

Albert Ali Salah · Alex Pentland ·  
Bruno Lepri · Emmanuel Letouzé  
*Editors*

# Guide to Mobile Data Analytics in Refugee Scenarios

The 'Data for Refugees Challenge' Study

 Springer

# Guide to Mobile Data Analytics in Refugee Scenarios

Albert Ali Salah · Alex Pentland ·  
Bruno Lepri · Emmanuel Letouzé  
Editors

# Guide to Mobile Data Analytics in Refugee Scenarios

The ‘Data for Refugees Challenge’ Study

With contributions by Patrick Vinck,  
Yves-Alexandre de Montjoye, and Xiaowen Dong

 Springer

*Editors*

Albert Ali Salah  
Information and Computing Sciences  
Utrecht University  
Utrecht, the Netherlands

Bruno Lepri  
Fondazione Bruno Kessler  
Trento, Italy

Alex Pentland  
Massachusetts Institute of Technology  
Cambridge, MA, USA

Emmanuel Letouzé  
Massachusetts Institute of Technology  
Cambridge, MA, USA

ISBN 978-3-030-12553-0      ISBN 978-3-030-12554-7 (eBook)  
<https://doi.org/10.1007/978-3-030-12554-7>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

*It's crucial to invest in data. Without good data,  
We're flying blind. If you can't see it, you can't solve it.*  
Kofi Annan, Nobel laureate (1938–2018)

Large-scale refugee movements are born out of some of the most drastic crises of our times, including wars, droughts, and collapse of national economies. Neither our advancing technology and welfare nor our increased ability to gather, process, and relay information can completely solve all the issues that are related to refugee movements. Modern societies are composed of tightly interlinked structures, including social services, education and healthcare infrastructure, job markets, and cultural and religious communities, each of which is affected by the forced displacement of people, both for the refugees and for the receiving countries. It is obvious that the solution of refugee problems will involve the collaborative and coordinated effort of many parties, and this, in its turn, can only happen with access to good and timely information, as well as interdisciplinary dialog between the stakeholders.

This book is an attempt to illustrate the possibilities offered by mobile data for improving the living conditions of refugees. It is born out of a big data challenge, called Data for Refugees, which was run between 2017 and 2019, with the partnership of Türk Telekom, Boğaziçi University, and TÜBİTAK. In this challenge, Türk Telekom opened a big mobile data set to researchers, and interdisciplinary teams from all over the world analyzed the data for gaining insights into refugee mobility. The Data for Refugees (D4R) Challenge officially ended in January 2019, with a closing workshop and award ceremony. This book collects a selected subset of project reports, edited for the present volume, and complements them with additional chapters.

The book is organized into four parts. Part I, “Big Data and Refugees” opens with a chapter that introduces the D4R Challenge, its history, the data shared with the participants, and a short summary of ethical considerations. The remaining chapters provide a comprehensive discussion of issues related to the usage of big data in refugee scenarios. While the book has a distinct focus on mobile data, this part goes beyond it and introduces perspectives and projects from international governmental organizations (i.e., UNHCR and UNICEF), tackles ethical and

privacy aspects, and includes one chapter on coding boot camps for refugees as a possible way of empowering refugees to process and use their own data.

The second and third parts of the book collect selected and revised reports from the D4R Challenge, and reflect its broad scope. Part II is dedicated to work on “Social Integration”, which was a major theme of the challenge. Part III collects works on “Labor, Education, Health, Safety”. This is a rough grouping; all topics are interlinked when it comes to refugees.

Part IV of the book, under the heading of “Conclusions”, contains three chapters for looking toward the future. The contribution by Letouzé describes the OPAL system, which is a viable and scalable approach for mobile operators to share data with researchers, international governmental organizations, and policy-makers. The chapter by Verhulst and Young describes Data Collaboratives and provides many examples of data usage for humanitarian purposes.

The final chapter of the book is written by the Challenge participants and was initiated during a discussion session at the closing workshop of the D4R Challenge. It summarizes the main findings and seeks to translate the knowledge gained from the Challenge into policy recommendations. It also critically discusses the risks involved in such recommendations. This chapter forms the basis of a white paper, written in Turkish and shared with the stakeholders in Turkey, in the hope of informing policy decisions for the benefit of refugees.

Taken together, we hope that this book will advance the field of mobile data analytics for refugee scenarios, illustrate many ways of using mobile data in conjunction with other relevant sources of data (e.g., satellite, census, social media data), and most importantly, provide incentives for and practical ways of allowing commercial companies with large amount of data to share them responsibly for humanitarian purposes.

This work is partially supported by an MIT-Boğaziçi University MISTI grant to Alex Pentland and Albert Ali Salah. We thank Türk Telekom, Boğaziçi University, TÜBİTAK, UNHCR, IOM, UNICEF, Istanbul & I, FBK, MIT Media Lab, and everybody who supported the Data for Refugees Challenge along the way.

Utrecht, The Netherlands  
April 2019

Albert Ali Salah  
Alex Pentland  
Bruno Lepri  
Emmanuel Letouzé  
Yves-Alexandre de Montjoye  
Xiaowen Dong  
Patrick Vinck

# Contents

## Part I Big Data and Refugees

<b>1</b>	<b>Introduction to the Data for Refugees Challenge on Mobility of Syrian Refugees in Turkey</b> .....	<b>3</b>
	Albert Ali Salah, Alex Pentland, Bruno Lepri, Emmanuel Letouzé, Yves-Alexandre de Montjoye, Xiaowen Dong, Özge Dağdelen and Patrick Vinck	
<b>2</b>	<b>Call Detail Records to Obtain Estimates of Forcibly Displaced Populations</b> .....	<b>29</b>
	David Pastor-Escuredo, Asuka Imai, Miguel Luengo-Oroz and Daniel Macguire	
<b>3</b>	<b>Mobile Phone Data for Children on the Move: Challenges and Opportunities</b> .....	<b>53</b>
	Vedran Sekara, Elisa Omodei, Laura Healy, Jan Beise, Claus Hansen, Danzhen You, Saskia Blume and Manuel Garcia-Herranz	
<b>4</b>	<b>Coding Boot Camps for Refugees</b> .....	<b>67</b>
	Jessica Francis	
<b>5</b>	<b>“Do No Harm” in the Age of Big Data: Data, Ethics, and the Refugees</b> .....	<b>87</b>
	Patrick Vinck, Phuong N. Pham and Albert Ali Salah	
<b>6</b>	<b>Pioneering Predictive Analytics for Decision-Making in Forced Displacement Contexts</b> .....	<b>101</b>
	Christopher Earney and Rebeca Moreno Jimenez	

## Part II D4R Challenge: Social Integration

- 7 Measuring Fine-Grained Multidimensional Integration Using Mobile Phone Metadata: The Case of Syrian Refugees in Turkey** ..... 123  
 Michiel A. Bakker, Daoud A. Piracha, Patricia J. Lu, Keis Bejgo, Mohsen Bahrami, Yan Leng, Jose Balsa-Barreiro, Julie Ricard, Alfredo J. Morales, Vivek K. Singh, Burcin Bozkaya, Selim Balcisoy and Alex Pentland
- 8 Towards an Understanding of Refugee Segregation, Isolation, Homophily and Ultimately Integration in Turkey Using Call Detail Records** ..... 141  
 Jeremy Boy, David Pastor-Escuredo, Daniel Macguire, Rebeca Moreno Jimenez and Miguel Luengo-Oroz
- 9 Using Call Data and Stigmergic Similarity to Assess the Integration of Syrian Refugees in Turkey** ..... 165  
 Antonio Luca Alfeo, Mario G. C. A. Cimino, Bruno Lepri and Gigliola Vaglini
- 10 Integration of Syrian Refugees: Insights from D4R, Media Events and Housing Market Data** ..... 179  
 Simone Bertoli, Paolo Cintia, Fosca Giannotti, Etienne Madinier, Caglar Ozden, Michael Packard, Dino Pedreschi, Hillel Rapoport, Alina Sirbu and Biagio Speciale
- 11 Quantified Understanding of Syrian Refugee Integration in Turkey** ..... 201  
 Wangsu Hu, Ran He, Jin Cao, Lisa Zhang, Huseyin Uzunalioglu, Ahmet Akyamac and Chitra Phadke
- 12 Syrian Refugee Integration in Turkey: Evidence from Call Detail Records** ..... 223  
 Tugba Bozcaga, Fotini Christia, Elizabeth Harwood, Constantinos Daskalakis and Christos Papademetriou
- 13 Assessing Refugees' Onward Mobility with Mobile Phone Data—A Case Study of (Syrian) Refugees in Turkey** ..... 251  
 Harald Sterly, Benjamin Etzold, Lars Wirkus, Patrick Sakdapolrak, Jacob Schewe, Carl-Friedrich Schleussner and Benjamin Hennig
- 14 Segregation and Sentiment: Estimating Refugee Segregation and Its Effects Using Digital Trace Data** ..... 265  
 Neal Marquez, Kiran Garimella, Ott Toomet, Ingmar G. Weber and Emilio Zagheni



**15 Measuring and Mitigating Behavioural Segregation as an Optimisation Problem: The Case of Syrian Refugees in Turkey** ..... 283  
 Daniel Rhoads, Javier Borge-Holthoefer and Albert Solé-Ribalta

**Part III D4R Challenge: Labor, Education, Health, Safety**

**16 Seasonal Labor Migration Among Syrian Refugees and Urban Deep Map for Integration in Turkey** ..... 305  
 Sedef Turper Alışık, Damla Bayraktar Aksel, Asım Evren Yantaç, Ilker Kayi, Sibel Salman, Ahmet İçduygu, Damla Çay, Lemi Baruh and Ivon Bensason

**17 Refugees in Undeclared Employment—A Case Study in Turkey** ..... 329  
 Fabian Bruckschen, Till Koebe, Melina Ludolph, Maria Francesca Marino and Timo Schmid

**18 The Use of Big Mobile Data to Gain Multilayered Insights for Syrian Refugee Crisis** ..... 347  
 Özgün Ozan Kılıç, Mehmet Ali Akyol, Oğuz Işık, Banu Günel Kılıç, Arsev Umur Aydınöğlü, Elif Surer, Hafize Şebnem Düzgün, Sibel Kalaycıoğlu and Tuğba Taşkaya Temizel

**19 Improve Education Opportunities for Better Integration of Syrian Refugees in Turkey** ..... 381  
 Marco Mamei, Seyit Mümin Cilasun, Marco Lippi, Francesca Pancotto and Semih Tümen

**20 Optimizing the Access to Healthcare Services in Dense Refugee Hosting Urban Areas: A Case for Istanbul** ..... 403  
 M. Tarik Altuncu, Ayse Seyyide Kaptaner and Nur Sevencan

**21 Characterizing the Mobile Phone Use Patterns of Refugee-Hosting Provinces in Turkey** ..... 417  
 Erika Frydenlund, Meltem Yılmaz Şener, Ross Gore, Christine Boshuijzen-van Burken, Engin Bozdog and Christa de Kock

**22 Refugee Mobility: Evidence from Phone Data in Turkey** ..... 433  
 Michel Beine, Luisito Bertinelli, Rana Cömertpay, Anastasia Litina, Jean-François Maystadt and Benteng Zou

**Part IV Conclusions**

**23 Leveraging Open Algorithms (OPAL) for the Safe, Ethical, and Scalable Use of Private Sector Data in Crisis Contexts** ..... 453  
 Emmanuel Letouzé

<b>24</b>	<b>The Potential and Practice of Data Collaboratives for Migration</b> .....	<b>465</b>
	Stefaan G. Verhulst and Andrew Young	
<b>25</b>	<b>Policy Implications of the D4R Challenge</b> .....	<b>477</b>
	Albert Ali Salah, M. Tarık Altuncu, Selim Balcisoy, Erika Frydenlund, Marco Mamei, Mehmet Ali Akyol, Kerem Yavuz Arslanlı, Ivon Bensason, Christine Boshuijzen-van Burken, Paolo Bosetti, Jeremy Boy, Tuğba Bozcaga, Seyit Mümin Cilasun, Oğuz Işık, Sibel Kalaycıoğlu, Ayse Seyyide Kaptaner, Ilker Kayı, Özgün Ozan Kılıç, Berat Kjamili, Huseyin Kucukali, Aaron Martin, Marco Lippi, Francesca Pancotto, Daniel Rhoads, Nur Sevencan, Ervin Sezgin, Albert Solé-Ribalta, Harald Sterly, Elif Surer, Tuğba Taşkaya Temizel, Semih Tümen and Ismail Uluturk	
	<b>Index</b> .....	<b>497</b>

# Contributors

**Ahmet Akyamac** Nokia Bell Labs, Murray Hill, NJ, USA

**Mehmet Ali Akyol** Middle East Technical University, Çankaya, Ankara, Turkey

**Antonio Luca Alfeo** University of Pisa, Pisa, Italy;  
University of Florence, Florence, Italy

**M. Tarık Altuncu** TRT World, Istanbul, Turkey;  
Imperial College London, London, UK

**Kerem Yavuz Arslanlı** Istanbul Technical University, Sarıyer/Istanbul, Turkey

**Arsev Umur Aydınöglü** Middle East Technical University, Çankaya, Ankara,  
Turkey

**Mohsen Bahrami** Sabanci University, Istanbul, Turkey

**Michiel A. Bakker** Massachusetts Institute of Technology, Cambridge, MA, USA

**Selim Balcisoy** Sabanci University, Istanbul, Turkey

**Jose Balsa-Barreiro** Massachusetts Institute of Technology, Cambridge, MA,  
USA

**Lemi Baruh** Koç University, İstanbul, Turkey

**Damla Bayraktar Aksel** Koç University, İstanbul, Turkey

**Michel Beine** University of Luxembourg, Luxembourg, Luxembourg

**Jan Beise** UNICEF Data Research and Policy, New York, NY, USA

**Keis Bejgo** Massachusetts Institute of Technology, Cambridge, MA, USA

**Ivon Bensason** Attorney at Law, İstanbul, Turkey

**Luisito Bertinelli** University of Luxembourg, Luxembourg, Luxembourg

**Simone Bertoli** Université Clermont Auvergne, CNRS, IRD, CERDI, Clermont-Ferrand, France

**Saskia Blume** UNICEF Data Research and Policy, New York, NY, USA

**Javier Borge-Holthoefer** Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC), Barcelona, Spain

**Paolo Bosetti** Fondazione Bruno Kessler, Trento, Italy

**Christine Boshuijzen-van Burken** Eindhoven University of Technology, Eindhoven, The Netherlands

**Jeremy Boy** UN Global Pulse, New York, NY, USA

**Tugba Bozcaga** Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, USA

**Engin Bozdog** 4TU Centre for Ethics and Technology, Eindhoven, The Netherlands

**Burcin Bozkaya** Sabanci University, Istanbul, Turkey

**Fabian Bruckschen** Freie Universität Berlin, Berlin, Germany; Knuper, Berlin, Germany

**Jin Cao** Nokia Bell Labs, Murray Hill, NJ, USA

**Damla Çay** Koç University, İstanbul, Turkey

**Fotini Christia** Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, USA

**Seyit Mümin Cilasun** Central Bank of the Republic of Turkey, Ankara, Turkey

**Mario G. C. A. Cimino** University of Pisa, Pisa, Italy

**Paolo Cintia** Department of Computer Science, University of Pisa, Pisa, Italy

**Rana Cömertpay** University of Luxembourg, Luxembourg, Luxembourg

**Özge Dağdelen** Türk Telekom, Istanbul, Turkey

**Constantinos Daskalakis** CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA

**Xiaowen Dong** University of Oxford, Oxford, UK

**Hafize Şebnem Düzgün** Colorado School of Mines, Golden, CO, USA

**Christopher Earney** UNHCR Innovation, Geneva, Switzerland

**Benjamin Etzold** Bonn International Center for Conversion, Bonn, Germany

**Jessica Francis** New York University, New York, NY, USA

**Erika Frydenlund** Old Dominion University, Norfolk, USA

**Manuel Garcia-Herranz** UNICEF Office of Innovation, New York, NY, USA

**Kiran Garimella** Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

**Fosca Giannotti** ISTI-CNR, Pisa, Italy

**Ross Gore** Old Dominion University, Norfolk, USA

**Banu Günel Kılıç** Middle East Technical University, Çankaya, Ankara, Turkey

**Claus Hansen** UNICEF Data Research and Policy, New York, NY, USA

**Elizabeth Harwood** IDSS, Massachusetts Institute of Technology, Cambridge, MA, USA

**Ran He** Nokia Bell Labs, Murray Hill, NJ, USA

**Laura Healy** UNICEF Data Research and Policy, New York, NY, USA

**Benjamin Hennig** University of Iceland, Reykjavik, Iceland

**Wangsu Hu** Nokia Bell Labs, Murray Hill, NJ, USA;  
Rutgers University, Newark and Camden, NJ, USA;  
New Brunswick, NJ, USA

**Ahmet İçduygu** Koç University, İstanbul, Turkey

**Asuka Imai** UNHCR, Dakar, Senegal

**Oğuz Işık** Middle East Technical University, Çankaya, Ankara, Turkey

**Rebeca Moreno Jimenez** UNHCR Innovation, Geneva, Switzerland

**Sibel Kalaycıoğlu** Middle East Technical University, Çankaya, Ankara, Turkey

**Ayşe Seyyide Kaptaner** Birkbeck, University of London, London, UK

**Ilker Kayı** Koç University, Sarıyer/Istanbul, Turkey

**Özgün Ozan Kılıç** Middle East Technical University, Çankaya, Ankara, Turkey

**Berat KJamili** Migport, Ankara, Turkey

**Christa de Kock** Stellenbosch University, Stellenbosch, South Africa

**Till Koebe** Freie Universität Berlin, Berlin, Germany;  
Knuper, Berlin, Germany

**Huseyin Kucukali** Health Directorate of Istanbul, Istanbul, Turkey

**Yan Leng** Massachusetts Institute of Technology, Cambridge, MA, USA

**Bruno Lepri** Fondazione Bruno Kessler, Trento, Italy

**Emmanuel Letouzé** Massachusetts Institute of Technology, Data-Pop Alliance, Cambridge, MA, USA;

Data-Pop Alliance, MIT Media Lab, and OPAL, Brooklyn, NY, USA

**Marco Lippi** DISMI, University of Modena and Reggio Emilia, Reggio Emilia, Italy

**Anastasia Litina** University of Ioannina, Ioannina, Greece

**Patricia J. Lu** Massachusetts Institute of Technology, Cambridge, MA, USA

**Melina Ludolph** Knuper, Berlin, Germany;  
Humboldt Universität Berlin, Berlin, Germany

**Miguel Luengo-Oroz** United Nations Global Pulse, New York, NY, USA

**Daniel Macquire** UNHCR Innovation, Geneva, Switzerland

**Étienne Madinier** Paris School of Economics, Paris, France

**Marco Mamei** DISMI, University of Modena and Reggio Emilia, Reggio Emilia, Italy

**Maria Francesca Marino** Università degli Studi di Firenze, Florence, Italy

**Neal Marquez** University of Washington, Washington, DC, USA;  
Max Planck Institute for Demographic Research, Rostock, Germany

**Aaron Martin** Tilburg University, Tilburg, The Netherlands

**Yves-Alexandre de Montjoye** Imperial College London, London, UK

**Jean-François Maystadt** University of Antwerp, Antwerpen, Belgium;  
Lancaster University, Lancaster, UK

**Alfredo J. Morales** Massachusetts Institute of Technology, Cambridge, MA, USA

**Elisa Omodei** UNICEF Office of Innovation, New York, NY, USA

**Caglar Ozden** The World Bank, Washington DC, United States

**Michael Packard** Georgetown University, Washington DC, United States

**Francesca Pancotto** DCE, University of Modena and Reggio Emilia, Reggio Emilia, Italy

**Christos Papademetriou** Department of Computer Science, Columbia University, New York, NY, USA

**David Pastor-Escuredo** Technical University Madrid and LifeD Lab, Madrid, Spain

**Dino Pedreschi** Department of Computer Science, University of Pisa, Pisa, Italy

**Alex Pentland** Massachusetts Institute of Technology, Cambridge, MA, USA

**Chitra Phadke** Nokia Bell Labs, Murray Hill, NJ, USA

**Phuong N. Pham** Harvard University, Cambridge, MA, USA

**Daoud A. Piracha** Massachusetts Institute of Technology, Cambridge, MA, USA

**Hillel Rapoport** Paris School of Economics, Paris, France

**Daniel Rhoads** Internet Interdisciplinary Institute, Castelldefels, Spain;  
Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC),  
Barcelona, Spain

**Julie Ricard** Data–Pop Alliance, New York, NY, USA

**Patrick Sakdapolrak** University of Vienna, Vienna, Austria

**Albert Ali Salah** Utrecht University, Utrecht, The Netherlands;  
Boğaziçi University, Istanbul, Turkey

**Sibel Salman** Koç University, İstanbul, Turkey

**Jacob Schewe** Potsdam Institute for Climate Impact Research, Potsdam, Germany

**Carl-Friedrich Schleussner** Climate Analytics gGmbH, Berlin, Germany

**Timo Schmid** Freie Universität Berlin, Berlin, Germany

**Vedran Sekara** UNICEF Office of Innovation, New York, NY, USA

**Nur Sevencan** TRT World, Istanbul, Turkey

**Ervin Sezgin** Istanbul Technical University, Sarıyer/Istanbul, Turkey

**Alina Sîrbu** Department of Computer Science, University of Pisa, Pisa, Italy

**Vivek K. Singh** Rutgers University, New Brunswick, NJ, USA

**Albert Solé-Ribalta** Internet Interdisciplinary Institute (IN3), Universitat Oberta  
de Catalunya (UOC), Barcelona, Spain

**Biagio Speciale** Paris School of Economics, Paris, France

**Harald Sterly** University of Vienna, Vienna, Austria

**Elif Surer** Middle East Technical University, Çankaya, Ankara, Turkey

**Tuğba Taşkaya Temizel** Middle East Technical University, Çankaya, Ankara,  
Turkey

**Ott Toomet** University of Washington, Washington, DC, USA

**Semih Tümen** Department of Economics, TED University, Ankara, Turkey

**Sedef Turper Alışık** Koç University, İstanbul, Turkey

**Ismail Uluturk** University of South Florida, Tampa, USA

**Huseyin Uzunalioglu** Nokia Bell Labs, Murray Hill, NJ, USA

**Gigliola Vaglini** University of Pisa, Pisa, Italy

**Stefaan G. Verhulst** The GovLab, NYU Tandon School of Engineering, New York, USA

**Patrick Vinck** Harvard University, Cambridge, MA, USA

**Ingmar G. Weber** Qatar Computing Research Institute, Doha, Qatar

**Lars Wirkus** Bonn International Center for Conversion, Bonn, Germany

**Asım Evren Yantaç** Koç University, İstanbul, Turkey

**Meltem Yılmaz Şener** Norwegian Center for Human Rights, University of Oslo, Oslo, Norway

**Danzhen You** UNICEF Data Research and Policy, New York, NY, USA

**Andrew Young** The GovLab, NYU Tandon School of Engineering, New York, USA

**Emilio Zagheni** University of Washington, Washington, DC, USA;  
Max Planck Institute for Demographic Research, Rostock, Germany

**Lisa Zhang** Nokia Bell Labs, Murray Hill, NJ, USA

**Benteng Zou** University of Luxembourg, Luxembourg, Luxembourg



**Part I**  
**Big Data and Refugees**

# Chapter 1

## Introduction to the Data for Refugees Challenge on Mobility of Syrian Refugees in Turkey



**Albert Ali Salah, Alex Pentland, Bruno Lepri, Emmanuel Letouzé,  
Yves-Alexandre de Montjoye, Xiaowen Dong, Özge Dağdelen  
and Patrick Vinck**

**Abstract** The Data for Refugees (D4R) Challenge was a nonprofit challenge initiated to improve the conditions of the Syrian refugees in Turkey by providing a special database to scientific community for enabling research on urgent problems concerning refugees, including health, education, unemployment, safety, and social integration. The collected database was based on anonymized mobile Call Detail Record (CDR) of phone calls and SMS messages of Türk Telekom customers. It in-

---

A. A. Salah (✉)

Utrecht University, Utrecht, The Netherlands and Boğaziçi University, Istanbul, Turkey  
e-mail: [a.a.salah@uu.nl](mailto:a.a.salah@uu.nl)

A. Pentland

Massachusetts Institute of Technology, Cambridge, MA, USA  
e-mail: [sandy@media.mit.edu](mailto:sandy@media.mit.edu)

B. Lepri

Fondazione Bruno Kessler, Trento, Italy  
e-mail: [lepri@fbk.eu](mailto:lepri@fbk.eu)

E. Letouzé

Massachusetts Institute of Technology, Data-Pop Alliance, Cambridge, MA, USA  
e-mail: [eletouze@datapopalliance.org](mailto:eletouze@datapopalliance.org)

Y.-A. de Montjoye

Imperial College London, London, UK  
e-mail: [demontjoye@imperial.ac.uk](mailto:demontjoye@imperial.ac.uk)

X. Dong

University of Oxford, Oxford, UK  
e-mail: [xdong@robots.ox.ac.uk](mailto:xdong@robots.ox.ac.uk)

Ö. Dağdelen

Türk Telekom, Istanbul, Turkey  
e-mail: [ozge.dagdelen@turktelekom.com.tr](mailto:ozge.dagdelen@turktelekom.com.tr)

P. Vinck

Harvard University, Cambridge, MA, USA  
e-mail: [pvinck@hsph.harvard.edu](mailto:pvinck@hsph.harvard.edu)

© Springer Nature Switzerland AG 2019

A. A. Salah et al. (eds.), *Guide to Mobile Data Analytics  
in Refugee Scenarios*, [https://doi.org/10.1007/978-3-030-12554-7\\_1](https://doi.org/10.1007/978-3-030-12554-7_1)

licated broad activity and mobility patterns of refugees and citizens in Turkey for 1 year. The data collection period was from January 1, 2017 to December 31, 2017. The project was initiated by Türk Telekom, in partnership with the Turkish Academic and Research Council (TÜBİTAK) and Boğaziçi University, and in collaboration with several academic and nongovernmental organizations, including UNHCR Turkey, UNICEF, and International Organization for Migration. This chapter describes the Challenge in detail, providing a history of its evolution, as well as a description of the data shared with the participants of the Challenge.

## 1.1 Introduction

After the Syrian Civil War started in 2011–2012, civilians in increasing numbers sought refuge in neighboring countries. By May 2017, Turkey had received over 3 million refugees—the largest refugee population in the world. Initially, a significant proportion of refugees (30% at one time) lived in government-run camps near the Syrian border. Some of these camps were later shut down, and the Directorate General of Migration Management announced the number of refugees living in these camps as 177,376, as of October 10, 2018. The official number of refugees living in camps was 228,251 at the end of December 2017.<sup>1</sup> Many have moved to cities looking for work and better living conditions. They face problems of integration, income, welfare, employment, health, education, language, social tension, and discrimination.

The Data for Refugees (D4R) Challenge<sup>2</sup> was a nonprofit project to ultimately improve the conditions of the Syrian refugees in Turkey by providing a special database to the scientific community for enabling research on some urgent problems. The challenge datasets opened to the community were based on anonymized mobile Call Detail Records (CDRs) of phone calls and SMS messages of Türk Telekom customers. These data indicate communication activities and mobility patterns in Turkey for 1 year. The D4R Challenge, called the Challenge hereafter, invited research groups across the globe to submit proposals, which had been carefully evaluated. Data were then opened to selected groups on strictly regulated terms. The five focus themes of the Challenge were health, education, unemployment, safety, and social integration, respectively. The project was initiated by Türk Telekom, in partnership with the Turkish Academic and Research Council (TÜBİTAK) and Boğaziçi University and in collaboration with several academic and nongovernmental or intra-governmental organizations, including UNHCR Turkey, UNICEF, and International Organization for Migration.

A Scientific Committee of international experts guided the execution of the project. A Project Evaluation Committee (PEC)<sup>3</sup> was formed with representatives from academia, government (i.e., ministries related to the Challenge), and nonprofit

---

<sup>1</sup><https://multeciler.org.tr>, accessed 13.11.2018.

<sup>2</sup><http://d4r.turktelekom.com.tr>.

<sup>3</sup><http://d4r.turktelekom.com.tr/presentation/project-evaluation-committee>.

organizations working in this area. This committee represented refugee interests, and its job was to ensure that the submitted research projects that are granted access to the data have clear goals, with foreseen benefits to the refugee population in Turkey and elsewhere. Access to the D4R data required PEC approval.

The general aims of the D4R Challenge were to

1. Contribute to the welfare of the refugee populations;
2. Gain insights into key issues, including safety and security, health, education, unemployment, social integration and segregation, mobility, and distribution of resources and infrastructure;
3. Help governments and international bodies model the dynamics of the refugee populations and to discover vulnerabilities (socioeconomic vulnerabilities, gaps in education and services, etc.);
4. Seed further projects, co-created with refugees, resulting in new applications, services, and innovative solutions for refugees in Turkey and elsewhere.

The lack of data on refugee mobility is a very important hurdle to the proper functioning of government services and international aid bodies. Innovative approaches attempted to overcome this included the use of satellite imagery to obtain information from the regions in crisis, with limited success [11]. The D4R project allowed, for the first time, the analysis of a large-scale mobile CDR database on refugees. The usefulness of such data has already been illustrated by the D4D Challenges [5, 14] organized previously, and the numerous projects completed on these challenges provided insights into data science for social good. Furthermore, the Challenge, by involving research groups from all around the globe, aimed to raise awareness for the refugee issues on a grand scale.

The possibilities that mobile CDR data afford for analysis of a broad set of problems are surveyed in [6]. Examples of projects conducted with similar data include analysis of disaster resilience [21], infrastructure planning [13], quantifying mobility effects on the spread of infectious diseases [4, 12], developing agent-based models for disease migrations [22], disease containment [10], analysis of community structures and sociodemographic indicators [23], detection of emergency events [9], poverty analysis [15], and mobility during holidays and religious festivals [17], to name a few.

The D4R Challenge had a distinct focus around refugee problems, and aimed to enable evaluation of refugee-related interventions and activities, including, but not limited to, educational activities, social gatherings, NGO actions, government infrastructure investments, etc. It also had the potential to provide insights into the analysis of residential segregation, population structures for specific geographical locations, and factors on social integration [19].

The rest of the present chapter is devoted to a short history of the Challenge (Sect. 1.2), to serve as an example to people who may consider running similar challenges, the description of the D4R data (Sect. 1.3), and the ethical and privacy guidelines that were made available to the participants at the onset of the Challenge (Sect. 1.4). This chapter extends the data description paper shared with the participants at the initiation of the Challenge [16].

## 1.2 History of the D4R Challenge

In this section, we summarize the chronological progression of the challenge its committees, and the evaluation criteria.

### 1.2.1 Time Line of the Challenge

The D4R Challenge was conceptually conceived in a meeting of researchers from TÜBİTAK BILGEM and Boğaziçi University on January 6, 2017. The purpose of the meeting was to develop joint project ideas for dealing with many aspects of the refugee crisis. Boğaziçi University had previously participated in the Data for Development (D4D) Challenges, and a mobile data challenge was proposed as a way of making data available to many research groups at the same time.

Following the initial idea, Türk Telekom was contacted and the possibilities of such a challenge, as well as the main approach proposed for addressing its ethical and privacy issues, were presented. A report written by the ethics committee of the Data for Development Challenge in Senegal served as a template for structuring the privacy measures [1]. At the same time, the Information and Communication Technologies Authority ICTA (in Turkish, BTK—Bilgi Teknolojileri ve İletişim Kurumu) was contacted to ensure that such a challenge could be initiated. Türk Telekom embraced the idea of the challenge, and asked for more detailed information on its accomplishment.

The Scientific Committee of the Challenge was formed between April and May 2017. A preliminary data description was prepared by this committee and presented to the legal team of Türk Telekom. It was important to ensure that no terms of the customer agreement were violated by the Challenge, and no personal information was shared in any way. In June 2017, a series of meetings were held with Türk Telekom teams to discuss the scope of the data and its anonymization. The data description was approved in July.

We have decided to collect data over 1 year to enable the observation of seasonal trends. To make the best use of the restricted refugee coverage, we requested the entire refugee base of Türk Telekom to be sampled. The refugee tag was obtained by different means (described in Sect. 1.3), which was purposefully noisy as an additional layer of protection. In September 2017, we have started forming the Project Evaluation Committee (PEC), inviting members from academia, NGOs, ministries, as well as members of the Syrian refugee community in Turkey (see Sect. 1.2.2).

A tentative schedule of the Challenge was formed in September. At the same time, the project procedure was presented to the Ethics Board of Boğaziçi University (INAREK). The Challenge, which aimed to open data of Türk Telekom to research groups outside Boğaziçi University with its own ethics committee to evaluate the project reports, was a very unusual project for the Ethics Board. The Board, chaired by Prof. Yaman Barlas, investigated the financial sourcing of the project, its partners



**Fig. 1.1** The D4R logo, designed by Marcom Ajans

and supporters, the procedure for project evaluation, and the data description before making a (positive) decision.

In the meantime, Türk Telekom worked on the website design. The D4R logo (see Fig. 1.1) was designed by Marcom Ajans. The door in the logo symbolizes the open door policy of the Turkish government in the Syrian refugee crisis. The open door could be seen as a symbol for sharing the data with the “outside”, a major commitment by any telecommunications company. Since the map of Turkey is roughly rectangular, it also illustrates the geographical location of Syria with respect to Turkey. The logo, D4R, and Data for Refugees are trademarked by Türk Telekom.

The dataset collection and anonymization took several months, and was completed in January 2018. Meanwhile, several organizations were contacted for their support in the project. Fondazione Bruno Kessler was the first institution to declare its official support in October 2017, followed by MIT Media Lab, Data-Pop Alliance, Istanbul and I, UNHCR, International Organization for Migration (IOM), and UNICEF. The most important concerns addressed at this stage, particularly by UNHCR, were that (1) the sharing of the data is permitted by the owners of the data, i.e., the refugees; (2) the data should be anonymized in a way to make it impossible to identify individuals in it; and (3) the data should not allow other governments to track specific refugees, even if they had other, complementary data about individuals. Our data collection ensured all these points.

The first call for proposals went out in January 2018, resulting in 102 proposals by the deadline in March, of which 33 were rejected, 59 accepted, and 10 deferred to PEC after the initial screening. The most important factor in the early rejection was the lack of scientific rigor (or content).

After the PEC evaluation, a total of 61 groups were granted access to the dataset, of which 31 groups were able to submit project reports by the Challenge deadline. The word cloud created from the titles of 61 proposals (see Fig. 1.2) illustrates the relative popularity of the main themes. Social integration was by far the most popular area selected by the project participants.

The 31 groups submitting project reports involved 163 researchers (118 male, 45 female) from three NGOs, three governmental or intergovernmental institutions, two research labs in commercial companies, and 57 academic institutions from 19 countries (see Fig. 1.3). One-third of the projects involved an expat Turkish researcher working outside Turkey. Six projects came from exclusively Turkish institutions, 18 projects from institutions outside Turkey, and seven projects involved collaborations



Fig. 1.2 A word cloud obtained from the titles of the successful project proposals

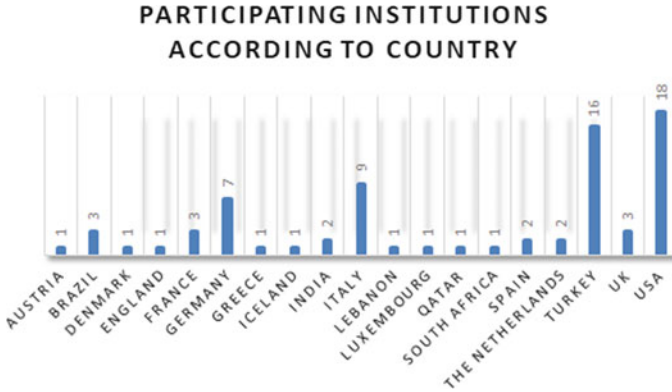


Fig. 1.3 The participating institutions according to the country

between Turkish institutions and those outside Turkey. Twenty-three projects, out of a total of 31, involved multiple institutions. Twenty-six projects were selected for presentations in the D4R Closing Workshop, and 17 groups were invited to submit revised reports for the present volume.

### 1.2.2 Committees

The day-to-day running of the Challenge was coordinated by the Scientific Committee, whose members were

- Albert Ali Salah (chair, Boğaziçi University and Utrecht University),
- Alex Pentland (Massachusetts Institute of Technology),
- Bruno Lepri (Fondazione Bruno Kessler),
- Emmanuel Letouzé (Massachusetts Institute of Technology and Data-Pop Alliance),
- Yves-Alexandre de Montjoye (Imperial College London),
- Xiaowen Dong (University of Oxford), and

- Patrick Vinck (Harvard Humanitarian Initiative).

This committee determined the scope of the dataset, drafted the ethical guidelines, as well as the project prescreening and evaluation criteria, and organized public dissemination activities.

The second and larger committee was the Project Evaluation Committee (PEC), whose task was to evaluate the project proposals to determine which groups will be granted access to the dataset, and to evaluate the project reports to ensure that no published content can potentially harm the refugee population. The latter is a broad criterion. In such a sensitive project, the results of analysis, even if scientifically rigorous, need to be framed carefully to make sure that they cannot be misused, intentionally or otherwise. The members of the PEC represented a wide range of interests, from academia, international nonprofit, nongovernmental organizations, refugee organizations, and the Turkish ministries:

- Senem Özyavuz (chair, Türk Telekom),
- Iyad Rahwan (Massachusetts Institute of Technology),
- Anahi Ayala Iacucci (Internews),
- Bülent Sankur (Boğaziçi University),
- Yildirm Bahadirlar (TÜBİTAK BİLGEM),
- Alex Rutherford (Massachusetts Institute of Technology),
- Claire Melamed (Global Partnership for Sustainable Development Data),
- Jean-Marie Garelli (UNHCR),
- Ahmad Garibeh (Istanbul & I),
- Geoffrey Charles Fox (Indiana University),
- Joséphine Goube (Techfugees),
- Firat Yaman Er (Türk Telekom),
- Phuong Pham (Harvard University),
- Mithat Büyükhan (Lifelong Learning General Directorate, Turkish Ministry of Education),
- Mazen AboulHosn (International Organization for Migration),
- Ömer Hakan Simsek (Turkish Medicines and Medical Devices Agency, Turkish Ministry of Health),
- Nona Zicherman (UNICEF),
- Manuel García-Herranz (UNICEF), and
- Vedran Sekara (UNICEF).

### ***1.2.3 Evaluation Criteria***

The projects submitted to D4R Challenge were screened by the Scientific Committee (SC) and the Project Evaluation Committee (PEC). The proposals that passed the initial screening were granted access to the dataset, upon submitting the signed User Agreement Form (See Appendix).

Criteria for passing the initial screening were as follows:



**Table 1.1** Summary of possible project evaluation outcomes

PEC recommendation	Description
Normal publication	No restriction applied. Can be invited to present in the Challenge workshop, in related sessions and is also fully eligible for prizes
Ask to consider adjustments	Requires some wording, visual, and/or content adjustment before decision. Without further adjustment, PEC asks Not to publish it
Do not publish	PEC asks Not to publish the document. This could be due to either of the following: <ul style="list-style-type: none"> <li>• The reports might be sensitive or too risky for a general publication, but could merit a presentation to the appropriate authorities. This will be evaluated on a case-by-case basis;</li> <li>• The methodology or conclusions do not confirm to scientifically rigorous methodology. Scientific correctness is a precondition for publication</li> </ul>

- A project proposal is submitted (in English), and all the team members who will access the data are individually identified;
- The project’s primary investigator has a permanent affiliation;
- The project uses the D4R data meaningfully;
- The project aims are aligned with the goals of the Challenge, do not represent a commercial interest, and do not endanger the privacy or well-being of individuals or groups.

The project reports were published publicly on the project website,<sup>4</sup> after evaluation by the D4R PEC and D4R SC. There were three possible outcomes for submitted reports (Table 1.1): (1) Normal Publication, for papers that treat the ethical issues correctly; (2) Ask to Consider Adjustments, for papers that require amendments and removal of sensitive material before publication. We asked the participants to be careful not to include statements that may harm the refugee population in any way or may promote negative sentiments about the refugee populations; and (3) Do Not Publish, either because the report is on a sensitive issue or because it is not scientifically rigorous and its conclusions are not warranted. Sensitive reports may be shared directly with related institutions or authorities. The PEC and the project proposers were designated to jointly decide on this, on a case-by-case basis. The user agreement stated that any further scientific publication based on the D4R data will be sent to the PEC/SC for evaluation, before the publication. A fixed period of 30 days is allotted to the PEC/SC to grant permission to such publications, after which tacit denial could be assumed. It was important to keep the publication options open beyond the duration of the Challenge, as rigorous scientific analysis of the data requires

<sup>4</sup><http://d4r.turktelekom.com.tr>.

significant commitment by the participating groups. It is natural that these groups would require permission to publish scientific papers (and dissertations) based on their findings. However, such publications also need to be screened before reaching the public, as the risks are no less after the Challenge is finished.

The following criteria were used for the evaluation of projects with regard to the awards in each category:

- **Relevance to the focus area:** The project addresses a major problem in the focus area and outlines how its results can be used to help solve the problem;
- **Methodology:** The project is assessed for its analytical rigor and empirical backing of its conclusions, and whether additional data sources are appropriately and creatively used in conjunction with the D4R data;
- **Public communication:** It is important to raise awareness about refugee issues. The project outcomes are assessed for the clarity of presentation, and the strength of its message, for instance, through good visualizations. Additional resources (web, videos, other multimedia material, etc.) that can make the results clear to the general public are welcome and are assessed.
- **Recommendations:** The Challenge outcomes and recommendations are compiled into an accessible white paper (written in Turkish) and shared with the related government agencies, NGOs, and other stakeholders.<sup>5</sup> This report is one of the most direct ways for the Challenge participants to improve refugee conditions. Subsequently, the reports are assessed for the relevance of their recommendations.

## 1.3 Description of D4R Data

This section describes the datasets shared within the D4R Challenge.

### 1.3.1 Brief Summary

The D4R Challenge is based on the successful Data for Development (D4D) Challenge series [5]. Three datasets are made available to the Challenge participants, along with external helper files. The main difference from D4D is that the D4R data contain a “Refugee” flag, which indicates (with a high probability) that the CDR belongs to a refugee customer. This flag is given to customers in the database that (1) have ID numbers given to refugees and foreigners in Turkey; (2) are registered with Syrian passports; and (3) use special tariffs reserved for refugees. The proportions of these three groups in the database are not equal; the first group has 42.87%, the second group 1.06%, and the last group 56.07%, respectively. None of these groups is guaranteed to include only and exclusively refugees, which serves as a layer of

---

<sup>5</sup>See the D4R project website for the white paper.

protection: It is not possible to identify with certainty that a particular CDR belongs to a refugee or not, since this indicator is noisy and it is only possible to derive patterns from aggregated records. In such data collection efforts, it is prudent to have some minimal noise in the sensitive labels (e.g., refugee vs. non-refugee). We list the datasets contained in the Challenge in individual subsections.

Turkey is party to the 1951 Geneva Refugee Convention, but only acknowledges “refugee” status for people originating from Europe. Syrian refugees are officially considered “temporarily protected foreign individuals.” We acknowledge that the term “refugee” is used as a blanket term in the dataset and includes migrants, asylum seekers, and foreigners (Syrian or otherwise) who have acquired a temporarily protected foreign individual ID number in Turkey (i.e., starting with 98 or 99). The dataset needs to be approached with these reservations in mind, and the analysis should carefully consider such biases in the data.

The D4R dataset is collected from 992,457 customers of Türk Telekom, of which 184,949 are tagged as “refugees”, and 807,508 as Turkish citizens. A total of 1,211,839 subscriptions are included. Of these, 980,697 belong to Turkish citizens, and 231,142 belong to refugees (we refer to these customers as refugees, but as mentioned before, there exists certain level of noise in this indicator). Some of the customers had multiple phone lines; each line corresponds to a single subscription.

Out of all the refugee-tagged customers, 75% are recorded as “male”, and 25% as “female”. There is clearly a gender bias in the ownership of the phone line. This does not mean that 75% of the phone lines are used by men, however. We have sampled the 807K Turkish customers with the same gender distribution.

Out of all the refugee customers, 45% are registered in Istanbul. This information is obtained from the address associated with the contract. Other major cities with refugee presence are Gaziantep, Izmir, Sanliurfa, and Mersin. To simplify comparisons, we have sampled the Turkish citizen customers mainly from the cities with registered refugee presence. Tables 1.2 and 1.3 show the distribution of customers and their registered cities (only top locations are shown). The distribution over all the cities of the country is provided to the participants in a separate file. This file shows the official number of refugees registered per city, the official city population in 2017 (excluding refugees), and the percentage of refugees with respect to the city population. Additionally, it shows the number of Türk Telekom customers used for the entire D4R data collection per city, broken into “refugee” and “citizen” counts.

The numbers of registered refugees and asylum seekers in Turkey according to registration dates can be obtained from the UNHCR website.<sup>6</sup> Another useful source of data is the TUIK census estimates for Turkish cities, according to the year.<sup>7</sup> This source indicates the population size and growth of each city between 2000 and 2017.

The usage of the D4R data requires caution in interpreting the representativeness of the data for the refugee population in Turkey. At the end of March 2017, there were 75,724,413 mobile customers in Turkey across all operators (94.9% penetration rate) [2]. Excluding machine to machine (M2M) and population of the age range of

<sup>6</sup>See <https://data2.unhcr.org/en/situations/syria/location/113>.

<sup>7</sup>See [http://www.tuik.gov.tr/PreIstatistikTablo.do?istab\\_id=1590](http://www.tuik.gov.tr/PreIstatistikTablo.do?istab_id=1590).

**Table 1.2** The distribution of customers tagged as refugees in the dataset and their registered locations. Numbers rounded to the third significant digit

Location	Number of customers	Percentage (%)
Istanbul	84,173	45.511
Gaziantep	14,898	8.055
İzmir	10,425	5.637
Sanliurfa	9,701	5.245
Mersin	9,660	5.223
Hatay	7,024	3.798
Ankara	5,580	3.017
Konya	4,718	2.551
Bursa	3,479	1.881
Outside Turkey	2,902	1.569
Other	32,440	17.540

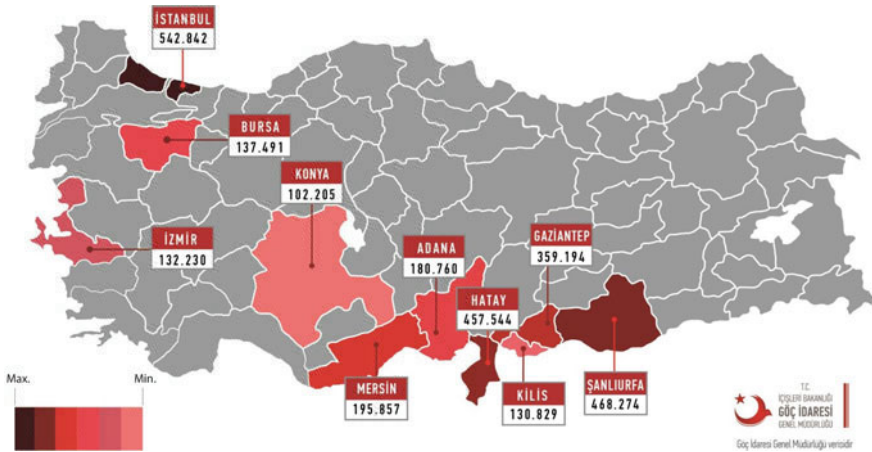
**Table 1.3** The distribution of customers tagged as Turkish citizens in the dataset and their registered locations. Numbers rounded to the third significant digit

Location	Number of customers	Percentage (%)
Istanbul	363,334	44.994
Gaziantep	80,655	9.988
İzmir	40,501	5.016
Ankara	40,443	5.009
Adana	40,415	5.005
Hatay	40,394	5.002
Konya	40,388	5.002
Antalya	40,367	4.999
Bursa	40,359	4.998
Sanliurfa	40,321	4.993
Mersin	40,242	4.983

0–9, the mobile penetration was 107%. According to data from the first 3 months of 2017, the mobile market share of Türk Telekom (Avea), from which the Challenge data were collected, was 24.7% [2].

We have used the entire refugee customer base (with the filtering conditions described previously); however, the market share of Türk Telekom also shows fluctuations according to the individual cities. Therefore, it is useful to look at official numbers of refugees distributed over the country. This is partly depicted in Fig. 1.4 according to figures<sup>8</sup> from the Ministry of Interior, Directorate General of Migration

<sup>8</sup>Only the top ten cities are shown. More detailed information can be obtained from [http://www.goc.gov.tr/icerik/migration-statistics\\_915\\_1024](http://www.goc.gov.tr/icerik/migration-statistics_915_1024).



**Fig. 1.4** The distribution of refugees in the country according to data from Ministry of Interior, Directorate General of Migration Management

Management in 2018. For the Challenge, we have supplied the official figures from March 2017, as mentioned before. In what follows, we describe the contents of the dataset in more detail.

### 1.3.2 Cell Tower Locations

The cell tower (i.e., base station) locations are provided in the file “Base\_Station\_Location.txt”. The file contains the following fields:

- **BTS\_ID**: The ID of the cell tower;
- **MX\_LAT1,MX\_LAT2,MX\_LAT3**: DMS latitude of the cell tower;
- **MX\_LONG1,MX\_LONG2,MX\_LONG3**: DMS longitude of the cell tower;
- **MX\_SAHAIL**: The registered city of the cell tower;
- **MX\_SAHAILCE**: The registered district of the cell tower;
- **MX\_POPAREA**: An unofficial note about the population type around the cell tower, used internally in Türk Telekom. It takes values of **RURAL**, **SUB\_URBAN**, **INDUSTRIAL**, **SEASONAL AREAS**, **DENSE\_URBAN**, **HOT SPOT**, **OPEN IN URBAN**, **AIRPORT**, **SUBURBANLOW**, **POPRURAL**, and **INDOOR**.

In some rare cases, the precise location information of the cell tower is missing, and only the city is indicated. The interpretation of the latitude and longitude follows the degree, minutes, seconds (DMS) syntax. For example, the district of Bartın in the city of Zonguldak is represented by these six numbers as follows: (41 25 43.1184 32 4 37.9344). This corresponds to 41°25'43.1184" N DMS latitude, and 32°4'37.9344" E DMS longitude.

### 1.3.3 District Locations

To disambiguate the cell towers, we provide a file that contains district coordinates. This file, named “district\_coordinates.csv”, has the following fields:

- CITY: Name of the city;
- DISTRICT: Name of the district;
- POPULATION\_2014: The official census population of the district in 2014;
- LATITUDE, LONGITUDE: The 2D (GPS) coordinates of the district;
- MX\_LAT1, MX\_LAT2, MX\_LAT3, MX\_LONG1, MX\_LONG2, MX\_LONG3: The DMS coordinates of the district.

A conversion script (such as <https://www.latlong.net/lat-long-dms.html>) can be used to convert the (latitude, longitude) variables into DMS coordinates. For example, the 2D coordinates of (41.428644 32.077204) for Zonguldak, Bartın, translate into the DMS coordinates given in the previous subsection.

### 1.3.4 Dataset 1: Antenna Traffic

The first database we provide includes 1-year site-to-site traffic on an hourly basis. This dataset contains the traffic between each site for a year. Calls between Türk Telekom (TT) customers and other service providers (SP) only have information about the TT side. For each record, total number and duration of calls are recorded in an aggregated fashion.

The database is split into voice and SMS partitions. For the voice partition, the file “Dataset 1\_2017XX.txt” contains the data for month XX, and there are 12 such files. Each file contains the following fields:

- TIMESTAMP: Day and hour considered in format DD-MM-YYYY HH (24 h format);
- OUTGOING\_SITE\_ID: The ID of the site the call originated from. Unknown stations are denoted as “-99” or “9999”;
- INCOMING\_SITE\_ID: The ID of the site receiving the call;
- NUMBER\_OF\_CALLS: The number of calls in the one-hour slot;
- NUMBER\_OF\_REFUGEE\_CALLS: The number of calls involving numbers tagged as “refugee”;
- NUMBER\_OF\_TOTAL\_CALL\_DURATION: The total call duration from all calls;
- REFUGEE\_CALL\_DURATION: The total call duration from calls involving numbers tagged as “refugee”.

Note that it is possible for a call to be labeled incorrectly, as we do not know the refugee status of the other party, when the call is to a different SP than TT.

For the SMS partition, the file “Dataset 1\_SMS\_2017XX.txt” contains the data for month XX, and there are 12 such files. Each file contains the following fields: