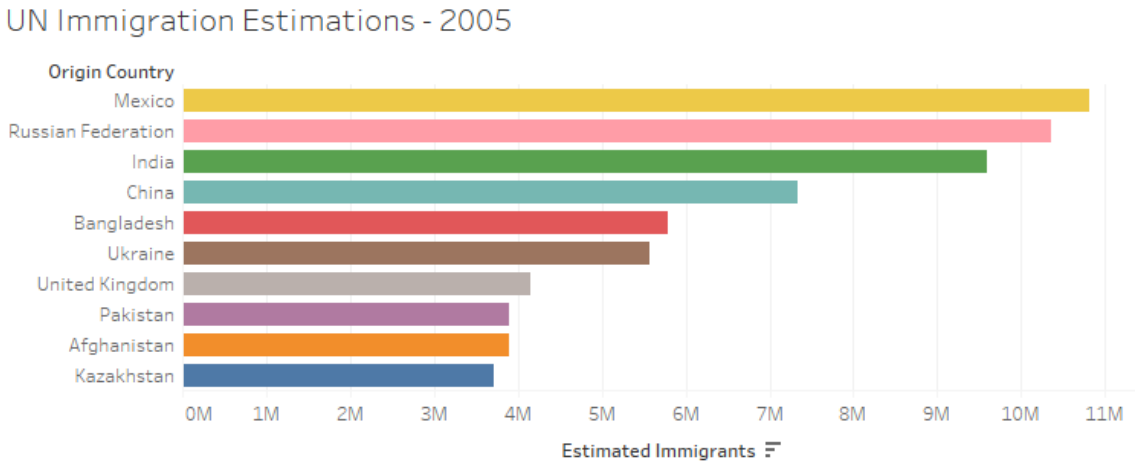

Figure 3.6 United Nations immigration estimates rankings - 2005

Unlike the emigration estimates, the rankings of the immigration estimates are more variant. It can be speculated that the main reason for this observation is due to immigration movements taking place as a result of a social or economic incident in the origin country. The top country for immigration changes for all year of data collection.

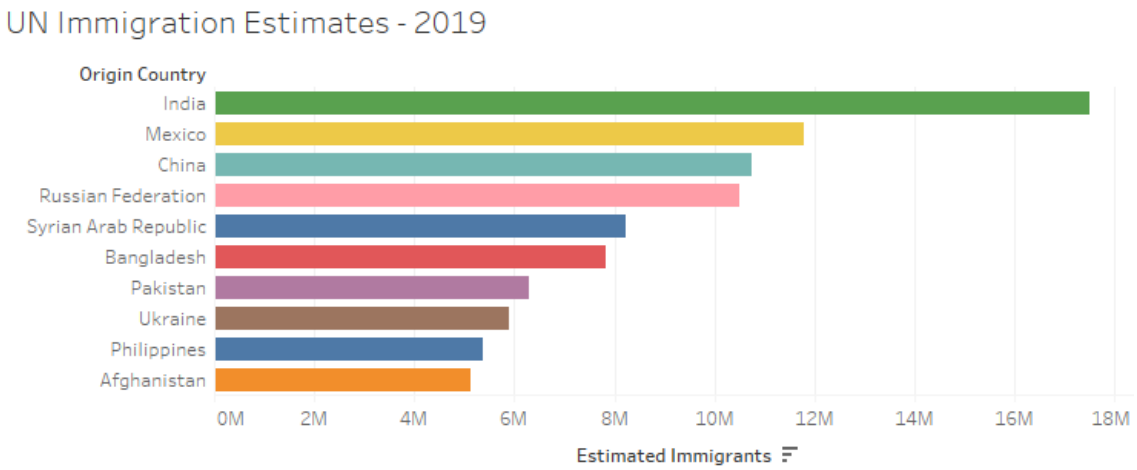

Figure 3.7 United Nations immigration estimates rankings - 2019

3.2 Eurostat Regional Metrics

Eurostat Regional Metrics are employed to serve regional metrics to be compared with data mostly aggregated at the country level. However, as datasets shared by Eurostat are also available at the country level, these metrics can be employed for both insights about countries and their regions. These metrics are employed in the Comparison Panel and the Ladder Plot components of the visual tool which is described in Chapter 5 comprehensively.

Regional Metrics shared by Eurostat are available at different spatial resolutions. These spatial resolutions are four different NUTS granularity levels. As described by Eurostat, NUTS refers to a standard to refer to geographic divisions of European regions. However, the standard of dividing regions is not imposed on countries. Rather, countries fit their existing regional division mechanisms to satisfy NUTS conditions[3]. This way regions with similar statistics such as population and GDP can be obtained. While NUTS 0 refer to country borders, these regions are more specific as the granularity increase. Regions in NUTS 3 refer to provinces or districts of countries[4].

[3]History of NUTS of Eurostat, (2021) Accessed May 2021 from the following URL: https://ec.europa.eu/eurostat/web/nuts/history

[4]NUTS Maps of Eurostat, (2021) Accessed May 2021 from the following URL: https://ec.europa.eu/eurostat/web/nuts/nuts-maps

Spatial Resolution	Number of Centers
NUTS 0	37
NUTS 2	332

Table 3.2 Number of NUTS regions present in utilized spatial resolutions for the visual tool.

The visual tool utilizes regional metrics shared by Eurostat at two different NUTS levels, which are NUTS 0 and NUTS 2. Metrics included in the visual tool are total population, population density, GDP, number of health personnel per individual, the ratio of people at the risk of poverty, and the employment rate of the young population. Datasets for these metrics are accessed from the regional databases of Eurostat[5]. These metrics are utilized in the Comparison Panel to divide migration estimates into regions and in the Ladder Plot to visualize the difference of metrics between the origin and destination of migration patterns.

In Table 3.2 below, number of NUTS regions for spatial resolutions utilized in the visual tool developed are shared. While NUTS 0 regions correspond to 37 countries in Europe, NUTS 2 regions refer to smaller regions in these countries.

3.3 Emigration Estimates From Facebook Marketing API

In this work, datasets obtained from Facebook's Marketing Platform are used as the primary source of innovative data. Originally, this platform of Facebook is intended for individuals planning to post targeted advertisements on Facebook. However, this platform can be utilized to obtain estimates of emigrants from Facebook with the condition to have Facebook accounts. The reason that gathering these estimations is possible is because of the available filtering options that the Marketing Platform allows its users for targeted advertisements. These filterings are available in three categories which are demographic, interests, and behavior categories. Demographic filtering options cover targeting parameters such as age and nationality, filtering options based on interests mostly cover Facebook pages that users engage with, and filtering options based on behaviors cover how users engage with Facebook.

Among available filtering options for behaviors, what makes generating emigration

[5]Database - Regions from Eurostat, (2021) Accessed February 2021 from the following URL: https://ec.europa.eu/eurostat/web/regions/data/database

estimates of regions is possible is the presence of filters as *"People used to live in country X."*. When these filters are used with other geo-locations, emigrants represented in Facebook for the region or country can be inferred. Hence, filterings such as *"People used to live in country X and currently live in Y."* can be performed to gather reach estimations of advertisements. In this study, emigration estimations from Facebook Marketing API rely on this type of filtering options.

Additionally, users can also operate on the Marketing API service of the platform to manage their advertisement programs. This way data with desired filters can be collected in a programmed or automated manner.

3.3.1 Data Collection

As also utilized by Palotti, Adler, Morales-Guzman, Villaveces, Sekara, Garcia Herranz, Al-Asad & Weber (2020), the python wrapper library pySocialWatcher[6] is used for gathering data from the Facebook Marketing Platform. As presented by Araujo, Mejova, Weber & Benevenuto (2017), the main functionality of the pySocialWatcher library is to manage API requests that need to be sent to Marketing API for given parameters.

As the Marketing API and wrapper library require a validated advertisement campaign in the Facebook Marketing Platform, a template campaign is created for this purpose. Establishing a campaign allows users to generate access tokens that can be used for sending requests to the API service. With an access token, API requests can be sent with desired filtering options. This way, reach estimates of a possible advertisement can be obtained. As the pySocialWatcher library requires JSON objects for setting filtering options, each API request needs the preparation of a specific JSON file for the reach estimates to be obtained. Below, an example JSON file is presented.

[6] https://github.com/maraujo/pySocialWatcher

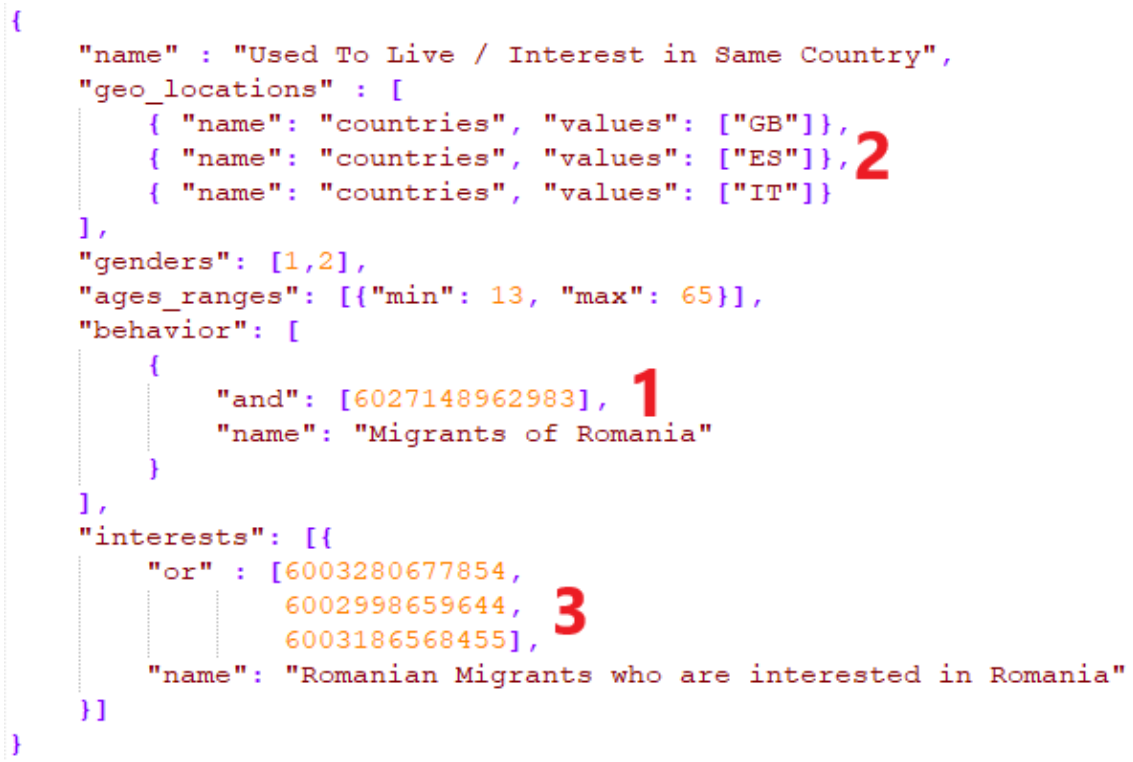

Figure 3.8 Example JSON file for API requests to be utilized with pySocialWatcher

In Figure 3.8 above, a sample JSON file to be utilized by the wrapper library is displayed. Point 1 indicates the behavior filter used and this filtering targets Facebook users who used to live in Romania. Point 2 indicates the target countries that their estimates are interested in. Lastly, Point 3 indicates the filtering options based on the interests of Facebook users. Codes with numbers that represent these interest codes are pages and entities in Facebook which are related to Romania. With such configurations, reach estimates are generated to depict the reach of advertisements to *Romanian individuals who are interested in Romanian subjects and currently live in Great Britain, Spain, or Italy.* Additionally, *genders* and *ages_ranges* parameters are used for targeting a specific age group, and male and female users of Facebook individually.

As the example above describes, country-to-country emigration estimates are obtained similarly. To this end, the origin country is selected for behavior *"People used to live in country X"* and reach estimates for this behavior is analyzed for all countries. The resulting estimates are used as country-to-country emigration estimates. These estimates are also utilized in the components of the Visual Tool covered in Section 5.

3.4 Transactional Data

This section describes the transactional datasets utilized in the case study to detect internal migration patterns of Turkey. It should be noted that the case study that employs the dataset discussed here is covered comprehensively in Chapter 4. To this end, transactional data shared by a private bank of Turkey is utilized. This section, unlike previous sections that first define the data source and then provide relevant statistics, presents and describes the data source with tables and figures.

The mentioned transactional data includes the credit card spending of customers of the bank between 2014 and 2015. The utilized dataset includes transactions of a sample of customers among all of the customers of the bank. In the dataset, transactions also include attributes such as amount, spending category, and coordinates of the location where the transaction took place. However, for the majority of the dataset location information for transactions is missing. Important statistics regarding the size of the dataset can be observed from Table 3.3. It can be seen that almost two-thirds of the transactions are missing location information. When the missing transactions are reflected on unique customers, although the dataset contains transactions of 102,893 unique customers, there are only 98,834 customers whose transaction coordinates are available. The dataset contains 94,803 unique POI information with latitude and longitude pairs. These POIs refer to vendors that customers perform their transactions. To supplement the transaction data, this work also uses an additional dataset for demographic information regarding customers. These two datasets are combined on masked customer IDs to both have demographic insights about customers and their transaction history.

Some of this demographic information is displayed below in an aggregated manner. Figure 3.9 demonstrates the distributions of gender and marital status in the sample. It can be seen that the majority of the sample consists of customers who are male. Again, the majority of the individuals in the sample are married.

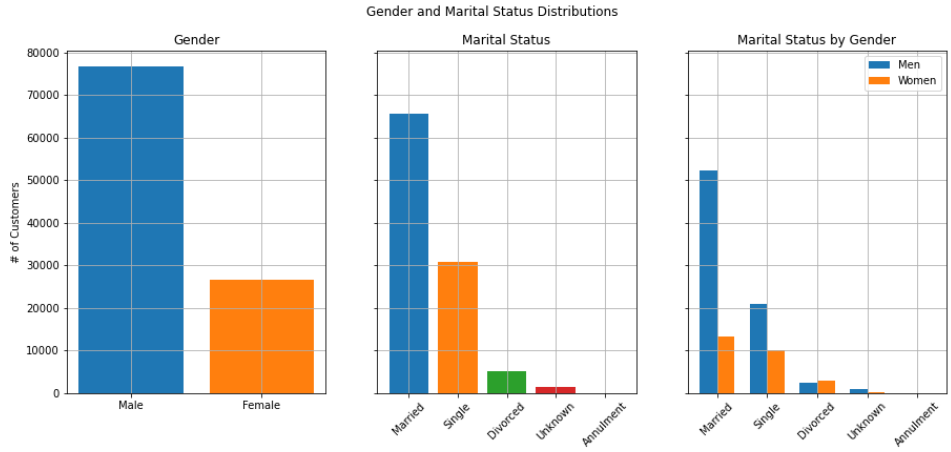

Figure 3.9 Distributions of gender and marital status in the dataset.

Again, thanks to the available demographic metrics of customers, education status and income of individuals in the sample can also be observed. Figure 3.10 displays the education status of individuals, mean income of individuals per education status, and education status by gender respectively. It can be observed that most of the customers of the bank included in the sample dataset are high school graduates or people with a university diploma. This results in individuals in the sample being more educated than the average education profile of the country. Due to this, the sample utilized in this work may not be a successful representative of the average customer profile of private banks of the country. This phenomenon is also covered in the Discussion & Future Work section in more detail.

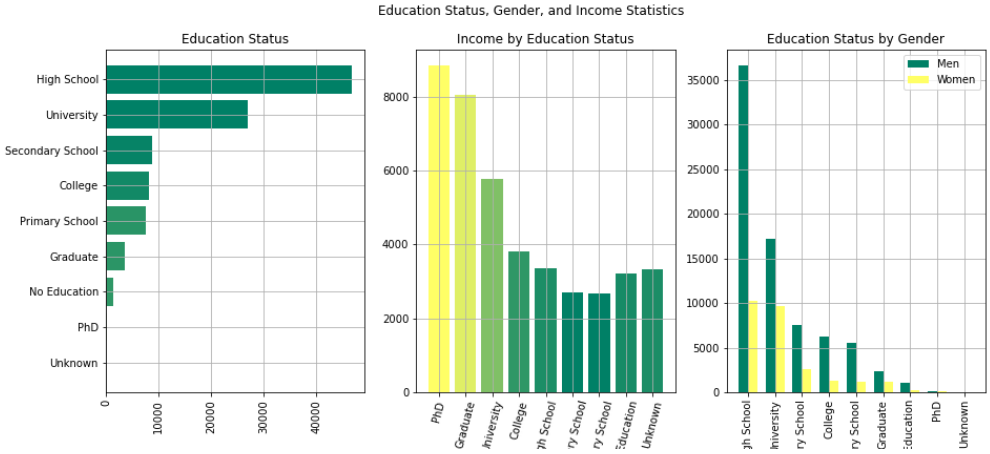

Figure 3.10 Distributions of education status, gender and income in the dataset.

As mentioned above, transactions of customers in the dataset contain the spending category information. Regarding these categories, Figure 3.11 can be observed. The majority of transactions are made in Food, Restaurants, Clothing, and Gas Stations. It is also visible that more than one million transactions do not have

Statistical Property	Numerical Value
# of transactions	9,334,625
# of transactions with coordinates available	3,729,193
# of unique customers	102,893
# of unique customers with transaction coordinates available	98,834
# of unique POI locations	94,811
# of unique POI locations with transaction coordinates available	94,803

Table 3.3 Statistical properties of utilized transactional data. # symbol is used to denote the word 'Number'.

category information. Categories of these transactions are labeled as Not Specified for a better representation.

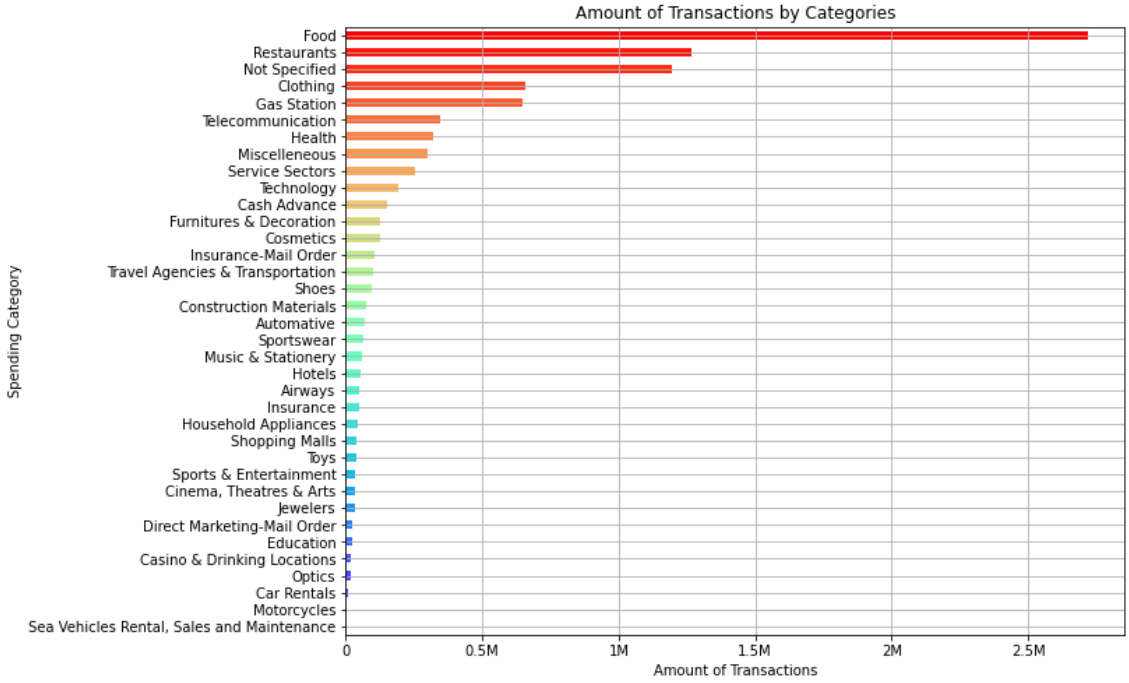

Figure 3.11 Amount of transactions per spending categories in the transactions dataset.

As demographic metrics of customers of the bank are available, Figure 3.12 displays the origin city of customers. In order to assign origins of customers, the initial branch information of the bank where the customers are registered is used. The results suggest that most of the customers are located in Istanbul. Cities such as Kocaeli, Ankara, and Izmir are some of the other cities that have the most customers.

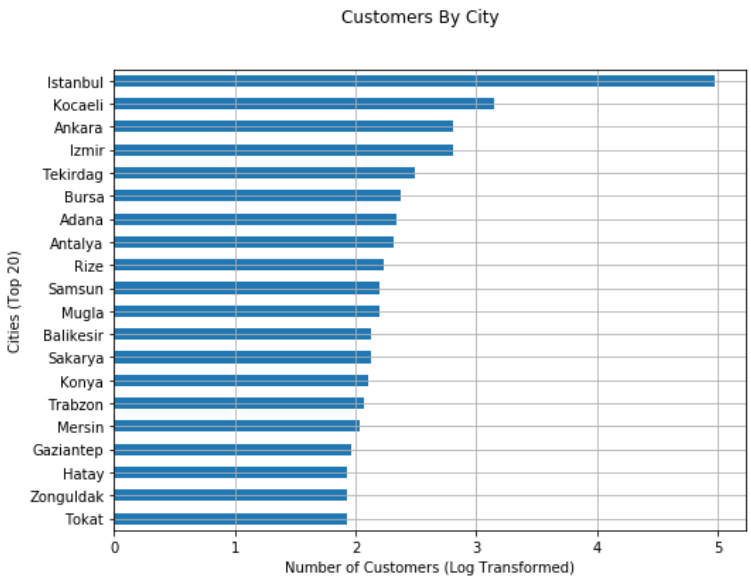

Figure 3.12 The first 20 cities with the most number of unique customers in the dataset. The chart displays the log transformed number of customers per city.

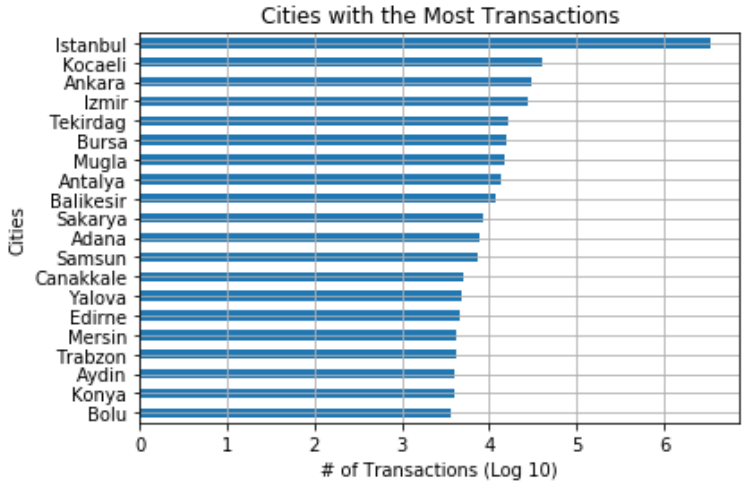

Figure 3.13 The first 20 cities with the most number of transactions (log transformed) in the dataset.

To combine the observations from Figure 3.12 with amounts from transactions customers perform from Figure 3.13, Figure 3.14 can be analyzed. The figure displays a heatmap of the sub-sample of transactions in the dataset with the size of 25,000. While the color scale in the heatmap is used to denote the density of transactions, amount of transactions are encoded with circle size. Similar to what Figure 3.11 suggests, majority of customers being from Istanbul cause the city to have the highest transaction density. Some dense areas are also present in the major cities of Turkey such as Ankara, Izmir, Kocaeli, and Muğla.

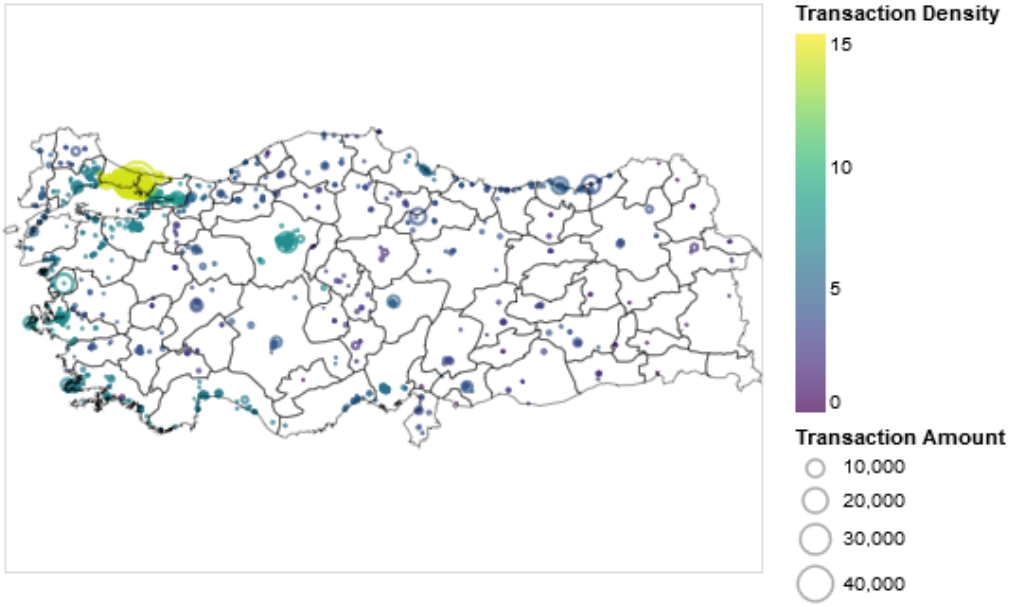

Figure 3.14 Heatmap of transactions from the dataset.

As mentioned earlier, the dataset described in this section is utilized for identifying possible internal migration patterns of individuals. As date of transactions in the datasets are also known, a period-based methodology is embraced to infer possible internal migration movements between cities. The reflection of this methodology is provided by defining periods in which customers stop their spendings in the initial city and start to perform spendings from a new city. The details of the methodology and reasoning on the implementation choices for the case study with this dataset is elaborately discussed in the Case Study With Transactional Data chapter.

3.5 Turkstat Internal Migration Estimates of Turkey

TURKSTAT Internal Migration Estimates of Turkey[7] datasets are utilized to validate the methodologies employed for inferring potential migration patterns with the case study on the transactional dataset described earlier in this chapter. As the data collection period of the transactional dataset spanned between 2014 and 2015, internal migration stock estimates by TURKSTAT for both years are utilized for validation purposes. Table 3.4 and Table 3.5 below display the origin destination

[7] Ülke İçi Göç of TURKSTAT , (2020) Acessed February 2021 from the following URL: https://tuikweb.tuik.gov.tr/PreTablodoalt_id1067

pairs with the most number of internal migrants.

Origin City	Destination City	Estimated Migration
Kocaeli	İstanbul	28,272
Tekirdağ	İstanbul	23,170
İstanbul	Tokat	19,388
İstanbul	Ankara	19,021
Ankara	İstanbul	18,775
İzmir	İstanbul	16,129
İstanbul	İzmir	15,559
Tokat	İstanbul	15,395
İstanbul	Kocaeli	14,952
İstanbul	Van	13,155

Table 3.4 TURKSTAT estimations of primary migration patterns of Turkey in 2014 with origin and destination city pairs.

Origin City	Destination City	Estimated Migration
Kocaeli	İstanbul	29,475
Tekirdağ	İstanbul	25,422
Ordu	İstanbul	21,420
Ankara	İstanbul	18,907
İstanbul	Ankara	18,066
Giresun	İstanbul	17,935
İzmir	İstanbul	17,124
Tokat	İstanbul	17,035
Bursa	İstanbul	14,215
İstanbul	Kocaeli	13,939

Table 3.5 TURKSTAT estimations of primary migration patterns of Turkey in 2015 with origin and destination city pairs.

Both tables display the first then origin destination pair cities with most estimated migrants. It can be observed that the primary patterns for both years show similarities. For both years, the leading migration patterns in the country is the internal migration from Kocaeli and Tekirdağ to Istanbul. These indicate that Istanbul, as the main center of economic activity of Turkey, draws individuals more from the neighboring cities. These patterns are followed by migration movements between major cities of Turkey such as Istanbul, Ankara, and Izmir.

4. CASE STUDY WITH TRANSACTIONAL DATA

This chapter discusses the case study conducted on the transactional dataset described comprehensively in the Transactional Data section of Chapter 3. The goal of the case study is to infer possible internal migration movements of individuals based on their geo-located credit card usage records. Section 4.1 will clarify the study design and the methodology for the case study. Section 4.2 will describe the implementation and will elaborate on the findings.

4.1 Study Design

As discussed earlier, the case study depicted in this work aims to infer possible internal migration movements of individuals who are customers of a private bank of Turkey. The way such study is possible is thanks to the availability of the geo-located credit card usage data of individuals. This dataset is discussed thoroughly in the Transactional Data section of Chapter 3. The dataset includes approximately 9,3 million transaction records of 102,893 unique customers of the bank that shared this dataset. As the geo-location of these spending behaviors is also available, this study processes the transaction histories of individuals to detect location changes that span a long period.

The methodology of this study can be articulated as the following. As the transactional data of customers span one year with geo-location coordinate attributes available, what is aimed to be achieved is the individuals who, at a certain time period, changed their city for a selected time period. This way these customers can create the list of individuals who potentially changed the city they lived in or resettled to another city.

There are studies with innovative data sources to infer possible migration patterns

with similar approaches. The study of Fiorio et al. (2017) utilizes Twitter data for a study with similar methodology. Methodology in their study is revolve around a successful selection of interval and duration that determines a location change of tweets of individuals. Figure 4.1 display the definitions of duration and interval in their study. As mentioned, these definitions serve to categorize the location of individuals from their geo-tagged tweets. While *duration* determines the each period of time to be considered to label the location of the individual, *interval* determines the time between these periods.

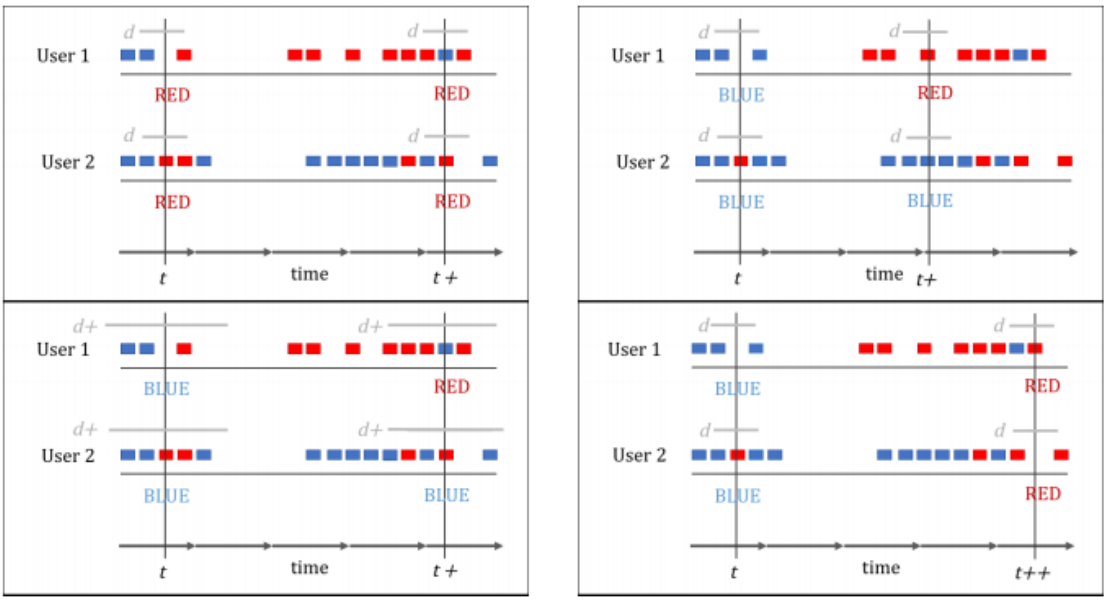

Figure 4.1 Concepts of duration (on the left) and interval (on the right) as discussed by Fiorio et al. (2017)

The recent study of Chi et al. (2020) share a framework to extract trajectories of movements of individuals. The provided framework relies on differentiating digital trace segments of individuals from different locations with a previously defined sufficient time interval. The discussed framework is then applied to two innovative digital trace data sources which are CDR data and Twitter data. Figure 4.2 is taken from the study to briefly summarize the framework described for similar studies. The study of Chi et al. (2020) showcase the performance of their segment based approach against frequency-based geo-labeling methods as well.

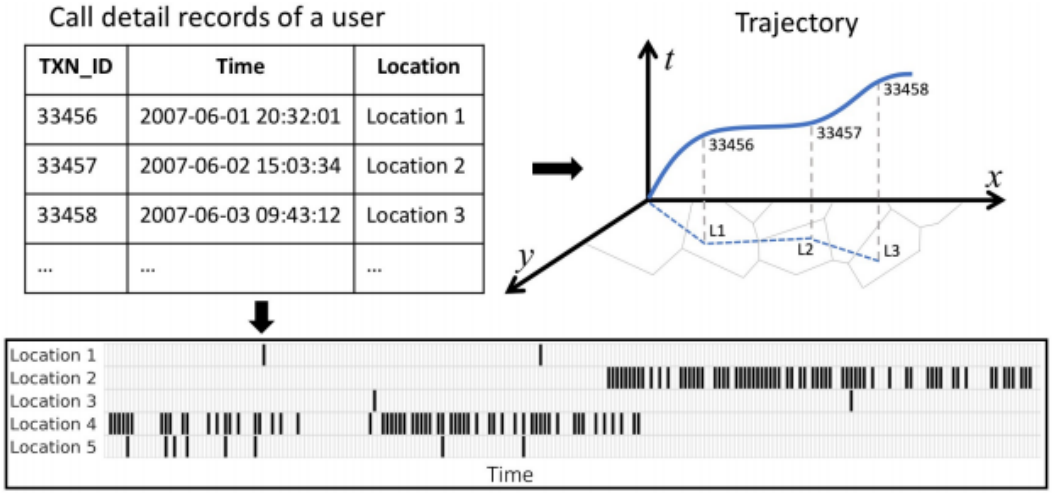

Figure 4.2 Constructing human mobility from raw trace data as discussed by Chi et al. (2020)

Similar to the studies discussed above, the methodology adopted for this case study aims to discover potential internal migration patterns by differentiating between segments of transactions originated from different cities. To this end, a variety of duration and interval periods are tested with the sample data being utilized. Because of the limitations of the sample data, only a handful of the duration values are selected to be applied for segmenting credit card transactions of individuals. The concept of the interval is not applied in the case study as the transactions in the sample only spanned one year. Instead, a continuous approach is adopted to infer resettlement of individuals.

This application of segmentation is described as scenarios for the case study. Due to mentioned limitations, three scenarios with different constraints on transactions of customers are evaluated. Section 4.2 will discuss these scenarios and their reflection on the case study in more detail.

4.2 Case Study

Before detailing the implementation and results of the Case Study, more insight about the dataset utilized and sample that represent the customers of this bank should be provided.

As discussed in earlier sections, datasets used in this study consist of credit card

transactions of a major private bank of Turkey recorded between 2014 and 2015. Again, as pointed out in the Utilized Datasets chapter, the dataset utilized in this work operate on a sample of the bulk of transactional data shared by the mentioned private bank. This way, customers in the sample dataset is a portion of the 33% of the adult population that owns a credit card[1]. The private bank that shared the dataset is stated to have a market share of 10 to 11% for the period between 2014 and 2015[2] [3]. Hence, it can be claimed that customers of the private bank amount for approximately 3% percent of the adult population who use credit cards in Turkey. The importance of this customer profile is stated by Lusardi & Mitchell (2011) to be resulting in better financial literacy.

The significance of the internal migration patterns of Turkey should also be stated in this section. It can be argued that internal migration is one of the primary social phenomena that affect the Turkish population dynamics. Filiztekin & Gökhan (2008) reveal prominent patterns about the history of internal migration patterns of Turkey. It is stated that almost 62% of the population of Istanbul are individuals who are from different cities.

Statistical Property	Numerical Value
Transactions with coordinates	3,729,193
Transactions with coordinates & Each customer $>$ 10 trx.	2,583,101
Unique customers with coordinates	98,834
Unique customers & $>$ 10 transactions with coordinates	42,139

Table 4.1 Details of Transactions Dataset with Introduced Filtering Options. The abbreviation 'trx.' denote transactions.

Before continuing to the definition of applied scenarios for the case study, a data filtering step is applied to the customers in the sample. One of these filtering options was to make sure that each customer to have at least ten transactions recorded in the dataset. This way, a standard for reliability was introduced. A threshold value that is bigger than the applied value could also be preferred. However, it was observed that introducing larger values for this threshold discarded too many unique customers in the dataset. Because of this, this value is kept at 10. Table 4.1 below display the result of this filtering option in the second and fourth row of it. The fourth row demonstrates the number of unique customers who have more than ten

[1]Global Findex Database of The World Bank, Accessed October 2020 from the following URL: https://globalfindex.worldbank.org/sites/globalfindex/files/2018-05/Global%20Findex%20Database.xlsx

[2]Commited to Sustainable Leadership of Akbank, Accessed October 2020 from the following URL: https://www.akbank.com/doc/Akbank_Investor_Presentation.pdf

[3]Executional Excellence in 2016 and Beyond of Akbank, Accessed October 2020 from the following URL: https://www.akbank.com/en-us/investor-relations/Documents/InvestorPres2Q16.pdf

transactions in the sample dataset.

A further constraint in transactions of individuals is also introduced. As the previous constraint provides reliability for the total number of transactions present for each customer, these additional constraints ensure that each customer has at least has the number of transactions from multiple cities as the selected threshold number. The resulting number of unique customers for different settings of this threshold value is shared in Table 4.2.

These filtering operations are reflected in the sample dataset by utilizing selection operations of Pandas library. The first filtering mechanism discussed above is introduced to the sample by selecting customers who had at least 10 transactions by using an aggregation operation. Then, only these customers were kept in the sample dataset with the selection mechanism of Pandas. Similarly, the second filtering operation is applied to the sample dataset by checking each unique customer in the dataset to see if customers satisfy the minimum threshold for each city they performed spending.

Applied Threshold	Number of Customers
# of customers with 5 different transactions for each city	2153
# of customers with 10 different transactions for each city	649
# of customers with 20 different transactions for each city	155
# of customers with 30 different transactions for each city	53

Table 4.2 Number of Unique Customers With Minimum Transactions per City. '#' symbol is used to denote the word 'number'.

4.2.1 Scenario Definitions

After this data filtering operations, it can be observed that the number of unique customers decreases as the required number of transactions from cities where these individuals generated them. Due to the resulting number of unique individuals when the threshold number of transactions is set as 30 being a very small sample, this filtering option is not utilized further in the case study. Scenarios described and discussed in this section mainly utilized a sample of customers obtained with other threshold settings. Below these scenarios are defined.

The scenarios described below aim to reflect an actual internal migration movement in the dataset utilized for this study. To this end, each scenario adds additional constraints and requirements on the transaction patterns of individuals. Because of

this iterative constraint structure, these scenarios are named Scenario 0, Scenario 1, and Scenario 2. While Scenario 0 only requires individuals to have transactions originated from different cities at certain periods, Scenario 1 and Scenario 2 requires transactions of customers to conform to the rule sets defined for them.

Scenario 0: As the first scenario to reflect internal migration movements, Scenario 0 assumes a homogeneous flux in transactions of customers. This means that, for this scenario, only customers whose location of transactions changed completely are considered. The homogeneity assumes that transactions for an individual in the sample start and end in the origin city before starting to take place in the destination.

Scenario 1: In this scenario, the rigid assumption on the transactions of individuals is aimed to be relaxed. The reason this is considered to be necessary is due to the presumption that the individual move from the origin city to the destination city once and only one time with the setting that Scenario 0 proposes. Scenario 1 relaxes this way of flow in transactions by allowing a period where transactions of individuals originated from different cities can overlap. Here, this overlapping period is set as two months as it was considered to be enough period to relocate.

Scenario 2: To increase the reliability in the transactions of customers, a further constraint is introduced for individuals to be considered in the sample. This constraint is the requirement of customers of the bank to have a transaction history that at least spans two months in the city that the individual relocates. With this constraint, it is aimed to discard individuals who traveled to different cities for short-term purposes and vacation purposes. This constraint also mostly rules out a common type of internal movement pattern in Turkey, which is temporary relocation to summerhouses.

After these definitions, for Scenario 0, resettlement to a new city for an individual is defined as having no transactions from the origin city of individuals taking place after transactions from the second city begin to take place. This definition is updated for Scenario 1 and Scenario 2 with the given descriptions. For a better understanding, Figure 4.3 below displays these scenario settings with visual helpers.

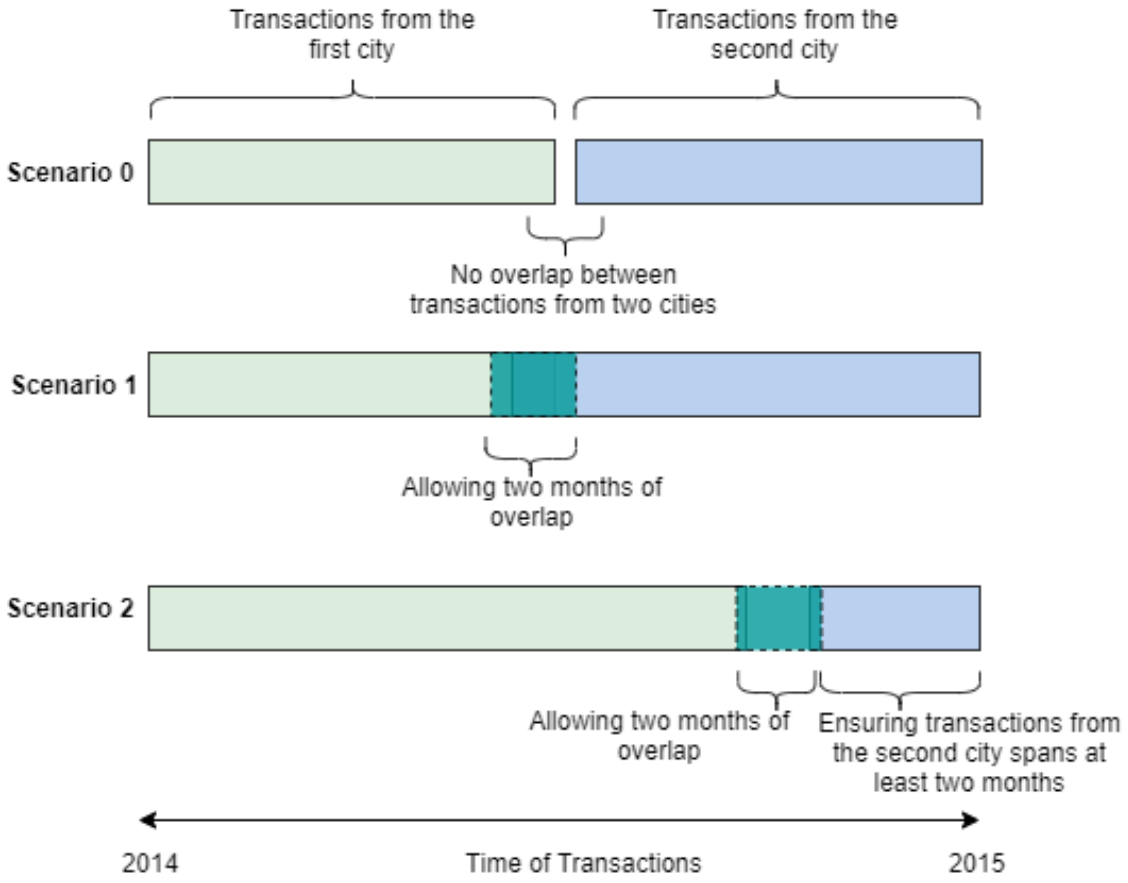

Figure 4.3 Visual expressions of scenarios defined in the section

Table 4.3 display the individuals who are considered settled to a new city in the period of the sample dataset utilized. As defined earlier, Scenario 0 denotes numbers categorized with the homogeneous flux of transaction assumptions. Scenario 1 refers to numbers categorized with 2-month tolerance for overlapping transactions. Scenario 2 refers to individuals categorized by introducing the minimum period needed for transactions from a city restriction as 2 months.

Minimum Transactions per City	Scenario 0	Scenario 1	Scenario 2
30	-	3	3
20	6	12	9
10	25	70	54
5	150	368	205

Table 4.3 Number of customers categorized as settled to a new city.

As also stated above, it can be observed that when the Minimum Transactions per City threshold is selected as 30, almost no individuals in the dataset can be considered as resettled to another city. It is because very few individuals in the sample dataset had this many transactions available to be analyzed. In addition to this observation, it can be seen that introducing a further constraint in Scenario

2 reduces the total number of individuals who are considered as settled to another city is decreased for all minimum transactions per city requirement.

As categories of these transactions are also available, these categories of customers can also be analyzed. Figure 4.4 and Figure 4.5 display aggregated transactions for spending categories in the dataset for minimum transactions requirement as 5 and 10 respectively. While Figure 4.4 display the spending categories of origin and destination cities of individuals for Scenario 0, Figure 4.5 display these categories for Scenario 2. The rankings in terms of the number of transactions performed in these categories can be seen from these figures.

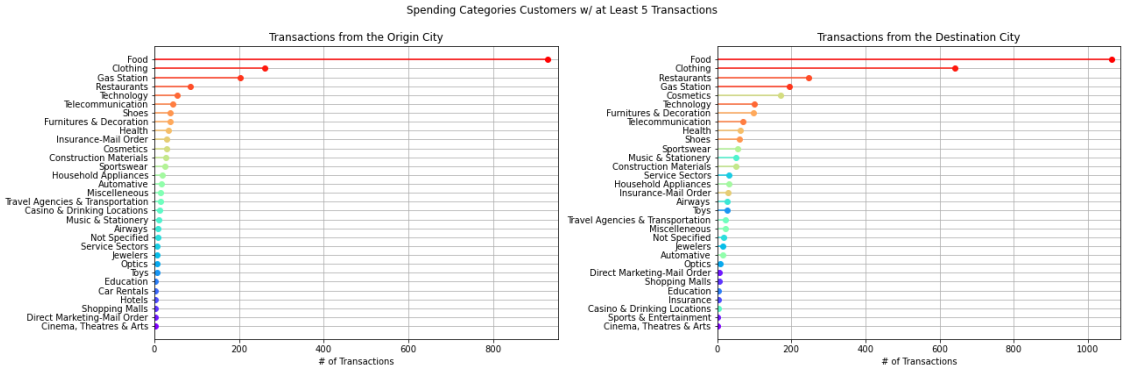

Figure 4.4 Spending categories of customers in origin and destination. Figures display the results of Scenario 0.

The most visible change in terms of rankings in Figure 4.4 is the increase of Cosmetics transactions in the destination cities. Increases in categories such as Service Sectors and Airways can be interpreted to be related to mobility and movement.

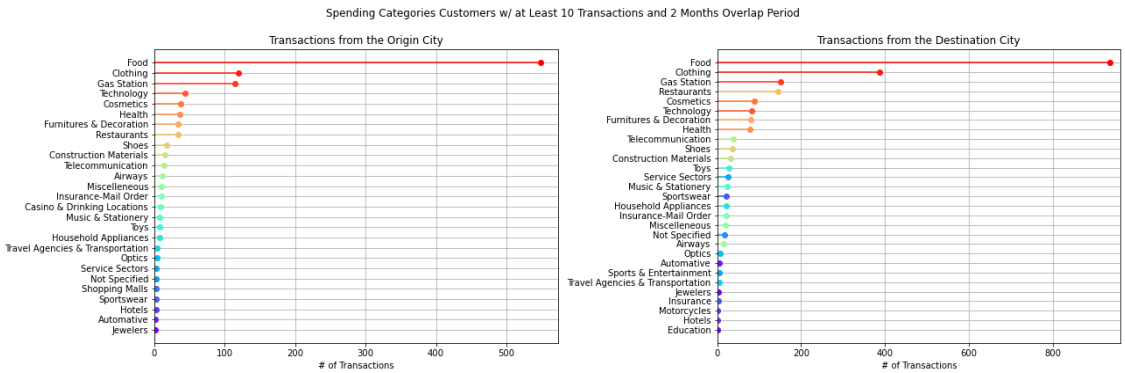

Figure 4.5 Spending categories of customers in origin and destination. Figures display the results of Scenario 2.

In Figure 4.5, the most visible change is the increase in the transactions for the Restaurants category. Similarly, for this scenario as well, the number of transac-

tions for Service Sectors increase. However, the number of transactions for Airways categories decreases, unlike the results for Scenario 0.

4.2.2 Diversity Metrics

As discussed in earlier sections and Chapter 3, having a transactional data source with many available attributes such as transaction amounts, transaction categories, and location information enables studying further metrics regarding these transactions. Since the study depicted in this chapter divides these transactions into origin and destination places, these metrics can also be studies for origin and destination cities in a comparative manner. This subsection includes the definition and discussion of three diversity metrics regarding transactions of individuals. These are named Categorical Diversity, POI Diversity, and Transaction Amount Diversity. These metrics are derived and implemented as a part of the case study.

Similar approaches to derive this metrics are also present in the study of Singh, Bozkaya & Pentland (2015). Such approach does not aggregate customers directly. Rather, spending metrics of individuals are generated prior to the aggregating. In this study, this order is reflected in these metrics as well. Aggregation is performed after finding diversity metrics of customers for origin and destination cities.

4.2.2.1 Categorical Diversity

Categorical Diversity is defined as the number of unique category occurrences over the number of transactions per customers who are considered as settled to another city. The formula for Categorical Diversity is given below;

$$D_{i_{cat}} = \frac{uniq(cat)_i}{t_i}$$

where $D_{i_{cat}}$ is the Categorical Diversity value for the customer i, $uniq(cat)_i$ is the number of unique categories for given transactions, and t_i is the number of available transactions for the given customer i. Resulting Categorical Diversity values of customers are in range (0, 1] as the number of unique categories can at most be equal to the number of transactions.

4.2.2.2 POI Diversity

POI Diversity is defined as the number of unique POI occurrences over the number of transactions per customer who is considered as settled to another city. The formula for POI Diversity is given below;

$$D_{i_{POI}} = \frac{uniq(POI)_i}{t_i}$$

where the $D_{i_{POI}}$ is the POI Diversity value for the customer i, $uniq(POI)_i$ is the number of unique POI occurrences for given transactions, and the t_i is the number of available transactions for the given customer i. Again, the resulting POI Diversity values of customers are in range (0, 1] as the number of unique POI occurrences can at most be equal to the number of transactions.

4.2.2.3 Transaction Amount Diversity

Transaction Amount Diversity is defined as the standard deviation of transaction amounts over the number of transactions per customers who are considered as settled to another city. The formula for Transaction Amount Diversity is given below;

$$D_{i_{TRX}} = \frac{\sigma_{t_i}}{t_i}$$

where the $D_{i_{TRX}}$ is the POI Diversity value for the customer i, σ_{t_i} is the standard deviation of given transaction amounts, and the t_i is the number of available transactions for the given customer i. Unlike the previous metrics, Transaction Amount Diversity is not bound in the range between 0 and 1.

4.2.3 Results of the Case Study

Collected results from different operations are shared in this subsection of the chapter.

Prior to the further findings, the frequency of origin and destination cities can

be observed from Figure 4.6. Results displayed in these figures are results of the Scenario 0 with threshold for the minimum number of transactions are set to 5.

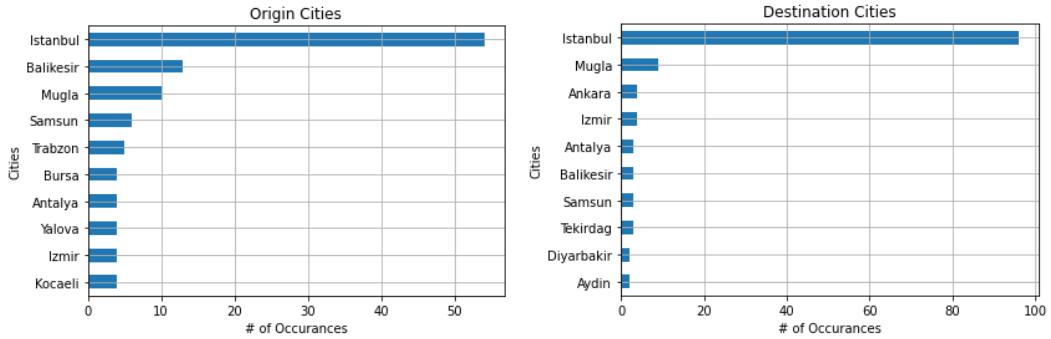

Figure 4.6 Origin and destination cities of customers. The first 10 cities for both origin and destination routes are visualized.

As previously discussed, the fact that the sample dataset utilized being Istanbul-centered is reflected in both the origin and destination cities of individuals. While origin cities with the most mobility of individuals include cities from Black Sea Region and cities close to Istanbul, destination cities mostly include cities from Egean Region.

The resulting number of individuals who are considered as settled to a new city can be observed from Table 4.3. For Scenario 0; 6, 25, and 150 individuals are considered this way for threshold values of minimum transactions per city set as 5, 10, and 20. As also present in Figure 4.6, the majority of mobility patterns of these 150 individuals revolve around Istanbul.

The number of customers considered to be settled to another city increases for all settings for the threshold value for transactions in Scenario 1. The main contributing factor of this observation is the introduction of a possible overlapping period for transactions of customers. This way, customers who could not be considered this way due to a short period of overlap are included in the group of individuals who settled in a new city too. The number of customers considered this way increased to 368 from 150 in Scenario 0. For other settings of threshold transactions per city, the number of individuals moved to a new city increased to 70 from 25 (when minimum transactions per city are set to 10) and 12 from 6 (when minimum transactions per city are set to 20).

On the other hand, the requirement of the transaction period in the destination city introduced in Scenario 2 decreases the number of individuals who are considered to be settled in a new city. Here, customers that do not have at least two months of transactions are discarded from being categorized as settled to a new city. The

numbers in Scenario 2 decrease to 9, 54, and 205 from 12, 70, and 368. These results are for when the threshold value for minimum transactions per city is set to 5, 10, and 20.

As defined earlier, additional diversity metrics of transactions of customers are analyzed for origin and destination cities as well. When the threshold for the minimum number of transactions needed from each city is set to 5, the average value of Categorical Diversity decreases from 0.402 to 0.372. When the same threshold is taken as 10, the mean value of this diversity metric decreases from 0.265 to 0.242.

Unlike the Categorical Diversity metrics, When the threshold for the minimum number of transactions needed from each city is set to 5, the average value of POI Diversity increases from 0.641 to 0.650. When the same threshold is taken as 10, the mean value of this diversity metric decreases from 0.475 to 0.514.

However, the Transaction Amount Diversity metrics decrease in the destination city as well. When the threshold for the minimum number of transactions needed from each city is set to 5, the average value of Transaction Amount Diversity decreases from 15.69 to 13.60. When the same threshold is taken as 10, the mean value of this diversity metric decreases from 11.8 to 8.32. The reason for a general decrease in the Transaction Amount Diversity decrease for both origin and destination city when the threshold minimum transactions are 10 is due to a decrease in the available transactions of customers.

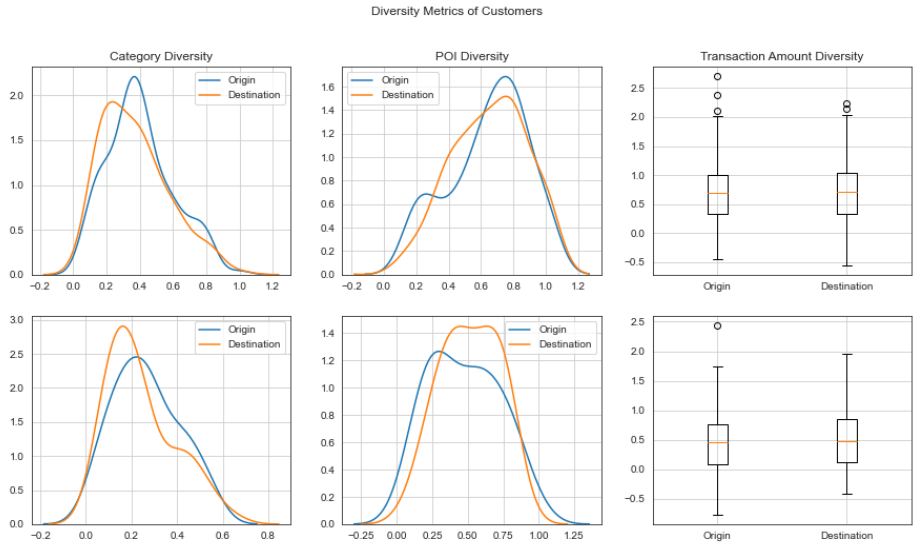

Figure 4.7 Diversity metrics of customers.

The distributions of these metrics of individuals are visualized in Figure 4.7. Figures on the upper row are obtained from customers with the threshold set as 5 trans-

actions for each city. Figures at the bottom are obtained from customers with at least 10 transactions for each city and 2 months of the overlapping period allowed. This means that the first row displays results for Scenario 0, while the figures in the latter row display results for Scenario 2. Similar trends are present for Category and POI diversity in both origin and destination cities. While the average Categorical Diversity decreases slightly after settling in a new city, POI diversity seems to be increasing slightly.

Figure 4.8 and Figure 4.9 visualizes the number of individuals considered as settled to a new city. The left choropleth for both figures display the frequency of origin cities, while the right choropleths display the frequency of destination cities. Figure 4.8 provides the findings of Scenario 0 and Figure 4.9 demonstrates the findings of Scenario 2.

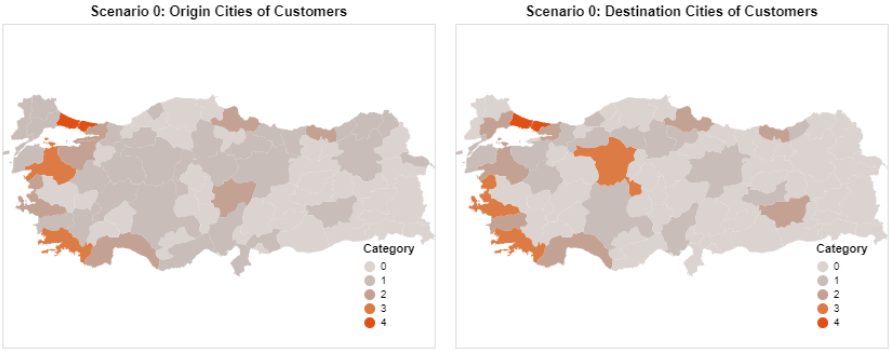

Figure 4.8 Choropleth maps of findings of Scenario 0. The figure on the left displays the frequency of origin cities, while the figure on the right displays the frequency of the destination cities.

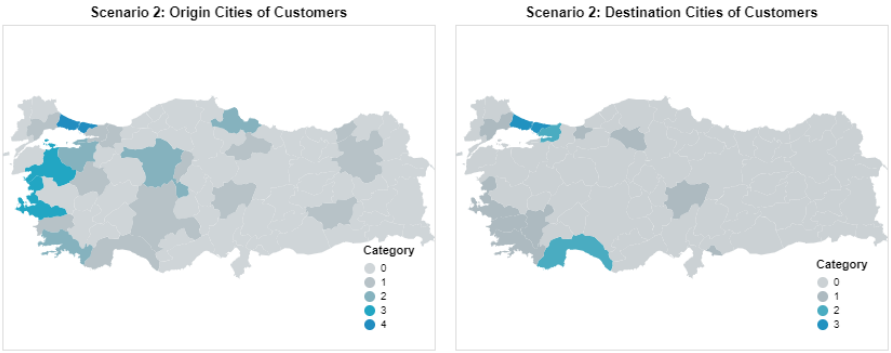

Figure 4.9 Choropleth maps of findings of Scenario 2. The figure on the left displays the frequency of origin cities, while the figure on the right displays the frequency of the destination cities.

The interpretations of the results and further discussion regarding the case study will be presented in the Discussion & Future Work chapter. The limitations of this case study will be discussed thoroughly in the mentioned chapter as well.

5. VISUAL TOOL

In this chapter, the proposed visual exploration tool is covered in detail. Section 5.1 covers the visual tasks, design rationals, and visual components the tool is composed of. These sections are then followed by Section 5.2 where use case scenarios are displayed and discussed with the developed tool.

5.1 System Design

The visual exploratory tool is developed with D3 by Bostock, Ogievetsky & Heer (2011) along with other web development technologies HTML5, CSS3, jQuery, and Javascript. D3 is also a Javascript library that is mainly utilized for interactive data visualizations. The tool adopts the unidirectional data flow principle with the library. It means that after the visual tool is launched, it is in a looping state for rendering based on user inputs.

User inputs in the tool can be in two different forms. The first form is through dropdown menu selections. Users can select the datasets and metrics to be visualized in the components. As will be discussed below, changing the dataset in Choropleth Component updates the dataset that will be utilized in Metrics Comparision and Ladder Plot components. The second form of input is through a selection with clicking. Visualized regions or countries can be clicked for selection in Chroropleth Component, Statistics Component, and Dataset Comparison Component to update the visual tool with the selected object.

The current implementation of the visual tool utilizes locally utilized datasets. For all components in the dataset, a two-layered data management approach is utilized. With this approach, the data to be visualized is first filtered and prepared by the corresponding wrapper function. Then, the data based on selections are sent to

functions that implement the rendering of visual components. This model is also displayed in Figure 5.1.

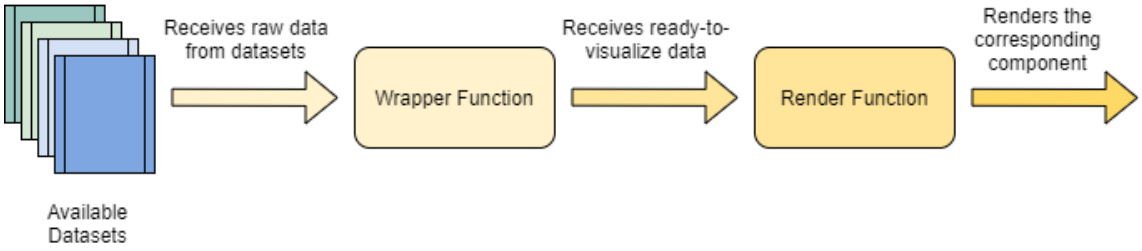

Figure 5.1 Described data model of the visual tool.

5.1.1 Visual Tasks and Design Rationale

As stressed and demonstrated at various times in this thesis, migration is a complex social phenomenon. To provide a sufficient understanding of migration flows and movements, efficient ways of utilizing data visualization methods are crucial. In addition to this, introducing more datasets to the study of migration patterns can contribute to a better understanding of the dynamics of the investigated phenomena.

Because of the points described above, a visual tool to be developed should support multiple datasets simultaneously with efficient visualization choices. Hence, the visual tasks that need to be addressed by the visual exploratory tool can be summarized as;

- **Visual Task 1:** Comparisons of different datasets for migration estimates,
- **Visual Task 2:** Comparisons of migration statistics of countries and regions,
- **Visual Task 3:** Introduction of demographic and socio-economic metrics for analysis of migration patterns.

5.1.2 Visual Components and Interactions

With the described visual tasks, design rationals, and research questions sought to be addressed by the developed tool, this subsection describes the components of the proposed system.

The user interface of the developed visual tool can be observed from Figure 5.2 below.

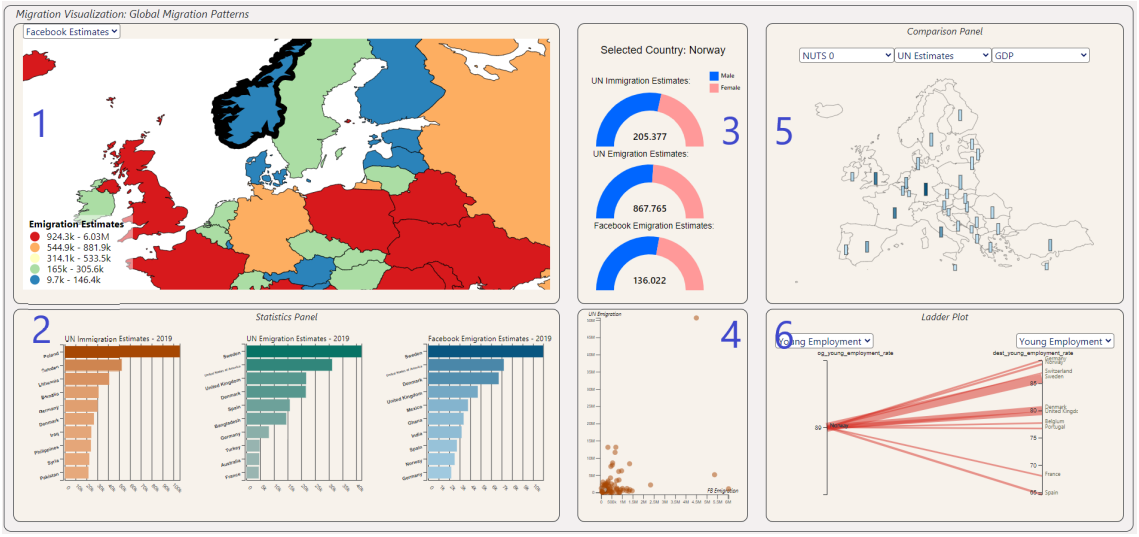

Figure 5.2 Components of the visual tool. 1. Choropleth component, 2. Statistics component, 3. Description component, 4. Dataset Comparison component, 5. Metrics Comparison Component, and 6. Parallel Plot component.

With the components presented above, the interaction schema of the visual exploration tool can be inspected from the Figure 5.3 presented below. As discussed in the System Design section, users can introduce two different input types which control the datasets or metrics to be utilized for generating visualizations and control the country or region selection.

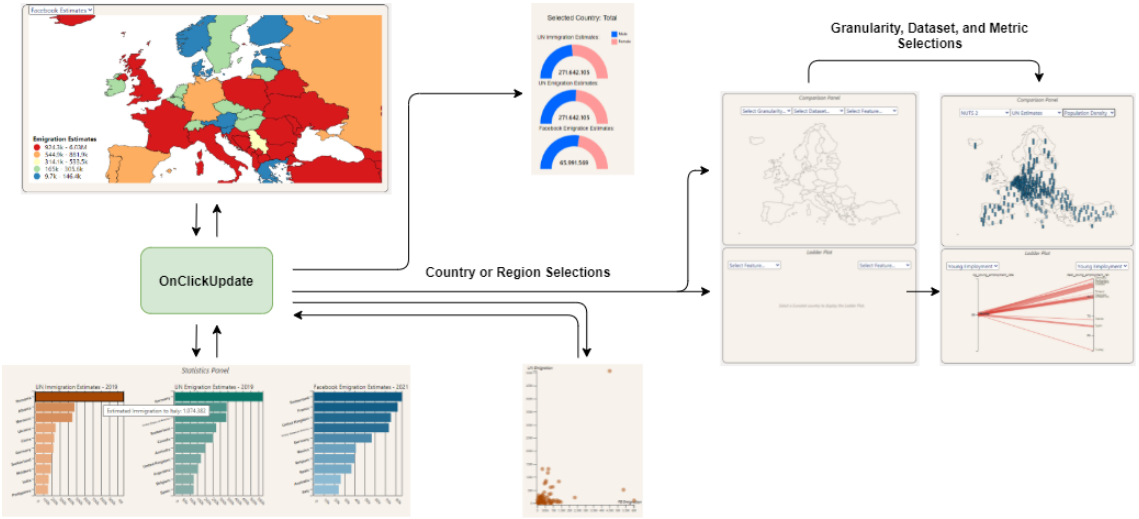

Figure 5.3 Interaction schema of the visual tool. Directed arrows denote the visual component to be updated after a selection is performed.

Below, components of the visual tool are discussed elaborately with the numerical order presented in Figure 5.2.

5.1.2.1 Choropleth Component

The Choropleth component displays the emigration and immigration estimates of countries or regions from available datasets. Categorization of estimates is obtained after selecting the quartile of estimates in the respective distribution of estimates. Countries or regions with missing immigration or emigration estimates are colored as the lowest category in choropleth visualizations.

Below, an exemplary choropleth map can be observed from Figure 5.4.

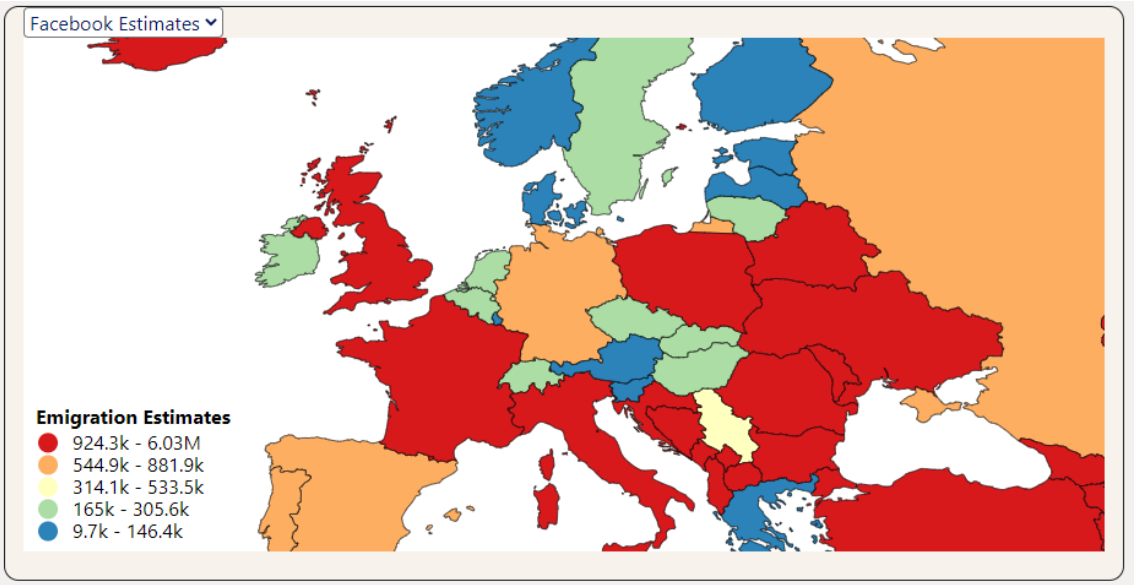

Figure 5.4 Choropleth map of Facebook emigration estimation statistics.

5.1.2.2 Statistics Component

This component aims to communicate the primary emigration and immigration trends of the selected region or country. To this end, horizontal bar charts are preferred in the component. All bar charts display the first ten routes for the defined migration directions for them. On hover, the bar chart displays the estimated statistics. An example resulting state after a selection can be seen in Figure 5.5.

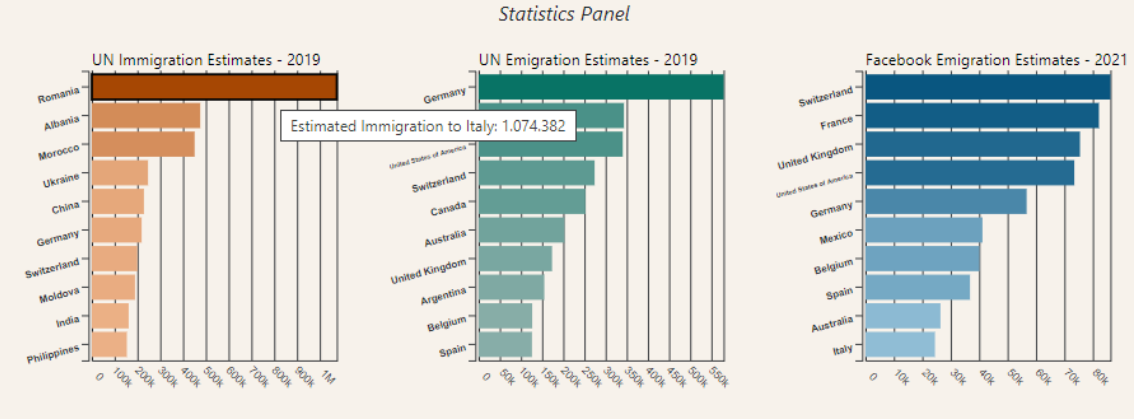

Figure 5.5 Statistics panel of the visual tool displaying emigration and immigration statistics of Italy.

5.1.2.3 Description Component

The description component reflects the selected object in the visual tool. This selected object is either a country or a region. Then, emigration originated from the selection, and immigration towards the selection statistics are visualized with half-donut charts. Statistics displayed are estimates aggregated by gender. If no country or region is selected by the user, the component display the total emigration and immigration statistics from the available datasets.

Below, an example snapshot of the component can be viewed from Figure 5.6.

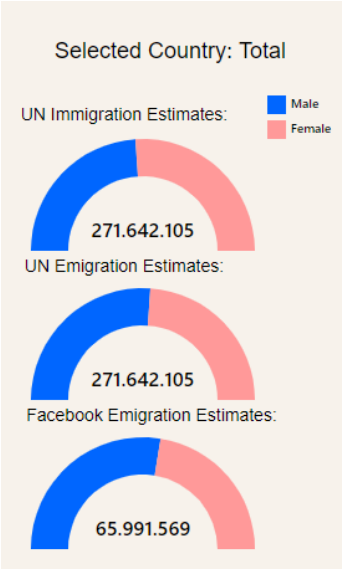

Figure 5.6 Description component of visual tool.

5.1.2.4 Dataset Comparison Component

Dataset comparison component visualizes migration estimates of countries or regions from datasets for comparison purposes. Having a medium for such comparisons enables the analysis of the relationship between utilized datasets in the visual tool.

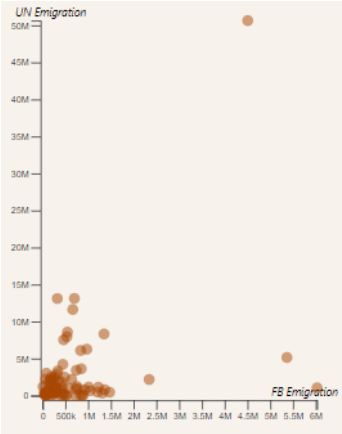

Figure 5.7 Dataset comparison component.

Figure 5.7 above is a sample of the component where United Nations and Facebook emigration estimates are being compared.

5.1.2.5 Metrics Comparison Component

The Metrics Comparison component is the primary component designed to integrate additional demographic and socio-economic metrics into the analysis of migration movements. Through the component, emigration statistics at lower spatial resolutions can be distributed to regions with higher spatial resolution. To this end, the visual component utilizes previously discussed NUTS regions. After the necessary country (or region and cities can also be selected with different use cases), dataset, granularity, and metric is selected, component distributes the emigration originated from the selected country or region into other regions.

The distribution of estimated emigration statistics into higher spatial resolutions are calculated with the following formula.

$$emig_{r_i} = emig_{C_i} * \frac{f_{i_k}}{\sum_{j=1}^{n} f_{j_k}}$$

where r_i is the region i in the country C and f_{i_k} is the value of the feature k for the region. Then emigration estimate from the selected country to the region i, which is $emig_{r_i}$, is the multiplication of the ratio of the feature of the region to the sum of the country and total emigration from selected country to country i.

As the current implementation of the visual tool utilizes metrics obtained from Eurostat, this component is only available for European countries when the analysis aimed to be conducted is at the country scale. However, with small adjustments to the implementation, the component can be adapted to studies with different spatial resolutions. The use case depicted with transactional data from Turkey utilizes a modified version of this component. In that version, the component is employed to aggregate city-to-city level emigration estimates into emigration estimates at NUTS 2 regions. The modified version of the visual component also uses circles for visual cues instead of rectangular displays.

Figure 5.8 demonstrates the component at its initial stage and after being activated by user inputs.

5.1.2.6 Parallel Plot Component

With the parallel plot component, emigration originated from the selected country or region can be analyzed with socio-economic and demographic indicators available in the visual tool. These indicators form the left and the right axes for a parallel plot visualization. Axes can be changed to analyze the available indicators of countries and higher resolution regions obtained from Eurostat.

Emigration estimations from the selected country or its regions are encoded as the stroke width of the lines in the parallel plot components. This way, the placement of countries of regions on axes ranks them to their corresponding metric, while the magnitude of the migration movement is visualized with the area. This way component compares metrics of origin and destination locations similar to ladder plots. The component can also be considered as a ladder plot with only two axes. Because of this resemblance, the component is named as Ladder Plot in the visual tool.

Figure 5.8 below showcases the Parallel Plot component and Metrics Comparison component at the initial stage of the visual tool and after the user introduces inputs.

After launching, at the initial stage of the visual tool, these components are inactive.

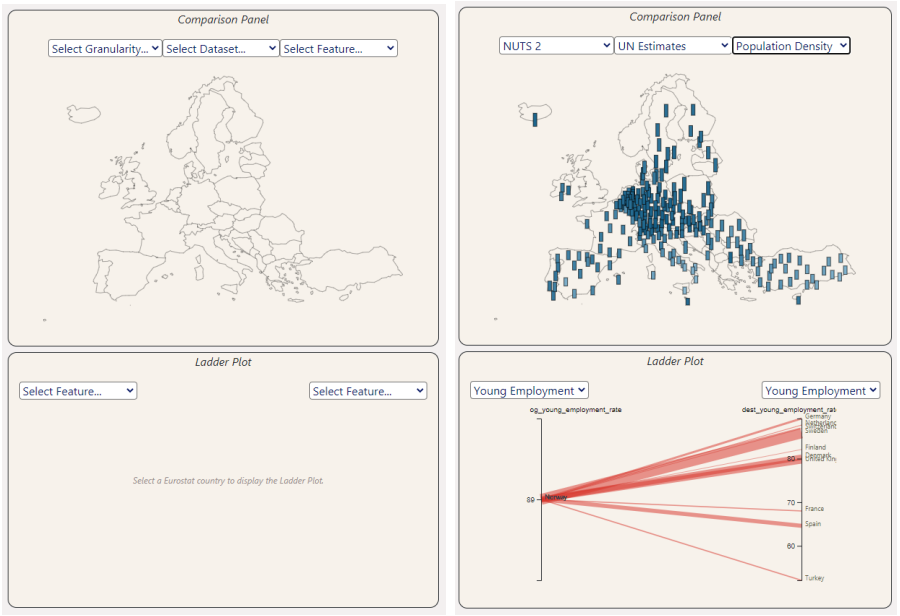

Figure 5.8 Initial (on the left) and active stages (on the right) of Parallel Plot and Metrics Comparison components.

The Parallel Plot component gets activated whenever the user makes a selection. The Metrics Comparison component gets activated when the user makes a region or country selection and selects spatial resolution, dataset, and metric inputs from dropdown menus.

After their descriptions, contributions of visual components to visual tasks defined earlier can be summarized as follows;

- Choropleth component, Description component, and Dataset Comparison component contribute for addressing the Visual Task 1 as they create mediums where multiple datasets are visualized together.
- Statistics component, Metric Comparison component can answer the Visual Task 2 as they allow users to compare statistics of different countries or regions.
- Metric Comparison component and Parallel Plot component serve as visual components that address the Visual Task 3. They introduce regional metrics obtained from Eurostat into the analysis of migration patterns.

5.2 Use Cases

This section covers two use cases achieved with the proposed visual exploratory tool. The first use case aims to compare United Nations migrant stock estimations with the emigration estimates from Facebook discussed earlier. This way, the reliability of the data obtained from the Facebook Marketing API can also be validated. The latter use case visualizes the findings of Chapter 4 to generate a better understanding of internal migration routes in Turkey.

5.2.1 Migration Patterns in United Nations and Facebook Marketing API Estimations

Previous sections covered the Facebook Marketing API as a potential data source for emigration estimates. Additional traditional data sources, which are United Nations migrant stock datasets, and regional metrics from Eurostat are also discussed in a detailed manner in the Utilized Datasets chapter. As available discussion for these data sources also suggests, these datasets can be utilized together for a better understanding of migration movements. Their analysis can also generate insights on factors that lead to these movement patterns. Additionally, Facebook estimates can be validated by comparing the statistics from innovative and traditional data sources.

To this end, a use case involving these datasets is created. The initial stage of the visual tool is already presented in Figure 5.2 with the same data sources. This use case displays country selection and updating Metrics Comparison and Parallel Plot components.

Snapshot on the top in Figure 5.9, display the stage of the visual tool after a country and metric parameters for Parallel Plot component selected. The user-selected country Norway in this case and through the Parallel Plot component analyzed the emigration estimates by also comparing young employment percentages of Eurostat countries.

The snapshot on the bottom displays the usage of the Metrics Comparison component. Again for Norway, the user distributed the emigration statistics to other European regions with higher granularity by using the component. In this case, the

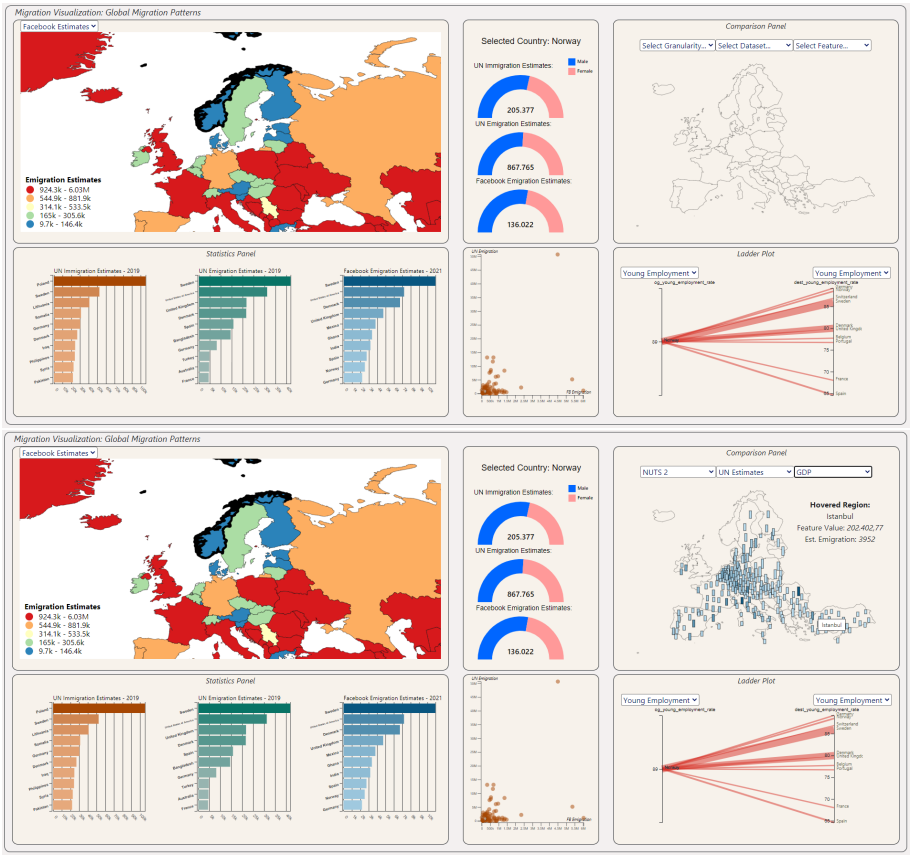

Figure 5.9 Case study with global emigration estimates with selections by the user.

user distributed estimations from United Nations data by using the GDP feature.

5.2.2 Internal Migration Patterns of Turkey

As depicted in Chapter 4, Case Study with Transactional Data aims to discover possible internal migration movements in Turkey. The findings of this case study create an opportunity to be compared with the official statistics published by TURKSTAT. The difference between scales of estimated migration raises doubts about the applicability of comparisons. However, migration trends present in both datasets can be compared. The developed tool can be employed to visualize these movement trends.

To reflect the findings in the visual tool, only one of the components is slightly changed. That is the way of representing the estimated amount of migration in the Metrics Comparison Component. Instead of a representation with rectangular visual cues, circles are utilized as the space available is sufficient to support the rendering of larger objects. The figures shared below demonstrate the use case.

Figure 5.10 demonstrates the initial stage of the visual tool after launching. As in the first use case, the Metrics Comparison Component and Parallel Plot Component await a selection of a city to be activated.

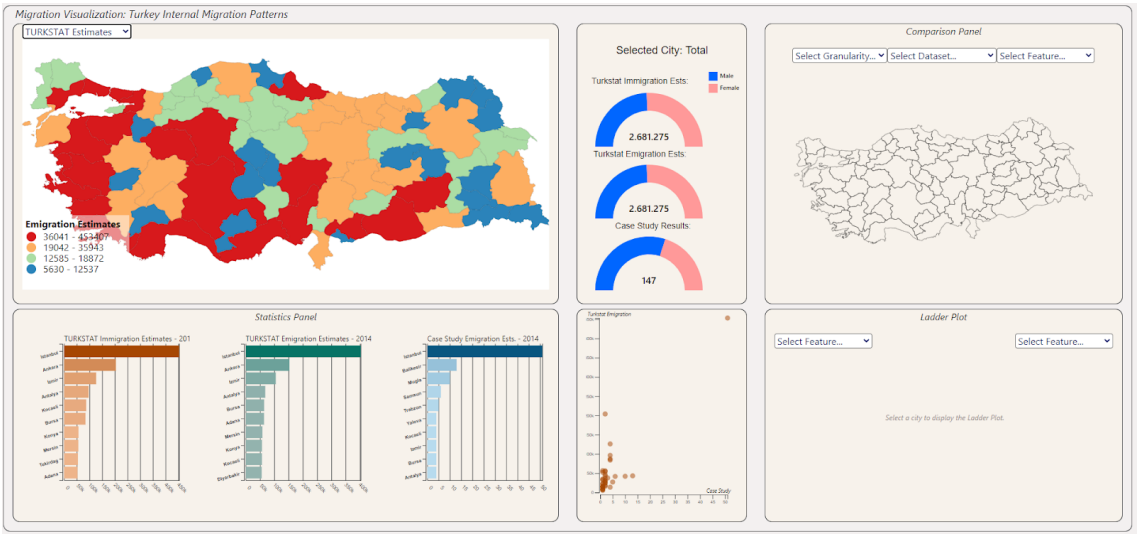

Figure 5.10 A screenshot of the initial stage of the visual tool with the datasets utilized for the transactional case study.

Figure 5.11 display the visual tool after a city is selected. As can be observed from the figure, the Metrics Comparison Component display emigration statistics of regions with circle objects.

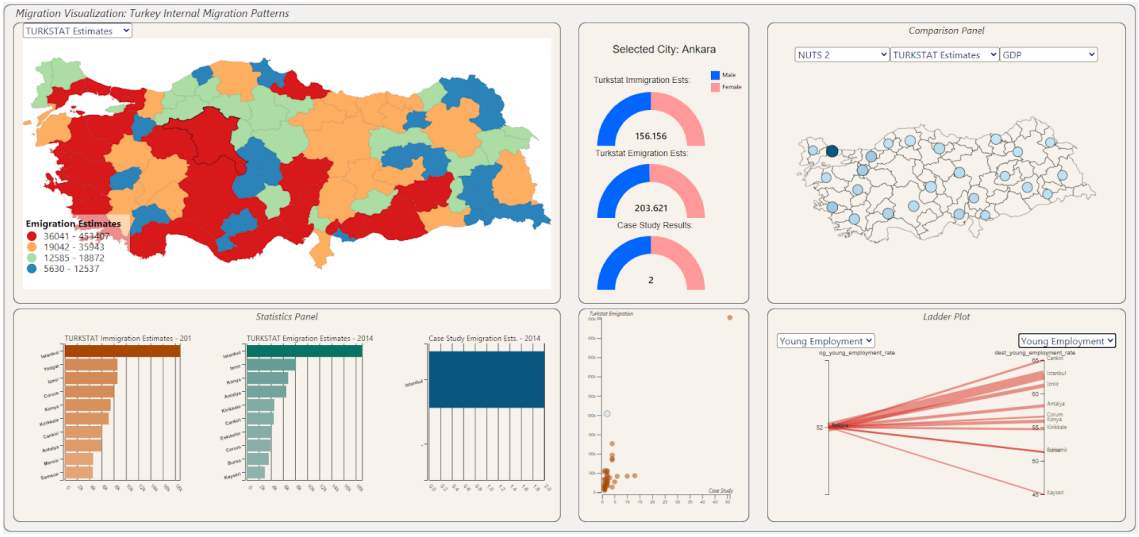

Figure 5.11 A screenshot of the case study after a city is selected by the user.

Similar to Figure 5.11, Figure 5.12 display the use case with a city selection. Here, the choropleth map is updated with the results from Scenario 0 of the case study.

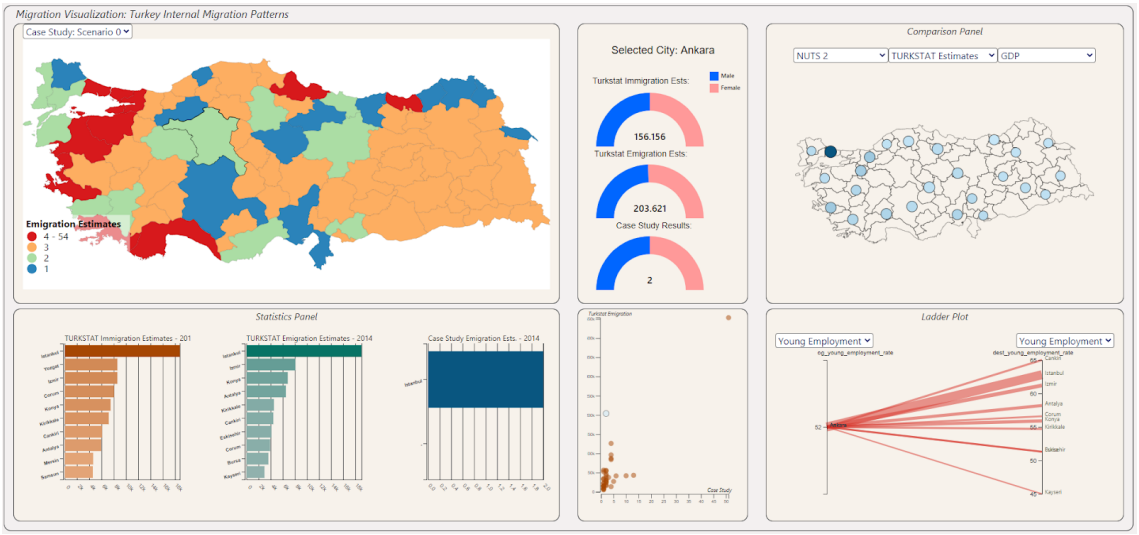

Figure 5.12 A screenshot of the case study after visualized choropleth map is changed by the user.

6. DISCUSSION & FUTURE WORK

In this chapter, the findings from both studies described earlier are thoroughly discussed. After these discussions, the limitations of innovative datasets utilized for each study are clearly stated. Section 6.1 will discuss the implications of the findings of the case study with transactional data. It will be followed by the insights from the emigration estimates of Facebook data in Section 6.2. Section 6.3 will cover the limitations of these datasets. Lastly, in Section 6.4, Future Work for both studies elaborated in Chapter 4 and Chapter 5 will be presented.

6.1 Findings of the Case Study with Transactional Data

As discussed in Chapter 4, the case study with transactional data aims to capture internal migration patterns of individuals who settled in a different city in the period of their transactions in the sample dataset. To this end, scenarios defined aimed to follow a trend that gradually increases the restrictiveness. With these gradually increasing constraints, it is aimed to filter out movements that are not migration movements but visiting a different city for work or holiday. These constraints are relaxed while switching to the definition for Scenario 1 from Scenario 0. However, this change is performed with a justification mentioning that migration movements may not take place with only one movement of individuals. This relaxation also increased in individuals who are considered as settled in a new city in Scenario 1 when compared to Scenario 0. As can be observed from Table 4.3, the number of individuals considered internal migrants increased to 368 from 150 in Scenario 0, and the minimum number of transactions threshold set to 5.

With the introduction of Scenario 2, the relaxation of the constraints on transactions of customers is decreased. It is in this scenario where the distinguishing between movements which are between cities but not migration and internal migration move-

ments are attempted. The necessity of individuals to have at least two months of expenditures from the destination location. Thus, for the utilized dataset, the scenario that captures the internal migration movements of individuals with the most accurate way can be argued to be Scenario 2.

This period could also be increased further. However, after testing with various parameters for this period, it was observed that longer periods filtered out many individuals from the scenario definition. Hence, it can be argued that the value for the necessary period in the destination city is selected after an empirical investigation. Although this period can be considered to be short, results indicate its effectiveness. As Table 4.3 shows, the number of individuals categorized as internal migrants reduced to 205 from 386 with the minimum number of transactions threshold set to 5. The estimated value decreased to 54 from 70 when the minimum number of transactions is 10.

For the period of transactions in the sample dataset, the total internal migration estimate reported by TURKSTAT indicate the estimated statistics as 2,720,438 in 2014 and as 2,619,403 in 2015[1]. When compared to the results of the case study, it can be seen that these estimations do not seem to be directly comparable in terms of scales. It should be reminded that the biggest estimate that is indicated from scenarios in the case study report 368 individuals as internal migrants. However, comparisons in terms of movement trends can still be performed. Table 6.1 demonstrate the origin and destination city pairs which are the most common in both TURKSTAT estimates and case study results.

Findings from Scenarios and Estimated Number of Migration							
Scenario 0 (From the Total of 150 City Pairs)		Scenario 2 (From the Total of 54 Ctiy Pairs)		2014 Estimates		2015 Estimates	
City Pair	Est. Value	City Pair	Est. Value	City Pair	Est. Value	City Pair	Est. Value
Bahkesir to Istanbul	13	Izmir to Istanbul	6	Kocaeli to Istanbul	28,272	Kocaeli to Istanbul	29,475
Muğla to Istanbul	10	Bahkesir to Istanbul	6	Tekirdağ to Istanbul	23,170	Tekirdağ to Istanbul	25,422
Istanbul to Muğla	9	Yalova to Istanbul	3	Istanbul to Tokat	19,388	Ordu to Istanbul	21,420
Samsun to Istanbul	6	Bursa to Istanbul	3	Istanbul to Ankara	19,021	Ankara to Istanbul	18,907
Trabzon to Istanbul	5	Istanbul to Antalya	3	Ankara to Istanbul	18,775	Istanbul to Ankara	18,066
Antalya to Istanbul	4	Istanbul to Kocaeli	3	Izmir to Istanbul	16,129	Giresun to Istanbul	17,935
Yalova to Istanbul	4	Ankara to Istanbul	3	Istanbul to Izmir	1,5559	Izmir to Istanbul	17,124
Istanbul to Izmir	4	Samsun to Istanbul	3	Tokat to Istanbul	15,395	Tokat to Istanbul	17,035
Bursa to Istanbul	4	Muğla to Istanbul	2	Istanbul to Kocaeli	14,952	Bursa to Istanbul	14,215
Izmir to Istanbul	4	Kocaeli to Istanbul	1	Istanbul to Van	13,155	Istanbul to Kocaeli	13,939
Istanbul to Ankara	4	Kırıkkale to Istanbul	1	Ordu to Istanbul	12,934	Istanbul to Tokat	13,844
Kocaeli to Istanbul	4	Antalya to Istanbul	1	Bursa to Istanbul	12,901	Istanbul to Izmir	13,237
Kayseri to Istanbul	3	Aydın to Istanbul	1	Istanbul to Balıkesir	12,494	Istanbul to Van	12,908
Istanbul to Tekirdağ	3	Tekirdağ to Istanbul	1	Istanbul to Ordu	12,437	Sakarya to Istanbul	12,047
Istanbul to Samsun	3	Konya to Istanbul	1	Sakarya to Istanbul	11,888	Samsun to Istanbul	10,924
Istanbul to Antalya	3	Diyarbakır to Istanbul	1	Istanbul to Giresun	11,814	Antalya to Istanbul	10,141
Kocaeli to Balıkesir	3	Erzurum to Istanbul	1	Istanbul to Bursa	11,132	Istanbul to Bursa	10,128
Gaziantep to Istanbul	2	Sakarya to Istanbul	1	Istanbul to Tekirdağ	11,101	Istanbul to Tekirdağ	10,104
Çanakkale to Istanbul	2	Istanbul to Istanbul	1	Antalya to Istanbul	10,887	Manisa to Izmir	9,479
Istanbul to Kocaeli	2	Istanbul to Çankırı	1	Istanbul to Sivas	10,403	Balıkesir to Istanbul	9,479

Table 6.1 Comparison between official statistics of migration estimates of Turkey published by Turkey Statistical Institute and findings of scenarios discussed in this section. For all of the 4 statistics, maximum 20 values are displayed.

These results can be compared against the whole population as well. In 2014 and

[1]İstatistik Veri Portalı of TUIK, (2021) Accessed June 2021 from the following URL: https://data.tuik.gov.tr/Kategori/GetKategoripÑufusveDemografi109

2015, the ratio of internal migrants to the population of Turkey are 0.034 and 0.032 respectively. When the resulting number of individuals considered internal migrants from Scenario 0 and Scenario 2 compared against the total number of unique customers in the sample dataset, these ratios are 0.0035 and 0.0048 respectively (when the minimum number of transactions expected from customers is 5). Hence, as depicted in Table 6.1, some of the migration trends are captured by the case study. The scale of migration movements however is not represented accurately.

Converting estimates in the form of origin-destination pairs also allows the analysis of the correlation between the findings of the case study and official statistics. Here, to investigate the relationship, Pearson correlation is used between case study estimates and internal migration estimates from TURKSTAT. Results of these correlation analyses can be observed from Table 6.2

Compared Set 1	Compared Set 2	Correlation Coefficient
Scenario 0	2014 Estimations	0.266
Scenario 0	2015 Estimations	0.225
Scenario 2	2014 Estimations	0.223
Scenario 2	2015 Estimations	0.246

Table 6.2 Correlation coefficients between Case Study findings and estimates published by Turkey Statistical Institute

It can be seen that for each comparison a week positive correlation is obtained. Hence, it can be claimed that there is some parallel between our findings and migration estimates from official sources.

Implications of generated diversity metric statistics are investigated as well. To conduct such a study, distributions of diversity metrics from origin and destination locations, which are only compared by visualizations previously, are compared with statistical tests. To this end, a t-test with independent sample assumptions is employed to test whether there are significant behavioral differences of customers in origin and destination locations. The resulting p values from test cases of Scenario 0 can be observed from 6.3.

Compared Distribution 1	Compared Distribution 2	P Value
Origin Categorical Diversity	Destination Categorical Diversity	0.194
Origin POI Diversity	Destination POI Diversity	0.754
Origin Transaction Amount Diversity	Destination Transaction Diversity	0.645

Table 6.3 T-tests results for comparing distributions of diversity metrics in origin and destination locations in Scenario 0. These results are obtained with minimum threshold is set as 5 transactions.

The same study with exact settings is also applied to the resulting metric distributions of Scenario 0. The resulting p values can be seen in Table 6.4.

Compared Distribution 1	Compared Distribution 2	P Value
Origin Categorical Diversity	Destination Categorical Diversity	0.403
Origin POI Diversity	Destination POI Diversity	0.380
Origin Transaction Amount Diversity	Destination Transaction Diversity	0.525

Table 6.4 T-tests results for comparing distributions of diversity metrics in origin and destination locations in Scenario 2. These results are obtained with minimum threshold is set as 10 transactions.

Although mean values of distributions for these metrics change between origin and destination cities, results indicate that these differences are not statistically significant.

Hence, to summarize the results of the case study, it can be stated that the primary origin and destination pairs for migration movements are reflected. However, findings in terms of the scale of the migration and behavioral changes of customers after migration movements cannot be argued confidently to be significant. These results need more support to be considered to be statistically significant.

6.2 Insights from Facebook Marketing API Estimations

The visual exploratory tool discussed in Chapter 5 serves as an environment where the emigration estimates from Facebook can be compared with the statistical data from United Nations estimations. Based on the Description component of the visual

tool in the first use case, the total estimate for emigration is stated as approximately 66 million from the Facebook data. The corresponding estimates from United Nations denote approximately 272 million. However, it should be reminded that the Facebook emigration estimates only cover 90 countries[2]. When these estimates compared to their matching counterparts in United Nations data, the United Nations emigration estimate for these countries is 227,579,942. It should be noted that this estimate is for 2019, whereas the Facebook estimates are obtained during May 2021.

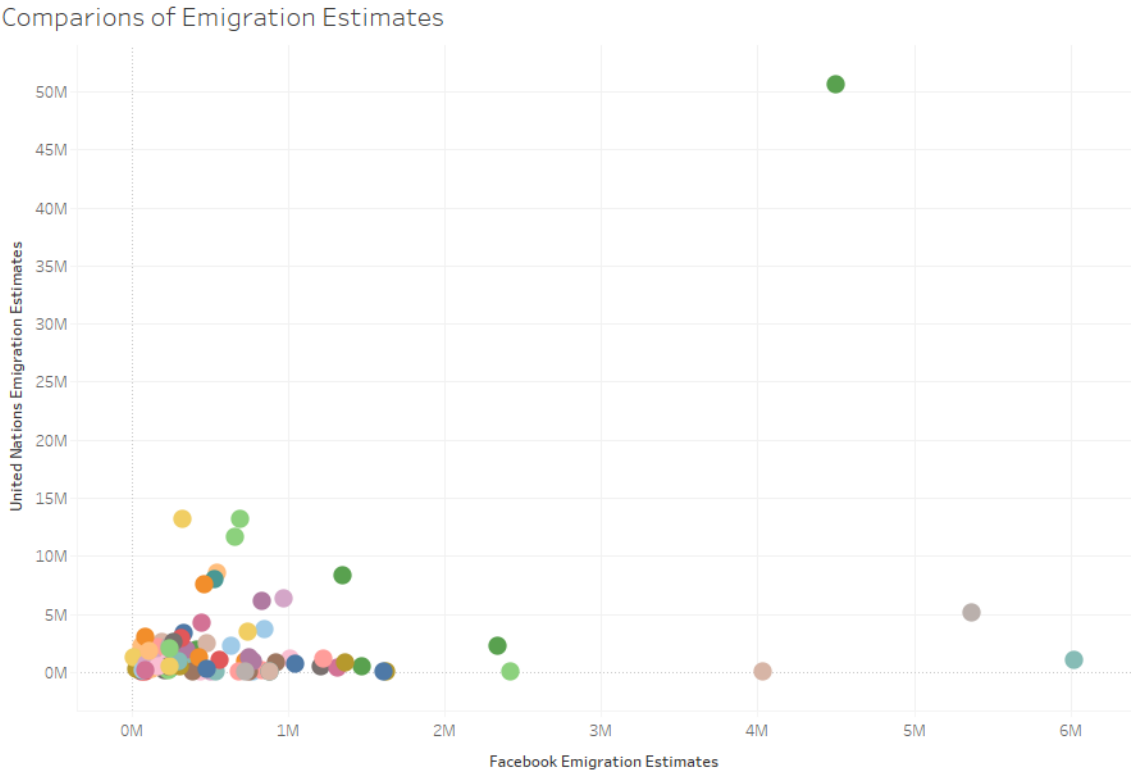

Figure 6.1 Comparison of emigration statistics of United Nations and Facebook.

When the values in the scatter plot are analyzed, for 29 countries Facebook emigration estimates overestimate the emigration statistic shared by United Nations. For the remaining 61 countries, the United Nations emigration estimate is higher than the Facebook estimates. The countries with a higher estimate of Facebook data are shared in Appendix A.

Because the data gathered from Facebook is also in origin-destination form, analyzing the transposed version of the dataset at hand can produce immigration emigration of countries. Although behavioral filters utilized for data collection are only available for 90 countries, API returns estimates for all countries as queries are also

[2]These 90 countries and their emigration estimates from Facebook Marketing API are listed in the Appendix A.

set this way. It means that immigration estimates from all countries in the world to 90 countries available for filtering can be achieved. It can be advantageous as the rest of the missing values can be attempted to be filled with various approaches such as collaborative filtering. Below, countries with the most immigration estimates in Facebook data are visualized in Figure 6.2.

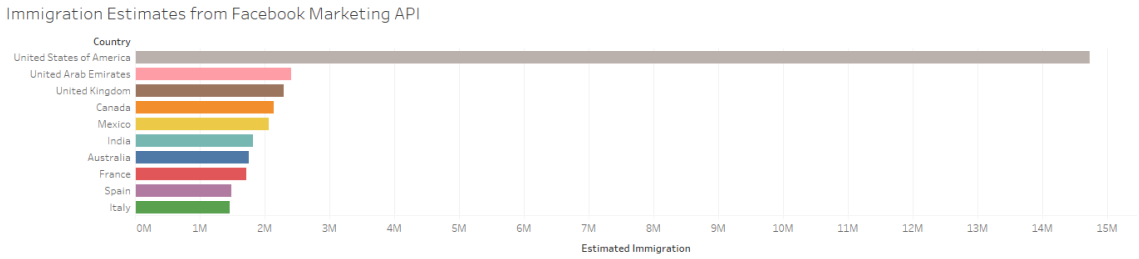

Figure 6.2 Converted immigration estimates of countries from Facebook Marketing Platform data.

Data obtained from Facebook have instances where the origin and destination countries are the same. When interpreted with the filters applied, having values other than zero for these instances state the *People used to live in country X, and live in country X*. Because of this, these statistics may imply that individuals who are filtered with it can be considered as previous migrants that returned to their country of origin. Below, in Figure 6.3, countries with the highest statistics are shared.

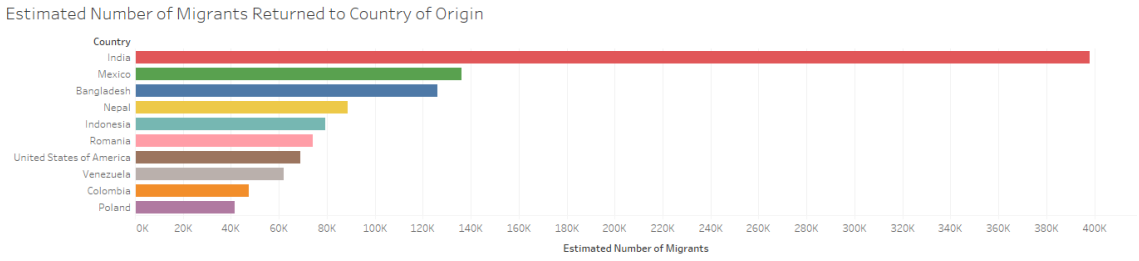

Figure 6.3 Countries with return migration estimates from Facebook Marketing Platform.

Significant observations can be gathered when emigration estimates from Facebook data are analyzed together with the population sizes of countries. Although some European countries such as Portugal, Romania, and Lithuania are not at higher rankings for total emigration, their Facebook emigration estimations to population ratios are some of the highest in the whole world. When this observation is investigated in social science literature, one can come across studies that indicate emigration as a common phenomenon in these countries. Justino (2016) categorizes Portugal as the country with the highest emigration rate, OECD (2019) states Ro-

manian diaspora to be the fifth largest migrant community in the world, the lack of workforce and brain drain in Lithuania claimed to be present[3].

6.3 Limitations of Utilized Datasets

This section discusses the limitations of the innovative data sources utilized in this thesis. Although their advantages and ability to support traditional data sources, these datasets also introduce some challenges to the analysis of movement or migration patterns studies. The section first lists the limitations of the transactional data used in this thesis. This discussion is then followed by the shortcomings of the Facebook Marketing API estimations.

6.3.1 Limitations of Transactional Data

It can be said that the primary limiting factors of the transactional data utilized stem from the size of the sample used and the lack of validation methods.

The original form of the utilized dataset had approximately 9,3 million transactions and 100,000 unique customers. However, filtering out transactions without geolocations and filtering out customers whose number of transactions was not adequate for a reliable study, these numbers decreased drastically. After filtering operations, scenarios described are performed approximately with 2,5 million transactions of 42,139 customers. The number of customers who had intercity movement patterns with enough representation with expenditure instances was much less as shown in Table 4.2 in Chapter 4.

Working with a small sample of customers from only one financial institution may lead to difficulties in obtaining the same results when the scenarios depicted are replicated with different datasets. The sample size and using data from only one institution may have also introduced two types of selection bias to the case study.

The first type of selection bias may be present because the sample of individuals

[3]Emigration-immigration statistics of Renkuosi Lietuva, (2020) Accessed at May 2021 from the following URL: https://www.renkuosilietuva.lt/en/emigration-immigration-statistics/

whose transactions are analyzed not representing the customer base of the financial institution. This bias can also be present because the sample utilized was also Istanbul-centric meaning that most of the customers were assigned to branches of the bank in Istanbul and the most of transactions are from Istanbul.

The second form of selection bias may be because the customer base of the financial institution not representing the demographic and socio-economic realities of Turkey. Due to this, replicating the study may not result in similar findings. Compared to the first form of bias, eliminating this bias argument is easier. It is because the private bank shared the dataset is stated not to have a specific occupational focus for its customers compared to some of the other banks in Turkey.

Apart from these possible bias discussed, the biggest challenge for such a study seem to be obtaining the transactional datasets in similar forms. Due to personal data and privacy regulations, financial institution may not be willing such datasets. Thus, replicating this study with other data from other organizations may not be possible.

6.3.2 Limitations of Facebook Marketing API Estimations

Besides the advantages and opportunities that Facebook data provide, there are also some negative aspects and shortcomings of the data obtained. These shortcomings are listed in this subsection.

The first form of shortcomings originates from the reliability problems with the Facebook estimations. The emigration estimates originally are obtained as daily and monthly estimations. The statistics utilized in this thesis uses daily estimations. The first reason that monthly estimates are not utilized is that the differences between daily and monthly estimates were too high. While the total emigration estimates for daily estimations are 66 million, monthly estimates state this as 126 million. The second reason those monthly estimations are not used is that most of the values for monthly estimates are scaled to higher values and rounded. Due to this, there are many repeating estimates for different origin-destination pairs.

Similar to the discussion of shortcomings of transactional data, the presence of selection bias is also stated to exist in the study of Acosta et al. (2020). As in the case of customers of a singular financial organization, users of Facebook are stated to be not representing underlying communities and countries.

How people are considered to be part of the results after the described behavioral filters utilized are also not documented adequately. The study of Zagheni et al. (2018) and Herdağdelen et al. (2016) state user reported city locations and countries that individuals list as they lived in their profiles contribute to the creation of these behavior filters. One of the implications of this uncertainty leads to the difficulty of matching appropriate migration types with the Facebook estimations.

Lastly, it is also challenging to conceptualize the emigration statistics obtained from Facebook data with a single definition of migration. As previously mentioned works suggest, individuals seem to be categorized with the behavior filters based on their past and present locations. However, this information alone does not hint too much about the migration movement these individuals performed. To elaborate more, utilized filters and estimated emigration statistics obtained do not imply that the underlying movements were regular or not. It creates a limitation as a case of uncertainty for this study. It is because other utilized datasets were statistics resulted from regular migration movement of individuals.

6.4 Future Work

More scenarios can be defined to resemble the internal migration scenario with transactional data even better. Also, thresholds set can be increased for a more reliable study if a transactional dataset of expenditures of customers from another organization can be employed. If possible, replicating the scenarios utilized in the case study could also validate or eliminate bias arguments with the other datasets.

For the case study with transactional data, more metrics from transactions of customers can be generated. Current diversity metrics utilize category, amount, and POI features of the sample dataset. There are more demographic features about the customers that can be utilized for generating additional metrics. The geographical extent of the spending behavior of customers can also be studied to introduce as a diversity metric. Additionally, generated metrics can be analyzed after aggregating for cities as well.

Having a dataset with high spatial resolution and category information for transactions can also allow further studies that aim to analyze categories of expenditures of individuals after their movement. Such a study can be informative about how individuals tend to prioritize their spending after movement patterns. Additionally,

insights can be produced for a better understanding of what towns, cities, or regions can offer to people that settle there. With such studies, the socio-economic well beings of individuals can be analyzed in their origin and destination locations.

The sub-national data collection from Facebook Marketing API is still an ongoing process for this thesis. After a significant amount of data is obtained, these estimations could also be introduced to the visual tool. This way, Facebook estimates at higher granularity can be compared with the estimates distributed to higher granularity in the visual tool.

Recent works utilize other forms of Facebook services for studies in social science domains. The study by Kuchler, Russel & Stroebel (2020) introduces the Facebook Connectedness Index and utilizes the dataset for demonstrating its relationship with COVID-19 spread maps. The same indicators are also employed for displaying its relation with the international trade flows in the study of Bailey, Gupta, Hillenbrand, Kuchler, Richmond & Stroebel (2021). As the indicators are available in various spatial resolutions, they can be integrated with the visual tool. Lastly, the utilized Facebook emigration estimates and Facebook Connectedness Index datasets can be used for predictive models to infer migration flows.

BIBLIOGRAPHY

Abel, G. J. & Cohen, J. E. (2019). Bilateral international migration flow estimates for 200 countries. *Scientific data*, *6*(1), 1–13.

Acosta, R. J., Kishore, N., Irizarry, R. A., & Buckee, C. (2020). Quantifying the dynamics of migration after a disaster: Impact of hurricane maria in puerto rico. *medRxiv*.

Alexander, M., Polimis, K., & Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the united states. *Population Research and Policy Review*.

Andrienko, G. & Andrienko, N. (2008). Spatio-temporal aggregation for visual analysis of movements. (pp. 51 – 58).

Andrienko, N. & Andrienko, G. (2013). Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, *12*(1), 3–24.

Araujo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. WebSci '17, New York, NY, USA. ACM.

Armstrong, C., Poorthuis, A., Zook, M., Ruths, D., & Soehl, T. (2021). Challenges when identifying migration from geo-located twitter data. *EPJ Data Science*, *10*.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R., & Stroebel, J. (2021). International trade and social connectedness. *Journal of International Economics*, *129*, 103418.

Blumenstock, J., Chi, G., & Tan, X. (2019). Migration and the Value of Social Networks. (13611).

Borjas, G. J. (1989). Economic theory and international migration. *International migration review*, *23*(3), 457–485.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309.

Buchin, K., Speckmann, B., & Verbeek, K. (2011). Flow map layout via spiral trees. *IEEE transactions on visualization and computer graphics*, *17*(12), 2536–2544.

Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, *142*, 102347. Special Issue on papers from "10th AFD-World Bank Development Conference held at CERDI, Clermont-Ferrand, on June 30 - July 1, 2017".

Card, M. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Carmon, N. (1996). *Immigration and Integration in Post-Industrial Societies: Theoretical Analysis and Policy Implications*.

Chi, G., Lin, F., Chi, G., & Blumenstock, J. (2020). A general approach to detecting migration events in digital trace data. *PLOS ONE*, *15*(10), 1–17.

Dubois, A., Zagheni, E., Garimella, K., & Weber, I. (2018). Studying migrant assimilation through facebook interests. In Staab, S., Koltsova, O., & Ignatov, D. I. (Eds.), *Social Informatics*, (pp. 51–60)., Cham. Springer International Publishing.

Fatehkia, M., Coles, B., Ofli, F., & Weber, I. (2020). The relative value of facebook advertising data for poverty mapping. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, (pp. 934–938).

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., & Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, *9*(1), 22.

Filiztekin, A. & Gökhan, A. (2008). The determinants of internal migration in turkey. *International Conference on Policy Modelling. EcoMod 2008*.

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017). Using twitter data to estimate the relationship between short-term mobility and long-term migration. (pp. 103–110).

Giurgola, S., Piaggesi, S., Karsai, M., Mejova, Y., Panisson, A., & Tizzoni, M. (2021). Mapping urban socioeconomic inequalities in developing countries through facebook advertising data.

Guo, D. & Zhu, X. (2014). Origin-destination flow data smoothing and mapping. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2043–2052.

Hannigan, A., O'Donnell, P., O'Keeffe, M., & MacFarlane, A. (2016). How do variations in definitions of "migrant" and their application influence the access of migrants to health care services? world health organization health evidence network synthesis report 46.

Herdağdelen, A., State, B., Adamic, L., & Mason, W. (2016). The social ties of immigrant communities in the united states. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, (pp. 78–84)., New York, NY, USA. Association for Computing Machinery.

Justino, D. (2016). *Emigration from Portugal: Old Wine in New Bottles?* Migration Policy Institute. info:eu-repo/grantAgreement/FCT/5876/147304/PT UID/SOC/04647/2013.

Karemera, D., Oguledo, V. I., & Davis, B. (2000). A gravity model analysis of international migration to north america. *Applied Economics*, *32*(13), 1745–1755.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, *8*(1), 1–8.

Kikas, R., Dumas, M., & Saabas, A. (2015). Explaining international migration in the skype network: The role of social network features. In *SIdEWayS '15*.

Kuchler, T., Russel, D., & Stroebel, J. (2020). The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook. Technical report, National Bureau of Economic Research.

Lewer, J. J. & Van den Berg, H. (2008). A gravity model of immigration. *Economics letters*, *99*(1), 164–167.

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, *109*(29), 11576–11581.

Lusardi, A. & Mitchell, O. S. (2011). Financial literacy around the world: an overview. *Journal of Pension Economics and Finance*, *10*(4), 497–508.

Migali, S., Natale, F., Tintori, G., Kalantaryan, S., Grubanov-Boskovic, S., Scipioni, M., Farinosi, F., Cattaneo, C., Benandi, B., Follador, M., et al. (2018). International migration drivers. *A quantitative assessment of the structural factors*

driving migration. European Commission JRC Science for Policy Report.

Moise, I., Gaere, E., Merz, R., Koch, S., & Pournaras, E. (2016). Tracking language mobility in the twitter landscape. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, (pp. 663–670)., Los Alamitos, CA, USA. IEEE Computer Society.

OECD (2019). *Talent Abroad: A Review of Romanian Emigrants*.

Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Garcia Herranz, M., Al-Asad, M., & Weber, I. (2020). Monitoring of the venezuelan exodus through facebook's advertising platform. *PLOS ONE*, *15*(2), 1–15.

Ramos, R. & Suriñach, J. (2016). A gravity model of migration between the enc and the eu: A gravity model of migration between the enc and the eu. *Tijdschrift voor economische en sociale geografie*, *108*.

Robillard, P. (1975). Estimating the od matrix from observed link volumes. *Transportation research*, *9*(2-3), 123–128.

Rodriguez, M., Helbing, D., Zagheni, E., et al. (2014). Migration of professionals to the us. In *International conference on social informatics*, (pp. 531–543). Springer.

Salah, A., Pentland, A., Lepri, B., Letouzé, E., Montjoye, Y.-A., Dong, X., Dağdelen, O., & Vinck, P. (2019). *Introduction to the Data for Refugees Challenge on Mobility of Syrian Refugees in Turkey*, (pp. 3–27).

Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization* (pp. 364–371). Elsevier.

Singh, V. K., Bozkaya, B., & Pentland, A. (2015). Money walks: Implicit mobility behavior and financial well-being. *PLOS ONE*, *10*(8), 1–17.

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., Guidotti, R., Bertoli, S., Kim, J., Muntean, C. I., et al. (2020). Human migration: the big data perspective. *International Journal of Data Science and Analytics*, 1–20.

Slingsby, A. & van Loon, E. (2016). Exploratory visual analysis for animal movement ecology. In *Computer Graphics Forum*, volume 35, (pp. 471–480). Wiley Online Library.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2018). Migration data using social media: a european perspective.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2019). Quantifying international human mobility patterns using facebook network data. *PLOS ONE*, *14*(10), 1–22.

Stephen, D. M. & Jenny, B. (2017). Automated layout of origin–destination flow maps: Us county-to-county migration 2009–2013. *Journal of Maps*, *13*(1), 46–55.

Stevens, S. S. (2017). *Psychophysics: Introduction to its perceptual, neural and social prospects*. Routledge.

Tobler, W. R. (1987). Experiments in migration mapping by computer. *The American Cartographer*, *14*(2), 155–163.

Tufte, E. R. (1986). *The Visual Display of Quantitative Information*. USA: Graphics Press.

Tufte, E. R., Goeler, N. H., & Benson, R. (1990). *Envisioning information*, volume 2. Graphics press Cheshire, CT.

Ware, C. (2004). *Information Visualization: Perception for Design: Second Edition.*

Willumsen, L. G. (1978). Estimation of an od matrix from traffic counts–a review.

Wood, J., Dykes, J., & Slingsby, A. (2010). Visualisation of origins, destinations and flows with od maps. *The Cartographic Journal*, *47*(2), 117–129.

Yang, Y., Dwyer, T., Goodwin, S., & Marriott, K. (2016). Many-to-many geographically-embedded flow visualisation: An evaluation. *IEEE transactions on visualization and computer graphics*, *23*(1), 411–420.

Zagheni, E., Garimella, V., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from twitter data. (pp. 439–444).

Zagheni, E., Polimis, K., Alexander, M., Weber, I., & Billari, F. C. (2018). Combining social media data and traditional surveys to nowcast migration stocks. In *Annual Meeting of the Population Association of America*.

Zhou, Z., Meng, L., Tang, C., Zhao, Y., Guo, Z., Hu, M., & Chen, W. (2018). Visual abstraction of large scale geospatial origin-destination movement data. *IEEE transactions on visualization and computer graphics*, *25*(1), 43–53.

APPENDIX A

Statistics of Facebook Marketing API Emigration Estimates

a. Countries with Emigration Estimates Available

Countries with emigration estimates from Facebook Marketing API available are;

Algeria, Argentina, Australia, Austria, Bangladesh, Belgium, Brazil, Cameroon, Canada, Chile, China, Colombia, Cuba, Cyprus, Czechia, Côte d'Ivoire, Dem. Rep. Congo, Denmark, Dominician Republic, El Salvador, England, Estonia, Ethiopia, Finland, France, Germany, Ghana, Greece, Guatemala, Haiti, Hong Kong, Hounduras, Hungary, India, Indonesia, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kenya, Kuwait, Latvia, Lebanon, Lithuania, Luxembourg, Malaysia, Malta, Mexico, Monaco, Morocco, Nepal, Netherlands, New Zealand, Nicaragua, Nigeria, Norway, Peru, Phillipines, Poland, Porto Rico, Portugal, Qatar, Romania, Russia, Rwanda, Saudi Arabia, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Tanzania, Thailand, Uganda, United Arab Emirates, United States of America, Venezuela, Vietnam, Zambia, and Zimbabwe.

b. Countries with Higher Facebook Emigration Estimates

Compared to their corresponding United Nations emigration estimates, countries with higher Facebook estimates are listed below. These countries are;

Algeria, Bangladesh, Brazil, Colombia, Cuba, Dominician Republic, El Salvador, Guatemala, Haiti, Hounduras, India, Indonesia, Jamaica, Lithuania, Mexico, Monaco, Morocco, Nepal, Nicaragua, Peru, Phillipines, Poland, Porto Rico, Romania, Sierra Leone, Slovakia, Sri Lanka, Venezuela, and Vietnam.

c. Emigration Estimates of Countries

Facebook emigration estimates of countries can be observed from Table A.1

Country	Emigration Estimate	Country	Emigration Estimate	Country	Emigration Estimate
Algeria	478798	Hong Kong	314034	Poland	1042266
Argentina	636493	Hounduras	687982	Porto Rico	775255
Australia	461015	Hungary	237837	Portugal	723409
Austria	110367	India	5364662	Qatar	65313
Bangladesh	2338441	Indonesia	1314919	Romania	1470601
Belgium	238699	Ireland	222418	Russia	660896
Brazil	1360982	Israel	345642	Rwanda	73036
Cameroon	239137	Italy	969957	Saudi Arabia	321646
Canada	531483	Jamaica	391159	Senegal	239588
Chile	296761	Japan	481348	Serbia	398811
China	562652	Jordan	332398	Sierra Leone	83203
Colombia	1223301	Kenya	291842	Singapore	167737
Cuba	726923	Kuwait	83576	Slovakia	209757
Cyprus	87106	Latvia	128150	Slovenia	53065
Czechia	164907	Lebanon	415731	South Africa	443373
CÃ´te d'Ivoire	264790	Lithuania	232045	South Korea	1008200
Dem. Rep. Congo	772336	Luxembourg	30831	Spain	833885
Denmark	95084	Malaysia	738682	Sri Lanka	503198
Dominician Republic	752098	Malta	55760	Sweden	180910
El Salvador	881834	Mexico	6026375	Switzerland	196407
England	1612686	Monaco	62172	Tanzania	146318
Estonia	67986	Morocco	827948	Thailand	847160
Ethiopia	433900	Nepal	1204643	Uganda	178970
Finland	100652	Netherlands	241018	United Arab Emirates	544840
France	1347022	New Zealand	283242	United States of America	4503753
Germany	696045	Nicaragua	442239	Venezuela	2417727
Ghana	305509	Nigeria	751304	Vietnam	1626389
Greece	9603	Norway	136022	Zambia	94857
Guatemala	878668	Peru	924280	Zimbabwe	401213
Haiti	533477	Phillipines	4032815		

Table A.1 Estimated emigration statistics from Facebook Marketing API.