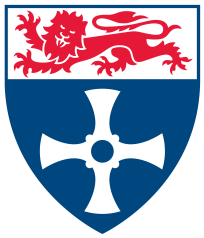
Geographical Veracity of Indicators from Mobile Phone Data

A Study of Call Detail Records Data in France



Maarten Vanhoof

School of Computing
Newcastle University

Thesis submitted in partial fulfillment of the degree of
Doctor of Philosophy in Computing Science

Supervisors:
Dr. Zbigniew Smoreda
Dr. Clement Lee
Prof. Dr. Patrick Olivier

June 2018

I dedicate this thesis to doctor Paul Vanhoof,
who taught us courage, compassion, and gratitude.

Declaration

I hereby declare that, except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work or is the outcome of work done in collaboration with others, as will be specified in the acknowledgements and in the sections of acknowledgement preceding each chapter.

Maarten Vanhoof
November 2018

Acknowledgements

Many collaborators have contributed to the preparation of this thesis or simply to the learning process a PhD really is. To me, our interactions are a manifold more valuable than the works we have been pursuing together. They have formed the cornerstone of my day-to-day motivation and inspiration for the past few years and I'd like to thank each and every one of them. Detailed acknowledgments on the contributions to each chapter will be given in the acknowledgement section at the beginning of each chapter. In no particular order, I value the collaborations with:

- (Soon to be Dr.) Maxim Janzen and Prof. Dr. Kay W. Axhausen from the Institute for Transport Planning and Systems at ETH Zürich on the use of mobile phone data to study domestic tourism trips.
- (Soon to be Dr.) Aare Puussaar from the Open Lab and Urban Observatory at Netwcastle University on the technology behind mobile phone data and the vision to use data for justice and social good.
- Dr. Luca Pappalardo, Dr. Lorenzo Gabrielli, Prof. Dr. Dino Pedreschi and Prof. Dr. Fosca Giannotti from the Knowledge Discovery and Data Mining Laboratory at the University of Pisa and the Information Science and Technology Institute of the Italian National Research Council on the extraction of insights from mobile phone data for mobility, well-being and city functionalities.
- Willem Schoors, Liane Henderickx, Prof. Dr. Gert Verstraeten and Prof. Dr. Anton Van Rompaey from the department of Earth and Environmental Sciences at KU Leuven on using mobility patterns from mobile phone data for large-scale comparison and for tourism destination detection.
- Prof. Dr. Santi Phithakkitnukoon of the Department of Computer Engineering at Chiang Mai University on exploring mobile phone data to understand the link between social interactions and human mobility.
- Prof. Dr. Thomas Ploetz from the School of Interactive Computing at Georgia Institute of Technology on machine learning and the home detection of mobile phone users.
- (Soon to be Dr.) Joris Beckers and Prof. Dr. Ann Verhetsel from the department of Transport and Regional Economics at the University of Antwerpen on the application of network theory in transport geography.

- Dr. Clementine Cottineau, Dr. Carlos Molinero, Prof. Dr. Elsa Arcaute and Prof. Dr. Mike Batty from the Centre for Advanced Spatial Analysis at University College London on the application of scaling laws and percolation theory on indicators and networks derived from mobile phone data.

- Iacopo Iacopini and Prof. Dr. Vito Latora from the School of Mathematical Sciences at Queen Mary University of London on the combination of spatial interaction models with movement patterns from mobile phone data.

- Dr. Sebastian Grauwin, Prof. Dr. Michael Szell (currently at Central European University) and Prof. Dr. Carlo Ratti from the Senseable City Lab at Massachusetts Institute of Technology on the analysis of interaction networks from mobile phone data

- Prof. Dr. Stanislav Sobolevsky from the Center for Urban Science + Progress at New York University for the application of network theory on interaction networks from mobile phone data and mobility networks in general.

- Craig Hannabus, Angie Hatting, Boyd Roberts and the entire team of Havas Boondoggle SA on the use of data visualisation in service design.

- Stéphanie Combes, Benjamin Sakarovitch, Marie-Pierre de Bellefon, Pauline Givord and Vincent Loonis from the Insititut National de la Statistique et des Etudes Economiques (INSEE) on the application of mobile phone data in official statistics: the validation of home detection, the use of machine learning to detect functional areas, and the spatial analysis of statistical indicators.

- Fernando Reis and Michail Skaliotis from the Eurostat big data task-force at the European Commission on the application of mobile phone data in official statistics.

- Dr. Alexis Bley, Dr. Iva Bojic, and Prof. Dr. Xu Yang (currently at Hong Kong Polytechnic University) from the Singapore-MIT Alliance for Research and Technology Center on the unveiling of communication and mobility segregation in cities based on mobile phone data.

Thanks

I'd like to thank Dr. Erez Hatna and Phil James for revising this thesis. Apart from their help, I'm lucky to have been guided by several supervisors throughout my academic career. Although often too short, I'm grateful for the times spent together and for the many lessons learned. Thanks to Prof. Dr. Gert Verstraeten for taking a chance on me and for your understanding the period I needed it most, to Prof. Dr. Santi Phithakkitnukoon for challenging me at the very beginning, to Prof. Dr. Thomas Ploetz for improving my academic writing and for the many courses taught, to Dr. Clement Lee for the amazing support during the last phase of the thesis and to Prof. Dr. Patrick Olivier for the never-ending source of energy.

Most importantly, I wish to thank Dr. Zbigniew Smoreda who, apart from being my supervisor, has become a true mentor and a valued friend. Merci Zbig, pour ta patience, tes blagues, ta vision sur la recherche, les choses banales, et les choses hyper-importantes (la bouffe et la politique, dans cette ordre). Enfin pour tous les moments partagés, ceux qui ont étés, et ceux qui arriveront encore!

I love to thank the many colleagues, in Open Lab and Orange Labs, who lightened my everyday and formed a source of inspiration. Thanks Aare, Stefania, Eilean, Cezary, Ralph, Kevin, Jean-Sam, Nico, Yoann, Rose, and many others. Merci particulièrement à Mathieu Sannié et Maryse Piart pour vos efforts administratives et pour la porte toujours ouverte.

A huge thanks also to my friends, whether newly-made or beautifully old-crafted. Here's to the ones that drink the beers, Baileys, Amaretto or champagne, watch GOT, take the trains, cross the seas, discuss the sports, and turn around despair. You lot take care of my welfare.

So much thanks to my mum and dad for the relentless efforts they put in raising, protecting, guiding, and motivating me. Praise, joy, and lots of sunshine to Tinne, Wouter en Robbe, and the rest of my great family too. For always making their home, our home.

Ultimately, all of my love goes to that stellar girl *Snoetje Chavez* for combining the best of all the rest. For the inspiration, the support, al je weetjes (vooral die over voetbal). Voor elk (skype-)gesprek, voor elke dag, for making Dolfijnias fly, en voor je lieve lach.

To the friends and family I left behind.
To familiar strangers that filled
my mind, my sight, each night
in Paris, each breeze in the North.
To all unknowns left unexplored.

Maarten Vanhoof

Abstract

The study of mobile phone data opens opportunities in many research domains and for many applications. One point of critique is that, within current analyses, mobile phone users are considered uniform and interchangeable. To counter this *social atom problem*, good research practice demands an increasing contextualization of research results, for example by confrontation with auxiliary datasets or geographical knowledge. The latter forms the starting point of this thesis. The main argument is that there exists a *spatial knowledge gap* when it comes to the use of indicators derived from mobile phone data. The presented studies assess the *geographical veracity* of indicators derived from Call Detail Record (CDR) data and the underlying methods used. Based on a CDR dataset of almost 18.5 million users in France captured during a 154-day period in 2007, they show how mobile phone indicators can be constructed for all individual users using *big data technologies*. Investigation then is on the performance, sensitivity to user choices, and error estimations of home detection methods, which form a primordial step for the aggregation of users in space. Next, a spatial analysis of the popular Mobility Entropy (ME) indicator is performed, revealing its bias to cell tower density, for which a correction is then proposed. Ultimately, the relations between mobile phone indicators, indicators from other data sources and city definitions in France are explored. The main contribution of the thesis is that it reveals multiple limits of the common practices, results, and interpretations that govern mobile phone data research. The presented studies challenge the veracity of mobile phone indicators in different, predominantly geographical, ways and open up discussion on what should be done to improve trustworthiness.

Table of contents

List of figures xix

List of tables xxiii

Nomenclature xxv

1 Setting the stage 1

1.1 Preface . . . 3

1.2 Aim of the Thesis . . . 5

1.2.1 Research Questions . . . 5

1.3 Context of the Thesis . . . 6

1.3.1 Mobile Phone Data at Orange Labs France . . . 6

1.3.2 Knowledge Network . . . 7

1.3.3 Technological Context and Methodologies . . . 8

1.4 Main Concepts and Outline of the Thesis . . . 11

1.4.1 Mobile Phone Data Research . . . 11

1.4.2 Mobile Phone Indicators . . . 11

1.4.3 Home Detection . . . 12

1.4.4 Diversity of Human Movement . . . 12

1.4.5 Urban Systems . . . 13

1.5 Scientific Contributions . . . 14

1.5.1 Publications and Communications . . . 14

2 Literature Review 19

2.1 Research and Mobile Phone Data . . . 21

2.1.1 Captured Data for Selective Research . . . 21

2.1.2 Research Fields, Paradigms, and Critiques . . . 22
2.1.3 Moving Forward . . . 24
2.2 Available Information in Mobile Phone Data . . . 26
2.2.1 Two Dimensions of Human Activity . . . 26
2.2.2 Calling Patterns . . . 26
2.2.3 Movement Patterns . . . 29
2.2.4 Temporal Patterns . . . 30
2.3 Studying Human Interactions at Nation-wide Scale . . . 31
2.3.1 Topological Properties of Contact Networks . . . 31
2.3.2 Community Detection on Contact Networks . . . 32
2.3.3 Spatial Communities in Contact Networks . . . 33
2.3.4 Spatial Interaction Models for Human Communication . . . 33
2.4 Studying Movement Patterns at Nation-wide Scale . . . 35
2.4.1 Statistical Properties of Movement Patterns . . . 35
2.4.2 Predicting Individual Movement . . . 35
2.4.3 Spatial Interaction Models for Human Mobility . . . 35
2.5 Studying Presence Patterns at Nation-wide Scale . . . 38
2.5.1 Seasonal Changes and Tourism . . . 38
2.5.2 Temporal Presence and City Structures . . . 39
2.6 Individual Indicators from Mobile Phone Data . . . 41
2.6.1 Indicators for Individual Calling Patterns . . . 41
2.6.2 Indicators for Individual Movement Patterns . . . 43
2.6.3 The Mobility Entropy Indicator . . . 44
2.6.4 Deploying Mobile Phone Indicators . . . 46
2.7 Identifying Home Locations from Mobile Phone Data . . . 50
2.7.1 The Role and Act of Home Detection . . . 50
2.7.2 Detecting Homes from Large-Scale Location Traces . . . 50
2.7.3 Detecting Homes from Mobile Phone Data . . . 51
2.7.4 Defining Decision Rules for Single-Step Approaches . . . 53
2.7.5 Validating Home Detection Methods . . . 54
2.7.6 Shortcomings of Current Home Detection Methods . . . 56

3 Creating Mobile Phone Indicators **57**

3.1 Introducing the Dataset . . . 59

3.1.1 The French Call Detailed Records (CDR) data . . . 59

3.1.2 Locations of Cell Towers in France . . . 60

3.1.3 Types of Events . . . 61

3.1.4 Magnitudes of the Dataset . . . 62

3.1.5 Handling the French CDR Dataset . . . 63

3.2 Constructing Mobile Phone Indicators . . . 64

3.2.1 Individual Measures from Call Detailed Records . . . 64

3.2.2 User-Territory Mapping and Aggregation . . . 65

3.2.3 Simple Model Construction . . . 66

3.3 Mobile Phone Indicators in France in 2007 . . . 67

3.3.1 Distributions of Mobile Phone Indicators at User Level . . . 67

3.3.2 Distributions of Mobile Phone Indicators at Cell Tower Level . . . 68

3.3.3 Temporal Patterns of Mobile Phone Indicators . . . 77

3.3.4 Spatial Patterns of Mobile Phone Indicators . . . 78

3.4 Relating Mobile Phone Indicators to Socio-Economic Indicators . . . 81

3.4.1 Socio-Economic Indicators . . . 81

3.4.2 Correlations between Human Mobility, Calling Behavior and Socio-Economic Indicators . . . 81

3.5 Discussion . . . 84

3.5.1 A Situated Dataset . . . 84

3.5.2 Sufficient Sample Sizes . . . 84

3.5.3 Objectivity of Territorial Aggregation . . . 84

3.5.4 Studying the Territory with Mobile Phone Indicators . . . 85

4 Evaluating Home Detection Performance **87**

4.1 Detecting Homes from CDR Data . . . 89

4.1.1 Five Home Detection Algorithms with Simple Criteria . . . 89

4.1.2 Deployment and Metadata of HDAs . . . 90

4.2 Measuring Performance of Home Detection at Nation-Wide Scale . . . 92

4.2.1 High-level Validation Based on Census Data . . . 92

4.2.2 Comparisons between HDAs . . . 95
4.2.3 Uncertainty Assessment based on Metadata . . . 95
4.3 Performance of Home Detection in France . . . 97
4.3.1 Numbers of Detected Homes . . . 97
4.3.2 Performance Measures at Nation-Scale Based on Census Census Data . 98
4.3.3 Individual Level Differences Between HDAs . . . 102
4.3.4 Spatial Uncertainties Related to the Home Decision . . . 103
4.4 Discussion . . . 106
4.4.1 A Large Performance Gap at Nation-Scale . . . 106
4.4.2 Criteria Sensitivity Plays out at Another Level . . . 108
4.4.3 Spatial Uncertainties and Future Work . . . 109

5 Sensitivities of Home Detection Performance 111
5.1 Investigating Sensitivities of Home Detection Performance . . . 113
5.1.1 Defining 9 HDAs with Simple Criteria . . . 113
5.1.2 Defining 23 Time Periods . . . 114
5.2 Assessing Performance and Sensitivity . . . 116
5.2.1 Performance Measures at Nation-Scale . . . 116
5.2.2 Spatial Patterns of Home Detection Performance . . . 116
5.2.3 Assessing Sensitivity . . . 117
5.3 Results . . . 119
5.3.1 Relations between User Counts and Population Counts . . . 119
5.3.2 Spatial Patterns of LogRatio . . . 120
5.3.3 Sensitivity of Performance to Time Period . . . 122
5.3.4 Sensitivity of Performance to the Duration of Observation . . . 124
5.3.5 Sensitivity of Performance to Criteria and Parameter Choice . . . 126
5.4 Discussion . . . 128
5.4.1 Magnitudes of Sensitivities . . . 128
5.4.2 Contributions . . . 129

6 Correcting Mobility Entropy 131
6.1 A Flawed Mobility Entropy Indicator . . . 133

6.1.1 Defining the Mobility Entropy Indicator . . . 133
6.1.2 Bias of the Mobility Entropy Indicator . . . 134
6.2 Correcting Mobility Entropy . . . 135
6.2.1 Defining the Corrected Mobility (CME) Indicator . . . 135
6.3 Relations between ME, CME, Urban Areas, and Other Indicators in France . . . 138
6.3.1 Calculating CME from CDR data . . . 138
6.3.2 Distributions of ME and CME in Urban Areas . . . 139
6.3.3 Relations with Socio-Economic and Environmental Indicators . . . 141
6.4 Patterns of Mobility Diversity in France . . . 142
6.4.1 Relations between ME, CME, and Cell Tower Density . . . 142
6.4.2 Spatial Patterns of (Corrected) Mobility Entropy . . . 143
6.4.3 Mobility Diversity in Urban Areas . . . 145
6.4.4 Relations with Socio-Economic and Environmental Indicators . . . 147
6.5 Discussion . . . 151
6.5.1 Plausibility of the CME Indicator . . . 151
6.5.2 A Changing View on Mobility Diversity in France . . . 151
6.5.3 Wider Relevance . . . 153

7 Scaling Relations of Mobile Phone Indicators 155

7.1 Outstanding Knowledge Gaps for Mobile Phone Indicators . . . 157
7.1.1 Urban Scaling Laws . . . 157
7.1.2 Urban Scaling Laws and City Definitions . . . 159
7.1.3 Relations between Indicators and Sensitivity to City Definitions . . . 161
7.2 Scaling Laws of Mobile Phone Indicators in France . . . 164
7.2.1 Significance of Scaling Laws . . . 164
7.2.2 Superlinear Scaling Regimes . . . 165
7.2.3 Sublinear Scaling Regimes . . . 166
7.2.4 Changing Scaling Regimes . . . 167
7.3 Relations between Mobile Phone Indicators and Income Measures . . . 168
7.3.1 Relations for All City Definitions . . . 168
7.3.2 Sensitivity to City Definitions Parameters . . . 172
7.4 Discussion . . . 178

7.4.1 Urban Scaling Regimes of Mobile Phone Indicators 178
7.4.2 Sensitivity of Relations with Mobile Phone Indicators to City Definition 180

8 Discussion, Relevance, and Conclusion 181

8.1 Discussion 182
8.1.1 Spatial Knowledge Gaps of Mobile Phone Indicators 182
8.1.2 Uncertainty on the Performance of Home Detection Methods 183
8.1.3 Spatial Patterns of Mobile Phone Indicators 186
8.1.4 Urban Scaling Laws of Mobile Phone Indicators 187
8.2 Transferability and Relevance of Findings 189
8.2.1 Transferability 189
8.2.2 Relevance 192
8.3 Conclusion 195
8.3.1 Answering the Research Questions 195

References 199

List of figures

2.1 Illustration of a contact network constructed from CDR data 28

2.2 Illustration of a movement pattern constructed from CDR data 29

2.3 Illustration of a Cell tower Movement Graph (CMG) for France based on CDR data 30

2.4 Spatial communities resulting from community detection on a Cell tower Call Graph (CCG) in the UK 34

2.5 Spatial communities from a Cell tower Call Graph (CCG) and results of three spatial interaction models: Gravity, Radiation and Hierarchy. 34

2.6 Illustration of the difference between continuous distance and the concept of intervening opportunities. 36

2.7 Illustration of the seasonal changes in population presence in France based on CDR data. 38

2.8 Results of machine learning experiments to reproduce French urban areas from temporal presence patterns from CDR data. 40

2.9 Illustration of a second-order contact network constructed from CDR data. . . . 42

2.10 Distributions of real entropy, temporally uncorrelated entropy and random entropy in a CDR dataset. 45

2.11 Relations between diversity of calling patterns and socio-economic rank for communities in the UK. 47

2.12 Distribution of scaling exponents (β) of multiple census indicators when calculated for different city definitions in France. 49

3.1 Cell tower locations in France 60

3.2 Framework to create and deploy mobile phone indicators 64

3.3 Percentile values of different mobile phone indicators 67

3.4 Distributions at cell tower level of indicators related to calling behavior. 69

3.5 Distributions at cell tower level of indicators related to contact networks. 71

3.6 Distributions at cell tower level of indicators related to movement patterns. . . . 73

3.7 Distributions at cell tower level of indicators related to home detection. 75
3.8 Distributions at cell tower level of three mobile phone indicators for different months. 77
3.9 Spatial distributions of mobile phone indicators related to calling behavior and contact networks. 78
3.10 Spatial distributions of mobile phone indicators related to movement patterns and home detection. 80
3.11 Relations between mobile phone indicators and socio-economic indicators in France. 82
3.12 Relations between mobility entropy, entropy of contacts and the European Deprivation Index (EDI) in France 83

4.1 Share of activities performed at the most, second most, and third most frequently used cell tower 91
4.2 Spatial pattern of the ground truth population count from census data 93
4.3 Correlation between user and population counts for two HDAs during two months. 99
4.4 Spatial patterns of population count (census data) and user counts for the MA algorithm 101
4.5 Similarity Matching Coefficients (SMC) for all pairwise combinations of HDAs 102
4.6 Distributions of spatial uncertainty (SU) values for all users for different HDAs and months 104
4.7 Relation between SU and CSM values 105
4.8 Effect on CSM when filtering for different values of SU 110

5.1 Durations and related time period used in the analysis 115
5.2 Overview of the methodology to assess the sensitivity of home detection performance at nation-wide scale to to criteria 118
5.3 Relation between Pearson's R and CSM values. 119
5.4 Spatial patterns of home detection performance. 121
5.5 Sensitivity of home detection performance to time period. 122
5.6 Spatial pattern of home detection performance for a summer and a non-summer period. 123
5.7 Sensitivity of home detection performance to duration of observation 125
5.8 Sensitivity of performance to criteria and/or parameter choice. 127

6.1 Illustration of the calculation of the ME 133

6.2 Illustration of the dependency on cell tower density for the calculation of ME values . . . 134
6.3 Spatial pattern of the correction factors c_i. . . . 137
6.4 Urban area classification for 2010 in France. . . . 140
6.5 Scatterplot of cell tower density, the ME, and the CME. . . . 142
6.6 Relations between ME and CME at cell tower level. . . . 143
6.7 Spatial patterns of the ME and CME values in France. . . . 144
6.8 Distributions of the CME values for different urban areas. . . . 146
6.9 Linear regressions of the ME and the CME by a selection of indicators at municipality level. . . . 148

7.1 A range of city definitions for different density and flow thresholds . . . 160
7.2 Urban scaling laws for the percentage of nocturnal calls and the mean duration of calls . . . 164
7.3 Urban scaling laws for the number of visited cell towers, mobility entropy and the number of contacts . . . 165
7.4 Urban scaling laws for the median and the standard deviation of the duration between mobile phone events. . . . 166
7.5 Urban scaling laws for the spatial uncertainty and the distance between L1 and L2. 167
7.6 Correlations between the EDI index of wages and a selection of mobile phone indicators for different city definitions . . . 169
7.7 Correlations between the Gini index of wages and a selection of mobile phone indicators for different city definitions . . . 170
7.8 Correlations between the segregation index of wages and a selection of mobile phone indicators for different city definitions . . . 171
7.9 Correlations between the entropy of contacts and the segregation index of wages in the city definition parameter-space . . . 173
7.10 Correlations between the radius of gyration and the EDI in the city definition parameter-space . . . 174
7.11 Influence of the commuting thresholds on the relation between the segregation index of wages and a selection of mobile phone indicators. . . . 176
7.12 Influence of the density thresholds on the relation between EDI and a selection of mobile phone indicators. . . . 177

List of tables

2.1 Example of records in a mobile phone dataset . . . 26
2.2 Example of Call Detailed Records. . . . 27
2.3 Mobile phone indicators based on calling behavior and contact networks . . . 41
2.4 Mobile phone indicators based on movement patterns. . . . 43

3.1 Example of Call Detailed Records in the French dataset. . . . 59
3.2 Recorded events in the French CDR dataset . . . 61
3.3 Shares of different events for one day in the French CDR dataset . . . 62
3.4 Number of events in the French CDR dataset . . . 63
3.5 List of calculated mobile phone indicators. . . . 65

4.1 Criteria for the five deployed home detection algorithms . . . 90
4.2 Total number of detected home for five different HDAs . . . 97
4.3 Pearson's R values for the relation between user and population counts. . . . 100
4.4 CSM values for the relation between user and population counts . . . 100

5.1 Description of the nine deployed HDAs . . . 114

6.1 Sensitivity of the correlation between CME and cell tower density to different scaling ranges (a,b) . . . 136
6.2 Urban areas in France . . . 139
6.3 Socio-economic and environmental indicators considered to study the relation with CME. . . . 141
6.4 Summary statistics of the ME and the CME values per urban area. . . . 145
6.5 Significance tests for the pairwise differences in CME distributions between urban areas. . . . 146
6.6 Results of single linear regression of the ME, and CME by selection of other indicators. . . . 147

6.7 Coefficients of a multiple linear regression between the ME and a selection of indicators 150

6.8 Coefficients of a multiple linear regression between the CME and a selection of indicators 150

7.1 Increase factors when doubling population size for different β values 158

Nomenclature

Acronyms / Abbreviations

CCG Cell tower Call Graphs / Contact networks between cell towers derived from CDR data.

CDR Call Detail Records

CME Corrected Mobility Entropy

CMG Cell tower Movement Graphs / Movement networks between cell towers derived from CDR data.

CRM Customer Relationship Management

CSM Cosine Similarity Measure

CSSP Computational Social Science Paradigm

D4D Data for Development Challenge

DDR Data Detail Record

GDPR General Data Protection Regulation

GIS Geographical Information System

HDA Home Detection Algorithm

INSEE Institut National de la Statistique et des Études Économiques, the French Official Statistics office.

L1 Cell tower considered the most plausible home location given a home detection algorithm

L2 Cell tower considered the second most plausible home location given a home detection algorithm

L3 Cell tower considered the third most plausible home location given a home detection algorithm

MAUP Modifiable Area Unit Problem

MCG Mobile Call Graphs / Contact networks between mobile phone users derived from CDR data.

ME Mobility Entropy

OD Origin-Destination matrix / Notation to describe the structure of Cell tower Movement Graphs (CMGs).

PhD Doctor of Philosophy / The doctoral study

SENSE Sociology and Economics of Networks and Services

SMC Simple Matching Coefficient

SQL Structured Query Language

UDF User Defined Functions

User Individual Mobile Phone User

Chapter 1

Setting the stage

What is life but a series of inspired follies!

George Bernard Shaw

Abstract

This chapter introduces the aim, research questions, context, main concepts, outline, and scientific contributions of the thesis.

Related Publications and Acknowledgements

- Section 1.4.5 is based on the introduction of the paper in preparation: C Cottineau, M Vanhoof, E Arcaute. *Urban scaling laws of mobile phone indicators and their relation to census data*. This introduction has been written by Dr. Clementine Cottineau and the PhD candidate but was reworked by the candidate to integrate in the chapter.

1.1 Preface

Recent advances in information technologies and data collection have led to a true data deluge. Indeed, the age of *big data* has arrived and now enables analysis at unprecedented scale, detail, and speed of individual and collective behavior, immense technical systems, or even complex natural phenomena. Clearly, the arrival of new data sources is impacting many research fields, creating opportunities for new empirical observations, theory formulations, and for epistemological debate. Now more than ever, the rapidly occurring changes demand scientific rigorousness. Empirical claims, whether made based on big or small data, need to be grounded in reproducible, preferably validated, methods. Scientific debate needs to remain open to all positions, whether inspired by theory, practice or morals. And, above all, a clear perspective needs to be retained on how the engagement of research communities with new technological and analytical possibilities is creating or dissolving barriers, limiting or empowering people, and inspiring or discouraging the youth of today.

The field of mobile phone data research sits in the middle of such developments. Ever since mobile phone datasets have become available, studies investigating their deployment have stirred interests in multiple research domains ranging from urban studies, transport studies, to epidemiology and disaster management. By now, it is clear how mobile phone data, in combination with advanced techniques and methods, enables the analysis of large-scale human behavior such as movement or communication.

Still, the exact contributions mobile phone data research is going to make to different research domains and real world problems remain unclear. After all, access to mobile phone data, as well as the infrastructure and know-how to treat them, is unequally distributed amongst researchers, mainly because of data ownership by private operators. Additionally, research and commercial applications are increasingly being restricted by privacy policies. Meaning that governments are stepping up to their legal role and might, potentially, enforce some kind of legal right to *data access* in the near future. Especially in Europe, the latter is being anticipated, amongst others by Official Statistics offices. How and whether such data access will be formalized is still unknown but it makes one wonder what added value mobile phone data can actually bring to governments and whether this would be worth the public expenditure. Optimizing business processes, elaborating big data methodologies, or exploring research questions is one thing, political decision making or territorial management is another; and it remains an open question to which degree the quality of mobile phone data and its methodologies is sufficient to support the latter.

Simultaneously, mobile phone data from the global south are massively being mobilized by a combination of, internationally active institutions, NGO's, operators and research groups. Their activities can, no doubt, be considered a sort of, say, *data colonialism*, sapping ownership and agency from local governments, enterprises, and populations. Regardless of the seizing of mobile phone data, however, it is undeniable that research results in this perspective are being leveraged to astonishing levels. Mobile phone data research on developing countries has helped in complementing (sometimes non-existent, often completely outdated) census data, in

informing epidemiological models, in revealing social structures, or in describing the extreme (urban) migrations many developing countries are undergoing.

Throughout all developments, both in the global south and north, runs the debate of *data ownership*. Discussions on ownership of digital data take many forms including questions on consent, claimability or even financial gain. One direction that, I believe, is of particular interest, is how the absence of ownership, combined with the inability to correctly assess and understand the way 'personal data' is being used, might create resentment towards all forms of big data analysis. Such resentment, whether by politicians, legislators, citizens or, worst case scenario, the youth in the global south, could eventually result in restricted views on the benefits big data analysis can bring to society. Or, even worse, to a situation in which only the limited few can benefit from the availability of practically everyone's data, and this regardless of their intents.

To avoid such situations education is essential, that is a certitude, and it should be focused on raising data literacy and erasing inequality therein. But research has an important role to play too. Apart from elaborating and pursuing visions of how mobile phone data can advance fundamental research and real world problems (something the field of mobile phone data research is quite well versed at), there is an aspect of correctness and nuance that becomes extremely important. This thesis seeks to make its own, small contribution towards the idea of correctness and nuance. It aims at pointing out some methodological practices related to mobile phone analysis with low validity, missing error estimation, unknown uncertainties or un-nuanced interpretations. As a research field, we simply cannot afford such practices to seep into our works right now, as they, quite certainly, will become the source on which future education, public expenditure, or even political decision making is based.

1.2 Aim of the Thesis

The aim of this thesis is:

to explore the veracity of indicators derived from mobile phone data, specifically from a spatial perspective.

The underlying reasons are:

i. (*The relevance reason*) It can be expected that the role of mobile phone indicators will grow. In the western world this is because of the ongoing integration of mobile phone data in official statistics. In the global south it is because of their unique position to inform on society-wide changes and to support sustainable development in a data scarce environment.

ii. (*The opportunity reason*) The technology to create mobile phone indicators for large samples of users has only recently become available. The implication is that very little work has explored their construction and interpretation. At the same time, privacy regulations in the western world are becoming increasingly restrictive. This leaves only a limited amount of time to investigate and validate methodologies in countries where validation data is of decent quality, before methodologies would travel to other countries where regulation might be looser, but validation less trustworthy.

iii. (*The scientific contribution reason*) Up to now, mobile phone data research has given very little attention to the analysis of spatial patterns, the estimation of spatial errors and uncertainty, or to the development of spatially inspired validation techniques. In the case of mobile phone indicators, spatial aspects form an important part of interpretation and validation and so their study forms a clear contribution to the research domain.

1.2.1 Research Questions

Keeping the aim in mind, the research questions of this thesis are.

- What is the state of the art of mobile phone data research, especially with regard to the creation of indicators from mobile phone data?
- Is it possible to create mobile phone indicators from a mobile phone dataset available for France?
- To what extent can we interpret mobile phone indicators? What do they describe? How are they distributed in space? Can we validate the spatial patterns of mobile phone indicators?
- Can we investigate the relation between mobile phone indicators and indicators from census data? What problems arise? What relations can be uncovered and how are these relations shaped geographically?
- What are the uncertainties that come with mobile phone indicators? Can we quantify what errors relate to them?

1.3 Context of the Thesis

This thesis was elaborated within an university-industry collaboration between Newcastle University and Orange Labs France. To understand, situate, and assess the contributions of this thesis, at least a partial understanding of this university-industry collaboration is needed. The following subsections provide for such an understanding. At the same time, they serve as a reminder that academic progress is always situated, even when academic communication is done as objectively as possible.

1.3.1 Mobile Phone Data at Orange Labs France

Access to large mobile phone datasets is difficult to come by. Generally speaking, mobile phone data is owned by private operators and operationalized within their premises only. This is partly for strategical reasons, partly to align with directives from legislation on privacy protection. The consequence is that analysis of mobile phone data is primordially done in-house and on the available infrastructure of individual operators. Research progress, be it fundamental or applied, is thus dependent on the willingness of private companies to pursue it, either by partnerships with external researchers, or by the development of their own research team.

In the case of Orange Labs France, there has been a long tradition of doing in-house research, both fundamental and applied. Nowadays, for instance, the Sociology and Economics of Networks and Services (SENSE) department at Orange Labs France is a well-respected research lab, mainly studying the sociology and economy of digital economies. In addition, and under leadership of Dr. Zbigniew Smoreda, the lab has been mobilizing mobile phone datasets for fundamental research, first in France and more recent also in Côte d'Ivoire and Senegal[1].

Despite having an entire in-house research faculty, Orange Labs France has long been known for its collaborations within France and around the world. In the case of SENSE and with respect to mobile phone data research, this collaborative nature can be exemplified by, for example, several research papers that were published together with different international research groups, by the involvement in the Netmob conference [2] or, as is the case for this PhD, by the enlistment of PhD-students associated with foreign universities.

In preparation of the manuscript, the PhD candidate has resided at the SENSE department of Orange Labs France in Paris from November 2013 until November 2016 in order to ensure access to a French mobile phone dataset and the available big data infrastructure. The advantages of this residence are not to be underestimated.

[1] The mobilization of the Côte d'Ivoire and Senegal mobile phone data has been done in the context of the Data for Development (D4D) challenges (www.d4d.orange.com). D4D challenges are prepared by Orange Labs and MIT and aim at stimulating research teams worldwide to find applications of mobile phone data that could support development in the considered country.

[2] The Netmob conference is considered the main conference on the scientific analysis of mobile phone datasets. It has been (co-)organized since 2010 by multiple parties including the MIT Media Lab, IEEE, Orange, and Vodafone Italy.

Working at Orange Labs has lowered barriers in the form of contractual agreements, legal issues, understanding of privacy regulations, infrastructure investments, technical learning curves and access to knowledge networks. During the period from January 2017 until July 2018, the PhD candidate resided in the Open Lab at Newcastle University for the preparation of scientific contributions and the thesis manuscript. During this period, there was no access to the data but for the occasional visit to Orange Labs.

1.3.2 Knowledge Network

The aspect of the knowledge network deserves elaboration because it touches upon the independence of the presented work.

One of the main reasons for Orange Labs to invest in a fundamental research department, such as SENSE, is to position themselves as a capable research partner for high-quality research. The advantages of being an active partner in high-level knowledge networks are numerous, including quality assurance, knowledge attraction and development, improved access to public funding, and media visibility. The interesting thing in this perspective is that, from the side of Orange Labs, development and maintenance of knowledge networks are not being demanded from PhD students, or at least not in the case of this PhD. Rather, efforts are made to facilitate researchers in the elaboration of a knowledge network as they see fit. For this thesis, this means that all research lines and collaborations with external partners, have been developed by the PhD candidate himself, although being actively facilitated by the Orange Labs (as well as by the Open Lab which is a key example of promoting international collaboration in science). In other words, the position of Orange Labs is to promote the independence of (young) researchers, which is a position to praise.

The active facilitation of collaboration, not in the least by allowing research partners in-house access to the data, has been invaluable for the development of this thesis. Because of the interdisciplinary and often exploratory nature of mobile phone data research, external domain expertise is of critical importance to properly situate and explore the contribution mobile phone data can make to research. In other words, the possibility to develop international research collaborations has added scope, expertise and academic quality to the works preceding (although not all presented in) this manuscript. For a full list of collaborations undertaken we refer the reader to the acknowledgements.

1.3.3 Technological Context and Methodologies

Big Data Systems at Orange Labs

Besides the elaborated partnerships, the presented work has, to a large degree, been influenced by the technological context. It is a practical reality that mobile phone data analysis (as well as any big data analysis) is limited by the availability of technical capabilities to treat large datasets, in terms of infrastructure, software, and know-how.

Regarding infrastructure, Orange Labs has provided for two big data systems in which large datasets can be stored, safely accessed, and treated.

A first system concerns a small computer cluster, that was set up by a team of the university of Pisa in the summer of 2013. The cluster was located and managed locally (meaning, next to the office of the administration) and offered access to the inner working of the cluster so that operations could be investigated, paused, or even killed by any user. Despite being versatile, the disadvantage of this local system was that it was limited in storage capacity and computational power. The analysis in chapter 6, for example, is limited to one and a half month of data for the simple reason that this local system could not store and treat more data.

Parallel to this local system, the wider Orange Labs group has been developing a (very) large centralized, big data system. Throughout the years, this central system has been extended continuously, facilitating more and more complex tasks to be performed. The sensitivity tests performed in chapter 5, for example, became practically feasible starting from spring 2016 when a major extension of the cluster was undertaken following a move of offices.

The increasing demand for big data analytics within Orange Labs has undoubtedly accelerated the development of the central system but has had major implications for the work flow of individual researchers too. Initially, around 2013-2014, the central system was used by a handful of researchers only. Allocation of resources happened on a first-come-first-served basis and was sporadically regulated by a priority system (for example, once every week a billing analysis was performed on the cluster, overriding all other actions and allocating most of the resources on the cluster). When the demand for the system grew and more users entered the system, access to the inner workings of the cluster was restricted and the allocation of resources was regulated by proportionally distributing capacity of the cluster to different subgroups of Orange Labs.

The major drawback of the central system, however, came when an in-house developed user (web)interface allowed users with little knowledge of big data to start performing operations on the system. Possessing little to no experience in assessing or managing the resources that are needed for specific big data operations, these new users increasingly hindered the system's workings, obligating the team behind the central system to take on much more debugging and maintenance tasks. Between April and June 2016, this proliferation of inexperienced users on the system has lead to a constant blockage of the central system, forcing the administrators to undertake drastic measures.

One of these measures is a clean-up of the system each night at 00.00 hours, meaning that all tasks that are still running by that time are killed irrevocably. The consequence is that analyses which could normally run for days or even weeks, now have to run within the duration of a day before getting killed on the system. This has forced users to split their scripts into smaller chunks, which had led to more storage of intermediate results on the system leading again to more blockage. But the main consequence is that it has led to users inevitably losing around eight hours of processing time a day because, for security reasons, one can only access the central system when being physically present in the offices of Orange Labs.

Software and Methodologies

Besides the limitations in computation time imposed by the available infrastructure, the methodologies of this thesis has been influenced by the availability of software too.

In general, methodologies to study big data will deploy pre-processing steps at a big data system, after which selected or aggregated information is investigated in more detail by traditional tools such as Python scripts or Geographical Information System (GIS) software. The analytical possibilities of the pre-processing steps are limited by the available software on the big data system, mainly because these software have only very recently been developed, and are still under continuous development.

In the case of the central system of the Orange Labs, the open-source tool *Apache Hadoop* is used to govern the big data system. Within the Hadoop framework, different software projects have been developed over the last decade to handle different types of actions on big data sources. One of these projects is *Apache Hive*, which develops database querying on big data in a similar way to Structured Query Language (SQL)-querying of small databases. This software has been available at the Orange cluster from the very beginning of the PhD but is limited as, basically, it only facilitates counting, summarizing or querying of the data.

A second software project available at the central system is *Apache Pig*. Apache Pig, and its associated language Pig Latin, makes it possible to create and run high-level programs on Hadoop and can be extended by means of User-Defined functions (UDF), which can be written in traditional high-level languages such as Java or Python. Most of the work presented in this PhD has been done based on Pig Latin scripts extended with UDFs written by the candidate. The creation of mobile phone indicators, for example, happened by means of a Pig Latin script feeding the available mobile phone data to an UDF that was developed to calculate one (or several) indicators, subsequently feeding these indicators back into the system. The general development of Apache Pig was, and still is, underway, adding new capabilities and fixing problems with every new release. Consulting the version releases of Apache Pig, for example, one can observe that during the 3-year period at Orange Labs, six new versions of Pig were released. This is just to evocate that the deployment of of Apache Pig on the Orange central system was not always straightforward and that sometimes analysis was simply restricted by the state of the software.

One of the main limitations of Apache Pig (and in general, quite a lot of software projects in Apache Hadoop) is that it does not handle iterations on the dataset well. In other words, algorithms that would deploy for loops, or that would need frequent querying of the data are very inefficient in Pig, making it very time consuming, for example, to be performing interactive data exploration or machine learning tasks in this framework. Exactly for this reason, the software project Apache Spark was started. Deploying a better way to distribute data over a cluster of machines during analysis, Spark has managed to overcome this problem and has, ever since, taken the big data analytics world by storm. Its 1.0 version being released in 2014 only, Spark has been operational on the Orange Labs central system fairly recently (end of 2015). As a consequence, very little analysis has been done within Spark except for some of the sensitivity tests in chapter 5. One of the many advantages of Spark is that it can be coded in traditional syntaxes, such as Python, rendering the learning curve less steep than, for example, the one for Pig Latin.

The development of Apache Spark offer perspectives for a multitude of methodologies on big data that were not possible before. Machine learning, which forms a main focus of the Spark development at the moment, but also network analysis (although to a limited extent only) and potentially elements of spatial analysis all seem within reach due to this new development. As stated before, this type of advanced analysis currently happens outside big data frameworks using traditional tools on aggregated or selected data. The case studies presented in this thesis are no different. Because of the complexity of the pre-processing step, the advanced analyses presented in this thesis are limited to basic applications of statistics, spatial analysis and geo-computation. Other more advanced analysis techniques, such as machine learning and network analysis have been explored (and published) throughout the course of the PhD too, but will not be discussed in-depth because they concern too big of a sidestep to the general narrative presented in this manuscript. These contributions are listed with an '*' in section 1.5. Most of them are open access available online, and all of them are available upon request to the PhD-candidate or one of the co-authors.

1.4 Main Concepts and Outline of the Thesis

Building upon the question on the geographical veracity of mobile phone indicators, the presented chapters touch upon different topics and research fields. Here, a short description of the main concepts and research fields involved will help in understanding the structure of the thesis and the contribution of the presented works.

1.4.1 Mobile Phone Data Research

Although mobile phone data research is a rather young research field (starting from about 2006 only), it touches upon a remarkable number of methods, applications and research domains. In chapter 2, some of the main findings of the state of the art in mobile phone data research are reviewed. It is shown how mobile phone data enable the study of large-scale communication, movement, and presence patterns, and different applications are discussed. Despite its many scientific contributions, we discuss how the field of mobile phone data research has predominantly evolved around a computational social science paradigm. One of the main critiques on this paradigm is that it lacks contextualization, both of research practices and results. As one example of this lack of contextualization, we argue that mobile phone data could benefit from a confrontation of methods and results with spatial thinking. This idea forms the core of the thesis, and its value will be demonstrated throughout the subsequent chapters.

1.4.2 Mobile Phone Indicators

Recently, there has been an increasing interest in deploying mobile phone data as a complementary source to census data [46]. In this perspective, multiple applications have established the potential of mobile phone data. For example, spatio-temporal patterns of mobile phone activity have been explored to estimate population presence on a national scale [51], to quantify commuting patterns, or to detect long-distance and domestic tourism trips [78, 72, 141].

Pushing analysis one step further, research has initiated the creation of mobile phone indicators from individual users' mobile phone activity. The idea behind is that based on mobile phone data, indicators for calling, movement and sometimes even purchase behavior can be constructed for large populations. One advantage of mobile phone indicators is that they can more easily be reproduced at high-resolution and in a much more timely way compared to standard census data. In addition, there is a growing body of work that confronts mobile phone indicators, such as the amount of contacts or the amount of visited cell towers, with census data hereby unveiling interesting relations between both. For example, [54] show a clear relation between regional calling patterns and economic development in the UK. Similar relations have been uncovered in other studies, such as the relation between between calling and purchase behavior and food security in a Central African country [49], or between several mobile phone indicators (call, movement and purchase behavior) and multiple census variables on education, demographics and purchase power in a Latin American country [57].

In chapter 3, we explore to which degree these ideas are applicable to a French mobile phone dataset. We discuss how different indicators can be created from mobile phone data and what actions need to be undertaken to render them comparable to census data. Succeeding in this, we explore the relation between mobile phone indicators and socio-economic indicators on income. Our work is the first to have created mobile phone indicators in France, opening up possibilities for the integration of mobile phone data in official statistics in unforeseen ways.

1.4.3 Home Detection

Consequent to their findings, studies that uncover relations between mobile phone indicators and census data have suggested the possibility of using mobile phone indicators to *nowcast* or even predict census indicators. The problem with such claims is that they are presumptuous given that there still exists a knowledge gap regarding the understanding and interpretation of spatial patterns and variations of mobile phone indicators, including their representativeness and the spatial error that comes with their construction.

One of the reason for this knowledge gap is in home detection practices. Home detection of mobile phone users is crucial, because it is a prerequisite step to aggregate users in space and, eventually, render mobile phone indicators comparable with census data. Although multiple studies have discusses home detection from mobile phone data, and although a large part of mobile phone data research actively deploys home detection, the last part of the literature review in chapter 2 argues that current knowledge on performance, sensitivity and error estimations of home detection is of poor quality, which questions the trustworthiness of deploying home detection on mobile phone data.

To elaborate this position, in chapter 4 and chapter 5, we develop methodologies that, respectively, assess the performance of Home Detection Algorithms (HDAs) and explore their sensitivities to user decisions and data availability. To the best of our knowledge, our studies are the first to point out the limits of current home detection practices on mobile phone data, to assess nation-wide performance in different ways, and to document different sensitivities, amongst others to user decisions.

1.4.4 Diversity of Human Movement

A second example of the spatial knowledge gap on mobile phone indicators is revealed in chapter 6. Here, the focus is on one mobile phone indicator: the so-called Mobility Entropy (ME), which is a measure for the diversity of individual movement patterns. The ME indicator is of particular interest because it has been used in the groundbreaking work of [126] to evidence that human mobility at individual level, when captured by mobile phone data, is predictable to a very large degree. The ME indicator is also of interest because it shows one of the strongest relations with income variables, as will be shown in chapter 3 and 7.

Still, an investigation of the spatial pattern of ME values in France revealed how the indicator is dependent on the density of cell towers, which is highly variable across areas in France. The consequence is that patterns of ME are biased towards areas with higher densities of cell towers, such as city centers, obscuring the comparison across areas and leading to false interpretations. The study in chapter 6 is the first one to point out this bias and propose a simple correction of the ME indicator to a Corrected Mobility Entropy (CME) indicator as a solution. The subsequent investigation of the differences in spatial patterns between ME and CME shows a substantial altering in interpretation, revealing suburban areas instead of urban centers to have the highest overall diversity of movement patterns. One merit of this chapter is that it highlights simultaneously how little attention typically is given to the interpretation of spatial patterns of mobile phone indicators and how validation of indicators at a nation-wide scale remains cumbersome.

1.4.5 Urban Systems

Because they capture human movement and presence at an adequate scale, mobile phone data have been extensively used to study urban systems [83, 137, 155, 33, 140]. One outstanding question is whether mobile phone indicators can be deployed for this purpose too. Remarkably, in this perspective, no studies have yet investigated urban scaling laws of mobile phone indicators. Even though urban scaling laws are being widely deployed to understand the relation between census indicators and city size, population, or population density. The study in chapter 7 is a first to study urban scaling laws of mobile phone indicators. Following recent findings that urban scaling laws are sensitive to city definitions [10, 42], investigation is on the effect of city definitions on scaling laws too.

Reckoning that urban scaling laws of mobile phone indicators and census data are sensitive to city definition, the last question is on how stable observed relations between the two would be over various city definitions. As discussed before, multiple studies have uncovered relations between mobile phone indicators and census data but, to the best of our knowledge, all of them are based on one city definition only. The last part of chapter 7 investigates, for France, how sensitive such relations are to city definitions; as high sensitivities would have considerable implications for their validity and interpretation.

1.5 Scientific Contributions

Before listing the tangible scientific contributions, it is appropriate to mention the (international) collaborations that preceded them. Besides the collaboration between Newcastle University and Orange Labs France, collaborations have been established with third academic or institutional partners, typically excelling in a domain relevant for a specific research question. These partnerships have nurtured the interdisciplinary nature of the PhD study (and thesis) and have guaranteed its academic quality. A list of all collaborations is given in the acknowledgments.

1.5.1 Publications and Communications

Work undertaken during the PhD has resulted in multiple scientific publications and communications. Relevant publications are acknowledged at the beginning of each chapter. Other contributions might not be discussed in this thesis, but are added for the sake of completeness. They are indicated with an asterisk (*).

Publications in Peer-Reviewed Journals

In chronological order.

[142] **M Vanhoof**, F Reis, T Ploetz, Z Smoreda (2018) Assessing the quality of home detection from mobile phone data for official statistics. Journal of Official Statistics, 2018, In Press.

[143] **M Vanhoof**, W Schoors, A Van Rompaey, T Ploetz, Z Smoreda (2018) Comparing Regional Patterns of Individual Movement Using Corrected Mobility Entropy. Journal of Urban Technology, 2018, In Press.

[21] *J Beckers, **M Vanhoof**, A Verhetsel (2018) Returning the particular: Understanding hierarchies in the Belgian logistics system. Journal of Transport Geography, 2018, In Press.

[72] *M Janzen, **M Vanhoof**, Z Smoreda, KW Axhausen (2018) Closer to the total? Long-distance travel of French mobile phone users. Travel Behaviour and Society, 2018, 11, pp. 31-42.

[141] ***M Vanhoof**, L Hendrickx, A Puussaar, G Verstraeten, T Ploetz, Z Smoreda (2017) Exploring the use of mobile phone data for domestic tourism trip analysis. Netcom, 2017, 31-3/4, pp 335-372.

[67] *S Grauwin, M Szell, S Sobolevsky, P Hövel, F Simini, **M Vanhoof**, Z Smoreda, A-L Barabási, C Ratti (2017) Identifying and modeling the structural discontinuities of human interactions. Scientific Reports, 2017, 7, 46677.

[98] L Pappalardo, **M Vanhoof**, L Gabrielli, Z Smoreda, D Pedreschi, F Giannotti (2016) An analytical framework to nowcast well-being using mobile phone data. International Journal of Data Science and Analytics, 2016, 1(2), pp. 75-92.

Publications in Submission at Peer-Reviewed Journals

- *M Janzen, **M Vanhoof**, KW Axhausen, Purpose imputation for long-distance tours without personal information. International Journal of Geo-information.
- *B Sakarovitch, P Givord, M-P de Bellefon, **M. Vanhoof**, Hello, where are you? Estimating residential population with mobile phone data: an exploration. Economics and Statistics.

Publications in Preparation for Peer-Reviewed Journals

- C Cottineau, **M Vanhoof**, E Arcaute, Urban scaling laws of mobile phone indicators and their relation to census data.
- **M Vanhoof**, C Lee, Z Smoreda, Sensitivities of home detection practices for mobile phone data.
- *C Molinero, **M Vanhoof**, E Arcaute, Percolation of mobile phone call graphs.

Book Chapters

[140] * **M Vanhoof**, S Combes, M-P de Bellefon (2017) Mining mobile phone data to detect urban areas. In Petrucci, A. and Verde, R., editors, Statistics and Data Science: New challenges, new generations, SIS 2017, pages 1005–1012, Firenze. Firenze University Press.

Book Chapters in Preparation

- **M Vanhoof**, C Lee, Z Smoreda (2020) Performance and sensitivities of home detection from mobile phone data. In: C Hill and P Biemer, editors, Big Data Meets Survey Science, John Wiley and Sons.
- C Cottineau, **M Vanhoof**, E Arcaute (unknown) Urban scaling of mobile phone indicators and their relation to census data. In: M Batty, editor, working title: The CASA book.

Full Papers Contributed to International Conferences

In chronological order:

- *M Janzen, **M Vanhoof**, KW Axhausen (2017) Purpose imputation for long-distance tours without personal information. 96th Annual Meeting Transportation Research Board (TRB 2017).
- ***M Vanhoof**, L Hendrickx, A Puussaar, G Verstraeten, T Ploetz, Z Smoreda (2016) Extracting domestic tourism trips from mobile phone data and exploring the use of mobile phones during trips. Mobilités et (R)évolutions Numériques, 15th colloque du GT Mobilités Spatiales, Fluidité Sociale (MSFS 2016).
- *M Janzen, **M Vanhoof**, KW Axhausen, Z Smoreda (2016) Estimating long-distance travel demand with mobile phone billing data. 16th Swiss Transport Research Conference (STRC 2016).
- **M Vanhoof**, F Reis, T Ploetz, Z Smoreda (2016) Detecting home locations from CDR data: introducing spatial uncertainty to the state-of-the-art. 3th Mobile Tartu (Mobile Tartu 2016).
- W Schoors, **M Vanhoof**, A Van Rompaey, T Ploetz, Z Smoreda (2016) Correcting mobility entropy from CDR data for large-scale comparison of individual movement patterns. 3th Mobile Tartu (Mobile Tartu 2016).
- *M Janzen, **M Vanhoof**, Z Smoreda, KW Axhausen (2016) Closer to the total? Long distance travel of French mobile phone users. 3th Mobile Tartu (Mobile Tartu 2016).
- L. Pappalardo, **M. Vanhoof**, G. Lima, P. Paolini, Z. Smoreda, D. Pedreschi, F. Giannotti (2014) Sociality, Mobility and Wealth indicators: a study in France based on big mobile phone data. 2nd Mobile Tartu (Mobile Tartu 2014).

Short Papers and Abstracts Contributed to International Conferences

In chronological order:

- M Szell, S Grauwin, S Sobolevsky, P Hovel, F Simini, **M Vanhoof,** Z Smoreda, A-L Barabasi, C Ratti (2018) Structural discontinuities of human interactions in space. Netsci (Paris, 2018).
- **M Vanhoof**, T Ploetz, Z Smoreda (2017) Geographical veracity of indicators derived from mobile phone data. Netmob (Milano, 2017).
- ***M Vanhoof**, S Combes, MP de Bellefon, T Ploetz (2017) Mining mobile phone data to recognize urban areas. New Technologies and Techniques for Statistics (NTTS) (Brussels, 2017).

- **M Vanhoof**, B Sakarovitch, MP de Bellefon, V Loonis (2017) Home detection algorithms and mobile phone data. New Technologies and Techniques for Statistics (NTTS) (Brussels, 2017).
- **M Vanhoof** (2016) Lessons from the trenches: using mobile phone data for official statistics. ESS Big Data Workshop (Ljubljana, 2016).
- *J Beckers, **M Vanhoof**, A Verhetsel (2016) Assessing hierarchical boundaries in big data logistics. Nectar Cluster 8 Workshop 'Big data: a new opportunity for urban transport and mobility policies' (Sevilla, 2016).
- *S Combes, MP de Bellefon, **M Vanhoof** (2016) Mining mobile phone data to recognize urban areas. European Forum for Geography and Statistics (EFGS) (Paris, 2016).
- *S Rubrichi, M Mirkolesi, **M Vanhoof**, S Zmoreda (2016) Modeling and simulating the spreading and containment of Ebola in Cote D'Ivoire using mobile phone data. NetsciX (Wroctaw, 2016).
- **M Vanhoof** (2015) Scaling relations in constructed CDR-indicators. Annual Meeting of the American Association of Geographers (AAG) (Chicago, 2015)

Invited Talks and Seminars

In chronological order:

- ***M Vanhoof** (2018) Big Data and Geography. Guest Lecture for the Bachelor Course on Geographical Representations (Ku Leuven), (Leuven 2018).
- ***M Vanhoof** (2018) Mobile phone data, networks, and space. Seminars of the Complexity Science Hub (CSH), (Vienna 2018).
- ***M Vanhoof**, C Lee (2017) Mobile phone data and spatial statistics. Seminars of the Center for Urban Science and Progress (CUSP), NYU (New York, 2017).
- ***M Vanhoof** (2017) One hell of a bumpy ride: Using mobile phone data to study spatial and social systems. Lab Lunch Talk of the Senseable City Lab, MIT (Cambridge, 2017).
- ***M Vanhoof** (2017) Mobile phone data and the study of human mobility. Institute for Planning Transport and Systems (IVT) Seminars, ETH Zürich (Zürich, 2016).
- ***M Vanhoof** (2016) Big Data: Opportunities for geographical research. Advanced Geography and Tourism Seminars, KU Leuven (Leuven, 2016).
- *Z Smoreda, **M Vanhoof** (2015) Cartographier les interactions sociales via les données des mobiles. Seminaire de Sociology and Economics of Networks and Services department (SENSe), Orange Labs France (Paris, 2015).

Popular Scientific Communications

In chronological order:

- *__M Vanhoof__, C Ratti, Z Smoreda (2016) Dessiner les frontières de l'interaction sociale. Orange blog de la recherche (2016).
- *C Ratti, Z Smoreda, **M Vanhoof** (2015) Dessiner les frontières de l'interaction sociale. Usages et Valeurs (2015).
- *C Ratti, Z Smoreda, **M Vanhoof** (2015) Drawing boundaries of social interaction. Uses and Values (2015).

Chapter 2

Literature Review

He's been studying the books until his eyes grew weary but his mind got sharp.

Character in Rome Total War II

Abstract

This chapter discusses the state of the art in mobile phone data research. In a first section, it discusses how current mobile phone data research is related to the emergence of the Computational Social Science Paradigm (CSSP) and formulates some main critiques on this development. A second section describes the type of information captured by Call Detail Record (CDR) data and the ways this information can be mobilized to study calling patterns, movement patterns, and presence patterns at a large geographical scale. The next three sections summarize the main findings and applications of current research on CDR data with respect to human interactions, mobility and presence for large populations. A sixth section shifts the focus to the description of individual calling and movement patterns by means of mobile phone indicators and the way they have been deployed to understand, for example, relations with census data. Ultimately, a seventh section elaborates a critical discussion on the limits of current home detection methods for CDR data, which form an important prerequisite step for many mobile phone data studies. A final, eighth, section summarizes the main findings and their relations with following chapters.

Related Publications and Acknowledgments

Ordered by contribution to the chapter.

- Section 2.7 is a slightly reworked version of the literature review presented in [142]. This publication is authored by the PhD candidate. The literature review was written entirely by the PhD candidate.
- Sections 2.6.1, 2.6.2 and 2.6.4 are based on the introduction, methodology and experiments in [98]. This publication is co-authored by the PhD candidate. The sections in the original paper were co-written by Dr. Luca Pappalardo and the PhD candidate. More extensive discussion, including additional figures can be found in the paper.
- Sections 2.3.4 and 2.4.3, are inspired by the experiments performed by the Phd candidate during the preparation of [67], a publication of which the candidate is co-author. More extensive discussion on the experiments and additional figures can be found in the paper.
- Section 2.6.3 is based on [143]. This publication is authored and written by the PhD candidate, but builds on the master thesis of Willem Schoors: Mobility entropy and space: A CDR-based study of entropy behavior in France, which was presented at the department of Geography, KU Leuven in 2016 under co-supervision of the PhD candidate.
- Section 2.5.2 borrows ideas and a figure, from the conference paper [140]. This conference paper was authored by the PhD candidate and was written in collaboration with Stéphanie Combes who set up the research design, performed the main part of the analysis, and co-edited the paper.
- Section 2.6.4 is inspired by the methodology section in the paper in preparation: C Cottineau, M Vanhoof, E Arcaute, *Urban scaling laws of mobile phone indicators and their relation to census data.* This section was co-written by Dr. Clementine Cottineau and the PhD candidate.

2.1 Research and Mobile Phone Data

2.1.1 Captured Data for Selective Research

In 1946, George Zipf published a study on the influence of distance on communication [157]. Although the main argument of the work is on the diminishing value of news over distance, the empirical analysis on telephone and telegraph messages between 30 cities in the USA might very well have been the first study using phone logs, and can be seen as an early precursor of the blooming field of mobile phone data research arriving only 60 years later.

By now, 72 years after the original publication, a lot has changed. Phones have become ubiquitous devices that are carried around and used for a multitude of tasks by a large share of the world population. The extreme, and still accelerating, changes in information technologies have opened a set of functionalities for (mobile) phones such as wireless communications, affordable GPS tracking, and mobile connection to the internet that have undoubtedly altered the way we communicate and behave, as individuals and as a collective.

Recent developments in information technology have also enabled the capturing, storage and analysis of digital traces, whether produced by mobile phones, sensors in a fridge, online social networks, or satellite imagery, just to name a few. In this era, the treatment of digital traces forms the cornerstone of what can be deemed a data-rich knowledge industry [30] in which the analysis of data for optimization, strategic insights, or sales purposes has become a money-making machine. This is true, at least, for the (few) players in the market that manage to successfully combine technical infrastructure, a rapidly changing technological scene and the necessary human knowledge to do so.

Besides their clear role in business, large collections of digital traces are also significantly impacting the research domain. One way this impact takes shape is by the political economy of i) access to (high-quality) datasets and ii) the production of know-how needed to extract insights from datasets.

Currently, the key to the former, data access, is predominantly in the hands of private companies that perform the bulk of all data-harvesting worldwide. Despite both increasing, the minimum of regulations imposed by a selected number of governments worldwide and the limited public awareness on *digital data rights* are indicative for the hegemony imposed by private companies when it comes to deploying 'their' data. In this perspective, it is an intriguing observation that public institutions, such as official statistics offices, have only very recently started to take part in the data access debate. Still, the situation is that private companies decide on who gets access to large datasets of digital traces and what type of research will be pursued.

Concerning the latter, in the case of mobile phone data, most operators of mobile networks, have taken one out of two positions. Either they have actively been developing partnerships with (top) research institutions to pursue innovation, develop a research community, and exchange know-how, or they have not (yet) engaged with research except when, maybe, selling data. In either case, the consequence is that data access and knowledge development is accumulated within selected circles of private-academic partnerships. Clearly, this leads to a widening gap between research institutes, as it empowers the happy few only to set research agendas or collect academic rewards with regard to the newly available information. Such uneven accessibility, of course, serves neither the long-term public interest nor the quality, reproducibility, or sustainability of academic work [80, 66]. One example of such 'winner-takes-it-all' scenario outside academia is the observation that US security services and some singular private organisations have developed very sophisticated monitoring and analytical methods that already go beyond just mobile phone data but link to many other data sources.

2.1.2 Research Fields, Paradigms, and Critiques

Besides impacting the relations between research institutions, the availability of large datasets of digital traces is also drastically changing the relations between research departments, disciplines and even paradigms [80, 75, 139, 19, 99]. This is a development that is augmented by a wave of subsidiary mechanisms to support these new, exciting lines of research and that has, perhaps consequently, also attracted criticism.

A Multi-Disciplinary Field

In the case of mobile phone data, the very first studies appeared in the period 2006-2008 and were in the fields of urban planning, tourism studies, and mobility studies with methodologies deploying basic GIS, data science, and data visualization techniques [105, 5, 4, 2, 108]. Although the mobility studies and, to a lesser degree, the urban planning lines of research have continued to develop over time, the majority of studies nowadays are taking on a physics perspective deploying methods mainly derived from network science (as is very well exemplified by the literature review in [24]). The interesting bit is that, because of the richness of information stored in mobile phone datasets, such physics-inspired applications have diverged over a multitude of research fields such as epidemiology, environment studies, sociology, economic geography, transport studies, development studies or even psychology. The consequence is that there has been an *intrusion* of quantitative, empirical approaches in research domains that were not necessarily prepared (in terms of skills, know-how, or data access) to engage or collaborate with this new, well-funded, line of research. Such development has led to less truly interdisciplinary works than could be expected, and to the paradigm clash discussed next.

The Computational Social Science Paradigm

The predominance of a physics approach to large datasets of digital traces, such as mobile phone data, has given rise to a *Computational Social Science Paradigm (CSSP)* [36, 40, 80, 75] and to the resurgence of *Social Physics* [99, 19, 139].

The CSSP is emerging from the fact that, before, research on human behavior has relied largely on one-shot, self-reported data; whereas now, new data sources provide information over extended periods of time and at high resolution on the structure and contents of our society. As such, emerging technologies and data sources facilitate unprecedented observations at whole different levels, ranging from large populations of individuals to collective behavior, which cannot but lead to qualitatively new perspectives on society [40, 80]. The main premise of the CSSP, then, is in the belief that a data-driven, interdisciplinary approach, merging social scientists with computer scientists, mathematicians and physicists, can lead to the empirical description, theory building and model construction of social phenomena in ways that were not possible before [40]. Within the CSSP, the involvement of network science, which offers methodologies for the description of large-scale interaction datasets, and complexity science, which offers models for the emergence of collective phenomena from individual actions, is easily understood. Despite steady advances in the methodologies that underly the CSSP as well as in its institutionalization, the CSSP still faces multiple obstacles, such as data hegemony and the ever-looming problem of privacy [80].

Critiques on Social Physics

One main critique on the CSSP, is that it is resurrecting the idea of Social Physics which is:

> "[Social physics is] marked by the belief that large-scale statistical measurement of social variables reveals underlying relational patterns that can be explained by theories and laws found in natural science, and physics in particular. This larger epistemological position is known as monism, the idea that there is only one set of principles that applies to the explanation of both natural and social worlds." [19, p. 1]

The idea of monism is not new, as it was already present in Ancient Greek philosophy (Sorokin (1928) in [19, p. 2], but the idea of Social Physics has gained modern relevance only after being introduced in the 19^{th} century by French philosophers Henri de Saint Simon and Auguste Comte as a positivist approach to study social phenomena after which the term itself was coined by Belgian astronomer and statistician Adolphe Quetelet in 1835 [19].

The reason why the idea of Social Physics, and thus in extension the CSSP, is being criticized is because it is deemed overly positivist, and overly reductionist. There are many ways to phrase this critique (see for example [75, 19, 139, 66, 28, 116, 115]) but the following paragraph sums it up quite well:

> "People do not act in rational, predetermined ways, but rather live lives full of contradictions, paradoxes, and unpredictable occurrences. How societies are organized and operate varies across time and space and there is no optimal or ideal form, or universal traits. Indeed, there is an incredible diversity of individuals, cultures and modes of living across the planet. Reducing this complexity to the abstract subjects that populate universal models does symbolic violence to how we create knowledge. Further, positivistic approaches willfully ignore the metaphysical aspects of human life (concerned with meanings, beliefs, experiences) and normative questions (ethical and moral dilemmas about how things should be as opposed to how they are) (Kitchin, 2006). In other words, positivistic approaches only focus on certain kinds of questions, which they seek to answer in a reductionist way that seemingly ignores what it means to be human and to live in richly diverse societies and places." [75, p. 8-9]

The Social Atom Problem

One aspect of the critique thus is that in the CSSP individuals are considered uniform and interchangeable; as *social atoms* [29] that are "incapable to understand and control the phenomena at the global level" [144, 15]. In reality, however, this assumption is not true. People are agents that interpret, incorporate, articulate, negotiate, neglect or are forced by a complex set of processes that play at the different scales, ranging from individual decisions to small-group behavior and collective phenomena, and that are always situated in time and space. Studies that fail to acknowledge this diversification and complexity are prone to ecological fallacy or even to the critique that they "work only at the price of simplifying the properties of micro-agents, the rules of interaction and the nature of macro-structures so that they conveniently fit each other" [144, p. 2]

2.1.3 Moving Forward

Without denying the advancements made possible by the CSSP take on mobile phone data, critiques are urging CSSP practitioners to recognize the limitations of their practices and results by demanding contextualization in the form of epistemological reasoning and confrontation with, for example, existing theories, small data studies, auxiliary datasets, or historical and spatial situatedness [20, 43, 75, 66, 28, 116, 21, 115]. In this thesis we take on such a confrontation. We actively engage with the CSSP, as it remains the main paradigm in which mobile phone data are currently studied and understood, but our aim is to critically assess CSSP practices by confronting them with spatial thinking. As such, we aim at investigating the *geographical veracity* of some mobile phone data research practices.

We consider the confrontation with spatial elements an interesting starting point for several reasons. Firstly, geographical location data form an inherent attribute of mobile phone data (and many other digital traces), but their potential for either in-depth investigation of results or critical assessment of methodologies have been largely ignored within the CSSP. Secondly, the use of geography forms an easy way to impose a first layer of context to mobile phone data research. Especially the concordance with census data, that possess more or less similar study areas and resolutions compared to mobile phone data, is interesting in this perspective. The fact that public institutions, such as official statistics offices, are also starting to explore this convergence between mobile phone data and official statistics, provides an extra relevance to our work. The third reason is the absence of estimates of spatial error and uncertainties in current studies using mobile phone data. Albeit important for validation efforts and insight in error propagation, it is remarkable that little discussion exists on this.

The next sections provide a literature review on the main characteristics and advancements of recent mobile phone data research within a predominantly computational social science setting before exploring the geographical veracity of mobile phone indicators (section 2.6 and chapter 3), and then specifically the validity of home detection (section 2.7 and chapters 4, 5), the spatial patterns of the mobility entropy indicator (section 2.6.3 and chapter 6), and the relation between mobile phone indicators and census data for different urban definitions (section 2.6.4, chapter 7).

2.2 Available Information in Mobile Phone Data

Research using mobile phone data is, to a large degree, dependent on the information available in mobile phone datasets. This section briefly introduces the types of information typically stored in mobile phone datasets. We show that mobile phone data capture two dimensions of human activity for large populations: social interactions by means of calling patterns and human mobility by means of movement patterns. Subsequently, we discuss how literature has been mining the available information to create insights on both dimensions and at large, often nation-wide scales.

2.2.1 Two Dimensions of Human Activity

At the very basics, mobile phone data are the registrations of all actions performed by mobile phone users (from now on plainly called: users) on the operator's cellular network. From a data(base) perspective, this means that each action on th network is stored by a tuple holding information on the timestamp, a user identifier, the event type (call, text, mobile data), the duration of the network activity and a cell tower id as illustrated in table 2.1.

Event	Timestamp	Cell Tower	User	Duration
...	...	...	...	...
Call	2007/10/01 23:45:00	15988	33647956872	3656s
Text	2007/10/02 01:12:04	2051	3367261532	125c
...	...	...	...	...

Table 2.1 Example of records in a mobile phone dataset

2.2.2 Calling Patterns

Here, already, one can make a distinction between two types of mobile phone data: CDR (Call Detailed Records) and DDR (Data Detailed Records). The former, which are by far the most deployed datasets in literature, only capture call and text events. The latter, on the other hand, also capture events related to the exchange of mobile data packages that for example allow users to surf the internet. One main difference between CDR and DDR data is that CDR data typically consists of smaller volumes of information. CDR data are only collected when a user makes or receives a call/text which, generally, does not happen that often. DDR data, on the other hand, are larger in volume; mainly because users tend to leave their 2G, 3G, 4G connections on for some time periods, resulting in large amounts of DDR data.

CDR data and Ego Contact Networks

Another important difference is that CDR data, normally store the intended, addressed user (table 2.2). In other words, if user A attempts to contact user B (by call or text), CDR data will store the identifier of both user A and user B. This is not the case for DDR data, as no specific person is addressed when user A turns on its mobile data connection. The consequence is that contact networks can be constructed from CDR data. Indeed, for each user A, a historical record of CDR data can be used to construct a full calling pattern or so called ego-network of contacts. In such an ego-network, all users B,C,D,...K that have contacted user A, or have been contacted by user A are denotes as nodes while edges between user A and the contacted users B-K are weighted by, for example, the amount of calls that have occurred between them.

Event	Timestamp	Cell Tower	Location Area	Initiating User	Receiving User	Duration
...	...	...	...	...	...	...
Call	2007/10/01 23:45:00	15988	00080177U8	33647956872	33649274861	3656s
Text	2007/10/02 01:12:04	2051	00000001D1	3367261532	33632415523	125c
...	...	...	...	...	...	...

Table 2.2 Example of Call Detailed Records

Mobile Call Graphs

Evidently, one can create ego-networks for all separate users in a CDR dataset. One can also overlay multiple ego-networks, creating one large-scale contact network. Figure 2.1, for example, shows an illustration of the overlaid ego-networks of multiple users. When done for all users in a CDR dataset, such large-scale contact networks offer a unique insight on the structure of social interactions in a large populations. In line with previous studies, we will call such large-scale contact networks Mobile Call Graphs (MCGs).

Cell Tower Call Graphs

Ultimately, one can investigate large-scale contact networks from the perspective of cell towers instead of that of the user. In a cell tower perspective, the contact network consists of cell towers as nodes with edges representing the amount of calls that have been observed between two cell towers. Note that it is possible to derive this information from CDR data because a record of an outgoing call from user A to user B done on cell tower X, can be related to the record of an incoming call from user B at cell tower Y, which enables the counting of calls between cell tower X and Y. Consequent to previous (and following) terminology, we will call this type of interaction networks Cell tower Call Graphs (CCGs).

Both ego call networks, mobile call graphs, and cell tower call graphs are used to study a first dimension of human activity that is captured by CDR data: human communication.

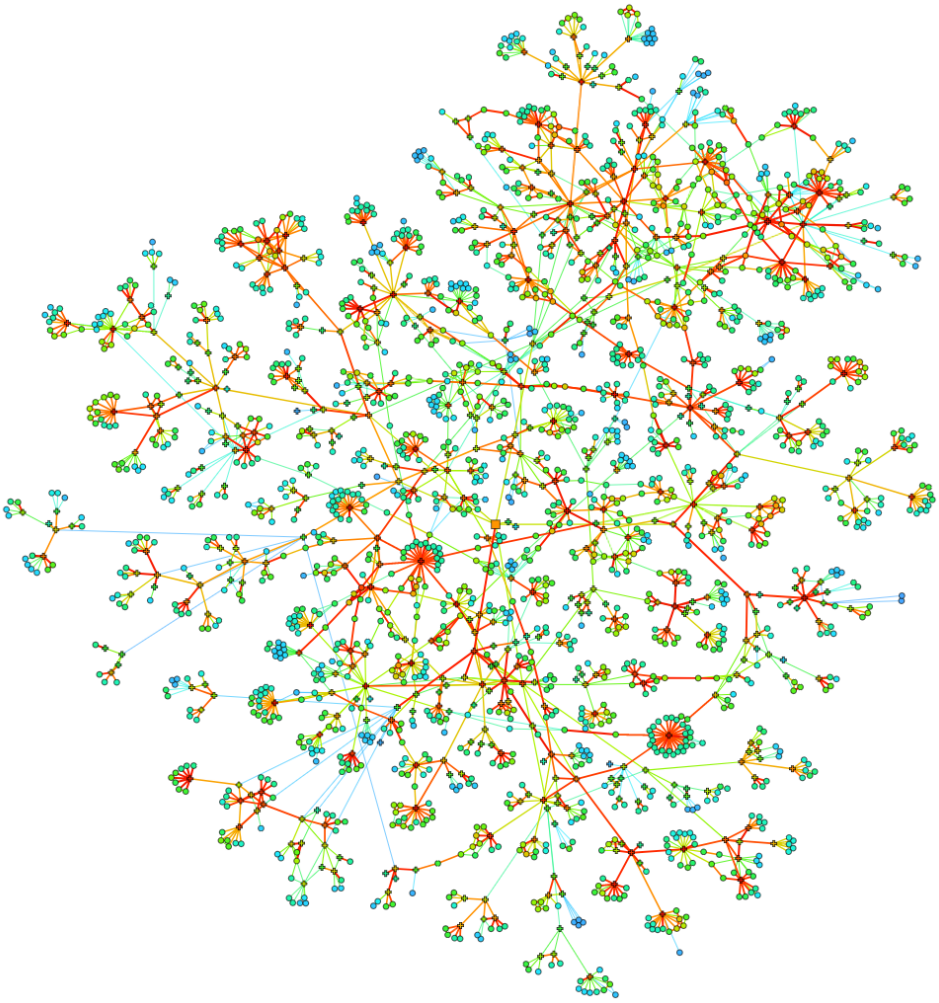


Fig. 2.1 Illustration of a contact network constructed from CDR data. The network is sampled by snowball sampling from one user (the orange square in the middle), and captures all other users that are within a distance of 5 nodes. For nodes denoted by '+' the entire ego-network is shown, for nodes denoted by 'o', only some of their nearest neighbors are visible, while the rest are outside the sample. Edges are colored from yellow (weak link) to red (strong link) based on the total duration of calls that occurred between two users. Source: [92].

2.2.3 Movement Patterns

A second dimension of human activities captured by mobile phone data (both CDR and DDR data) is human mobility. Most mobile phone data store information on the cell tower that is deployed. In general, network operators possess the location of each cell tower in their network, as well as an estimation of the covered area by each cell tower (often estimated by means of the Voronoi polygons created from the cell tower location points). This means that based on a historical record of mobile phone activities the different areas a single user has visited over time, thus a movement pattern, can be reconstructed.

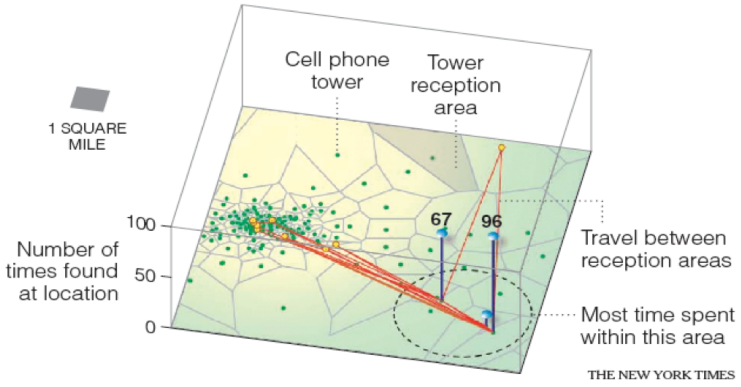


Fig. 2.2 Illustration of a movement pattern for one user as constructed from CDR data. Note the differences in cell tower density between regions, as well as the reference to meaningful locations based, in this figure, on the number of activities. Source: Illustration reworked by the New York Times from [65].

It is important to note that, when based on CDR data, the resolution of such movement patterns is dependent on two factors. First, the temporal resolution of the movement pattern is entirely dependent on the number and distribution of mobile phone activities performed by a user. Users that do not call or text for several days will not have CDR records during this period, leading to a sparser movement pattern. Secondly, the spatial resolution of the movement pattern is dependent on the location of cell towers and thus the size of the covered areas per cell tower. Places that have a high density of cell towers, such as city centres, have a higher spatial resolution compared to, for example, rural areas where cell tower density is typically low. An illustration on how CDR data capture the movement pattern of a single user is given in figure 2.2.

It is common practice in literature to study movement patterns by means of networks. The network that describes movement patterns of the users have cell towers as nodes (as opposed to call networks which have users as nodes) and edges are made up of displacements. Displacement, from this perspective, is defined as the consecutive observation of a single user (say A), on two cell towers (say X and Y). In other words, if user A has been active at cell tower X at time t, and the very next observation at time t+i is observed at cell tower Y, then a displacement has been made by user A between cell towers X and Y, which will implement an edge between X and Y in the movement network. Note that this definition of displacement is regardless of the magnitude of i, meaning that even if there are multiple days between the observations at cell tower X and Y (offering user A plenty of time to roam around the country), the only displacement that is captured is the one between cell tower X and Y.

Similar to ego-networks of calls, the movement networks of individual users can be aggregated into on large-scale movement network. Remember that movement networks have cell tower as nodes and displacement between cell towers as edges. Similarly, a large-scale movement network consists of nodes representing cell towers and edges being weighted by, for example, the sum of displacements as performed by all users in the CDR dataset [70]. Consequent to previous nomenclature, we will continue to call such networks Cell tower Movement Graphs (CMGs). Figure 2.3 illustrates a CMG for France based on the aggregation of movement patterns from the most visited to the second most visited cell towers for all users in a CDR dataset (the same dataset as we will deploy throughout the thesis). It is easy to observe that CMGs have the potential to capture movement patterns of large populations and for large territories, making them extremely useful tools in the large-scale study of human mobility.

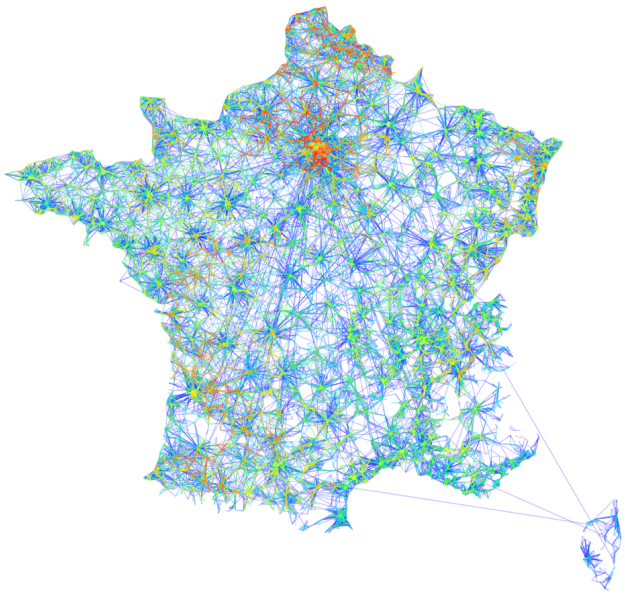


Fig. 2.3 Illustration of a Cell tower Movement Graph (CMG) for France based on one month of CDR data (September 2007). Edges are the aggregated number of users that have a most visited and second most visited cell tower between two locations, and are colored by count (blue=low, red=high). Figure made by Dr. Carlos Molinero in collaboration with the PhD candidate.

2.2.4 Temporal Patterns

Related to the movement patterns, an important aspect for the analysis of CDR data is the frequency of location (cell tower) visits. Although CDR data are insufficient to define a user's staying time within the areas covered by a cell tower (there only is a record when a call or text is initiated), the frequency and recurrence of activities performed by a user on specific cell towers offer proxies for staying time, which in turn offer possibilities to derive meaningful locations for individual users, and maybe even annotate them. In section 2.7, for example, we explore how this information can lead to the detection of home locations from users.

2.3 Studying Human Interactions at Nation-wide Scale

2.3.1 Topological Properties of Contact Networks

In the very beginning, interests in large-scale contact networks extracted from CDR data were stirred by the realization that MCGs (section 2.2.2) capture information on a much larger scale than ever before (questionnaires typically reach a maximum of 10^3 users, where CDR data collect data in the order of 10^6 users [92]), which enable to simultaneously study the local and the global structure of nation-wide communication [93]. Over the years, this realization has led to extensive research on the topological properties and statistical physics of MCGs [53, 89, 92, 117, 34, 117, 24].

In the following paragraphs, three topological properties are discussed that are found to characterize MCGs: a broad degree distribution, a small diameter, and a high clustering coefficient at the local level [53, 63, 117, 89, 92, 24, 79].

Broad Degree Distributions

The degree of a node in a network is the number of other nodes it is connected with. In terms of call networks, the degree of a node is the amount of contacts a user has. Investigating the distribution of degrees over all the nodes in MCGs typically shows a right-skewed distribution that can be approximated by power-law distributions [79, 53], or other distributions such as exponential forms or the Double Pareto LogNormal distribution [117]. The interpretation of these broad, right-skewed distributions is that, in MCGs, many users have a low degree, while a few users have very high degrees. In other words, many mobile phone users will have a limited number of contacts while a few users will have a large number of contacts. The consequence is that there exists no clear way to summarize the degree of all users in a MCG because the broad distribution implies large statistical fluctuation around the average [24]. The MCG network is said to be scale-free [17], which is an indication that the network represents a complex system in which individual, local actions lead to emerging large-scale phenomena that cannot be described by local properties only.

Small Diameters

Another property of MCGs is the small diameter, a property relating to the so called small-world phenomenon (also popularly known as the six-degrees of separation) [148]. The small-world phenomenon, as for example described in [138], is the observation that in our connected world the fastest way to connect one person to any other person runs only over a limited number of *in-between contacts*. In the original experiments, which were carried out by Milgram and Traver in the USA at the end of the 1960s, the average number of in-between contacts to connect a person in Boston or Nebraska to any person in Massachusetts was only 5.2 [138].

The small-world phenomenon can easily be translated to MCGs, as it is the number of nodes that lie between one node and any other node in the network; or thus the average distance between all the nodes in the network (with distance expressed by the number of in-between nodes). In MCGs, this number is found to be surprisingly low, too usually scaling logarithmically with the total number of nodes [89].

Local Clustering

One last property of MCGs, is their high (local) clustering coefficients, which are a quantification of the probability that two nodes that are connected to one node, are connected between themselves too. For contact networks, high clustering coefficients mean that contacts B and C of a user A are more likely to be contacts amongst themselves (so B calling C) than any two users I and J in the network are contacts [63]. Starting from high local clustering coefficients, the idea arises that calling patterns might reveal that users are grouped together in small cliques, groups, and eventually larger communities. This intuition turns out to be true as MCGs are found to be locally dense, displaying small cliques with strong ties and connections within local dense communities that are made up of fewer, and weaker ties [94, 79, 47, 93]. This observation has led to the development of community detection algorithms which aim to reveal the larger communities that make up MCGs, all in the quest for insights on nation-wide communication patterns.

2.3.2 Community Detection on Contact Networks

Community detection aims to reveal the large scale organization of MCGs. A community, in terms of contact networks, is generally defined as a subgroup of users that are in closer contact with each other than they are with others. Applying community detection on MCGs is a challenging task because of their large size. Different algorithms have taken the task at hand, such as the Combo algorithm [123], the Louvain-la-Neuve algorithm [25] or the Infomap algorithm [112], but all differ in their specific definition of a community (as is typically implied by means of an optimization function in the algorithm) and consequently in obtained results (see for example the differences in result between the Combo and Louvain-la-Neuve algorithm whenn applied to a UK MCG in figure 2.4).

Additionally, the wide distribution of degrees, combined with high cluster coefficients in MCGs, gives rise to tree-like structures which pose substantial challenges for existing community detection algorithms [133, 24]. The consequence is that resulting communities are dependent, to a large degree, on the chosen detection method. And that their *validity* is only given by the plausibility of researchers' interpretations of the obtained results [24]. Still, knowledge on the general organization of MCGs offers interesting insights, especially when combined with external data sources which can aid in understanding and characterizing communities [24].

Potentially the most well-known example of this is the analysis of a Belgian MCG, where communities resulting from the Louvain-la-Neuve algorithm coincide exceptionally well with the existing language barrier between the French-speaking and Dutch-speaking parts of the population [25]. Building on this finding, [79] have investigated the influence of distance on calling patterns between users. Unsurprisingly, they found the probability that two users have called or texted each other is inversely proportional to the square of the distance between them [79]. While communities resulting from community detection algorithms might provide insights on the structure of MCGs, they offer little information on the processes creating these structures. After all, the linguistic divide in Belgium observed by analyzing the MCG is also due to French and Dutch speaking Belgians living in different parts of Belgium.

2.3.3 Spatial Communities in Contact Networks

The influence of distance on calling patterns has inspired researchers to investigate the spatial characteristics of contact networks. Here, however, the focus is on the interactions between cell tower areas, or CCGS equivalently, instead of between individual users or MCGs equivalently (see also section 2.2.2).

An intriguing aspect of CCGs is that, when subjected to community detection, they tend to produce spatially homogeneous communities [106, 123, 124, 79, 67], as can also be observed in figure 2.4 where the Combo and Louvain-la-Neuve algorithm were deployed on a UK CCG (source: [123]). In other words, grouping cell towers by means of the strength of their calling interactions results in communities which, when mapped, are situated together in space. This spatial grouping effect is exaggerated by the very methodology of community detection in the sense that most community detection methods force each node to be in one community only, while the reality is that they belong in different degrees to different communities simultaneously. Still, the resulting communities and the boundaries between them have the capability to reveal how political, socio-economic or cultural boundaries impact human interactions [67]. They form a reflection of the shared social capital between regions which is interesting for, for example, regional economic policy or metropolitan planning.

2.3.4 Spatial Interaction Models for Human Communication

By clustering at the macroscopic level, community detection offers a partial view on the interactions between regions only, mainly because identified communities typically correspond to centers of high activity (large cities), and their related hinterland [67, 106, 124]. The regularity in communication activities that lie at the basis of these clusters, and that consists of both short-range and long-range interactions, have been the focus of interaction models. Much like Zipfs' empirical effort to quantify the underlying principle of diminishing relevance of news over distance [157], interaction models for human communication try to capture (and reproduce) a quantifiable principle that would enable to study how high-level spatial communities relate to low-level communication patterns.

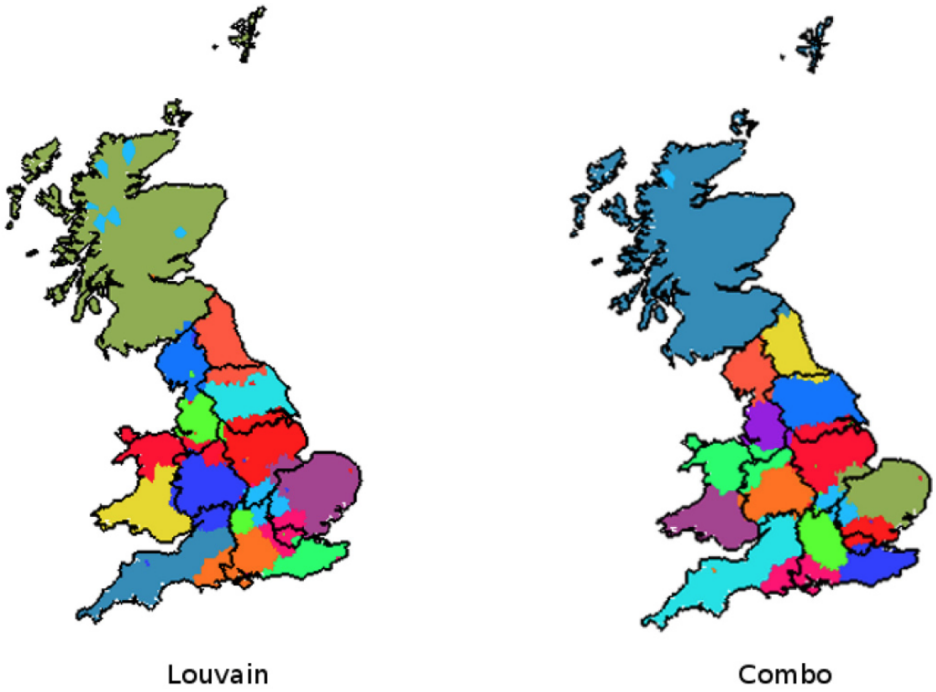


Fig. 2.4 Resulting spatial communities (colored) after performing the Louvain-la-Neuve and Combo community detection algorithm on a Cell tower Call Graph (CCG) in the UK. Source: [123].

Multiple interaction models have been deployed to better understand the patterns in CCGs, amongst which are the well-known gravity model [76, 55] and radiation model [122, 67]. Many of these models produce acceptable results when modeling large scale communication networks [67]. Recently, however, a new and simple model has challenged the idea of continuous distance used in other models. By replacing continuous distance with a discrete distance measure based on the boundaries of detected spatial communities, the hierarchy model has outperformed other models for multiple CCGs in different countries, showing the importance of approaching space as a complex product that is lived and performed by users rather than reducing it to a distance metric [67]. Figure 2.5 shows the results of performing community detection either on a CCG dataset in France or the modeled interactions by means of the gravity, radiation and hierarchy model, showing that the results from the hierarchy model align better with the observations from the CCG. Note that, the CDR data that was used to create this CCG for France is the same dataset as will be deployed throughout this thesis.

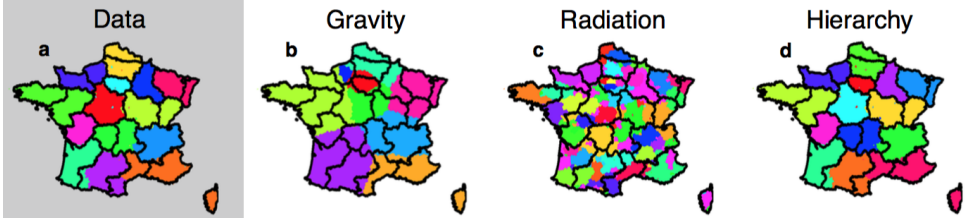


Fig. 2.5 Resulting spatial communities (colored) after performing the Combo community detection algorithm on (a) a source CCG from France and results of multiple interaction model: gravity (b), radiation (c) and hierarchy (d) that try to model the communication patterns observed in (a). Source: Figure created by PhD candidate in collaboration with Dr. Sebastian Grauwin, head author of [67].

2.4 Studying Movement Patterns at Nation-wide Scale

2.4.1 Statistical Properties of Movement Patterns

Similar to topological properties of contact networks (section 2.3.1), research has revealed statistical properties of movement patterns captured by CDR data. Main findings with regards to individual movement patterns are power-law distributions for the amount of displacements (expressing the heterogeneity of individual movement patterns across a population),as well as visitation frequency distributions and, accordingly, staying time distributions that are dominated by few locations. [7, 65, 126, 125, 131, 145]. The latter two properties indicate that people tend to spend long times in a small amount of locations, which points to the predictability of human movement patterns that will be discussed in more detail in section 2.6.3. They are also related to the possibility to derive meaningful locations such as the home location as discussed in section 2.7.

2.4.2 Predicting Individual Movement

Inspired by statistical properties of individual movement patterns, predictive models for individual movement have been developed. These models were first basing on traditional random-walk models, such as Levy-flight or Brownian Motion, but have over time evolved into models, such as the explorer-returner model, that better account for the statistical characteristics of individual movement patterns [125, 97, 145]. This evolution was necessary, as previous models to predict individual movement were found to be insufficiently capable of reproducing movement patterns at different spatial scales. One reason for this is that human movement as derived from digital traces is scale-related meaning that, for example, properties differ between long-range mobility and short-range mobility, or between urban movement and intra-urban movement [88, 8, 90].

Another reason is that mobility derived from digital traces is rarely studied in-depth and that neither the movement patterns of sub-populations nor the movement patterns related to singular modes of transportation are found to reproduce the statistical properties of human mobility observed for entire datasets of digital traces [145]. While it is not surprising that this observation is in line with the social atom critique on the CSSP, the consequence is that questions persist on the effect of scale, inter-modality, urban structure, and the limitations of existing datasets to study individual human mobility or the mobility of sub-populations [18].

2.4.3 Spatial Interaction Models for Human Mobility

Notwithstanding the anomaly between properties of individual, subpopulation, and general human mobility, spatial interaction models have been trying to describe the latter. In analogy with the interaction models for human communication (section 2.3.4), spatial interaction models have been developed to estimate the number of displacements between locations that are described in CMGs (section 2.2.3).

There exists many models that try to explain the underlying mechanisms of general human mobility. For example [145] make a division between seven sorts of models: descriptive, exploration of new locations and preferential returning, the effect of few dominant trips, hierarchical traffic systems, spatial heterogeneity, radiation models and aggregation of individuals without scaling properties, concluding that:"These models can reproduce parts of the empirical findings. Nevertheless, it is difficult to identify common rules from these models, and thus it remains controversial what drives the emergence of these abnormal properties in human mobility." [145, p. 2].

The general structure of spatial interaction models for human mobility is that they derive the number of displacements between locations as a function of local attributes (such as population density or income per capita) and a measure of distance. If needed, the spatial distribution of the number of displacements can be based on an entropy maximization principle first proposed by [151] that can be subjected to various constraints, such as the number of people living in the locations, proxies of attraction for location, or travel costs for users [151, 152, 18].

With respect to the distance measure, [18] makes a distinction between models that use continuous distance and models that use the concept of intervening opportunities. The former, much like Zipf's original work [157], assumes the number of displacements between locations as a decreasing function of distance. The latter, of which the radiation model is a main example [122], assumes that the number of displacements between any two locations is in relation with the number of *potential destinations* between both locations. The conceptual difference between both is illustrated in figure 2.6.

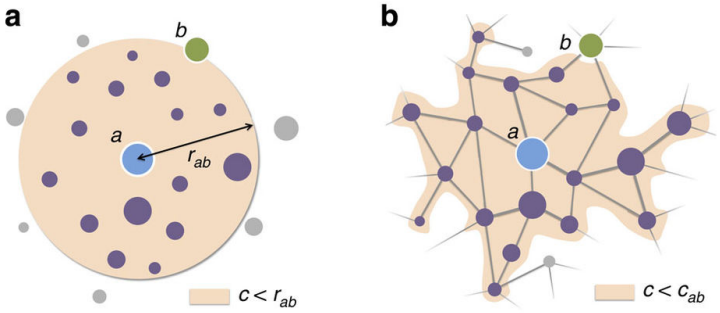


Fig. 2.6 Illustration of the difference between (a) the notion of distance, expressed here in the radius r_{ab} and (b) the notion of intervening opportunities, expresses here by the network based cost c_{ab} as deployed in spatial interaction models for mobility. Source:[110].

Applications

The availability of large-scale movement patterns in combination with spatial interaction models for mobility has opened up a multitude of applications. In transport studies, there exists an entire research field on travel behaviour [146] and traffic management and monitoring [130] based on mobile phone data. Complementary, MCG and spatial interaction models have been used to quantify and predict travel demand within and between cities [110, 74, 52, 155, 9] or to derive information on types of mobility that are hard to survey, such as long-distance travel [72, 82].

In epidemiology, estimations of movement patterns by empirically collected CMGs or constructed by interaction models, have served as a major input to meta-population models that simulate nation-wide disease spread and that offer insights on both temporal and spatial patterns of disease propagation [101, 113]. After all, the probability that users from an infected area come in contact with users from other areas is a determining factor in the (spatial) dynamics of infectious diseases [81, 87, 113]. Another application is in urban planning where MCGs are used to understand different functionalities of cities or large metropolitan areas. Filtering MCGs to specific timeframes, for example, facilitates the investigation of commuting patterns, which in turn can be compared between cities [78].

Shortcomings

One of the shortcomings of spatial interaction models for mobility is that they do not offer a framework on how local attributes relate to individual behavior. One main reason is that the displacement numbers between locations, which are captured in CMGs and form the objective of most spatial interaction models, can very well be aggregated from individual movement patterns but the inverse, the disaggregation to individual users, is not possible [18]. As such, differentiation between individual users based on their own characteristics or the relation with local attributes is impossible, again giving rise to the social atom problem mentioned in section 2.1.2.

Complementary to this shortcoming is the observation that there has been very little attention to behavioral and contextual aspects of human activities, both for movement and calling patterns, even though several authors have expressed considerations in this direction [35, 73, 131]. In the case of movement patterns, for example, there is a need to recognize the complexity of individual decision making (which is becoming increasingly flexible and fragmented [50]), and how this results in observed movement patterns. One promising line of research in this direction, is on the importance of an intra-personal perspective on mobility [135, 100].

Based on CDR data, which offer the unique opportunity to combine both movement and calling patterns, recent research has produced several indications for the link between social interaction and human mobility at large scale [100, 52, 135, 8]. Still, insights remain limited on what factors really drive movement and what role social interaction plays herein, especially at the individual level.

2.5 Studying Presence Patterns at Nation-wide Scale

2.5.1 Seasonal Changes and Tourism

Because CDR data register activities of large populations at high temporal and spatial resolution they capture temporal and spatial patterns of user presence, which in turn can be used to estimate population densities [51, 111]. The study of presence patterns and changing population densities has different applications at several scales. In [51], for example, seasonal differences in population presence between the main holiday period (July and August) and working periods (September and June) in France reveal the extent to which a large share of the French population is mobile during holidays (figure 2.7). Such and similar bird-eye views on large population presences and their changes over time offers unprecedented insights in patterns and processes that shape a territory.

Unsurprisingly, seasonal changes in population presence has formed the cornerstone for tourism studies by CDR data. In Estonia, for example, a long tradition of studying tourism based on presence patterns exists. Studies range from the investigation of seasonal tourism spaces, over the quantification of tourism destinations, up to the assessment whether mobile phone data can replace or complement tourism surveys [2, 3, 121, 107]. Similar directions have been investigated in different countries and it is no surprise that many operators[1] offer paid services for the mapping of population presence for large territories.

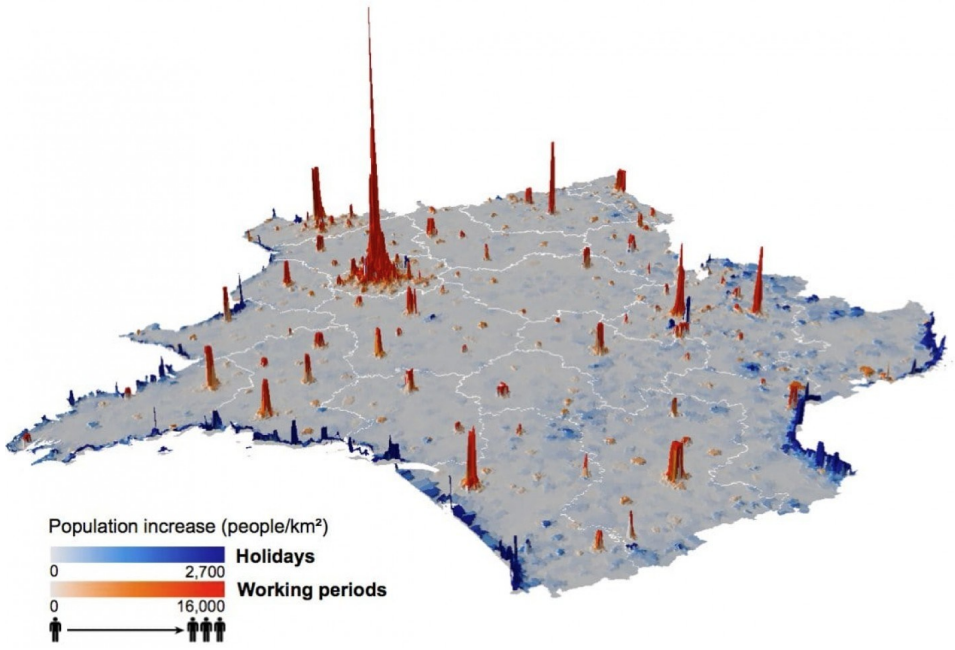


Fig. 2.7 Illustration of the seasonal changes in population presence in France based on CDR data. The CDR dataset is the same as will be used in this thesis. Source: Figure from Washington Post, created by Catherine Linard, co-author of the original paper: [51].

[1] From past experience, we can name at least three operators or affiliated companies that offer such services: Fluxvision for Orange in France, Positium for multiple operators in Estonia, and Proximus in Belgium.

2.5.2 Temporal Presence and City Structures

On the scale of cities, presence patterns can highlight city structures. In [83], for example, the focus is on the most crowded places in 31 Spanish cities (so-called hotspots) and their temporal patterns of user presence derived from CDR data. By investigating temporal presence patterns at these hotspots, the authors distinguish different categories of cities, ranging from mono-centric to poly-centric cities. Also at city level, the correlations of user presence between places has led to an increased understanding of the working of cities. In [137], for example, a technique to uncover correlations in temporal user presence between locations is discussed and it is shown how such correlations can detect connections between places such as the Charles de Gaulle airport and train stations in Paris. Such insights are of great value for urban planning and policy as they describe the day-to-day use of spaces and the connections between locations both of which, in traditional surveys, are hard to come by.

At a more regional scale, the growth of cities, and with it the extension of urban agglomerations, has become characteristic for contemporary times. In many countries, both developing and developed, urban growth has been sparsely monitored with consecutive population counts and urban area zonings being published by official statistics once every few years only [140]. In this context, population presence figures from CDR data can contribute in a meaningful way because they capture more recurrent observations of the territory.

Along this line of thinking, research has revealed how machine learning techniques can bridge between spatio-temporal presence patterns and land use, opening opportunities for the identification of functional areas in cities based on presence patterns [136, 1, 127, 140]. In [140], for example, the authors explore different machine learning procedures to recreate current urban area zonings in France (as produced by official statistics) based on the temporal signature of CDR activities at cell tower level. They find their techniques to perform well in areas with higher cell tower and user densities, such as urban centres or suburbs, but less in more isolated areas (see also figure 2.8) making them conclude that:

> "... our results encourage us to promote the use of mobile phone data as an alternative source for producing recurrent urban area zoning between official but less frequent releases. Specifically, We are quite optimistic on using supervised classification to, for example, show patterns on the emergence of urban centers or the progression of urban areas. We reckon, however, that assessments regarding the urbanization of rural and isolated areas should remain cautious as our classification tasks underperformed there." [140, p. 1012]

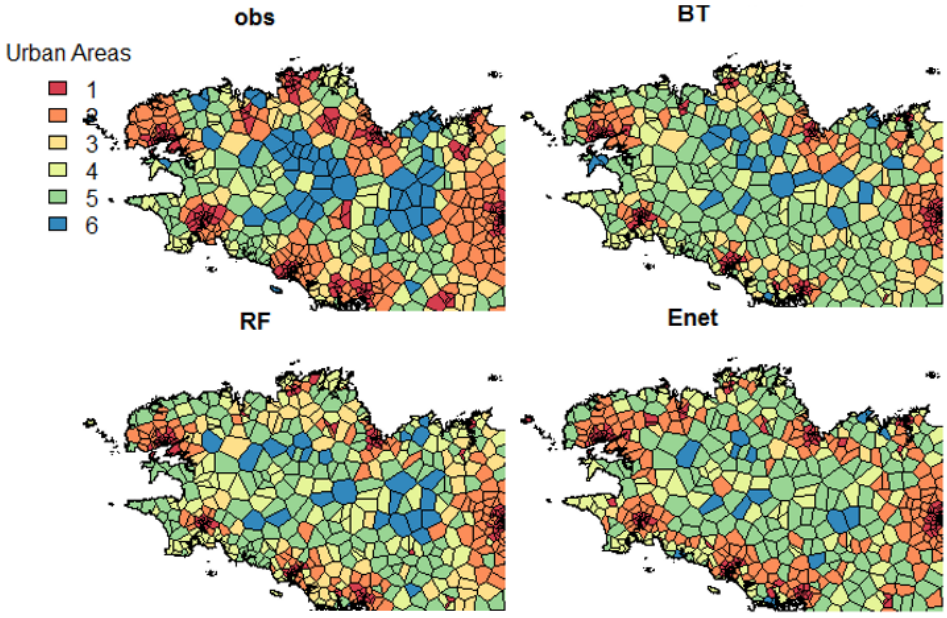


Fig. 2.8 Results of machine learning experiments to reproduce the official French Urban Area Zoning (obs) for the region of Normandy based on machine learning techniques (Random Forests (RF), Boosting Trees (BT), and Elastic Nets (Enet) that used temporal patterns of population presence at cell tower level as input. The official Urban Area classes are: Major urban centers (1), surrounding of major urban centers (2), multi-polarized municipalities in a large urban area (3), medium and small urban centers and their surroundings (4), multi-polarized municipalities (5), and isolated municipalities (6). Source: Figure created by Stéphanie Combes in co-operation with the PhD candidate for [140].

The idea that CDR data can, to a certain extent, be used to monitor changes in urban systems carries great potential for developing countries where urban growth is occurring extremely rapid and proper census data often is absent. Specifically because of this absence of census data, however, population presence patterns derived from CDR data in developing countries are in need of proper validation mechanisms like surveys or confrontation with other data sources such as satellite images [147]. In this perspective, the Worldpop project [2] does an amazing job collecting and combining both data sources and actors to arrive at dynamic population mapping in developing countries [132]. Amongst many other applications, their works, including temporal presence patterns of mobile phone data, help to better understand large-scale population movement because of political conflict in Ivory Coast [23], to reveal migration due to extreme weather events in Bangladesh [85], or to estimate poverty patterns in Senegal [129].

[2] http://www.worldpop.org.uk/, see also: [132].

2.6 Individual Indicators from Mobile Phone Data

Besides the analysis of large-scale patterns of communication, movement, and population presence, CDR data can also describe individual level patterns by means of summary measures, which are called *mobile phone indicators*. The following sections discuss which mobile phone indicators can be constructed from CDR data and how they have been deployed in research.

2.6.1 Indicators for Individual Calling Patterns

The open-source Python toolbox named Bandicoot [3], offers a good starting point to explore indicators that can be potentially derived from mobile phone data [48]. Table 2.3 lists some indicator with respect to calling patterns. Most of the *calling indicators* can be easily calculated and interpreted for individual users. For example, the number of calls, the duration of calls, and the inter-event time between calls can be created for individual users and, by time period (e.g. weekday and weekend) or by levels of reciprocal behavior of contacts.

Mobile phone indicator	**Description**
Number of calls	Number of calls made or received
Number of active days	Number of distinct days in which calling was observed
Percentage nocturnal calls	Percentage of calls made between 7pm and 9am
Duration of calls	Summary statistics of duration of all calls from one user (mean, median or standard deviation)
Inter-event time	Summary statistics of the duration between consecutive calls from one user (mean, median or standard deviation)
Number of contacts	Number of distinct contacts interacted with
Interaction per contact	Summary statistics of the amount of interactions per contact (mean, median or standard deviation)
Entropy of contacts* (eq.2.1)	Entropy measure of calls to contacts

Table 2.3 Mobile phone indicators based on calling behavior and contact networks

The computation and interpretation of calling indicators becomes exponentially complex when one tries to calculate them for second-order networks, meaning that one would extend the computation of an indicator to activities of all contacts of a single user (a second order network for a user u is illustrated in figure 2.9). This increasing complexity forms one of the main reasons why mobile phone indicators are typically calculated at the individual level only. Note that, on the other hand, grouping of individual indicators values is rather easy and therefore forms a main practice for treating mobile phone indicators.

[3] Bandicoot was developed at MIT in a collaboration with, among others, Orange Labs. See http://bandicoot.mit.edu/

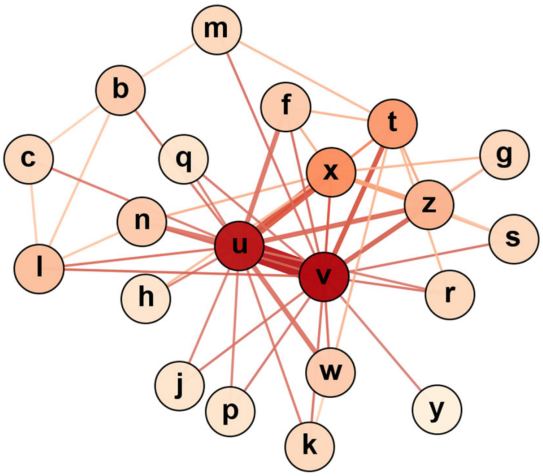


Fig. 2.9 Illustration of a second-order contact network constructed from CDR data. The network is based on the activities of a single user u and all its contacts (v,w,x,...). Nodes represent users, edges indicate reciprocated calls between the users, the thickness of the edges is proportional to the total number of calls between the users. Source: [98].

Entropy of contacts

The entropy of contacts deserves some more attention as it is a more complex measure. It was first introduced in [54] and expresses the diversity of an ego-network of calls pattern (such as the one illustrated in figure 2.9 but then taking into account only the interactions that start or end at user u). Based on an ego-network, the entropy of contacts is a measure that balances how often different contacts were called (for example, user v in figure 2.9 has been called the most by user u) and how evenly calls are distributed over all contacts (hence, the proportions of user u's calls to the different contacts in its ego-network: v,x,q,n,h, and so on). If all contacts in an ego-network are called an equal amount of times, the entropy of contacts will peak, meaning that the diversity of the call pattern is highest and, conversely, that its predictability is lowest. Note that the entropy of contacts is similar to the Mobility Entropy which will be discussed in section 2.6.3. The formal definition of the entropy of contacts is given by:

$$Entropy\,of\,contacts = -\frac{\sum_{i \in I} p(i) \log p(i)}{\log N} \tag{2.1}$$

where I is the set of all contacts of a user, $p(i)$ is the probability that the user is contacting a single contact i when active, and N is the total number of activities of one user.

2.6.2 Indicators for Individual Movement Patterns

In their most recent literature review on using mobile phone data for travel behavior studies, [146] recognize multiple indicators for individual movement patterns such as displacement, the radius of gyration, the mobility entropy (called Shannon's Entropy in [146]), eccentricity and the convex hull. Table 2.4 summarizes such *movement indicators*, their interpretation and some references that have deployed them. While most indicators are easily interpretable, the ones that are not, such as the radius of gyration or the mobility entropy, will be discussed in more detail in the following sections.

Most movement indicators can be calculated from a time-ordered list of cell phone towers on which a user has been active during an observation period which is routinely captured by CDR data. In this perspective, it is worth remembering that CDRs suffer different types of bias [69, 104], such as: (i) the position of a user is known at the granularity level of cell towers only; (ii) the position of a user is registered only when a cell or text event occurs; and (iii) events per user are sparse in time, that is, the time between consecutive events in CDR is known to follow a heavy-tailed distribution [16, 65]. In other words, since individuals are inactive most of their time, CDRs reconstruct only a subset of the mobility of an individual. Several works have studied the bias in CDR for reconstructing mobility patterns by, for example, comparing patterns with GPS data [96, 97]. Most studies agree that the encountered biases in CDR data do not significantly affect the study of human mobility patterns [98], although it must be noted that such studies often filter out CDR users based on the number of available events. In either way, one has to be aware that the individual mobility trajectories constructed by CDR data are only partial descriptions of individual movement and dependent on the amounts (and distribution in time) of events available in the dataset.

Mobile phone indicator	**Description**	**References**
Number of visited cell towers	Number of visited cell towers	[150, 128]
Displacement	Distance between two consecutively visited cell towers	[84, 37, 38]
Radius of Gyration (eq. 2.2)	Diameter of observed movement with respect to a center of mass	[156, 45, 65, 26, 100, 149]
Mobility Entropy (eq. 2.3)	Diversity of visited cell towers in a movement pattern	[156, 145, 126, 125, 39, 13, 44, 98]
Eccentricity	Shape of the movement pattern	[156]
Diameter of the Convex Hull	Maximum distance between any two visited cell towers	[45]

Table 2.4 Mobile phone indicators based on movement patterns

Radius of Gyration

The Radius of Gyration is one of the most common indicators for individual movement patterns from CDR data [150]. It measures the volume of mobility and is defined as the spatial spread of the cell towers visited by a user relative to his/hers center of mass, which is defined as the mean point of all the visited cell towers:

$$Radius\,of\,Gyration = \sqrt{\frac{1}{N}\sum_{i\in L} n_i(r_i - r_{cm})^2} \tag{2.2}$$

where L is the set of cell towers visited by the user, n_i is each cell tower's visitation frequency, $N = \sum_{i\in L} n_i$ is the sum of all the single frequencies, r_i and r_{cm} are the vector coordinates of cell tower i and of the center of mass, respectively.

Typically, the statistical distribution of the Radius of Gyration over a large population follows a truncated power law distribution, meaning that the majority of people move within a limited territory only, whereas a minority occupies a much larger territory [146]. Such findings, of course, accord with power laws for the amount of displacements discussed in section 2.4.1 and which relate to the heterogeneity of movement patterns across large populations.

2.6.3 The Mobility Entropy Indicator

The Mobility Entropy (ME) indicator is of special relevance to this thesis as it forms the subject of the investigation presented in chapter 6. Essentially, ME is a quantification of the diversity of a single user's movement pattern. The standard definition of mobility entropy is based on on the entropy concept developed by Claude Shannon in his seminal work: "A Mathematical Theory of Communication" [119]. The intuition of the mobility entropy is as follows: when moving through space and time, a user distributes his/her events over a set of cell-towers. The mobility entropy then forms a quantitative measure reflecting how many different cell towers were visited by one user while simultaneously taking into account how evenly the events are distributed among those cell towers. [126] propose three different ways of calculating mobility entropy: the random entropy, the temporally-uncorrelated entropy and the real entropy. For all three measures, locations are based on the visited cell-towers. However, the probabilities of being at a certain cell-tower are calculated differently:

1. Random entropy: probabilities are equally distributed over all cell towers visited by a user.
2. Temporally-uncorrelated entropy: probabilities are weighted by the number of times a user visited a cell-tower.
3. Real entropy: probabilities are weighted by the the staying times of a user in the area of a cell-tower. In addition, the temporal sequence of visits is also taken into account.

For the case of CDR data, calculating the staying time of a user in a cell tower area, as is needed to calculate the real entropy, is often not possible given the sparse temporal resolution of most users in the CDR dataset. The probabilities of the temporally-uncorrelated entropy, on the other hand, are perfectly obtainable from CDR data. The random entropy, ultimately, neglects to take full advantage of the available information in CDR datasets. For these reasons, in this thesis we define the Mobility Entropy (ME) as the temporally-uncorrelated entropy given by:

$$ME = -\sum_{i=1}^{n} p_i \log_{10}(p_i) \qquad and \qquad p_i = \frac{e_i}{\sum_{i=1}^{n} e_i}, \tag{2.3}$$

where ME is the mobility entropy of a user, i is a cell tower visited by the user, n is the total number of towers visited by the user, p_i is the probability of the user visiting tower i, and e_i is the number of mobile phone events by the user at tower i. The calculation of ME values from mobile phone data is also illustrated in figure 6.1 in section 6.1.1.

Predictability of Movement Patterns

When calculating the random entropy, temporally-uncorrelated entropy (or thus mobility entropy) and real entropy from CDR data for 45,000 users during a 3-month long period, [126] find that the distributions of different types of entropy vary fundamentally (figure 2.10 a) . Such a finding is logical, as it means that the gain of information that can be obtained by having an extra observation is depending on the type of observation made (staying time, number of visits, boolean of visit). As is evident from figure 2.10, the distribution of the real entropy measure in the population [126] peaks around a value of 0.8, which indicates that, on average, the uncertainty on which cell-tower a user is going to visit next is $2^{0.8} = 1.78$, thus fewer than two locations. Using Fano's inequality to translate entropy values to a measure of potential predictability, [126] arrive at a surprising 93% of potential predictability (figure 2.10 b, peak of Π^{max}), which implies that "despite the apparent randomness of the individuals' trajectories, a historical record of the daily mobility pattern of the users hides an unexpectedly high degree of potential predictability." [126, p.1020]

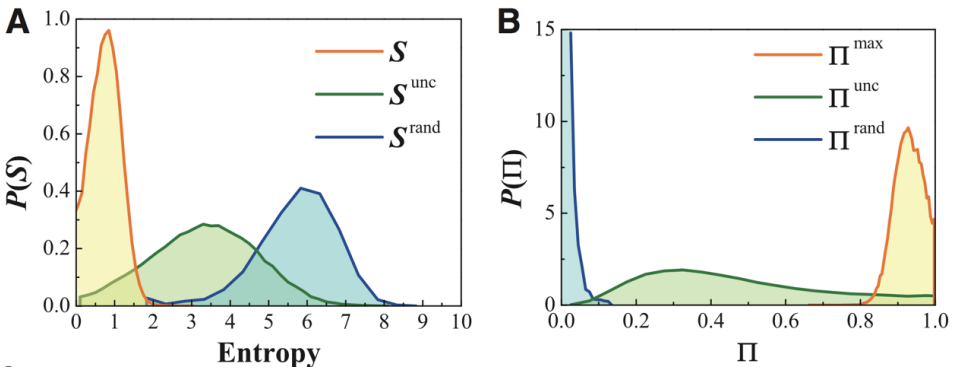


Fig. 2.10 (a) Distributions of real entropy (S), temporally uncorrelated entropy (S^{unc}) and random entropy (S^{rand}) for all 40,000 users in [126]. The distributions of the related measures for potential predictability are given in (b). Π^{max} corresponds with the real entropy (S). Source: [126].

This 93% of potential predictability is even more astonishing when one takes into account how peaked the distribution of the potential predictability for the real entropy is (figure 2.10 b), implying that this high predictability is similar for all users, regardless of their movement patterns [24]. This finding is in line with [65] who show that despite different magnitudes of radius of gyrations, movement patterns of mobile phone users show similar traits.

What is important, however, and what also becomes evident from figure 2.10, is that when information on the movement patterns is restricted to the temporally-uncorrelated entropy (= mobility entropy) or the random entropy; the peak of potential predictability is not reached anymore and, at least for the mobility entropy, a large heterogeneity in the population is observed. This means that, as CDR data do not enable the calculation of the real entropy for all users, researchers have to settle for the analysis of a mobility entropy measure that is showing a large diversity amongst users which forms a challenge when investigating nation-wide patterns, as will be done in chapter 6.

2.6.4 Deploying Mobile Phone Indicators

The Need for Home Detection

Mobile phone indicators are typically calculated at the individual level, and for large populations. Mobilizing this data for research purposes, such as the comparison with census data or the mapping of spatial patterns, requires that users can be allocated and aggregated in space. This requirement is partly due to privacy restrictions. Many countries, including France, prohibit the cross-over of personal information (as could be gathered in customer service datasets) and mobile phone datasets. As a consequence, when investigating mobile phone users, no socio-economic information on individual users is available, and neither are their addresses. To overcome this problem, research has turned to estimations of home locations based on the mobile phone records themselves: *home detection*. In the next section (section 2.7), literature on home detection methods is extensively reviewed, as it forms the main subject of our investigations in chapters 4 and 5. For now, it suffices to understand that once a home for each user in a CDR dataset is detected it becomes possible to allocate and aggregate users spatially, which is a prerequisite step to render mobile phone indicators comparable with census data that is aggregated at census grids.

Relations with Census Data

Based on home detection, there exists a considerable amount of studies that have confronted mobile phone indicators, such as the amount of contacts or the amount of visited cell towers, with census data, unveiling interesting relations between both. [98], for example, show how at commune level in France mobility entropy is related directly to the European Deprivation Index (EDI).

In a similar fashion, relations between mobile phone indicators and census indicators have been uncovered in other works, such as the relation between regional calling patterns and economic development in the UK [54] (figure 2.11), between calling and purchase behavior and food security in a Central African country [49], or between several mobile phone indicators (call, movement and purchase behavior) and multiple census variables on education, demographics and purchase power in a Latin American country [57].

Consequent to the uncovering of relations between mobile phone indicators and census data the possibility of using mobile phone indicators to *nowcast* or even predict macro-economic and socio-economic aspects of populations has repeatedly been suggested [14, 86, 62]. The advantage of mobile phone indicators in this perspective is that they can be produced more easily, at high-resolution, for large populations, and in a much more timely way compared to standard census data.

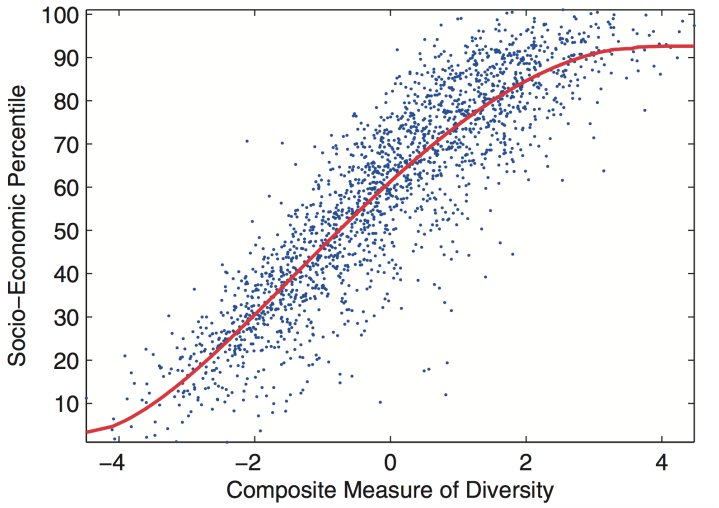


Fig. 2.11 Relation between network diversity and socio-economic rank for 32,482 communities in the UK. The measure for network diversity is constructed as a composite (using principal component analysis) of the Shannon entropy and Burt's measure for calling activities between communities. The socio-economic percentiles are based on the 2004 UK government's Index of Multiple Deprivation (IMD), a composite measure of relative prosperity. A fractional polynomial was fit to the data (red line).Source [54]

Understanding Spatial Patterns of Mobile Phone Indicators

The claim that mobile phone indicators can complement census data or nowcast socio-economic tendencies is a bit presumptuous, especially because there still exists a large knowledge gap with respect to the understanding and interpretation of spatial patterns and variations of mobile phone indicators, as well as with respect to their representativeness and the spatial error that comes with their dependence on home detection methods. As for the reliability of home detection methods, this will form the focus of section 2.7 and the investigations in chapters 4 and 5. Regarding the understanding of the spatial patterns of mobile phone indicators, it is most remarkable that not a single work in literature was found discussing the spatial patterns of mobile phone indicators. In chapter 6, we are the first to investigate the spatial pattern of the mobility entropy indicator

(in France), showing how important insights on spatial patterns are to the assessment of mobile phone indicators.

Scaling Laws of Mobile Phone Indicators

Equally remarkable is that little work in literature was found on the scaling laws of mobile phone indicators, although for census indicators they have been widely studied [10, 42, 103, 77, 22]. Scaling laws describe the variation of indicators with city size (often expressed in terms of population size or population density) and are calculated according to equation 2.4:

$$Indicator = \alpha * P^{\beta} + \varepsilon, \tag{2.4}$$

where α is a normalization constant, P is the population, β is the scaling exponent of interest and ε is a small error term.

Depending on the scaling exponent (β), the relationship between indicator and city size is either superlinear ($\beta > 1$) meaning that there is a non-linear growth of the indicator with city size, in this case the indicator values grow faster than the city population and thus high values are concentrated in larger cities. Or it is sublinear ($\beta < 1$), again indicating non-linear growth of an indicator with city size, only this time with the indicator values growing slower than the city population. Or it is linear ($\beta = 1$) meaning that indicator values scale proportional to city population. As such, scaling laws form a powerful summary of the variation of urban attributes with city size [42, 22, 10], making them easily deployable in, for example, policy decision making.

When the considered indicator is an absolute quantity such as, for example, the number of calls, the obtained scaling exponent β is considered in relation to 1. When $\beta \approx 1$, indicator values scale proportional to city population. When $\beta < 1$, the scaling law reveals a sublinear regime, meaning that there is a non-linear growth of the indicator with city size, in this case the indicator values grow slower than the city population. When $\beta > 1$ a superlinear regime is observed, again indication non-linear growth of an indicator with city size, only this time with the indicator values growing faster than the city population.

One of the few works that have studied scaling laws for mobile phone data is [114]. They show that the number of contacts and the total communication activity from CDR data in Portugal grow superlinearly with the number of urban dwellers, whilst the local clustering coefficient of ego-interaction networks scale linearly. Another exception is [76], at least, if one considers their proposed gravity model as a combination of three scaling laws using the population of origin, the population of destination, and the distance between calls as explaining variables. A second exception might be [83] who evaluate the scaling law of the number of *hot-spots* in a city derived from mobile phone data. Hot-spots, however, can be argued to be more of an urban phenomena than an indicator describing behavior of individual mobile phone users

One intriguing, recent finding with respect to urban scaling laws for census indicator is that they can be sensitive to city definition [10, 42]. This means that, when defining cities differently, the obtained scaling exponents β might vary, possibly altering interpretations. Figure 2.12 shows the distribution of β for different census indicators in France when calculated for 4914 different city definitions in France. Although for some indicators the distribution is narrow around a single value, indicating that β is independent from city definitions, multiple indicators depict a wide distribution of β, indicating superlinear, sublinear or linear relations depending on the city definition. In the case of the latter situation, where β is dependent on city definition, the validity of the scaling law as a trustworthy summary measure is severely challenged. The sensitivity of scaling laws from mobile phone data to city definition has not yet been studied before, except for [114], who investigate scaling laws for three different city definitions in Portugal: statistical cities, larger urban zones, and municipalities.

Concluding, the investigation of urban scaling laws for mobile phone indicators remains an outstanding challenge but, when addressed, should incorporate a sensitivity test to different city definitions. The investigation in chapter 7 will try to answer whether mobile phone indicators are sensitive to city definitions or, in other words, whether mobile phone indicators are dependent on population size or not.

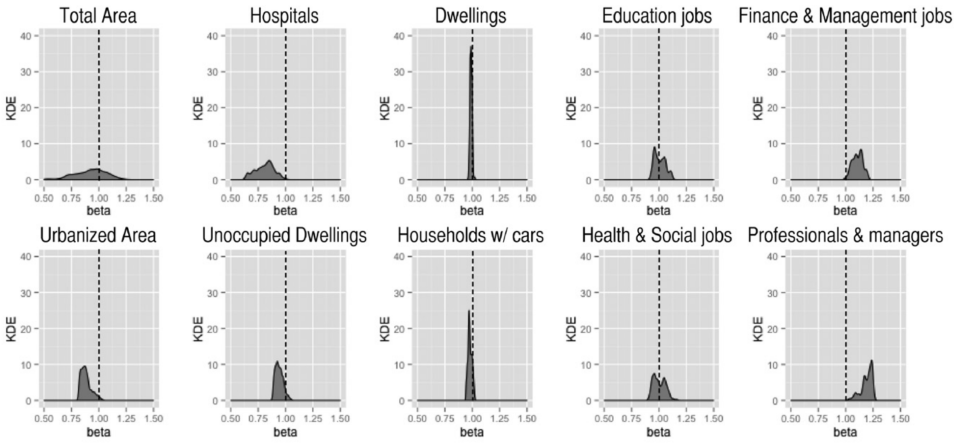


Fig. 2.12 Distribution of scaling exponents (β) of multiple census indicators when calculated for 4914 different city definitions in France. Source [42]

2.7 Identifying Home Locations from Mobile Phone Data

2.7.1 The Role and Act of Home Detection

In mobile phone data research, common to many, if not all studies, is the need to identify the home location of mobile phone users before proceeding to more advanced analysis. For example, knowing the place of residence is a prerequisite to studying commuting patterns, which in turn fuel mobility and epidemiological models [134, 113]. Besides its relevance within mobile phone data analysis, knowledge of home location also forms the crucial link between mobile phone data and other data sources such as census data, making it a key enabler for the integration of information.

In practice, identifying a user's home location means that a single cell tower is allocated as the assumed home location. This means that *homes* are detected at the spatial resolution of cell towers. The assumption then is not that a user lives at that exact cell tower location, but rather in the area covered by the tower. It is remarkable that even though home detection now forms a cornerstone of mobile phone research, home detection methods are often obscured in literature, meaning that details on their exact application, related uncertainties, perceived performance and even validation processes have seen little investigation or, even worse, are rarely communicated.

In the following sections, we show why current home detection practices are problematic. We show how, over time, methodologies for home detection have been simplified to single-step approaches using decision rules that are based on simple, predefined criteria of what defines a 'home'. Such methods are questionable because the possibilities to validate are limited, and there is a lack of knowledge on their sensitivity, specifically with respect to criteria choice. Our empirical work with the French CDR data in chapter 4 and 5 will exemplify some of the problems we raise, allowing us to put them in a more practical context and to outline their consequences in more detail.

2.7.2 Detecting Homes from Large-Scale Location Traces

Given the enormity of the datasets that capture geo-located traces of users, automated methods for the detection of homes, or other meaningful places such as the workplace, of individual users have been developed in literature. Here, it is necessary to distinguish between continuous location traces (e.g. GPS data) and non-continuous location traces (e.g. mobile phone data, credit card transactions, or check-ins through location based services and online social networks) where the latter do not provide a similar high-volume, high-resolution capture of location traces in time or space compared to the former.

Although our main interest is to outline the deficits in the methods used for non-continuous location traces, we will first shortly review some characteristics of automated home detection from continuous location traces, which helps to better contextualize developments for non-continuous location traces.

Detecting Meaningful Places from Continuous Data

The analysis of continuous location traces has been the focus of early developments in the automated identification of meaningful places. Related work typically used small-scale datasets, most commonly from continuous GPS traces but also from Bluetooth, or Wi-Fi positioning [153, 120]. The general methodology used to identify meaningful places from continuous location traces consists of a two-step approach.In the first step, location traces are clustered in space (and sometimes in time) in order to detect important places. Techniques for clustering continuous location traces range from manual GIS analysis [153, 64] to automated, unsupervised analysis using, for example, k-means clustering [12], non-parametric Bayesian approaches [91], or fingerprinting of the radio environment [68]. In a second step, the important places identified are then annotated as meaningful places (such as home, work, recreation area). Annotation can be done either through interpretation, for example by expert judgment, by surveying the user that produced the traces, or through automation, mainly by means of time-space heuristics [91].

Detecting Meaningful Places from Non-Continuous Data

In contrast to the above-mentioned continuous location traces, the use of non-continuous location traces has recently become very popular. The identification of meaningful places from non-continuous location traces poses substantial challenges, most notably due to the less frequent observations and the larger spatial resolution in which observations are captured (e.g. CDR data are only captured at the location of the cell tower used). These challenges, however, are outweighed by the presumed advantages associated with the larger coverage, in terms of users, timespan and spatial extent of the data sources [73, 78]. The next sections focus on how to identify one meaningful place - the user's home - using CDR data as one prominent example of non-continuous traces. Still, the methods and problems described extend to other meaningful places and to all datasets covering non-continuous location traces.

2.7.3 Detecting Homes from Mobile Phone Data

Two-Step Approaches for Non-Continuous Data

As with the two-step approaches for continuous traces, initial methods to detect home locations from CDR data also clustered location traces into important places before annotating them as meaningful places. For example, in [71] individual traces from CDR data are clustered using Hartigan's leader algorithm. Clusters are then annotated into meaningful places by means of a logistic regression model that is trained on data from 18 persons for which ground truth was available. Next, and for each user, the cluster with the highest score on the logistic regression model is chosen to be the presumed home area. Similar two-step approaches have been deployed in other studies for larger number of users ranging from 7989 to as much as 2 million cell phone users [37, 9].

The advantage of two-step approaches is that they make good use of the individual data available within location traces and, as such, relate more to the intuition on human movement. The disadvantage is that deployed methodologies are often complex, not uniform over different studies, and therefore not easily reproducible. A perfect example of an un-reproducible methodology can be found in [37]. Here, in a first step, the authors deploy a model based on a clustering method that optimizes a likelihood function that is expressed as the joint probability density function of the *sightings* (=locations) and parameterized by a vector of weights specifying the probability of a location's belonging to a particular cluster. Secondly, a logistic regression model is used to distinguish between activity and travel clusters that is based on a variable expressing the shape of the traces distributed within a cluster and a variable expressing the amount of sightings in one cluster as compared to those in other clusters. The resulting clusters, in a third step, then feed into a set of behavior-based algorithms, of which we will spare the reader the explanation, to detect types of locations visited, amongst them, the home location. The point is that, even though the results of two-step approaches might prove to be more accurate compared to simple single-step approaches ([37] for example, show their results to be outperforming the single-step approaches from [100, 32]), their complexity limits applicability to different areas or datasets, thus making them less interesting for reproduction in other studies.

Single-Step Approaches for Non-Continuous Data

Because of their limited transferability, two-step approaches for non-continuous location traces are quickly giving way to single- step approaches that are now widely deployed in literature [31, 33, 100, 78]. The difference between two-step and single-step methods is that the latter skips the clustering into important locations and thus acts directly on individual cell towers instead of groups of cell towers. One of the reasons for switching to single-step approaches is that the standard clustering methods used in the two-step approaches make it difficult to construct consistent spatial traces when combined with non-continuous location traces. Nevertheless, the main drawback of this switch to a single-step approach is that the spatial pattern of the location traces is largely neglected, as only single cell tower annotation is targeted. This increases the uncertainty of fixing home location because standalone or irregular events at individual cell towers may be sufficient to undermine the method.

In practical terms, detecting a home in a single-step approach is done by using a decision rule that is based on an a-priori definition of home - what we will continue to call the *home criterion* - in order to produce a list of one or several cell towers that could be the home location. A standard example of a home criterion for the case of CDR data is: 'home is where calls are made during the night'. The problem with single-step approaches is that such decision rules are being applied as heuristics, meaning that one general rule is applied to the location traces of all users even though a different set of decision rules for different users could potentially lead to better results for the simple reason that different users behave differently.

In terms of home detection, applying heuristics implies that meaningful places (like the home) can be described similarly for all users in the dataset, regardless of the user's characteristics as observed in their movements and calling patterns. It seems logical that the deployment of home criteria can only be done when proper evaluation and validation of their results has been carried out, or when clear evidence exists for the use of a specific criterion or decision rule. For this reason, the following section will discuss how to define decision rules for one-step home detection methods and which criteria to use.

2.7.4 Defining Decision Rules for Single-Step Approaches

Simple Decision Rules

The core challenge for single-step home detection is in defining a decision rule that is simultaneously capable of i) distinguishing between different important places and ii) annotating the correct home location. Most research employs simple decision rules that are either based on information from official statistics or rely on precedents found in literature. When examining the existing decision rules in research literature, the most popular are: time-based limitations for the night ('home is the location that has the most activity between x pm and y am'), time-based aggregations ('home is where the most distinct days, or weekend-days are spent), as well as spatial groupings ('home is the location with the most activity in a spatial radius of x km) [31, 100, 78, 58, 134].

For example, [31] use the highest distinct number of observations between 6pm and 8am to derive home locations while basing the time interval on statistics from a Boston dataset drawn from the American Time Use Survey. Almost all studies using simple decision rules rely on census data. They depend either on specific surveys and questionnaires to define the criteria deployed, [31] or, for high-level validation, on aggregated population density data [100] or commuting figures [78].

Complex Decision Rules

A few studies have elaborated more complex decision rules for home detection. The seminal work of [6], for example, uses a tree-based approach that combines a set of criteria including distinct days of activities on a cell tower, the starting times of calls, deviations of starting time of calls, and durations of calls for a training set of 14 people for which the ground truth was known. The decision rules, as defined by the classification tree, were consequently deployed to all users in an Estonian dataset (as heuristics in other words), raising the question of how representative a training set of 14 people could possibly be for a large population.

The problem of small training sets was alleviated in [59] who used a training set of 5000 users to construct a complex decision rule for home detection. Deploying a Genetic Algorithm technique, they focus on finding the best combination of temporal criteria to denote home locations in an emerging economy. Their best performance is a correct prediction of around 70% for a subset of 50% of the users. Users were filtered on the basis of having at least a 20% difference in the percentage of total calls between the first and second eligible cell tower. The complex decision rule they use to obtain this result is to select the cell tower logging the most activity during the nights of Friday, Saturday, Sunday, Monday and Tuesday from 5:15 pm to 8:30 am.

The individual ground truth data in [59] are retrieved from users' contracts with the provider. This data is not available in most countries due to legal obligations to anonymize users or bans on linking individual information to CDR data. As a consequence, [45] try to derive a temporal decision rule but this time without a training/validation dataset at individual level. Applying an unsupervised k-means algorithm to the temporal activity patterns of frequently used cell towers in Portugal, they find clusters that are either interpretable as temporal patterns related to presence at home, at work, or not interpretable at all. Consequently, their decision rule to detect home locations is based on these temporal patterns interpreted as home presence. Compared to [59], one of the drawbacks of their approach is that they did not construct their criteria based on individual observations. This raises the question as to the degree to which such criteria are realistic for different subsets of users.

In a way, the subset representativeness problem persists for all single-step approaches, regardless of whether their decision rules are defined in a complex or simple way. If the same decision rules is applied to all phone users, careful investigation into the effect at individual level, or at population subset level should be carried out, in order to know the degree to which generalization favors or disfavors subsets of users. In other words, if decision rules are applied generically, in-depth validation of the single-step approaches is important.

2.7.5 Validating Home Detection Methods

The use of a particular decision rule, whether derived from a census, borrowed from literature or defined by training sets, is often based on comparing population counts from mobile phone data with census data. However, such high-level validation does not offer a direct evaluation of performance at individual user level, nor does it allow researchers to compare between cases. In fact, assessing the performance of different decision rules by comparing the resultant population counts with census data is, strictly speaking, a rather limited alternative solely justified by the absence of individual level validation data.

The absence of validation data at individual level is a common problem in published research, and is therefore often taken for granted but has several consequences. First and foremost, it impedes the creation of evaluation metrics that can assess the performance of home detection at individual level. Such an individual level evaluation could allow us to better understand the workings of different decision rules on a specific dataset and user subsets, which in turn could enable a comparison between different decision rules, datasets, users and areas.

Second, the absence of validation data at individual level is the very reason why single-step approaches apply decision rules heuristically. In the absence of individual level validation data, it is impossible to understand which decision rules works best for any individual user. Consequently, case-adjusted, adaptive algorithms cannot be developed. This implicitly forces researchers and practitioners to adhere to a one-size-fits-all solution in order to be clear and consistent.

It is worth noting that, currently, high-level validation is still assumed to be a good solution in the absence of individual level validation data. In particular, two observations stand out. Firstly, census data is often used for high-level validation. For example, comparisons for small geographical areas can be made between the counts of home locations identified from mobile phone data and the aggregated counts of peoples' residential locations obtained from censuses. This is a very opportunistic, if not naive, validation attempt as census data has never specifically been gathered to serve this purpose and little or no information exists on how, for example, different spatial delineations or the distorted market shares of mobile phone operators could influence this kind of validation.

Secondly, it is noteworthy that no studies have used high-level validation to compare the performance of different decision rules on CDR data. Nor are there studies that evaluate the sensitivity of high-level validation to criteria choice. This absence is probably due to high-level validation that is not informative enough to properly understand the differences between criteria, decision rules, and their performances. The consequence is that we are far from obtaining a consensus on which criteria are best, or on how to construct optimal decision rules. In fact, we are far from understanding the strengths and weaknesses of different home detection methods altogether and it forms an interesting realization that part of the unknowns we face are due to the limits of high-level validation typically pursued.

2.7.6 Shortcomings of Current Home Detection Methods

In conclusion, we find that a clear framework is missing to allow us to understand the performance, uncertainty and sensitivity of the criteria choice or decision rule development, especially at individual level, when using non-continuous location traces to detect home location. Despite their widespread use, no clear reasoning exists as to why single-step approaches should be chosen over two-step approaches. Nor does a consensus exist on which criteria should be used, or how decision rules for a given dataset should be optimized. Similarly, it is striking that no work investigates the sensitivities of single-step approaches to criteria choice. Additionally, we find that the validation of large-scale home detection methods is severely limited because of the absence of ground truth data at individual level. As a result, current assessments of home detection methods are based on high-level validation, but the trustworthiness and exact contribution of this practice is rather dubious.

Chapter 3

Creating Mobile Phone Indicators

To measure is to know.

Lord Kelvin (William Thomson)

Abstract

This chapter details the creation of mobile phone indicators from a French CDR dataset. First, the French CDR dataset and its main characteristics are introduced. Second, an analytical framework is proposed, elaborating the different steps that lie between (big) data source and territorial indicators of human mobility and calling activities. An exploratory analysis investigates the different distributions and spatial patterns of several mobile phone indicators. Third, relations between mobile phone indicators and census data are briefly investigated upon, illustrating the potential to use mobile phone indicators to nowcast nation-wide measures of, for example, deprivation. In the wider perspective of the thesis, this chapter serves as an introductory chapter to the French CDR dataset, to the main characteristics of derived indicators, as well as to the several questions addressed in subsequent chapters.

Related Publications and Acknowledgments

- The structure of this chapter largely accords with [98], a publication that is co-authored by the PhD candidate. Compared to the original paper, the chapter extends the analysis to more indicators and time periods, elaborates more on the value distributions and spatial distributions of mobile phone indicators, and adds more depth to interpretations.
- Acknowledgements go to Giovanni Lima, Pierpaolo Paolini, and Cezary Ziemlicky for running the very first experiments on creating mobile phone indicators in France, and to Dr. Luca Pappalardo, Dr. Lorenzo Gabrielli, Prof. Dr. Dino Pedreschi, Prof. Dr. Fosca Giannotti, and Dr. Zbigniew Smoreda for pursuing the project that made this study possible.

3.1 Introducing the Dataset

This chapter explores how to create indicators from a French CDR dataset. First, however, it is important to get an insight in the information that is available in this French CDR dataset. The same dataset will be used in the analyses of subsequent chapters.

3.1.1 The French Call Detailed Records (CDR) data

CDR data capture information from mobile phones which, in Western countries, can be assumed to belong to individual mobile phone users. The captured information, as was discussed in section 2.2.1, enables insights in two dimensions of human activities: calling patterns and movement patterns.

To extract insights, CDR data needs to be treated first. The raw data records stored in CDR datasets in itself are not too informative and, because of the sheer volume of data, seldom organized in a way that is needed for analysis (for an example of raw CDR data, see table 3.1). In addition, the type of information that is available in CDR datasets is neither uniform between operators, nor between countries, nor over time. Between operators, for example, data might be captured differently for various technical reasons. Between countries (as well as between regions and individual users), and over time, the very usage of mobile phones may differ because of, for example, different adaptation rates of new technologies. The consequence is that every analysis on CDR needs to be complemented by proper prior knowledge on the construction of the dataset and the context in which the data were collected.

The French CDR dataset has been extensively studied [124, 51, 67, 137, 141, 140]. It was collected by the network operator Orange during a 154-day period between 13 May 2007 and 15 October 2007. The dataset itself consists of information on the type of the event, the timestamp on which the event took place, a cell tower identifier, a location area identifier, an initializing user identifier, a receiving user identifier and the duration/length of each event (table 3.1).

Event	Timestamp	Cell Tower	Location Area	Initiating User	Receiving User	Unit
...	...	...	...	...	...	...
VO	2007/10/01 23:45:00	15988	00080177U8	33647956872	33649274861	3656s
SI	2007/10/02 01:12:04	2051	00000001D1	3367261532	33632415523	125c
...	...	...	...	...	...	...

Table 3.1 Example of two Call Detailed Records stored in the French 2007 dataset

Anonymization

The dataset only stores information related to Orange subscribers, whose numbers range from 18.4 to 18.5 million active users during the available time period. User identifiers are pseudo-anonymized, meaning that a user identifier has the format of an existing cell phone number but is irreversibly decoded from the real phone number. User identifiers remain unchanged throughout the entire dataset, facilitating traceability of individual users throughout time, making it one of the largest CDR datasets in the world that offer access to individual user data during such a long time period.

3.1.2 Locations of Cell Towers in France

The spatial resolution of information in the French CDR dataset is dependent on the locations of the 18,275 cell towers in the Orange network (figure 3.1). Cell tower locations are provided by Orange. Their positioning is demand driven. As such, more cell towers are found in areas where mobile phone usage is higher, such as in city centers, or alongside main transport axes (figure 3.1).

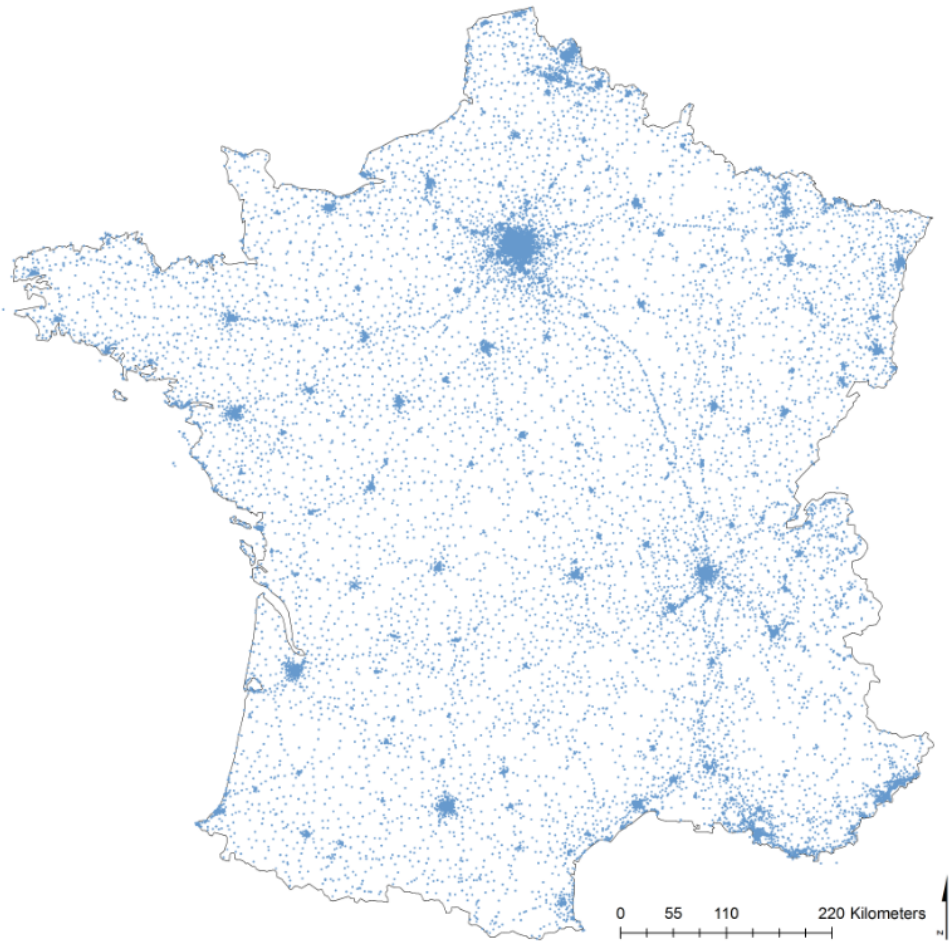


Fig. 3.1 Locations for all 18,275 cell towers in the French CDR dataset anno 2007 (each dot is one cell tower).

3.1.3 Types of Events

The CDR 2007 dataset stores different events. Table 3.2 gives a short overview of the recorded events.

Event	Duration	Explanation
VO	seconds	Voice outgoing
VI	seconds	Voice incoming
SO1	characters	Text outgoing
SI1	characters	Signal server incoming
SO2	characters	Signal server outgoing
SI2	characters	Text incoming

Table 3.2 Recorded events in the French CDR dataset

Calling Events

Concerning call events, notations are simple: *VO* stands for outgoing calls while *VI* stands for incoming calls. For one call, both the outgoing and the incoming event are thus captured as a different record in the dataset, leaving two traces that are each identified by two known users. This makes it easy to investigate reciprocity in contact networks as it is possible to derive who called whom, and who was called by whom.

It is, on the other hand, more difficult to reconstruct what is an unique call, as the incoming call entry and the outgoing call entry will have different durations and timestamps in the dataset. This is easy to imagine: when one person calls another, the second person will never pick up the phone at exactly the same moment. This slight difference in time makes it difficult to connect outgoing and incoming entries in a large database setting as there is no common field that connects both entries. The construction of unique calls between users grows even more difficult when the called person does not pick up the phone. In this case, there will be an outgoing call entry in the dataset but not an incoming call entry, as no action from the receiver has taken place. So even if one would try to reconstruct unique call correspondence, there will be outgoing calls that cannot be linked to an incoming call, and there is no clear criteria to filter such outgoing calls. Luckily, for every call entry, the identifier of both users is registered in the dataset, making it possible to reconstruct interaction between users even when it is not possible to reconstruct the exact time and duration of individual calls.

Texting Events

Concerning text messaging the situation is more complex. The problem is that text messaging is governed by interfering servers. In short, when a user sends a text, it is received by a server, which will, eventually, forward it to the intended receiver. The time lag between both events (sending and receiving) can be substantial depending on the receiver's availability (e.g. receiving phone is not active) and server capacity (e.g. new year's eve).

The consequence is that, in the CDR dataset, outgoing and incoming text events have an interfering server (in what follows called: the text-center) as their, respectively, receiving or initiating user identifier. This makes it impossible to reconstruct contact networks from text events, rendering them meaningless for the investigation of social networks or spatial networks of communication. Regarding locations, the text events in the French CDR dataset have an additional limitation. For technical reasons, only SO1 and and SI1 events have cell tower identifiers and thus location information, meaning that about half of the text events are useless to study movement patterns of users.

Table 3.3 offers the shares of the different events for one day in the dataset while indicating the possibility to study either contact networks or movement patterns with the available information. For Wednesday 1 October 2007, 54.01% and 79.35% of the collected data can be used to, respectively, study contract networks and movement patterns. Percentages vary slightly but are similar for other days in the dataset.

Event type	Records (in millions)	Share (%)	Contact networks (%)	Movement patterns (%)
VO	41.52	32.54	32.54	32.54
VI	27.40	21.47	21.47	21.47
SO1	7.76	6.08	-	6.08
SI1	24.57	19.26	-	19.26
SO2	19.90	15.60	-	-
SI2	6.44	5.04	-	-
Total	**127.59**	**100**	**54.01**	**79.35**

Table 3.3 Shares of events in the French CDR dataset for Wednesday 1 October 2007 and their availability to study contact networks or movement patterns.

3.1.4 Magnitudes of the Dataset

Number of Users and the Orange Market Share

The number of distinct and active users in the French CDR dataset is about 18.5 million, depending on the duration of observations considered (table 3.4). With mobile phone penetration being estimated at 86% in France in 2007 [11], and given a population of 63.9 million inhabitants during the observed period[1], that is around 33.6% of all French mobile phone carriers and 28.5% of the total population. As is the case with penetration rates [11], market shares are expected to vary regionally and locally, but those shares are unknown.

Number of Events

The number of records in the French CDR dataset that can, and will, be used in the analyses of this thesis total to around 15.4 billion. A breakdown per month of the number of distinct users, call events (VO and VI) and text events (SO1 and SI1) used in the analysis is given in table 3.4.

[1] This is the average of the monthly estimates for the period between May and October 2007 as obtained from the French National Statistics Website (www.insee.fr)

Remember that for text events only SO1 and SI1 events are useful because they have location information but that they can not be used to construct contact network.

Observed numbers in May and October are lower, because of shorter time periods of observation during these months, but proportional compared to other months. In general, shares of calls and texts that will be used in the analysis remain approximately constant over time, with about 66% of events being calls and 34% being texts. During August, a drop in both call and text events can be observed compared to other entire months in the dataset. The temporal changes in the absolute number of observed events are a useful reminder that CDR data is embedded in daily practices and therefore, subject to change. Such changes may occur on a day-to-day basis, a week versus week-end cycle or throughout the year, rendering it important to embed interpretation of the data and its characteristics in their historical context.

Month	Distinct users	Call events	Share	Text events	Share
May	17.8	1248.3	65.9%	646.6	34.1%
June	18.4	2141.9	66.9%	1061.0	33.1%
July	18.5	2040.5	66.6%	1022.3	33.4%
August	18.4	1824.6	65.3%	967.8	34.7%
September	18.5	2013.8	67.3%	979.0	32.7%
October	17.7	1020.6	68.2%	475.2	31.8%

Table 3.4 Number of distinct users and events per month in the French CDR dataset as will be used in the analyses. All numbers are in millions, shares are in percent. Call events are VO and VI events, text events are SO1 and SI1 events.

3.1.5 Handling the French CDR Dataset

Storing over 15.4 billion entries, the French CDR dataset needs a fairly advanced big data infrastructure to mobilize and analyze the available data (see also section 1.3.3. In Orange, this technical task was done by an Apache Hadoop system, an open source framework able to manage (very) large datasets and to govern the storage and distribution of calculations over a cluster of computers. Interactions with Hadoop were facilitated progressively over time by Hive, an SQL dialect to query data from Hadoop; by Pig Latin, a high-level language to create programs on Hadoop; and ultimately by Spark, a cluster computing framework that acts as a driver to invoke parallel operations based on, for example, Python programs. Regardless of the deployed language, retrieving and analyzing data from the system requires the comprehension of relative database logic as well as a good estimation of the computational costs of required operations in order to execute work effectively. Once the data has been queried, analyzed and stored to ones' needs (preferably in smaller subsets), data science operations with traditional tools such as Python or Geographic Information Systems (GIS) software can be deployed on local computers to engage more dynamically with (subsets of) the CDR data.

3.2 Constructing Mobile Phone Indicators

Throughout the rest of this chapter, the construction and deployment of mobile phone indicators from CDR data will be discussed. Based on the analytical framework proposed in [98] (figure 3.2), indicators are constructed for all individual users in the French CDR dataset. When territorially aggregated, the statistical and spatial distributions of indicators values can be investigated and relations with other datasets such as census data can be investigated.

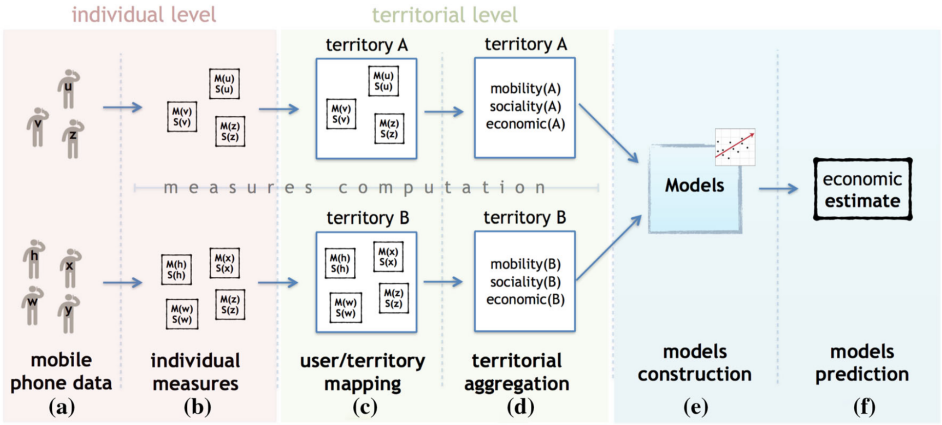


Fig. 3.2 Framework to create and deploy mobile phone indicators. Source: [98].

3.2.1 Individual Measures from Call Detailed Records

The creation of mobile phone indicators that aim at summarizing CDR users' movement patterns, contact networks or calling behavior is discussed in section 2.6. For the French CDR dataset, multiple indicators are calculated for all based on the CDR data for each month in the dataset (May-October 2007). In total, 23 indicators are calculated per user per month, summing up to a total of about 2.5 billion calculated indicator values. This task is performed by means of a Pig script extended with one User-Defined-Functions (UDF) for each indicator. In other words, the big data system is tasked with collecting information per user per month, then sending this information to a function that calculates the desired indicator value.

Calculated indicators can be categorized in 4 different domains, each capturing a different aspect of CDR data: calling behavior, contact networks, movement patterns, and home detection related indicators (table 3.5). Some of the more complex indicators are already defined and discussed in section 2.6. Their equations are referenced in table 3.5. Concerning the construction of indicators for movement patterns, it is worth remembering that CDRs suffer from different types of bias, such as: (i) the position of a user is known at the granularity level of cell towers only; (ii) the position of a user is registered only when a cell or text event occurs, and thus is related to some of the indicators describing calling behavior; and (iii) the amount of events per user are sparse in time, that is, the time between consecutive events in CDR is known to follow a heavy tail distribution [16, 65]. In other words, CDRs allow us to reconstruct a user's actual movement patterns only partially.

Domain	Mobile phone indicator	Description
Calling behavior	Number of calls	Number of calls made or received
	Duration of calls	Summary statistics of call durations (mean, median, or standard deviation)
	Inter-event time	Summary statistics of the time between consecutive calls (mean, median, or standard deviation)
	Active days	Number of active distinct days
	Percentage nocturnal calls	Percentage of calls made between 7pm and 9am
Contact networks	Number of contacts	Number of contacts interacted with
	Interaction per contact	Summary statistics of the amount of interactions per contact (mean, median, or standard deviation)
	Entropy of contacts (eq.2.1)	Entropy measure of calls to contacts
Movement patterns	Number of visited cell towers	Number of cell towers used to make calls
	Radius of gyration* (eq.2.2)	Radius of gyration of the movement pattern
	Mobility Entropy* (eq.2.3)	Entropy measure of visited cell towers
Home detection	Calls at home*	Number of calls made at the presumed home cell tower
	Percentage calls at home*	Percentage of calls made at the presumed home cell tower
	Distance between L1 and L2*	Distance between most plausible and second most plausible home cell tower
	Spatial uncertainty*	Uncertainty measure of the detection of home cell tower

Table 3.5 List of calculated mobile phone indicators. Indicators with an asterisk will be discussed in more depth in subsequent chapters.

3.2.2 User-Territory Mapping and Aggregation

An important step in mobilizing mobile phone indicators is user-territory mapping, and the subsequent territorial aggregation (figure 3.2 c,d). User-territory mapping is a prerequisite step in mobile phone research and is typically performed by means of home detection algorithms (section 2.7). Home detection on the French CDR dataset will be discussed in much more detail in chapters 4 and 5. For this chapter, it suffices to know that the *maximum actions* algorithm was used, which chooses a cell tower as the presumed home location of a user based on the highest amount of calls a user has performed over the course of a month.

Mapping users to cell towers is a prerequisite step for territorial aggregation of users and their related indicator values. In the case of mobile phone indicators, such aggregation is preliminary done at cell tower level (grouping all users that have the same cell tower as presumed home location), and is necessary for two reasons. The first reason is merely computational. Performing analysis on around 18.5 million data points (the amount of users in the French CDR dataset) is time consuming and cumbersome if not impossible depending on the type of analysis. Territorial aggregation reduces this number. In the case of the French CDR dataset 18,275 cell towers were active, meaning that territorial aggregation diminishes the amount of observation points with a factor of 1,000 (18.5 million to 18.2 thousands), enabling a much more dynamic treatment of the dataset.

A second reason is that territorial aggregation enables us to cross-reference information with other data sources, such as census data. This does not mean that the spatial resolution of two datasets always accord with each other. Most of the time, the contrary is true, forcing the translation of the spatial delineation of one data source to that of another, which in itself forms another action of spatial (dis)aggregation. To accord information between census data and the French CDR dataset, for example, a translation between the spatial delineation of cell tower locations (figure 3.1) and the census grid needs to be performed. In some cases, such as with the validation dataset created in chapters 4 and 5, geo-located census data can be aggregated at the cell tower levels, avoiding the cumbersome translation step.

In this chapter, territorial aggregation is performed on the municipality level defined by French official statistics (*communes* in French). Translation from cell tower level to municipality level is done based on the locations of cell towers. Cell towers and their associated users that are within the same municipality are grouped together. As will be discussed in depth in chapter 7, obtained results on, for example, the relations between mobile phone indicators, city characteristics, or census data are not independent from territorial aggregation practices. Territorial aggregation, in other words, cannot be deemed objective and contributes to the uncertainty related to the use of mobile phone indicators.

3.2.3 Simple Model Construction

An ultimate step in the analytical framework is the relation with other data sources (figure 3.2 e). To illustrate this step, the analysis elaborates on simple relations between mobile phone indicators and income data at municipality level. The aim of this exploration is not to go in detail on the uncovered relations, but rather to illustrate the final step of the framework. An in-depth investigation of the relation between mobile phone indicators and socio-economic census data for different territorial aggregations is performed in chapter 7.

3.3 Mobile Phone Indicators in France in 2007

3.3.1 Distributions of Mobile Phone Indicators at User Level

Several characteristics of the mobile phone use captured in the French CDR dataset become apparent when investigating the distributions of mobile phone indicators at user level. In figure 3.3 the percentile values are given for different indicators calculated for all users during September 2007. One observation is that the median numbers of activity or visited places are moderate, but that variation between users is high. For example, a median user in the dataset made 65 calls in 30 days, or about 2 calls a day, which is not an extremely high number. Variation between users, on the other hand, is rather high with highest percentiles leading to 263 calls or more and lower percentiles to 8 calls or less. In terms of visited cell towers, median users visit about 12 cell towers per month. Again, this is not an extremely high number, although these numbers do not indicate the frequency in which cell towers are visited. Still, in terms of reconstructing movement patterns, users with 12 or less visited cell towers in a month might not render the best movement patterns possible. Such observations are an indication for the degree to which all CDR data analyses are restricted by the spatial and temporal resolutions of mobile phone usage patterns by users.

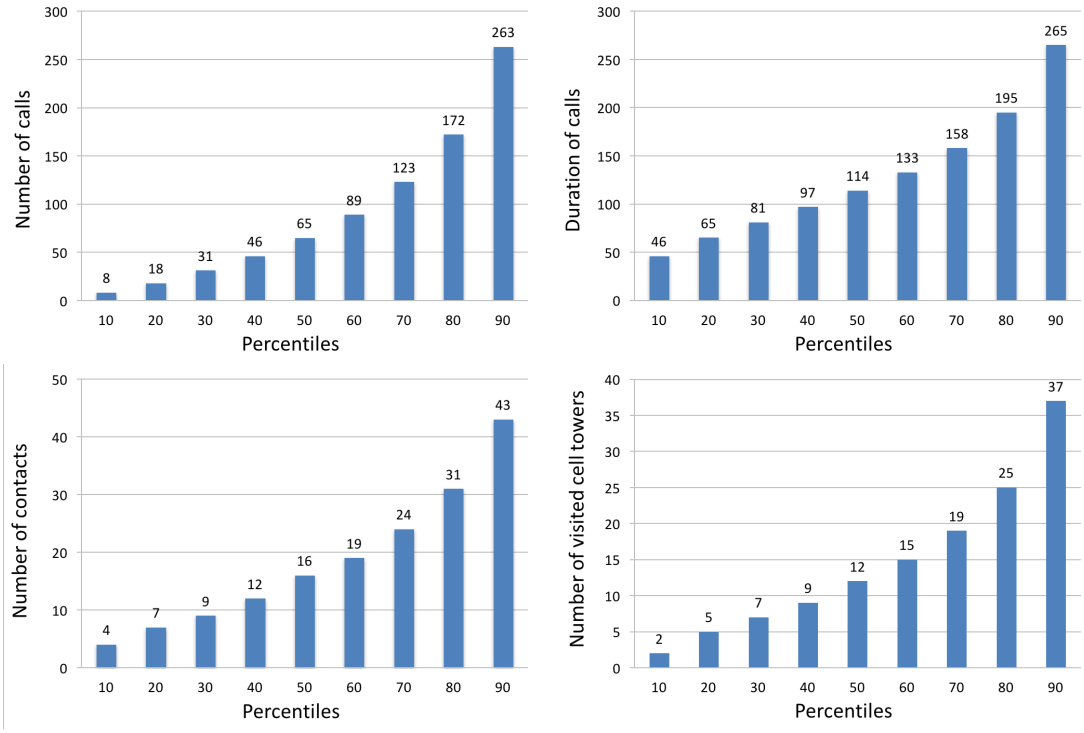


Fig. 3.3 Percentile values of different mobile phone indicators based on values for all users in the French CDR dataset calculated for September 2007.

3.3.2 Distributions of Mobile Phone Indicators at Cell Tower Level

Applying user-territory mapping by using of the *max actions algorithm* (section 3.2.2), territorial aggregation of mobile phone indicators at cell tower level becomes possible. Figures 3.4, 3.5, 3.6 and 3.7 show the distributions of different mobile phone indicators at cell tower level. They describe the median (50th percentile), 90th, 75th, 25th and 10th percentiles of indicator values at each cell tower. Figures are made more interpretable by ranking cell towers based on their median value. Because user counts between cell towers are not equal (cell towers have a different amount of users presumed to be living in their coverage areas) and because ranking is different for each indicator, figures are accompanied by a figure expressing the user counts at different ranks. The investigation of user counts is necessary as it can explain outlier values in the distributions of mobile phone indicator values due to small sample sizes. To avoid small sample sizes, cell towers with less than 100 users were omitted in the analysis (resulting in a total of 15,000 cell towers that are ranked instead of the about 18,275 cell towers in the territory). In addition there is one cell tower with a user count of about 10,000 users which produces outliers for many indicators. This cell tower is located at the Parisian Airport of Orly and was omitted also. The suspicion is that it captured cell phones used by the staff, and only during work, at the airport.

Indicators related to calling behavior

For the indicators related to call behavior in figure 3.4, a large similarity between cell towers is observed regarding the amount of calls and the average duration of calls. Only a limited amount of cell towers display significantly large or small values. These outliers can, partially, be explained by low user counts observed for involved cell towers indicating that, perhaps, the sample size at cell tower level is not always ideal to capture the general pattern. Regarding the amount of days that the users were active, the variation is larger and less uniform between cell towers, especially when looking at the median values. It is remarkable to note that, median values range, in general, between 10 and 20 active days, implying that median users at different locations in France would be active on the cell phone only every other day in September 2007.

For all calling related indicators, the variation between users within cell tower areas (so between user that have the same home location) is rather large as is expressed by the 10th, 25th, 75th, and 90th percentiles. For the amount of calls and the duration of calls, lower echelons of the within cell tower distributions are rather close, but large difference can be observed between the most active cell phone users and their median counterparts. For the number of calls, for example, the 90th percentile users, in general, perform about 300 calls per month more and with calls lasting about 150 seconds longer than a median user in their cell tower. The differences between users become most apparent for the amount of active days. For all cell towers, regardless of their median value (and thus rank), the highest percentiles are active almost all days in a month, whereas the lowest percentiles are active only a few days a month. The inter-personal differences, in other words, are much larger than the locational differences, which can be expected from something such generic as calling behavior, but which has clear implications when studying related indicators.

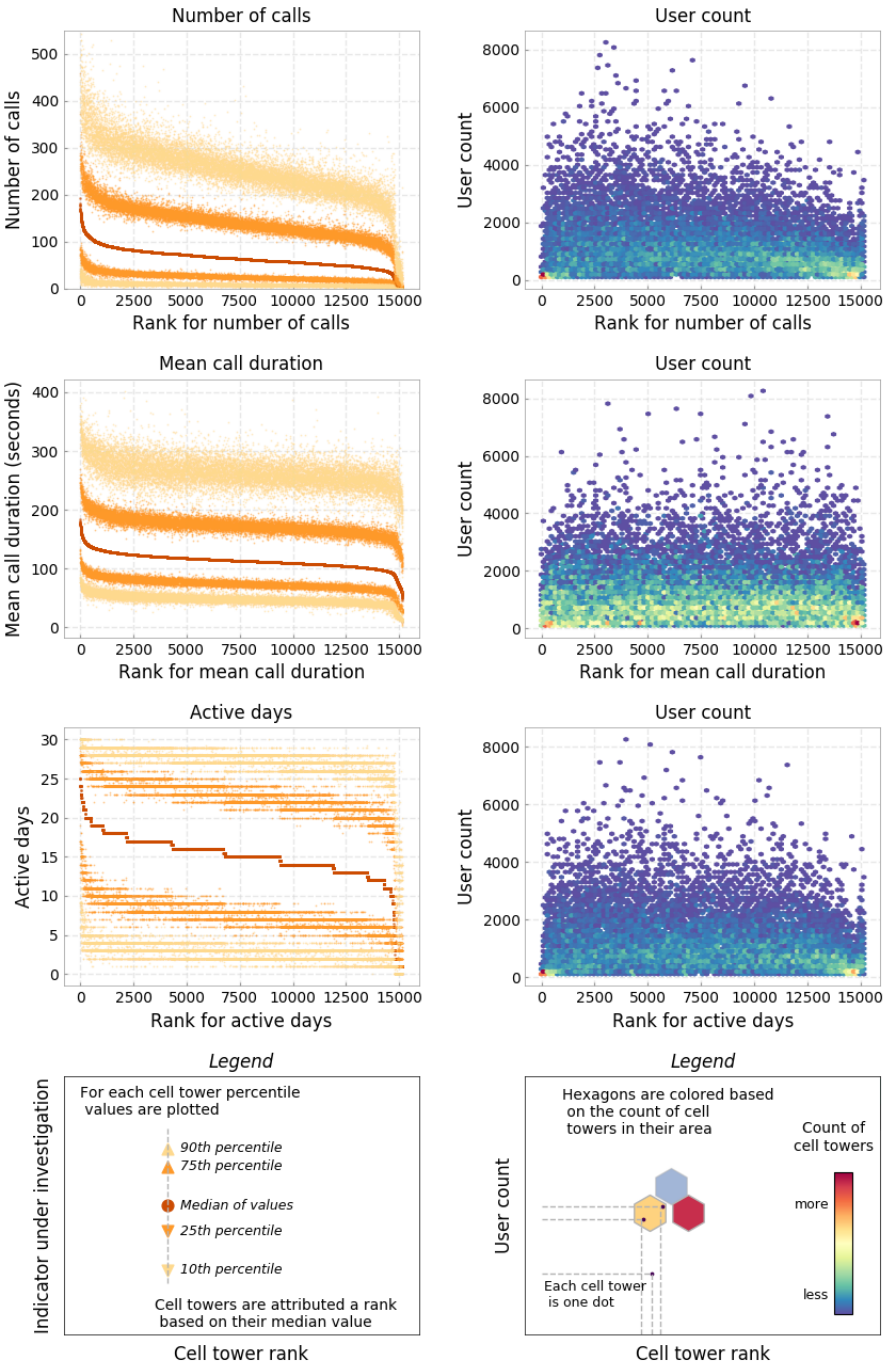


Fig. 3.4 Distributions at cell tower level of mobile phone indicators related to calling behavior. Indicators are calculated for all users in the French CDR dataset and are aggregated at cell tower level by means of the maximum action algorithm. Cell towers with less than 100 and more than 1000 users are omitted. Hexagons in the user count plots are colored by the amount of cell towers that are found within their area. The number of hexagons in the x-axis direction of each user count subplot is set to 80.

Indicators related to contact networks

Focusing on the contact networks of users in figure 3.5, indicators such as the amount of contacts and the average number of calls per contact depict little surprises. Indicators on contact networks display little variation between cell towers and a quite large variation within cell tower areas, are thus between users. Extreme cases that are observed for both the highest and lowest ranks can partially be attributed to sample size (user counts) as ranks of outliers are also related to ranks where lower user-counts were observed. In general, the figures show that median users in France during September 2007 contacted about 19 other peers and performed an average of 4 interactions with each peer.

One interesting observation with regard to the entropy of contacts is that, regardless of the rank of the cell towers, the percentile values of the within cell tower distributions are symmetrically distributed with respect to the median values. This property is related to the definition of the indicator (equation 2.1 in section 2.6.1) and is indicative of the potential of the entropy measure to render complex information (in this case the spreading of calls over different numbers of users) comparable over many different situations. A similar observation can be made for the mobility entropy indicator in figure 3.6.

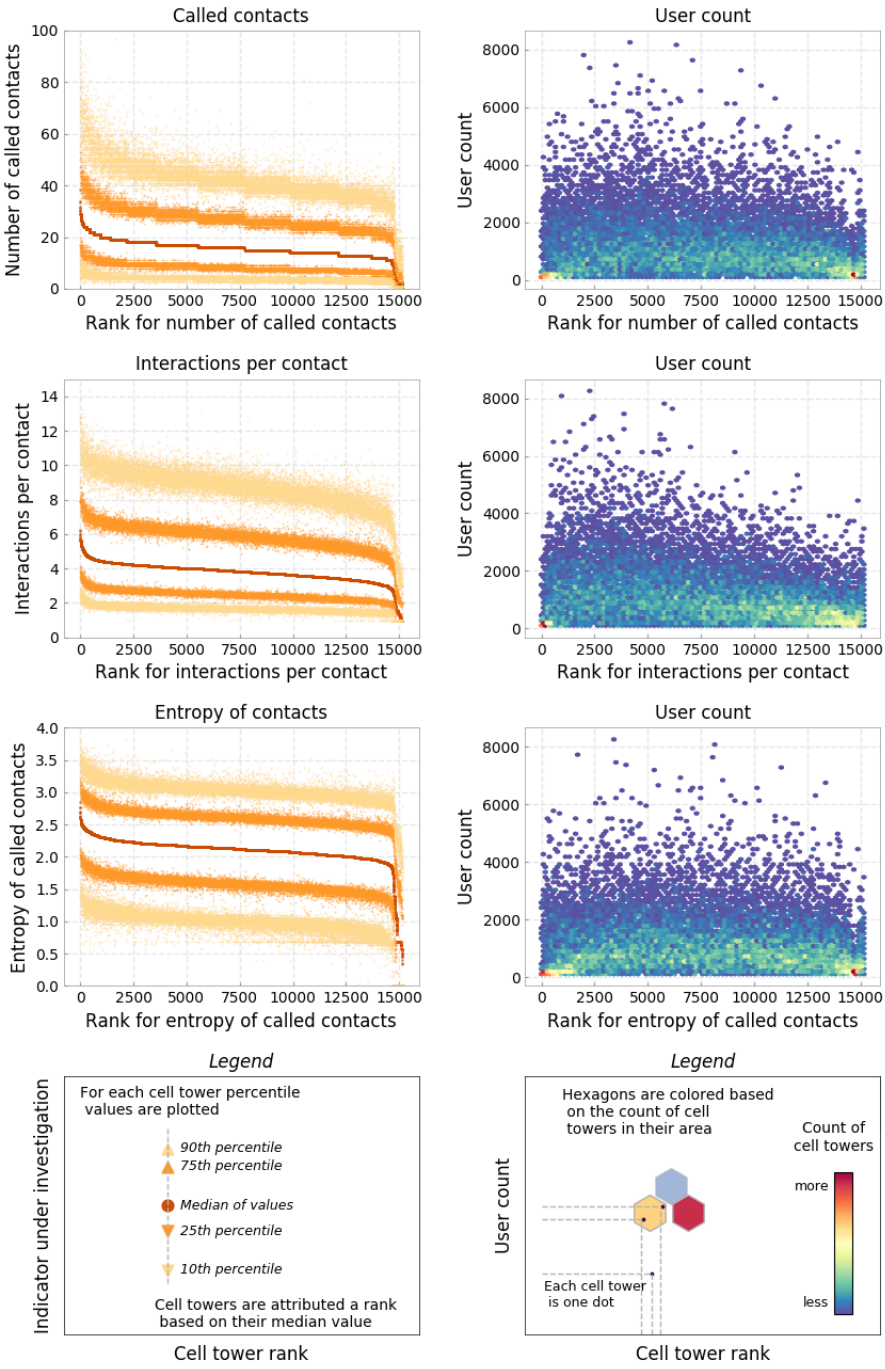


Fig. 3.5 Distributions at cell tower level of mobile phone indicators related to contact networks. Figure setup is equal to figure 3.4.

Indicators related to movement patterns

For indicators that describe movement patterns, as shown figure 3.6, larger variations between cell towers are observed. The amount of visited cell towers and the mobility entropy show large variations between cell towers with some cell towers having median values as high as the 90th percentile values of other cell towers. Regarding the amount of visited cell towers this is not at all surprising given the heterogeneous spatial distribution of cell towers in France (figure 3.1). It is only logical that users with home locations estimated in high density cell tower areas will have higher amounts of visited cell towers. What is interesting is that the mobility entropy, which expresses the diversity of users' movement patterns (equation 2.3 in section 2.6.2), displays a large variation between cell towers. This might be due to differences between movement patterns between locations but chapter 6 will argue that it is also partly due to the traditional definition of mobility entropy being biased towards cell tower density.

The distributions of the radius of gyration, a popular measure for the volume of individual movement (equation 2.2 in section 2.6.2), are peculiar too. What is remarkable is the high variability of within cell tower distributions, especially between cell towers that have similar ranks. This can be observed by, for example, the 75th and 90th percentile point clouds that overlap throughout the figure, meaning that cell towers with similar median values (similar ranks) can have 75th and 90th percentile values, respectively, that are similar. The figure on the user counts rules out that this effect is due to the sample size within cell towers (that is, the distribution of the sample sizes over the rank is similar to that of the other indicators). This means that the variation of the radius of gyration within cell tower areas is high, for reasons unknown. One potential reason is that the processes which drive variation in the radius of gyration are active at a very local level. Another possible reason is that the radius of gyration simply is an inconsistent measure to describe movement patterns when derived from CDR data but, again, reasons for this are unknown.

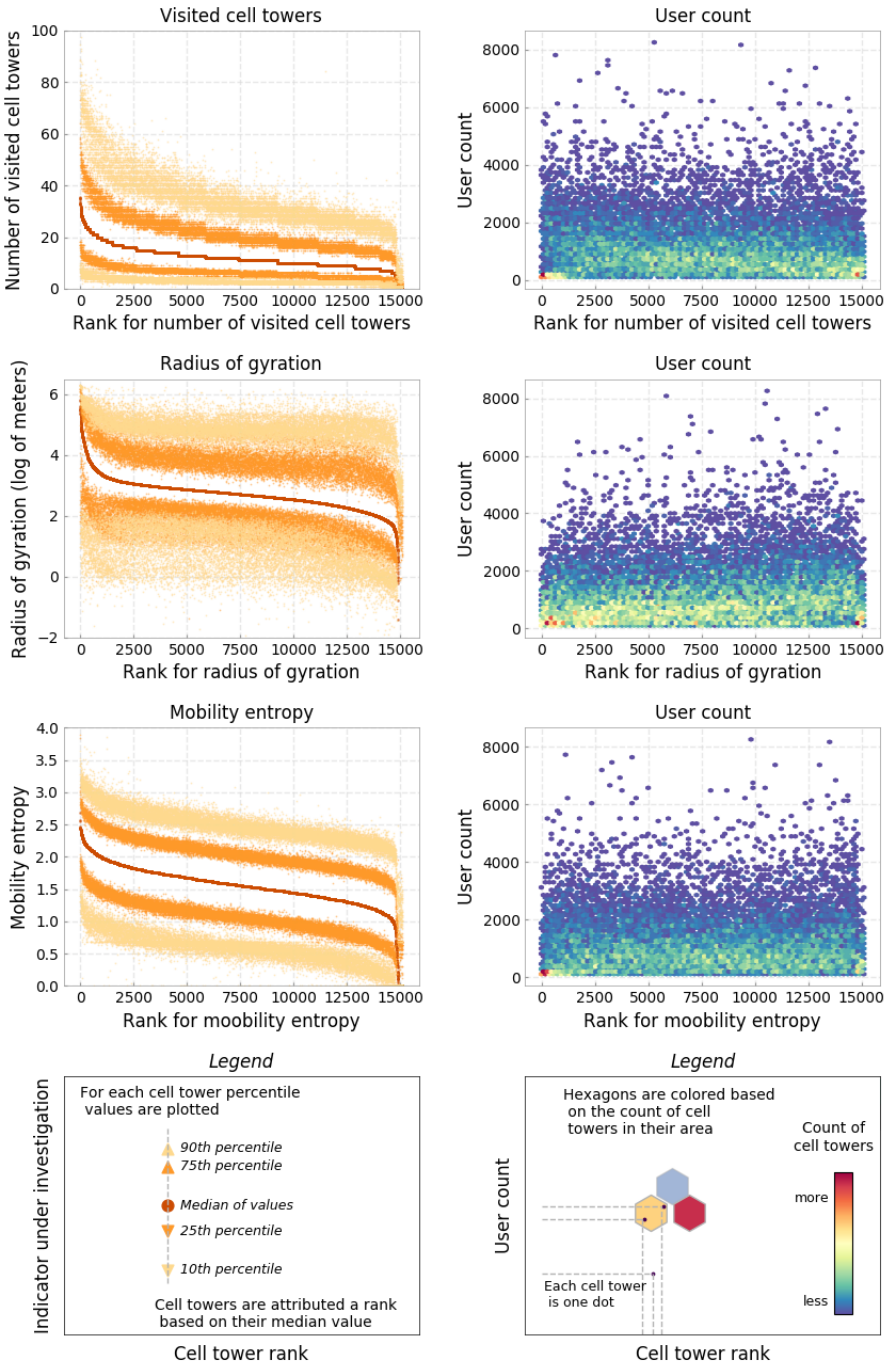


Fig. 3.6 Distributions at cell tower level of mobile phone indicators related to movement patterns. Figure setup is equal to figure 3.4.

Indicators related to home detection

A final set of indicators are shown in figure 3.7 and are related to home detection, which in this analysis is performed by the *maximum action algorithm*. The distributions of the number of calls made at home seems to differ little from the number of calls made in figure 3.4. Strangely, the variation between cell towers for the percentage of calls performed at home is larger and shows a different pattern. For the most extreme cell towers, and such effect does not seem to be attributable to small user counts, the median values reach close to 100%, meaning that at those particular cell towers, median users perform all their calls at the presumed home cell tower. Most cell tower medians, however, range between 60% and 40%, with variations on the percentage of calls made at home being rather high: the interquartile range is around 30% and the 10th-90th percentile range is about 60%).

Given the use of the maximum actions algorithm, the home of a user in this analysis is defined as the cell tower on which most activities have been performed. In other words, the distributions of the percentage of calls at home in figure 3.7 show the percentage of activities users have performed at the most used cell tower. One of the conclusions that could be drawn from this figure is that median users perform around half of their calls on the same cell tower, leaving relatively little room for visits to other cell towers, and thus movement patterns to be captured. If the most used cell tower indeed corresponds to users' home locations, then it can be said that users display a quite large variation in the percentage of calls they perform at home, rendering the question to which degree do home detection algorithms deal with this variability?

The wide variation in the percentage of calls at the most used cell tower also raises questions on the (un)certainty of home detection. Looking at the lower bound of the cell tower distributions in figure 3.7; that is the 10th percentile, one can observe that only 20% to 30% of calls are made at the most visited cell tower, leaving users with 80% to 70% of activities at other cell towers. For these users, in other words, there might be several cell towers that have similar (lower) percentages of calls, and so the question raises how a home detection algorithm with discern between them.

Something similar is true for users that have very high percentages of calls at their most used cell tower. Some 90th percentiles in figure 3.7 reach as high as 80% or even 90% of calls made at the most used cell tower. Users performing 90% of their calls at one cell tower, leave very little room for home detection methods to consider other cell towers. In such cases, it is very likely that multiple home detection algorithms will render the same cell tower as home, simply because it is one of the few a user has been active at. Having multiple home detection methods finding the same cell tower as presumed home, however, does not necessarily mean that this cell tower is the best home location estimation. Clearly, an argument such as: "this cell tower is deemed a home location because it is the only cell tower for which we have observations", is not very convincing. The use of cell towers by all users, in other words, is a given characteristics for the CDR dataset but implies an uncertainty for home detection as will be discussed in much more detail in chapters 4 and 5.

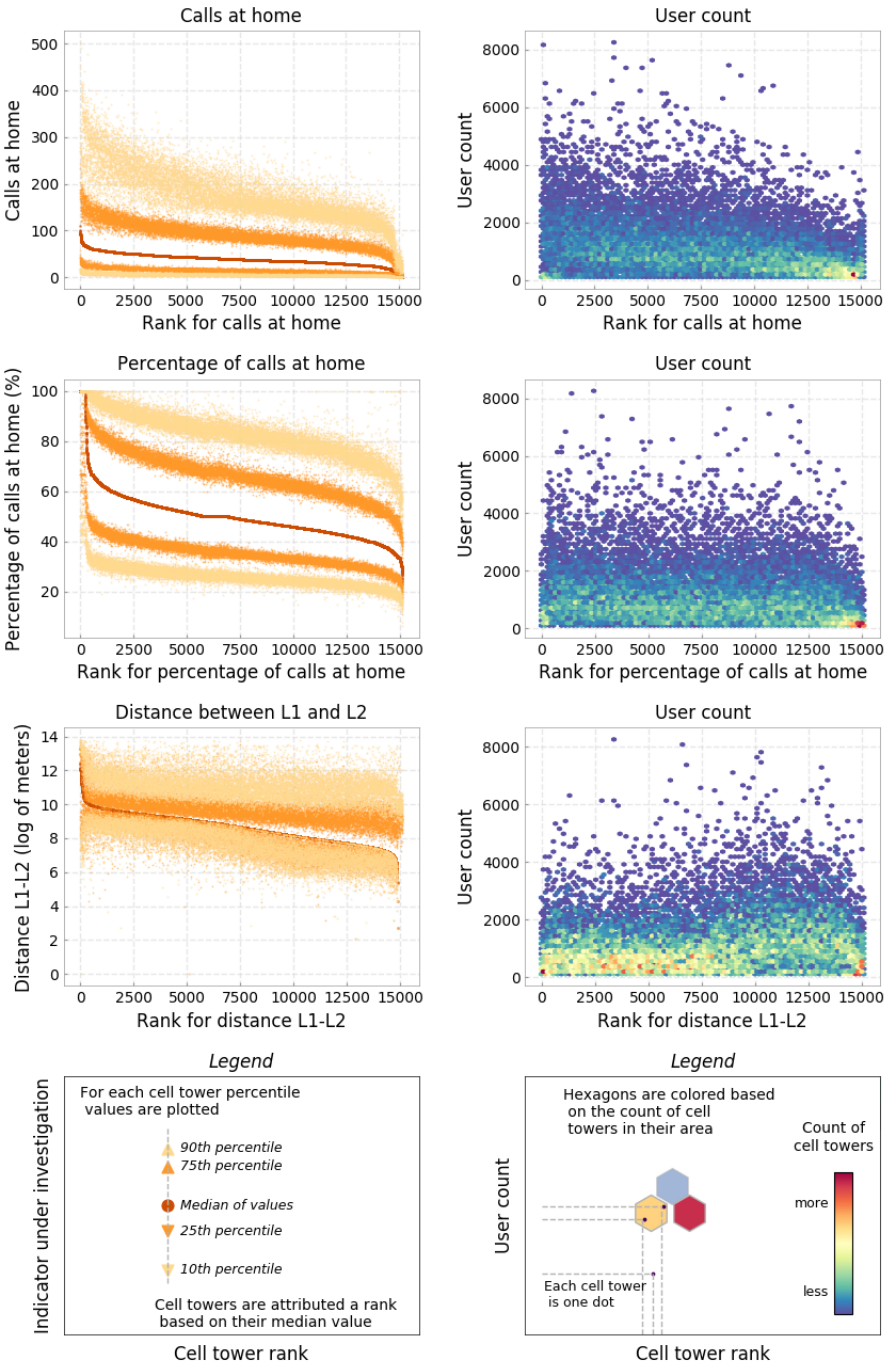


Fig. 3.7 Distributions at cell tower level of mobile phone indicators related to home detection. Figure setup is equal to figure 3.4.

Ultimately, the distributions of the distance between the cell tower that was used most, denoted by L1, and the cell tower that was used the second most, denoted by L2, in figure 3.7 are intriguing at least. Similar to the radius of gyration in figure 3.6, the L1-L2 distance shows a very high variation between cell towers with similar ranks of the within cell tower variation. Compared to the radius of gyration, the L1-L2 distance is even more prone to this finding, with point clouds of the 10th and 25th, and the 75th and 90th percentile intervals overlapping. Even more remarkable is that the 75th percentile values, especially for the lower ranks, are close to median values. And the most remarkable is that 25th percentile values, irregardless of the rank, are close, if not similar to median values. The L1-L2 distance, in other words, is remarkably equal for users living at the same cell tower but clearly different between cell towers. One potential explanation is that the L1-L2 distance is very locally defined by the density of cell towers. Based on the median values in figure 3.6, typical distances between L1 and L2 range from 2.9 km (e^8) to 22 km (e^{10}), which seem plausible distances to be dictated by differences in cell tower densities between locations in France.

The consequence is that measuring and comparing short movement at nation-wide scale is restricted at the lower bound, by the spatial resolution of cell towers and their differing densities. Observing that, for many users, the displacement between the two most actively used cell tower accords to short movement raises the question on how objective indicators such as the radius of gyration (which takes into account the distance and the center of mass between all cell towers as a starting point) really are. Such measures might, to a certain degree, describe structural and locational differences in the spatial resolution of the cell tower network rather than behavioral differences between residing populations. A similar question can be posed with regard to the derivation of commuting patterns. Multiple studies have been labeling the L1 and L2 locations from home detection as, respectively, home and work, consequently deriving commuting patterns from them. Although such labeling could be correct, if L1 and L2 locations are typically close to each other, the retrieved figures concerning the distance of commuting might partly reflect structural measures of cell tower density instead of actual commuting distances.

3.3.3 Temporal Patterns of Mobile Phone Indicators

Distributions of most mobile phone indicators at cell tower level are found to be invariant over time when calculated per month. The distribution of the percentage of calls at home in figure 3.8 shows an example of this invariance for the the months June, July, and August (figure 3.8), with the month of September shown in (figure 3.7). Other indicators depict a similar invariance to time but are not shown. Two indicators, however, form an exception and do have changing patterns of statistical distributions over time: the radius of gyration and the distance between L1-L2. Both are depicted in figure 3.8 and shown different patterns in July and August compared to June, and September (as can be observed in figure 3.6 and 3.7, respectively). One plausible reason for this temporal change in distributions is holiday movement of the French population during summer months July and August. This will be discussed further in chapter 5.

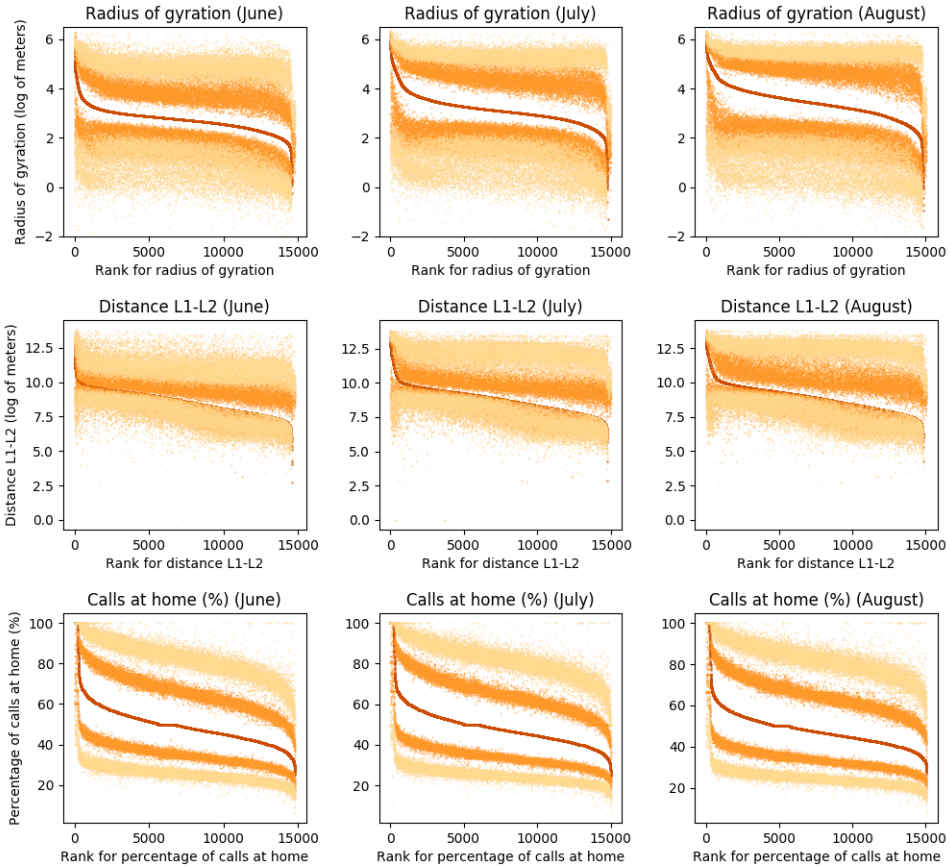


Fig. 3.8 Distributions at cell tower level of three mobile phone indicators for different months. Figure setup is equal to figure 3.4.

3.3.4 Spatial Patterns of Mobile Phone Indicators

The statistical distributions of mobile phone indicator values only offer a partial view on the millions of observations available. Another view can be obtained when investigating spatial distributions. In this case user-territory mapping is a prerequisite too as aggregation at cell tower levels assigns a location for the value distributions. Figures 3.9 and 3.10 map the average indicator value for each cell tower.

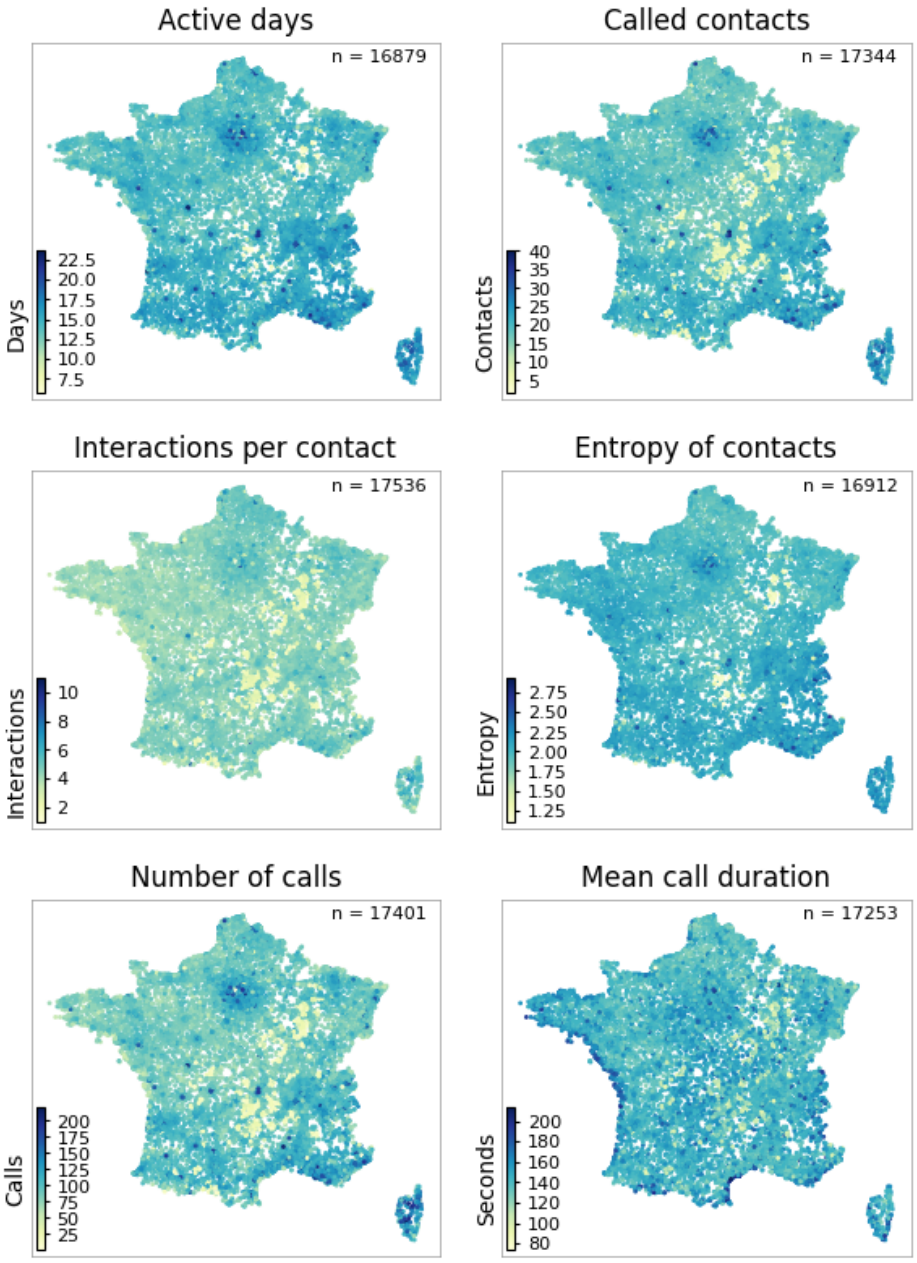


Fig. 3.9 Spatial distributions at cell tower level of mobile phone indicators related to calling behavior and contact networks. Indicators are calculated for the month September, for all users in the French CDR dataset, and users are aggregated at cell tower level by the Maximum Action algorithm. Cell towers with an average value that is higher, or lower, than 3 standard deviations from the nationwide average are omitted.

One observation is that indicators related to calling behavior and the contact network do not depict clear spatial patterns. The mean call duration, for example, is distributed pretty uniform across the country, with higher values being found in remote and coastal areas. One potential explanation could be that domestic tourists visiting these areas call longer, but this has to be investigated. Concerning the volume of calls, the spatial patterns is consistent for different indicators such as the number of calls, the number of contacts, and the number of active days. In general, mobile phone activity is higher in urban areas, lower in the center-east part of the country (predominantly rural areas) and there is a gradient from north to south with the south of France showing slightly higher cell phone activity.

The indicators that relate to user movement display more pronounced spatial patterns (figure 3.10). The radius of gyration, for example, is clearly higher in touristic areas than elsewhere in the country. High values in touristic areas are either due to long distance commuting, or to domestic tourists that get detected a home location at touristic places but for which the actual place of residence is rather far away. Given its definition, the radius of gyration is largely related to the distance between L1 and L2, and so it is no surprise that their spatial patterns accord to a large degree.

A second spatial pattern is found between the number of visited cell towers and the mobility entropy (figure 3.10). In this case, however, the mobility entropy would by definition by expected to eliminate a relation between both. And so the interpretation of the pattern should be that the diversity of individual movement (captured by the mobility entropy) is higher in city centers and surrounding areas which, coincidentally, are also the locations in which users visit more cell towers, mainly because the density of cell towers in these areas is higher. In chapter 6, this interpretation will be proven wrong when a bias is uncovered in the calculation of mobility entropy with regard to cell tower density.

The most surprising spatial pattern concerns the amount of calls at home, which in the case of the *maximum action algorithm* is equal to the cell tower that was used most by a user. Concerning the absolute number of calls performed at the most used cell tower, clear regional differences are revealed with Northern regions, such as Nord-Pas-de-Calais, Champagne, Alsace, Bretagne, and Normandy, showing higher values than southern regions or Paris. Concerning the percentage of calls performed at the most user cell tower, there is a uniform distribution across the entire country, except for some rural regions in the center-east of the country that show very high values.

Focusing on these rural regions, one can observe their atypical behavior for many indicators, both related to calling behavior and movement patterns. These regions are characterized by lower amounts of activities and limited movement. The most straightforward interpretation is that users in these regions use the cell phone significantly less than other users in France. Combined with the observation that cell tower density in these areas is low, meaning that the distance between neighboring cell towers is large, another interpretation can be that movement patterns of the users in these areas are insufficiently captured by CDR data.

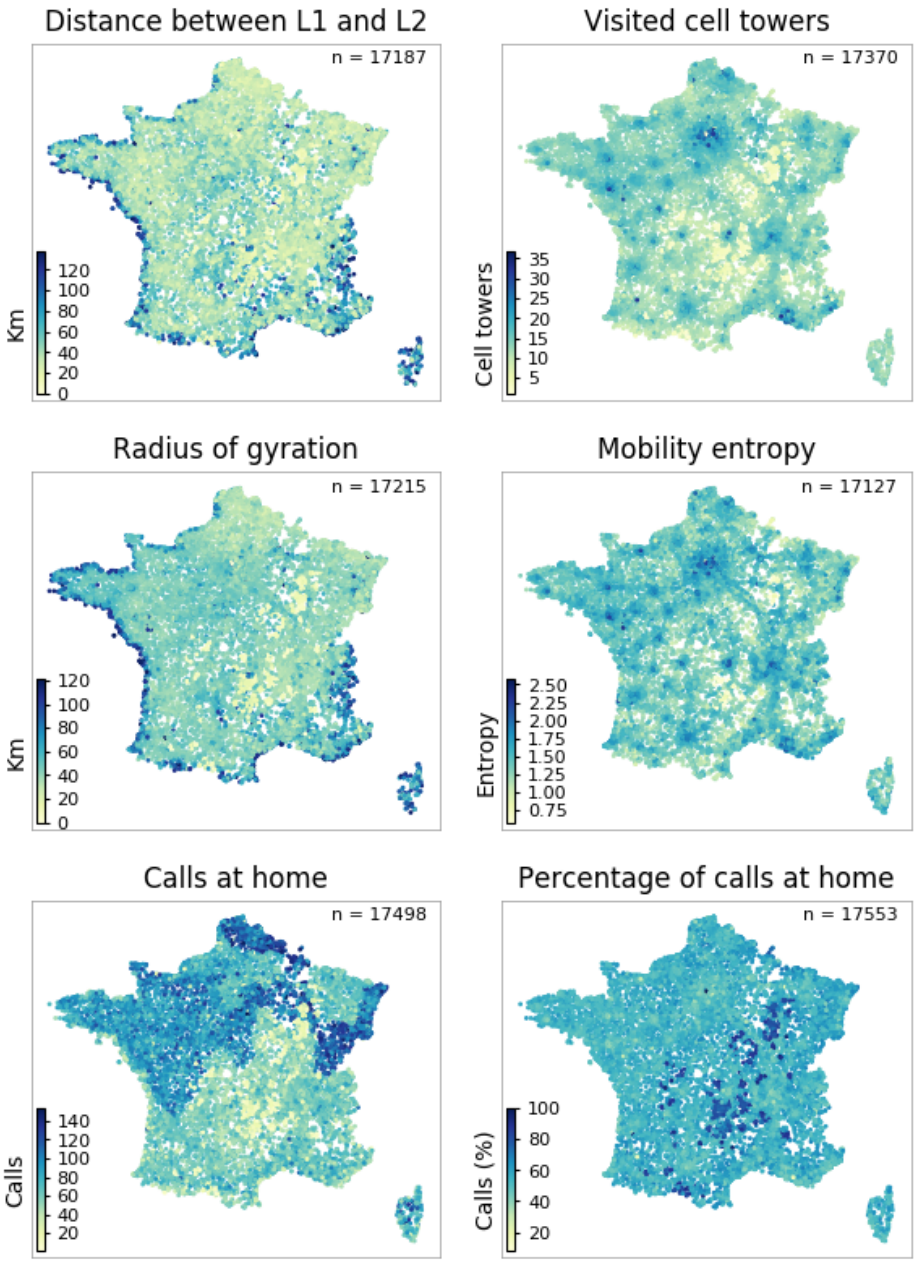


Fig. 3.10 Spatial distributions at cell tower level of mobile phone indicators related to movement patterns and home detection. Figure setup is equal to figure 3.9.

3.4 Relating Mobile Phone Indicators to Socio-Economic Indicators

3.4.1 Socio-Economic Indicators

Aggregating mobile phone indicators at cell tower level to municipality level in France enables the investigation of their relation with socio-economic indicators from census data. Two socio-economic indicators for the year 2007 are collected: per capita income, and the European Deprivation Index (EDI) [102]. Per capita income measures the average wealth of individuals living in a municipality, that is, the sum of the incomes gained by the residents of a given municipality divided by the number of residents. Per capita income offers only a very limited view on the socio-economic status of a territory and so the EDI is investigated also. The EDI reflects experienced poverty, and combines indicators such as overcrowding, no access to a car or electric heating, unemployment, and low education level into a single score by a linear combination with predefined weights for different countries in Europe [102]. The EDI, in other words, is a composite index for subjective poverty, the higher its value, the lower the well-being of the municipality is. Preliminary validation showed a high association between the French EDI and both income values and education level in French municipalities, partly supporting its ability to measure socio-economic development and well-being [102].

3.4.2 Correlations between Human Mobility, Calling Behavior and Socio-Economic Indicators

Figure 3.11 shows the relations between mobile phone indicators and socio-economic indicators. Two main findings stand out. First, the number of contacts is not correlated with the socio-economic indicators, while the radius of gyration is. Second, findings suggest that mobility entropy is a better predictor for socio-economic development than the entropy of contacts. The latter relation, was also revealed by [54] for regions in the UK. For the mobility entropy, clear tendencies appear: when the mean mobility entropy of municipalities increases the EDI decreases, while per capita income increases. For the entropy of (called) contacts, the relation with EDI is lower and no correlation with per capita income is found.

Figure 3.12 provides another view on the relations between mobility entropy, entropy of contacts, and EDI at municipality level. Municipalities are split into ten deciles according to the values of EDI. For each decile, distributions of the mean mobility entropy and mean entropy of contacts for the municipalities in that decile are plotted. For mobility entropies, when the deciles of EDI increase, the mean of the distribution increases, as does the variance, highlighting a change of the distribution in the different decile groups. This finding is consistent with the observations in figure 3.11. Conversely, for the entropy of contacts, distributions do not depict a significant change when increasing EDI deciles.

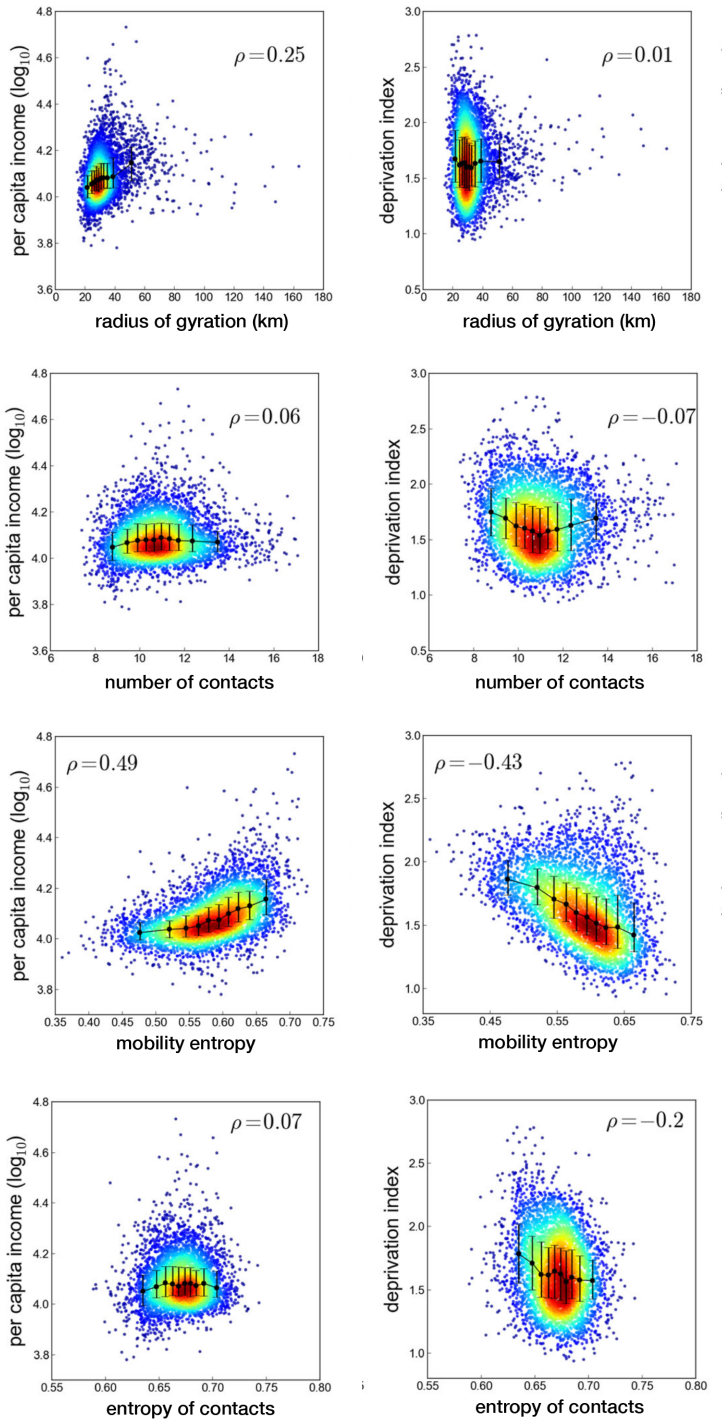


Fig. 3.11 Relations between per capita income, European deprivation index, radius of gyration, number of contacts, mobility entropy, and entropy of contacts at municipality level: The color of a point indicates, in a gradient from blue to red, the density of points around it. Municipalities are split into ten equally-sized groups according to the deciles of the measures on the x axis. For each group, the mean and the standard deviation of the measures on the y axis are calculated and plotted (black error bars). Municipalities with less than 1,000 official residents were omitted from the analysis. ρ indicates the Pearson correlation coefficient. In all the cases the p-value of the correlations is <0.001. Source: figure adapted from [95].

One should, however, be cautious when interpreting figure 3.12: although for the mobility entropy a trend can be seen, median values of the individual deciles still fall within the bulk of the distributions of all other deciles. Even though differences between the distributions at different EDI deciles could be significantly different (mainly because of their large sample sizes), the variation of distributions is high, indicating that a translation of the high-level trend to individual cell towers is not straightforward. Still, the observed variation of the mobility entropy distribution in the different EDI deciles is an interesting finding when compared to previous works such as [126], that suggest mobility entropy to be stable across different subpopulation defined by personal characteristics like gender or age group. Figures 3.11 and 3.12 suggest that the diversity of human mobility aggregated at municipality level in France is better associated with socio-economic indicators than with socio-demographic characteristics discussed in other works.

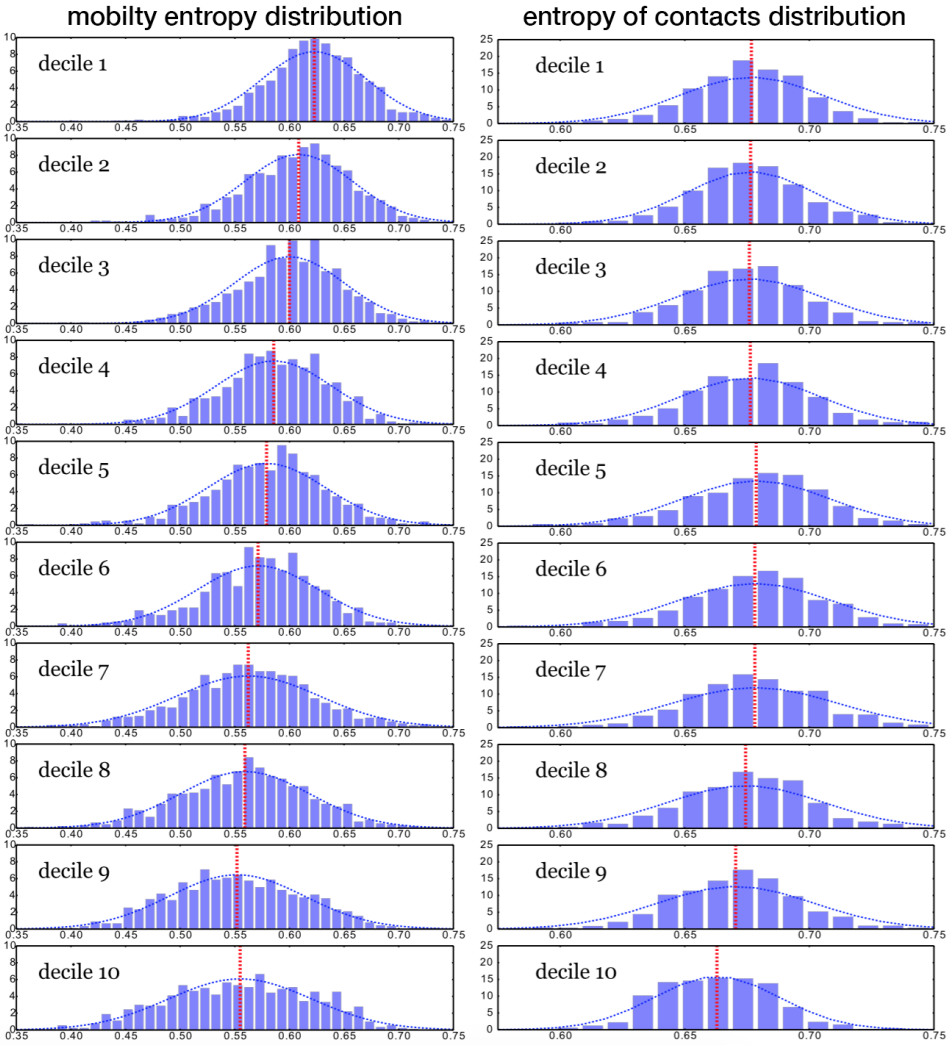


Fig. 3.12 Distributions of (a) mobility entropy, and (b) entropy of contacts for different deciles of EDI. Municipalities are split in ten equal-sized groups according to the deciles of the EDI values. For each group, distributions of the mean mobility entropy and entropy of contacts per municipality are plotted. The blue dashed curve is a fit of the distribution, while the red dashed line is the mean of the distribution. Municipalities with less than 1,000 official residents were omitted from the analysis. Source: figure adapted from [95].

3.5 Discussion

3.5.1 A Situated Dataset

In the fist part of this chapter, characteristics of the French CDR dataset and its collection process are described. These characteristics are specific, in the sense that they are determined by the way the operator collects the data, the characteristics of the user population, the historical context, etc. As such, characteristics will differ between datasets and influence the comparability of findings in all analyses. For the French CDR dataset, a good example of this *situatedness*) are the text events. Although shown to make up for about 46% of the records in the dataset (table 3.3), due to technical reasons text events can only partially be used to construct movement patterns and not for constructing contact networks. This restriction might not be the case for CDR data captured in different systems, but equally the share of text events in the Orange CDR dataset could very well change over time, challenging the transferability of findings.

3.5.2 Sufficient Sample Sizes

One aspect that balances out the situated nature of CDR data is the extremely large population for which information is captured. When translated into indicators, this information offers diverse distributions (both statistically and spatially) enabling the study of human behavior patterns at unprecedented depth and with high statistical significance. The creation of indicators reveal wide distributions at user level, for example, with respect to the amount of calls, or the number of contacts called (figure 3.3). Similarly, distributions at cell tower level show differences within cell tower coverage areas (18,275 small areas in France), as well as between them, and between wider regions, showing the wide scope of analyses that is possible with CDR data (figures 3.4, 3.4, 3.6, 3.7, 3.9, and 3.10). Only sporadically, outlier indicator values at cell tower level are suspected to result from too small sample sizes, showing the real magnitude of the collected data.

3.5.3 Objectivity of Territorial Aggregation

The extreme size of the French CDR dataset imposes the need for aggregation, as indicators for all users in the dataset (around 18.5 million) are simply un-treatable when it comes to more advanced analysis. The territorial aggregation at cell tower level based on home detection is proposed (both in the literature and in this chapter) and partially solves the problem. Territorial aggregation as compared to, for example, random aggregation, offers two main advantages. First, it permits indicators to be investigated spatially, which in turn enables spatial analysis. Secondly it facilitates the combination with other spatial datasets such as census data, land-use maps or satellite images.

Despite solving a technical problem and opening analytical possibilities, there are downsides to territorial aggregation too. For starters, it is difficult to assess how correct the territorial aggregation of CDR users actually is. More often than not, home detection methods are deployed to aggregate individual users to cell towers, but the validity of this approach is badly understood. Chapters 4 and 5 will focus on the validity of home detection for the French CDR dataset. Additionally, questions remain on further steps of territorial aggregation such as the translation from cell tower level to municipality level necessary to relate mobile phone indicators with census data. Are analyses sensitive to this two-step territorial aggregation or not? Are the observed relations between mobility entropy and income per capita, as observed in figures 3.11 and 3.12, independent of territorial aggregation or not? Chapter 7 will explore this question in more depth.

3.5.4 Studying the Territory with Mobile Phone Indicators

Ignoring the uncertainty related to territorial aggregation, an outstanding question is whether the observed (spatial) distributions of mobile phone indicators in figures 3.9 and 3.10 could help to better investigate or understand the French territory. Clearly, the relations between mobile phone indicators and census data, as illustrated in figures 3.11 and 3.12, suggest potential, but questions on the trustworthiness of mobile phone indicators need to be posed, especially since no validation studies on mobile phone indicators have (yet) been carried out.

For example, the considerations expressed on the locality of movement patterns observed in CDR data, their relations to cell tower density, and how this might influence spatial patterns of the radius of gyration and L1-L2 distances measures in figures 3.6, 3.7 and 3.8, are illustrative to why caution might be due. In chapter 6, an investigation on the influence of cell tower density to the calculation of the mobility entropy indicator will be carried out, revealing a clear bias that is challenging the trustworthiness of its observed (spatial) distributions (figure 3.6 and 3.10), as well as the uncovered relation between mobility entropy and per capita income or EDI in figures 3.11 and 3.12.

Mobile phone indicators, in other words, show great potential for scientific research mainly due to their magnitude of observations, but several questions remain especially regarding the transferability and reproducibility of findings, their own validity and the influence of related practices such as territorial aggregation on findings.

Chapter 4

Evaluating Home Detection Performance

> Home is a place where you grow up wanting to leave, and grow old wanting to get back to.
>
> *John Ed Pearce*

Abstract

This chapter investigates the performance of five Home Detection Algorithms (HDAs) with simple criteria when applied to the French CDR dataset introduced in chapter 3. Based on a unique validation dataset, different performance measures for home detection at the nation-wide scale are created and findings between different HDAs are discussed. In addition, insights on the performance of HDAs are drawn from their pairwise comparisons, as well as from the investigation of metadata related to their inner workings. The results presented are the first to approximate home detection performance for CDR data at the nation-wide scale. They show the limits of nation-wide population estimations from CDR data and lead to discussion on the estimated error and uncertainties. Such discussion is important as home detection forms a prerequisite step for many (research) applications based on CDR data (section 2.7).

Related Publications and Acknowledgments

- The structure of this chapter is based on [142], a publication authored by the PhD candidate. The analysis and writing of this paper was done by the PhD candidate.
- Different parts of the analyses in this chapter have been presented by the PhD candidate. Contributions in the form of full papers or abstracts were made to the Mobile Tartu 2016 conference, the ESS big data workshop 2016 and the New Techniques and Technologies for Statistics (NTTS) 2017 conference. These contributions are authored by the PhD candidate. They can be found online but have not officially been published.
- Acknowledgments go to the INSEE team: Stéphanie Combes, Marie-Pierre de Bellefon, Benjamin Sakarovitch, Pauline Givord, and Vincent Loonis for the construction of the ground truth dataset, to Fernando Reis, Michail Skaliotis and Dr. Zbigniew Smoreda for their help in establishing the partnership between INSEE and Orange Labs France.

4.1 Detecting Homes from CDR Data

In section 2.7, a literature study reviewed the detection of meaningful places based on CDR data, with a focus on automated home detection. It was argued that beyond the works performed on small-scale continuous traces, the majority of the works on CDR data faces several methodological challenges. Most notably, it was remarked that Home Detection Algorithms (HDA) for large-scale data lack ground-truth data at the individual level to develop learning methods or evaluation criteria. As a consequence, the majority of HDA use simple and implicit criteria for the semantic annotation of user traces on which no consensus or assessment of sensitivity exists in literature. Additionally, it was observed that assessments of errors for HDA are restricting themselves to nation-wide comparison with census data, since no framework has been developed that targets lower levels in the absence of the ground truth. All of this fundamentally limits the utility of current home detection methods and limits the potential of CDR data for systematic exploitation.

In this chapter, multiple ways are proposed to evaluate the performance of home detection at the nation-wide scale. Home detection is performed on the French CDR dataset and performance evaluated according to the proposed assessments. The results form a first, systematic evaluation of the performance of single-step HDAs when applied to CDR data at the nation-wide scale and point out several issues with regard to high-level validation, unknown local market shares, sensitivity to time period and criteria, absence of individual level data, etc.

4.1.1 Five Home Detection Algorithms with Simple Criteria

The literature study in section 2.7 revealed how current HDAs have converged to single-step approaches that apply complex or simple decision rules to all users in a dataset. Decision rules themselves consist of one criterion or multiple criteria, in which a definition of *home* is elaborated. A current gap in literature is that it is unclear to which degree such criteria influence the performance of HDAs, at the nation-wide scale, but also at individual level or the level for subset of users. Such a gap is due to the absence of individual level validation data, obliging researchers to use high-level validation only. Another reason is the simple lack of a systematic investigation on the effect of criteria choice when applied to one or, preferably, multiple CDR datasets.

To perform a systematic investigation for the French CDR dataset, five HDAs are constructed, each incorporating one criterion only (leading towards HDAs with simple decision rules only). The five criteria are selected from the investigation of literature [6, 59, 71, 31, 134, 37, 100, 45, 78] and have been deployed independently [134], combined [6], within simple decisions rules [100], within complex decision rules [45, 59], but always in single-step approaches. They are explained in table 4.1. Note that [27] use similar criteria when assessing home detection for a credit card transaction and a Flickr dataset. This means that the relevance of these criteria goes beyond the case of mobile phone data serving other datasets with non-continuous location traces too.

Criteria	Acronym	Description: 'home is cell tower where:'
Maximum Amount	MA	Most activities occurred
Distinct Days	DD	The maximum active days were observed
Time Constraints	TC 19-9	Most activities occurred between 19.00 and 9.00
Space Constraints	SC	Most activities occurred on all cell towers within a perimeter of 1km
Time and Space Constraints	TC-SC 19-9	Most activities occurred between 19.00 and 9.00 on all cell towers within a perimeter of 1km

Table 4.1 Description of deployed HDAs

4.1.2 Deployment and Metadata of HDAs

HDAs are executed for all users in the French CDR dataset in a similar way to the calculation of mobile phone indicators (section 3.2.1). HDAs are coded as UDFs coupled to a Pig script taking care of the data allocation for each user. Home detection was performed for all users in the dataset and for all different months (May to October 2007), resulting to a total of 546,840,595 detections (on average 18.2 million users in the dataset $\times$ 6 months $\times$ 5 HDAs).

Developing HDAs as UDFs has the advantage that metadata on the working of the algorithms can be derived and stored. The amount of calls and the percentage of calls made at the presumed home locations, as well as the distance between L1 and L2 used in the previous chapter (chapter 3) form examples of the metadata. The most important pieces of metadata, besides the actual resulting cell tower annotated as the home, are the second and third most plausible cell towers for home as defined by the different algorithms for each user. For the remainder of this thesis, these cell towers will be referenced as, respectively, L2 and L3, with L1 being the actual detected home location (cell tower) by a particular algorithm. Note that, in the previous chapter (chapter 3), this notation, has already been deployed in the *distance between L1 and L2* indicator.

Given that home detection needs to favor one cell tower over all potential others, it is worth exploring the importance of cell towers in terms of observed activity for users in the dataset. For each user, the share of activities performed in each of their top 3 most frequently used cell-tower was calculated. The percentiles of the shares for the whole French dataset are given in figure 4.1. The importance of the cell tower with most observations, the L1, is clearly visible. Its shares range from 40% (5th percentile) to 100% (95th percentile) with a median of 64%. Shares of L2 and L3 are substantially lower with medians of 23% and 10% and a 95th percentile of 41% and 27%. These observations support the decision to limit the analysis to three plausible locations only and are similar to [45] who found 95% of the users in a Portugal CDR dataset have fewer than four frequent locations.

The observations in figure 4.1 imply that, for the French 2007 dataset, HDAs operate in a rather restricted space, in the sense that an average user performs a large share of his/hers activities at a limited amount of cell towers. Home detection, in other words, is limited in the first place by the available data for all users in the dataset. In this perspective, good knowledge on both the performance of home detection and the differences between HDAs is important. It might, for example, very well be that different HDAs are forced toward similar home detection results simply by the nature of the data. But this does not mean that such detection is necessarily correct. Quantification of performance and of differences between algorithms, in other words, are complementary sources of information when it comes to the assessment of validity of home detection practices.

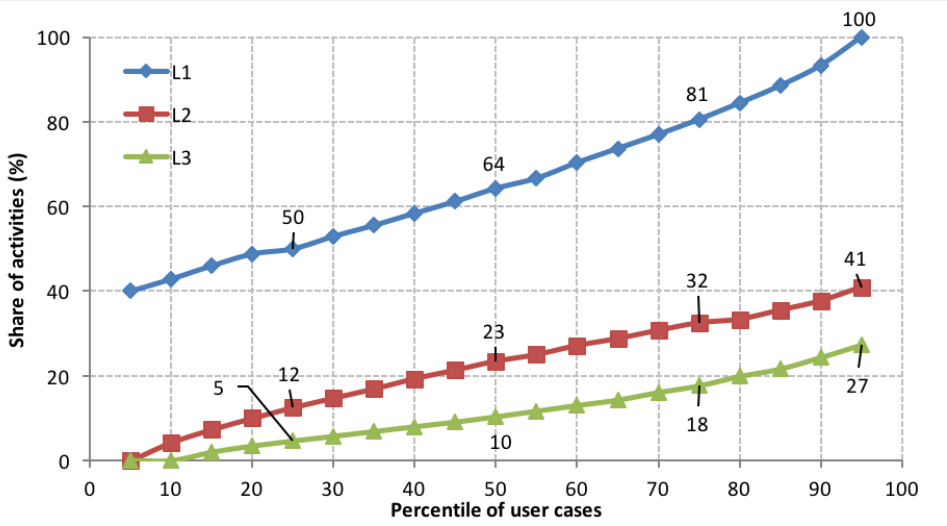


Fig. 4.1 Percentage of activities performed on the most, the second most, and the third most frequently used cell towers for different percentiles of users in the French CDR dataset. Values for the 25th, 50th (median), 75th and 95th percentile are given.

4.2 Measuring Performance of Home Detection at Nation-Wide Scale

Due to the absence of individual level validation data, assessing the performance of home detection at nation scale is not straightforward. The following subsections describe three different way to perform assessment, namely high-level validation based on census data, comparisons between HDAs, and uncertainty assessment based on HDA metadata.

4.2.1 High-level Validation Based on Census Data

The most deployed validation methods for home detection are comparisons to high-level census data. This section discusses one potential validation method, namely the comparison of population estimates that form the outcome of all HDAs, with ground truth population counts gathered by census data. To fully understand the high-level validation methods, a clear description of the ground truth dataset is provided and different performance measures are discussed.

Constructing the Ground Truth Dataset

In collaboration with INSEE (the French Official Statistics office), a validation dataset was constructed in the form of a nation-wide, population dataset aggregated at the level of the Orange cell tower network. To construct this validation dataset, residential population numbers were mobilized, that are collected by the Public Finances Directorate General (DGFIP) from revenue declarations, the housing tax and the directory of taxable individuals in the form of geo-localized individual or household residential locations. Subsequently, the French National Statistics office calculated aggregated population counts at the Orange network level by aggregating home locations to the nearest cell tower. The spatial pattern of the aggregated population count can be observed in figure 4.2.

The advantage of constructing a ground truth dataset at the spatial resolution of the cell tower network is that it avoids the spatial translation of statistical sectors to cell tower coverage. Such a translation is, given the spatially non-homogeneous distribution of cell-towers, complicated and prone to errors [59]. In addition, as a result of the demand driven distribution of cell towers, the cell tower resolution offers a much higher resolution for densely populated areas, compared to standard statistical sectors. As such the advantage is twofold: it consists of both a higher resolution of information and a higher accuracy.

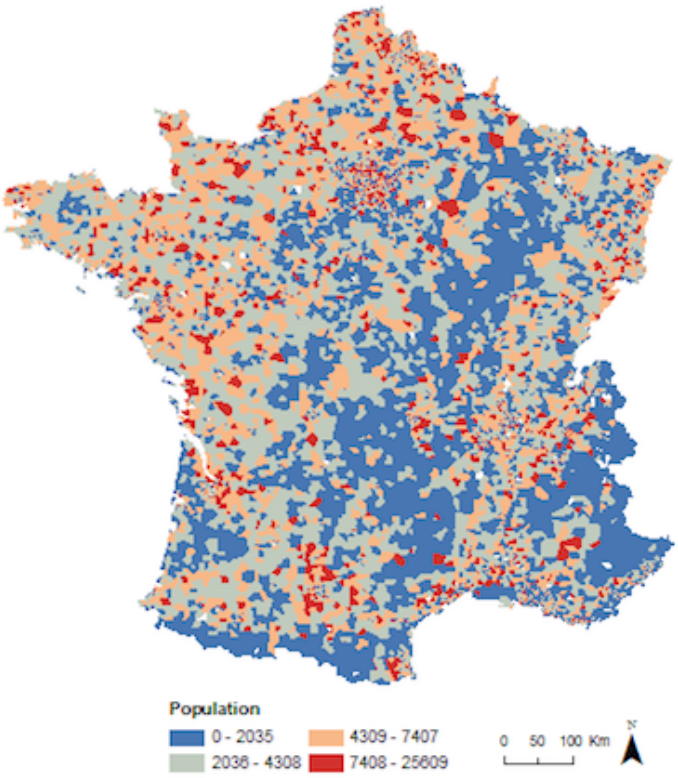


Fig. 4.2 Population counts of the validation dataset at the Voronoi polygons of the Orange cell towers.

The disadvantage of the ground truth dataset is that the large-scale, high resolution home location information could only be made available for the year 2010, while the mobile phone dataset was collected in 2007. Still, the use of this validation dataset is opted for over the low resolution, publicly available census data that are updated every year, mainly to avoid the spatial translation problem explained above. In this perspective, in the analysis of performance, the validation dataset is mainly used for relative comparison with the different HDAs, meaning that no absolute validation is attempted. As such, the assumption made by using this validation dataset is that, at a nation-wide level, relative population patterns do not change drastically within three years. Finally, it is worth noting that ground truth dataset, as well as all other publicly available census data, represents information on an entire year period and is therefore not capable of capturing the higher temporal resolutions which will be discussed when turning to home detection based on mobile phone data.

Measures of Performance

To measure the performance of each HDA, the user counts at cell tower level are compared to the population counts in the ground truth data. In other words, a vector $\vec{x}$, which denotes the user count of one algorithm for all cell towers, is compared to a vector $\vec{y}$, which denotes the validation population count for exactly the same cell towers. Both vectors $\vec{x}$ and $\vec{y}$ thus have an equal length representing the 18,273 cell-towers in the Orange network.

One important element in the comparison between user and population counts is the unknown local market share of the Orange operator. The consequence is that, because of an unknown spatial distribution of the 28% sample of Orange users (section 3.1.4), similarity assessment between vectors $\vec{x}$ and $\vec{y}$ cannot be absolute. Performance measures are therefore defined based on the relative similarities between both vectors.

Pearson Correlation Coefficient

A first, and straightforward way, to quantify (dis)similarities between vectors $\vec{x}$ (user counts) and $\vec{y}$ (population validation count), is by calculating their Pearson correlation coefficient (Pearson's R):

$$Pearson's\,R(\vec{x},\vec{y}) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}}, \tag{4.1}$$

where $\vec{x}$ and $\vec{y}$ are the vectors, respectively, of the user counts (from HDA) and the population counts (from the validation dataset) for all cell towers in the French CDR dataset and i represents one cell tower. $\bar{x}$ and $\bar{y}$ represent the average values of the vectors $\vec{x}$ and $\vec{y}$, or thus the average user count and population count for all cell towers.

Values of Pearson's R range between -1 and 1 indicating, respectively, (perfect) opposition and similarity. Pearson's R values larger than 0 indicate a positive association between both vectors, whereas values smaller than 0 indicate negative association. Of course, the Pearson correlation coefficient is only a general measure of the relation between both vectors. As such, a visual investigation of the point cloud of both vectors forms an additional tool in understanding their relation.

Cosine Similarity Measure

A second way to compare similarities between user and population counts is by the Cosine Similarity Measure (CSM). The CSM evaluates the general angle between two vectors of the same length (thus avoiding absolute similarity assessment) and expresses it in degrees (°):

$$CSM(\vec{x},\vec{y}) = \left| cos^{-1}\left(\frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\cdot\|\vec{y}\|}\right) * \frac{180}{\pi}\right|, \tag{4.2}$$

where $\vec{x}$ and $\vec{y}$ are the vectors, respectively, of the user counts (from HDA) and the population counts (from the validation dataset) for all cell towers in the French CDR dataset. A CSM value of 0° denotes the highest possible similarity between both vectors (there is no angle between both vectors), 90° indicates the lowest similarity and 180° degrees refers to opposite orientation.

4.2.2 Comparisons between HDAs

Besides evaluating the similarity between HDA outcome and high-level census data, a second way to derive insights on the performance of HDAs is by investigating the differences between multiple HDAs. In general, the assessment of such differences inform on the importance of criteria choice and the related sensitivity of performance to such choice.

Similarity Matching Coefficient

Regarding the differences between HDAs, one intriguing question is whether, for the same individual user, different HDAs would detect different home locations (L1 cell towers) or not. By means of the Simple Matching Coefficient (SMC) it is possible to assess the degree to which two different algorithms detect the same home locations for the individual users in the dataset [27].

$$SMC(HDA_A, HDA_B) = \frac{\sum_{i=1}^{N} \delta(Home_{A,i}, Home_{B,i})}{N}, \tag{4.3}$$

where $i = 1, 2, \ldots N$ denotes the N-users evaluated, and $\delta(Home_{A,i}, Home_{B,i})$ is the Kronecker delta which is equal to 1 when the home detected by algorithm A user $i$is identical to the home detected by algorithm B for the same user, 0 otherwise. The SMC ranges between 0 and 1 and can be interpreted as the percentage of individual cases for which both algorithms detected the same home locations. When calculating the SMC, all cases where at least one of the algorithms failed to detect a home location (e.g., when no observations were left after implementing a time constraint) were omitted.

4.2.3 Uncertainty Assessment based on Metadata

A third and final way to assess HDA performance is by investigating the metadata related to the uncertainty of the home detection decision (section 4.1.2). The intuition behind this type assessment is straightforward. Investigating the plausibility to which a second (L2) and third cell tower (L3) are eligible leads to an estimation of the uncertainty that relates to the decision to choose the most eligible cell tower (L1) as home. Such estimation at user level is useful since most of the time no ground truth data is available at the individual level (neither for learning, nor for validation) and, as a consequence, the magnitude of the potential error on the decision between L1, L2, and L3 is unknown.

Spatial Uncertainty

One way to calculate the uncertainty that comes with the L1 decision is by measuring the *Spatial Uncertainty* (SU). The idea behind is that if the distances between the three top locations (L1, L2, and L3) are small, the spatial error of annotating the wrong location as home will remain small, and vice versa.

To construct the measure of spatial uncertainty, we compare the plausibilities of different cell towers to be home by calculating their ratios of the amount of observations that have passed the criterion by the deployed algorithm. If the criteria of an algorithm, for example, demand the highest amount of distinct days for detecting home, then the comparison of plausibility will be on the amount of distinct days at the different considered locations. We explicitly incorporate the spatial extent of the uncertainty by inserting the absolute distances between considered locations. Distances between locations (e.g., uncertainty because of long distance traveling) and differences in spatial resolutions of observations (e.g., high-density cell-tower areas versus low-density) will therefore both resonate in the proposed spatial uncertainty measure:

$$SU_n = \sum_{\{i,j,\ldots\}} \frac{p_i}{p_n} \times \frac{d(n,i)}{2}, \tag{4.4}$$

where SU_n is the spatial uncertainty for detecting a home at cell tower n (in meters), $\{i, j, \ldots\}$ are alternative cell towers for home in descending order of p_i, p_n and p_i are the numbers of considered observations, given the criteria in the algorithm used, in respectively cell tower n and i, and $d(n,i)$ is the distance between cell towers n and i (in meters).

Here, the calculation of SU will be limited to the detected home locations only, i.e., n=1. Even though the formula also facilitates calculation of the spatial uncertainty related to the decision on the L2, L3, or any other eligible cell tower. Following the observation that most activities of French users happen at three cell towers only (figure 4.1) the calculation of SU will also be limited to two other plausible locations, limiting the set of $\{i, j, \ldots\}$ to $\{i, j\}$ only. Doing so, the SU measure that will be used becomes:

$$SU_{L1} = \frac{p_{L2}}{p_{L1}} \times \frac{d(L1,L2)}{2} + \frac{p_{L3}}{p_{L1}} \times \frac{d(L1,L3)}{2}. \tag{4.5}$$

Imagining a simplified example in which the MA algorithm (which considers the total amount of activities) is used to detect the home of user x who has a trace of calling 10 times at location A, 5 times at location B, and 1 time at location C and in which distances between location A, B and C are all 1 km. Based on the criteria used by the MA algorithm, location A is L1, B is L2 and C is L3. The according SU for home detection at location A (L1) then becomes:

$$SU_A = \frac{5}{10} \times \frac{100}{2} + \frac{1}{10} \times \frac{1000}{2} = 300\ (meters). \tag{4.6}$$

The spatial uncertainty of a cell tower n is thus influenced by the distance to all other considered cell towers and by the share of evaluated observations in these cell towers compared to the amount of evaluated observations in cell tower n. A smaller share of evaluated observations in other cell towers and smaller distances to these other cell towers both result in smaller spatial uncertainties, indicating a higher plausibility that home detection is done at the correct location.

4.3 Performance of Home Detection in France

4.3.1 Numbers of Detected Homes

Each HDA described in table 4.1 was applied to all users and months (May to October, 2007) in the French CDR dataset. As such, a maximum of 109.4 million home detections could be performed by each HDA (about 17.7 and 18.5 million users per month table 3.4). This is a maximum, because some HDAs are restricting the CDR observations they take into account, potentially leading to users that do not have any observations, and thus no detected homes. For example, a user x that did not perform calls or texts during the nighttime in month y, will have no observations that can be used in the TC algorithm and will therefore have no detected home for that month. Similarly, the amount of eligible cell towers can vary between HDAs. Following figure 4.1 that showed that large share of users' activities are performed on a limited number of cell towers, the question then is to which degree different HDAs have different numbers of eligible cell towers. Table 4.2 summarizes the number of L1 (so the number of users that have a detected home), L2 and L3 locations that have been detected by the different algorithms for all 109.4 million potential home detections.

HDA	Cases with L1	Cases with L2	Cases with L3
MA	109.4 (100%)	102.2 (93.5%)	96.1 (87.9%)
DD	109.4 (100%)	102.2 (93.5%)	96.1 (87.9%)
TC 19-9	98.4 (100%)	78.0 (81.3%)	65.0 (66.1%)
SC	109.4 (100%)	102.0 (93.5%)	94.7 (86.6%)
TC-SC 19-9	98.4 (100%)	78.4 (79.6%)	62.3 (63.3%)

Table 4.2 Total numbers of detected homes (L1) and amount of cases that had plausible L2 or L3 locations for different HDAs when applied to all users and for all months (May-October 2007) in the French CDR dataset. Numbers are in millions, percentages are given in brackets.

Table 4.2 reveals some clear differences between the HDAs. Because of their definition, the TC and TC-SC algorithms are restricting the amount of observations, leading to a small percentage of the users in the dataset for which no homes can be detected (1.6%). Furthermore, these algorithms also have a limited number of cases for which L2 and L3 cell towers can be detected. The MA and DD algorithms, as expected by their definition, have the same amount of cases with eligible L1, L2, and L3s. The number of cases with L2 and L3 in the SC algorithms are slightly lower because of the grouping of nearby cell towers, but not fundamentally different.

Reflecting on the number of cases for which at least three eligible cell towers are found (63.3% to 87.9% depending on the algorithm), table 4.2 also show how loose the used criteria for home detection are. They form an indication why the assessment of spatial uncertainty is deemed necessary. In other words, most of the time, the deployed HDAs are not capable of pinpointing one eligible cell tower only. As such, a decision between multiple eligible cell towers needs to be made, introducing potential error or thus uncertainty on the decision given the absence of validation data.

4.3.2 Performance Measures at Nation-Scale Based on Census Census Data

Relations between user and population counts

Figure 4.3 shows the relation between user and population counts for the MA and TC 19-9 algorithms in August and September, but results for other HDAs and time periods show similar relations. Three elements deserve highlighting.

Firstly, the majority of cell towers have rather low population and user counts (between 0 and 2000 persons). This effect is partly due to high densities of cell towers in urban areas.

Secondly, cell towers are typically centered around the 0.28 line (dashed), which aligns with a 28% overall market share of the Orange operator, but divergence is high. This is indicated by the means (black dots) and the standard deviations (vertical error bars) of sub-groups of cell-towers grouped according to the deciles of the population counts. *Overestimation* typically occurs for cell towers with lower ground truth population, while *underestimation* is more related to cell towers that have a higher ground truth population. Without more contextual information on the cell towers such as location or typical usage by mobile phone users, it is impossible to properly account for this pattern.

Thirdly, one can clearly distinguish a group of cell towers that have very low user counts. This group seems to be evenly distributed over the different ground truth population counts and can be considered an artifact of data collection. They correspond to cell towers that were (temporally) inactive during certain observation periods because they were, for example, out of action for some time.

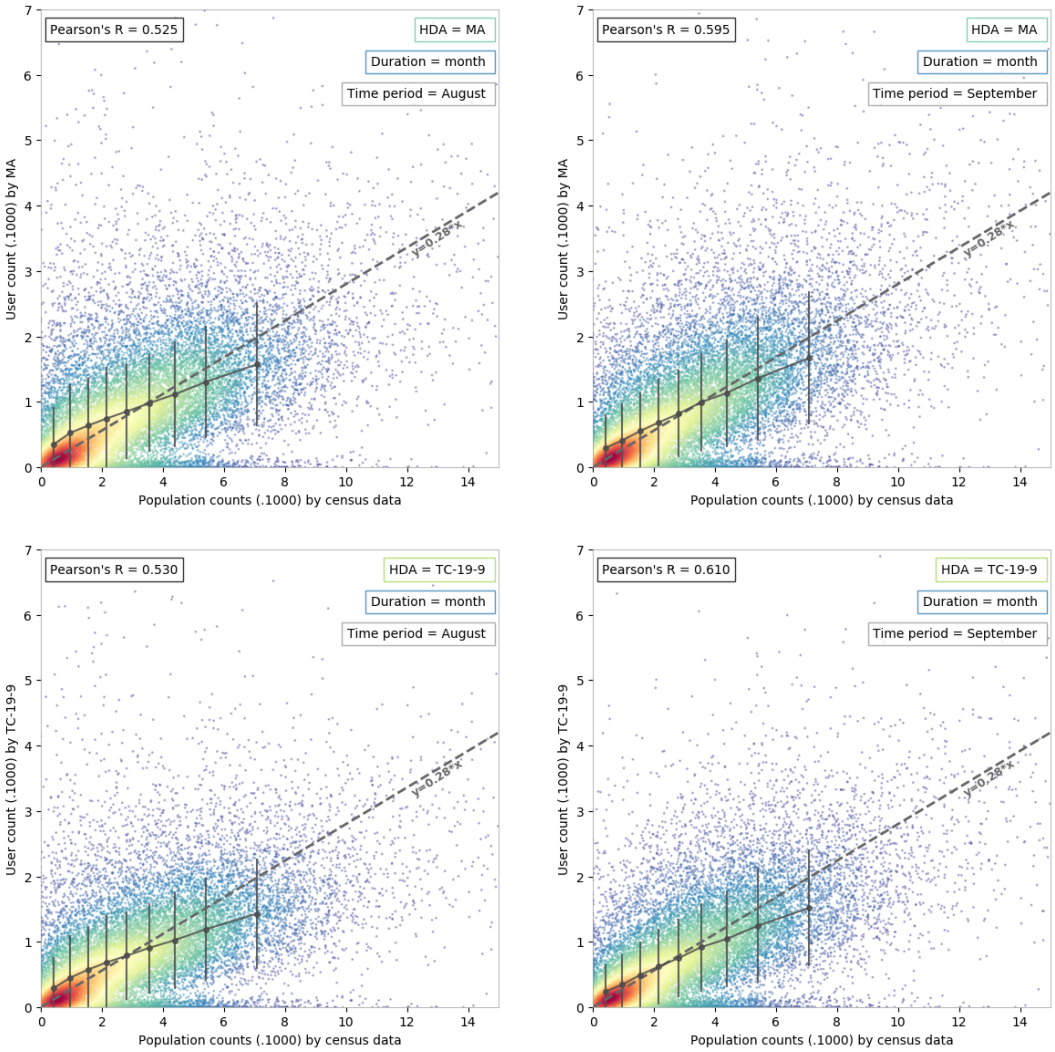


Fig. 4.3 Scatterplot of population counts (census data) and user counts two HDAs (MA and TC-19-9) and two months (August, September). Each dot represents one cell tower, and is colored by density of the dots in its surrounding, in a gradient from blue (low density) to red (high density). Black error bars represent values of the groups of cell towers based on the deciles (except the 90-100 decile which is not plotted) of the population counts (x-axis). For each group of cell-towers, the mean and standard deviation of the user counts (y-axis) were calculated and plotted by means of the error bars (middle point= mean, whisker=standard deviation)

Performance Measures

The Pearson's R and CSM values for all HDAS are reported in tables 4.3 and 4.4, respectively. Values range between 0.52 and 0.62 for Pearson's R and between 35° and 38° for CSM, which indicates a moderate performance as a perfect fit between user and population count would result in a Pearson's R of 1 and a CSM value of 0°. The two measures are showing similar patterns, although they do not display a one-to-one relation (see also section 5.3.1).

	Pearson's R					
	May	June	July	Aug.	Sept.	Oct.
MA	0.591	0.590	0.548	0.525	0.595	0.595
DD	0.599	0.598	0.560	0.538	0.603	0.602
TC	0.601	0.607	0.562	0.530	0.610	0.608
SC	0.581	0.585	0.552	0.535	0.588	0.575
TC-SC	0.581	0.593	0.563	0.528	0.594	0.569

Table 4.3 Pearson's R values for the relation between user and population counts.

	CSM (in °)					
	May	June	July	Aug.	Sept.	Oct.
MA	35.9	35.8	36.9	37.7	35.3	35.5
DD	35.5	35.4	36.3	37.1	34.9	35.2
TC	35.2	34.9	36.2	37.4	34.4	34.7
SC	36.5	36.2	36.9	37.4	35.8	36.6
TC-SC	36.5	35.8	36.3	37.8	35.5	37.1

Table 4.4 CSM values for the relation between user and population counts

For most months, the DD and TC 19-9 criteria perform best in replicating the population pattern of the validation dataset, followed by the MA criterion. For all months, the criteria that involve grouping in space perform worst, even though the applied criterion (1 kilometer) is not that far. A clear reason or good hypothesis for this is absent.

Differences between months are similar for all algorithms, with lower CSM values for June and September, and higher values for May, July, August and October (tables 4.3, 4.4). A possible explanation for high SMC values for May and October is the limited number of days on which data are available (18 and 14 days, respectively). All algorithms show a drop in performance during July and August, potentially because of the French population perform domestic tourism trips during the summer period [51, 141, 142]. The TC and TC-SC criteria have the largest drop in performance during summer months. This sensitivity to observation period questions their adaptation in literature. In addition it is an interesting observation that the differences in performance between algorithms are smaller than those between months. Future deliberation on the performance of HDAs should therefore take into account the sensitivity to the chosen time period, an aspect that is investigated in more depth in chapter 5.

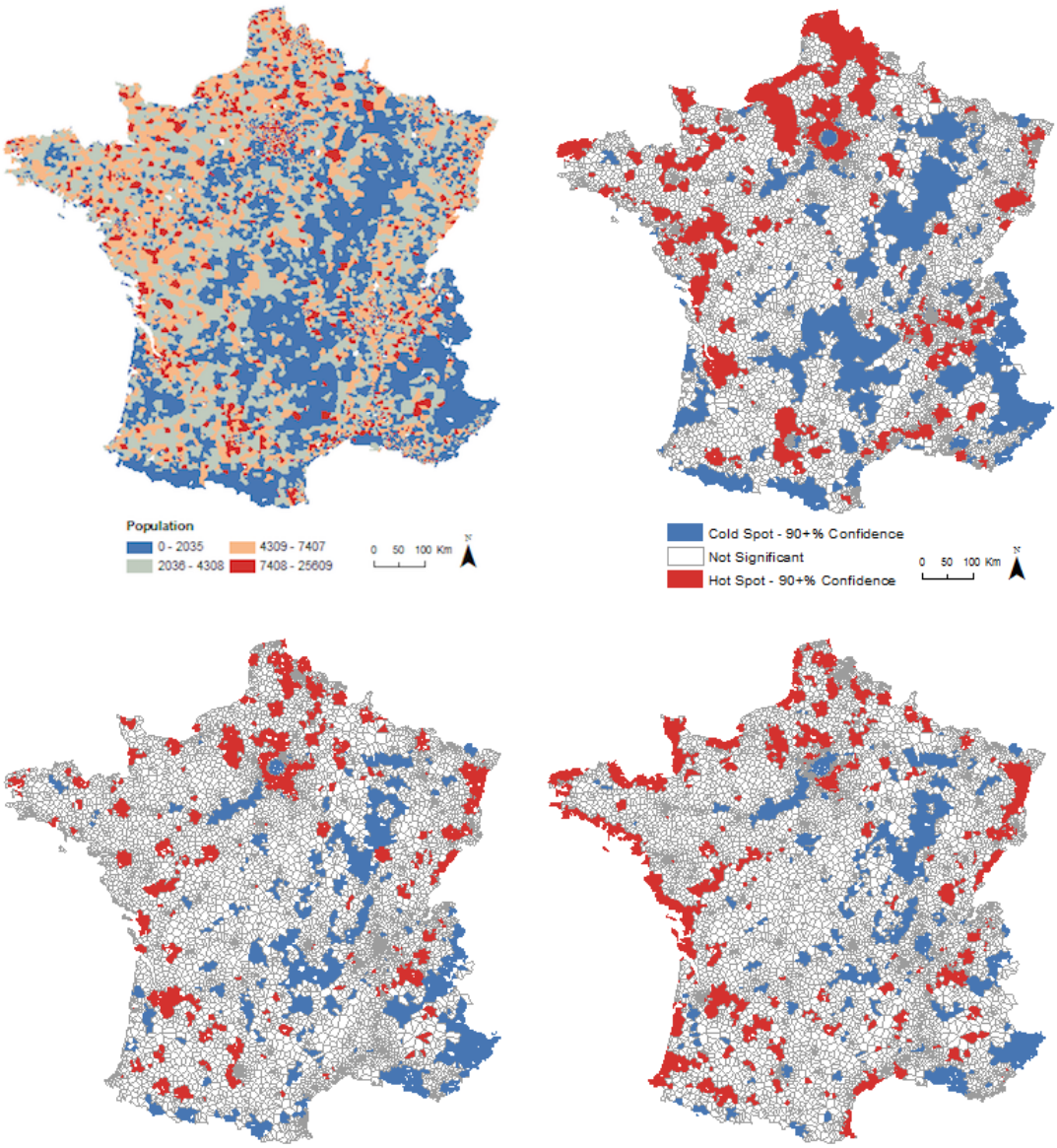


Fig. 4.4 Population counts of the validation dataset, as also shown in figure 4.2 (top left). Hotspots (red) and coldspots (blue) defined by the 90+% interval of the Getis-Ord Gi* statistic for the population counts of the validation dataset (top right), for the user count by the amount of activities algorithm in June (bottom left), and for the user count by the amount of activities algorithm in August (bottom right). All maps are made up of the Voronoi tessellation of the Orange cell towers.

Spatial patterns of user and validation counts

At first sight, differences in performance between HDAs seem not that large compared to their differences with perfect matching. For example, differences between algorithms or months of 1° or 2° (for CSM) seem little compared to the 35° that needs to be bridged to arrive at the intended 0° representing perfect matching. Still, when investigating the spatial patterns of user counts that accord to small difference in CSM values, it becomes clear they are rather significant. Figure 4.4, for instance, shows the spatial patterns, emphasized in *hot* (marked red) and *cold* (marked blue) spots based on the Getis-Ord Gi* statistic [61], of the detected homes in June and August by the MA algorithm. The difference in CSM values between June and August is 1.08° but results in a very different spatial pattern. One reason for this is the summarizing nature of the Pearson's R and CSM measures, meaning that different data can have similar correlations. Another reason is the sheer volume of the data, implying that local differences, even when large, do not translate in drastically altering general measures.

4.3.3 Individual Level Differences Between HDAs

A second way to assess the performance of HDAs at nation-wide scale is by comparing their results to one another. The Similarity Matching Coefficient (SMC) between two algorithms (section 4.2.2) summarizes the percentage of users for which two HDAs detect equal home locations. Pairwise SMC values were calculated for all combinations of the five HDAs and all months (figure 4.5).

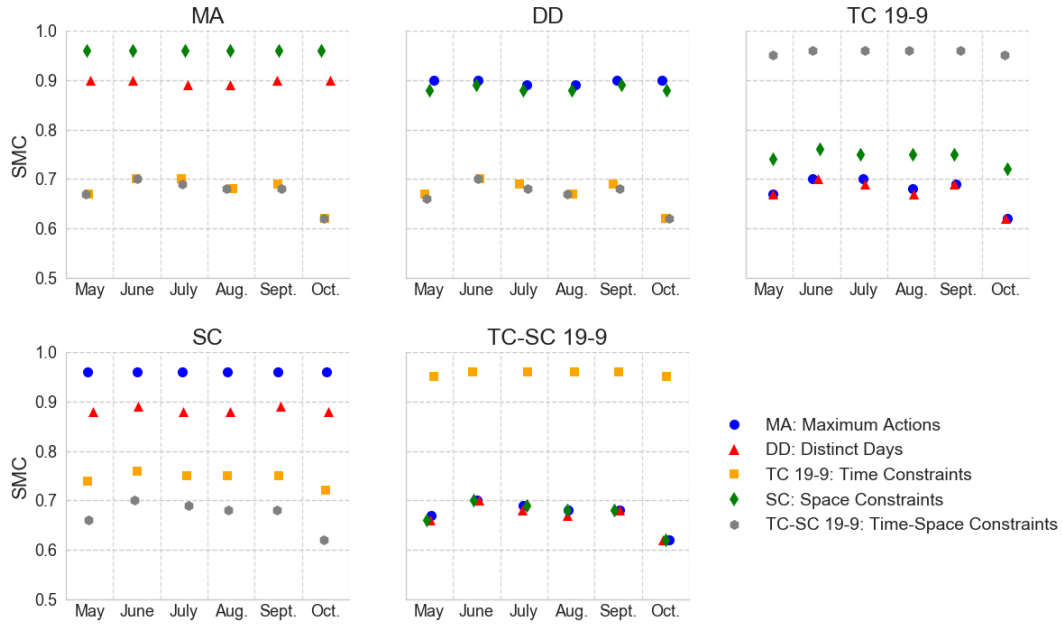


Fig. 4.5 Similarity Matching Coefficients (SMC) for all pairwise combinations of HDAs, for each month in the dataset. SMC values express the ratio of users for which two HDAs detect the same home.

Observed SMC values range between 61.5% and 96.4% of the detected homes, resulting in discordance rates between 40% and 4% (1-SMC). For the French CDR dataset, 40% and 4% respectively translate to 6.8 and 0.6 million users that had a different detected home. The patterns of (dis)similarities between algorithms are clearly visible. Algorithms that incorporate time-constraints (TC and TC-SC) discord to high degrees with algorithms that count amount of activities (MA), distinct days (DD), or perform spatial grouping only (SC). Additionally, MA, DD, and SC algorithms show high degrees of pairwise accordance amongst themselves. It is remarkable that, regardless of the time periods, the TC 19-9 criterion results in different detected homes for 30% to 40% of the cases compared to all other criteria, except for the TC-SC criterion. Sparser observations and different spatial behavior during nighttime could be an explanation.

4.3.4 Spatial Uncertainties Related to the Home Decision

A third, and last, way to assess the performance of HDAs is by investigation the Spatial Uncertainty (SU) related to their L1 decision (section 4.2.3). The median SU values of all users for different HDAS and months are shown in figure 4.6. Values range between 2.5 km and 7 km, suggesting that generally, spatial uncertainty on the detected homes is moderate and the decision between L1, L2, and L3 is rather local. The MA criterion depicts the lowest SU values. Remarkably, the SU of MA and TC algorithms are almost equal during non-summer months. However, during summer, SUs of the TC criteria increases drastically, whereas the SUs of the MA only rise moderately. Again, this poses questions regarding use of time-constraints criteria, especially during summer months.

Despite having one of the lowest CSM values, the DD criterion has the highest SU values during non-summer months. This is a clear indicator that using distinct days can be risky as it might favor secondary and tertiary locations. Distributions for the MA and the DD algorithms are also shown in figure 4.6, revealing the long tail distribution of SU values between users with extremes of the 90th percentile running as high as 120 km. For a large share of users, in other words, the decision between L1, L2, and L3 is not a local one, suggesting that wrongful home detection would results in a large spatial error for these users.

The higher SU values in summer suggest a change in the nature of observed traces. Home detection is more uncertain due to higher distances between plausible locations and different calling patterns at these locations; a change poorly dealt with in existing algorithms given their growing discordance with the validation dataset during July and August. In addition, low SU values in May and October suggest no change in the nature of traces and so the observed slightly lower Pearson's R values (table 4.3) and slightly higher CSM values (table 4.4) for these months cannot be explained in a similar way to summer months.

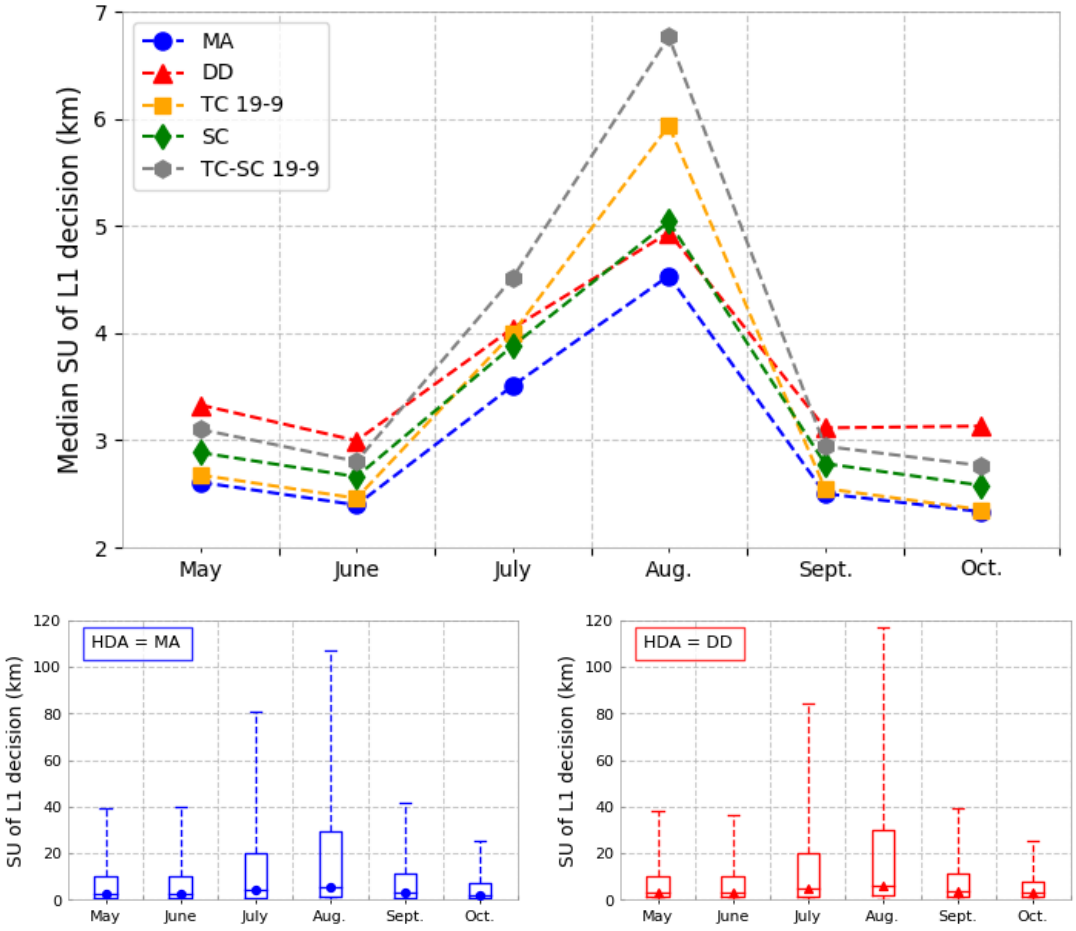


Fig. 4.6 (Top) Median values of spatial uncertainty (SU) for the L1 decision calculated for all users based on five algorithms during different months. Distributions of SU values for different months for the MA algorithm (bottom left) and the DD algorithm (bottom right). Boxes indicate the interquartile range, whisker the 10th and 90th percentile. Distributions of other HDAs show similar magnitudes

Relation between Spatial Uncertainty and General Performance

The similarity in temporal patterns between SU and CSM values suggests that spatial uncertainties at the individual level could be linked to the nation-wide performance of the algorithms. Figure 4.7 illustrates this relation and shows a strong correlation between both measures ($R = 0.53$ to $R = 0.72$ depending on the omittance of outliers for May and October). This is an interesting finding as it opens the door for a data-driven assessment of home algorithms, so diminishing the role of external validation datasets, even though one should be extremely cautious when interpreting the correlation between two summary measures.

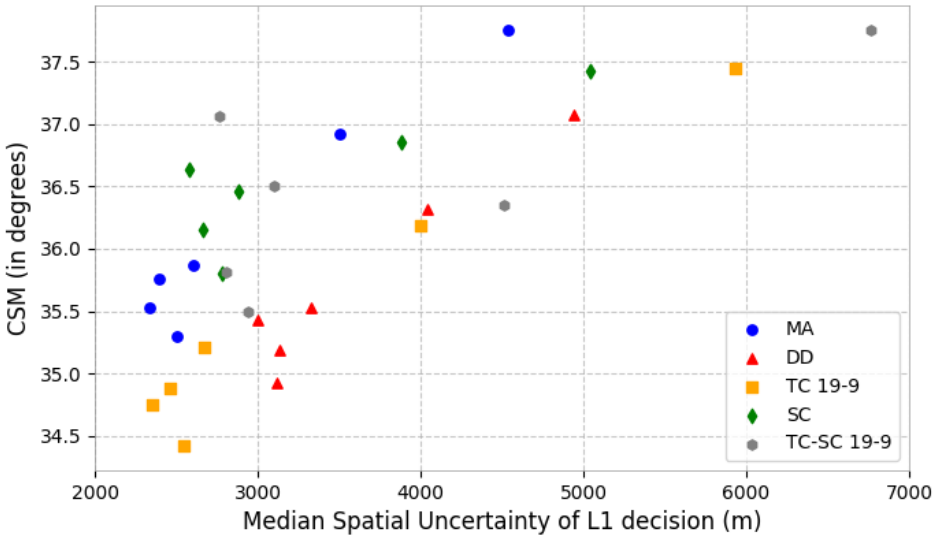


Fig. 4.7 Scatter of median SU values and CSM values for all algorithms and all months in the dataset.

4.4 Discussion

In this chapter, three ways to evaluate the performance of home detection at nation-wide scale were proposed and applied to five HDAs with simple criteria when deployed on the French CDR dataset.

4.4.1 A Large Performance Gap at Nation-Scale

The first way concerned the traditional comparison with census data, although this time it is based on a unique validation dataset. Comparing the obtained user counts from the HDAs with validation population counts at cell tower level, high-level accordance measures such as the Pearson's R or CSM were calculated. For the five proposed HDAs, performance was found to be moderate at best, with Pearson's R values ranging between 0.52 and 0.62 and CSM values ranging between 34° and 38° only, showing a large *performance gap* between obtained results and agreement with the validation dataset.

Structural Reasons for the Performance Gap

There are several possible reason for such a performance gap, but all are hard to investigate without additional data sources or validation data at user level. One reason might be the effect of the (unknown) local market shares on the performance measures. Although the nation-wide market share of Orange is known (or can easily be calculated from the amount of users in the dataset), the spatial distribution of this market share over the country might contribute to, at least, parts of the observed performance gap. The wide deviations around the 0.28 line in figure 4.3, are indicative of the existence of variations in the local market shares, but without additional data on the (historical) local market shares, their influence on the general performance measures remains unknown. Still, the small range of performance measures observed for all HDAs (about 4° for CSM) in comparison to the magnitude of the performance gap (about 35° for CSM) seems to suggest that this local market share, in combination with other reasons, might significantly affect nation-wide performance measures. One avenue for future research is to assess the influence local market shares can have on nation-wide performance measures for home detection.

A second reason might be the variation in mobile phone usage amongst populations. In a way, the regional differences in mobile phone usage can influence performance measures in a similar way as variations in local market shares would. Clear regional differences of the amount of visited cell towers, or the amount of active days, for example, are discussed already in chapter 3. They are also likely to indicate regional differences in the potential for performing (correct) home detection, regardless whether this potential is expressed by the number of users for which a home can be detected given a criterion, the number of eligible cell towers, or the number of records available per user. Given the absence of individual level data, such regional differences in the potential for (correct) home detection are hard to assess, as is their influence on nation-wide performance measures.

Other reasons for the existence of the performance gap can also be considered. A difference in the definition of home between census data and home detection from mobile phone data is one of them. Artifacts of data collection and treatment are another. For example, in mobile phone data research, as was also the case for this investigation, the coverage areas of cell towers are often estimated by means of Voronoi polygons. Such estimation entails errors with regard to the real life situation, which undoubtedly introduces a loss in performance when measured at nation-scale but whose effect is hard, if not impossible, to quantify.

Sensitivity to Time Period

Focusing on the performance of the five HDAs within their own performance measure range, it forms an interesting observation that the largest deviations in performance were found over time, regardless of the deployed HDA. As pointed out before, this temporal variation is probably due to holiday movements that are known from literature [1] [51, 141, 142] and that are also suggested by the spatial patterns in figure 4.4. Still, it remains surprising how sensitive the five HDAs, and in extension single-step home detection approaches, can be to such a holiday effect.

One major consequence of this finding is that it reveals the importance of insightfully defining the used time period for home detection at nation-wide scale. The intriguing realization in this perspective is that, when supporting on nation-wide validation only, it is impossible to foresee the influence of the time period on performance, thus choosing an appropriate time period, without sensitivity testing. The reason is that activities performed by users that influence home detection performance both positively or negatively (such as long-distance travel) cannot be properly defined because of the absence of individual level data. As such, although CDR data of individual users probably possess signals on the feasibility of performing home detection for a particular user, it is not possible to assess, at a given time, how many users have higher or lower potential to correctly detect homes.

The consequence is that testing of the sensitivity to time period is essential when performing or assessing home detection at nation-wide scale. However, such sensitivity to time period is seldom described in literature. Additionally, results suggest performance to be sensitive not only to the chosen time period, but also to the duration of the time period. As pointed out before, the performance of home detection in May and October is lower compared to other non-summer months such as September or June (tables 4.3, 4.4 and figure 4.6), while their differences in median SU values are limited only. One potential explanation is that the limited duration of the May (18 days) and October period (14 days) influences home detection. Here too, it is remarkable that little investigation exists on the needed duration of CDR data for home detection. As a consequence, chapter 5 will elaborate an in-depth investigation on the sensitivity of home detection to, amongst others, time period and time period duration.

[1] In 2015, 88.1% of all tourism trips performed by French people were estimated to be domestic tourism trips; making it one of the largest shares in Europe [141].

4.4.2 Criteria Sensitivity Plays out at Another Level

Regarding the performance of different HDAs, both Pearson's R as CSM values seem to suggest that the TC 19-9 algorithm performs best, followed closely by the DD algorithm (tables 4.3, 4.4). One clear disadvantage of the TC 19-9 algorithm is its higher sensitivity to time period, especially when compared to the DD algorithm. Interpretatively, this difference can easily be understood. If sensitivity of performance to time periods is indeed associated with holiday movement, then it is not surprising that the DD algorithm is less susceptible to it because of its definition. As can be observed from figure 3.4 in chapter 3, the median amount of distinct days on the most used cell tower for users in the French CDR dataset is around 15 days. Broadly speaking, this implies that for cell towers at a holiday destination to be chosen as home cell tower by the DD algorithm, quite a number of distinct days need to be obtained, making it less susceptible to holiday movement compared to the absolute number of activities at nighttime as used in the TC 19-9 algorithm.

At first sight, differences in performance between HDAs do not seem that pronounced. This is a false interpretation. At the nation-wide level, for example, small differences in the values of general performance measures were shown to translate to substantial differences in the spatial pattern of detected homes. Similarly, SMC values between pairwise HDA revealed how large shares of the users (between 4% and 40%) have different detected homes for different algorithms, with patterns being more or less insensitive to time period.

A remarkable finding in this perspective is the large discordance between the TC and TC-SC criteria, and the MA and DD criteria. Even though their overall performance are the best, the DD and TC 19-9 algorithms can detect a different home for up to 40% of the users. This finding clearly shows the limits of performance measures created at nation-wide level to understand and evaluate home detection criteria. Unlike the unknown effects of local market shares or the sensitivity to time periods that is partly imposed by characteristics of the dataset, the sensitivity of home detection performance to criteria choice is dependent on the researcher's choice only. As such, the relevance of a more in-depth investigation on criteria choice in HDAs becomes apparent, especially since this topic, once again, has received insufficient attention in literature. As a consequence, chapter 5 will elaborate an in-depth investigation on the sensitivity of home detection to criteria choice, with a special focus on the parameters of the TC algorithm which, in this chapter, has been fixed to 19.00 and 09.00 hours but which, obviously can be varied.

4.4.3 Spatial Uncertainties and Future Work

A final way to assess the performance of HDA is by investigating their related metadata. More specifically, the Spatial Uncertainty was constructed to express the potential spatial error related to the L1 decision, given that L2 and/or L3 locations were also eligible based on the deployed criteria. Considering the extent of the French territory, the magnitude of SU values, with user median values between 2 and 7 km (figure 4.6), does not seem that large implying that HDA decisions are rather local. However, a high variation between users is observed when it comes to SU values with SU for the L1 decisions reaching as high as 40 km, or even 80 km in summer for the 90th percentiles of users (figure 4.6). In such cases, the potential magnitude of error does make a significant difference. One example of this is in the cross-over with census data. Imagine a case for which individual users, based on their detected home location, are attributed an income from census data in order to, for example, study segregation patterns based on their movement patterns captured in CDR data. Home detections with a potential error of 10, 40, or even 80 km, might very well mean that a user is attributed to a wrongful census area, resulting in incorrect results.

The real merit of the SU, however, is that it challenges the misconception of home detection being a discrete choice, with one definitive cell tower as a result. This might very well be the case in reality, but in the absence of individual level data, single-step HDA are depending on a probabilistic choice between different options, and there relates a (spatial) uncertainty to this choice. Being capable of estimating this uncertainty without the need for additional data sources opens up several possibilities. For one, since SU and CSM values are correlated (figure 4.7), one could think about using SU at the individual level as a proxy for performance at nation-scale, hence avoiding the need for validation data.

Another possibility could be to use SU values as a filter, thus discarding users that depict too high of SU values. Results of early experiments in this direction are shown in figure 4.8, indicating the potential to improve nation-wide performance by filtering individual users, especially during summer time. The latter is not surprising given that the SU takes into account a factor of distance and thus is sensitive to long-distance movement of users during holiday season.

The main problem with such experiments, however, is that they are not viable without individual level validation data that can calibrate the relation between SU and wrongful home detection. After all, even though spatial uncertainty on the home decision might be high, this does not necessarily mean that the decision was wrong. In addition, other metadata, such as the home locations of peers or information on the temporal presence patters at a given cell towers, might also capture aspects that relate to correct or wrongful home detection. In other words, the relation between metadata and correct home detection needs to be extended much further than just the SU measure in order to really advance this line of research, potentially once arriving at HDAs that are adaptive to the user traces presented to them.

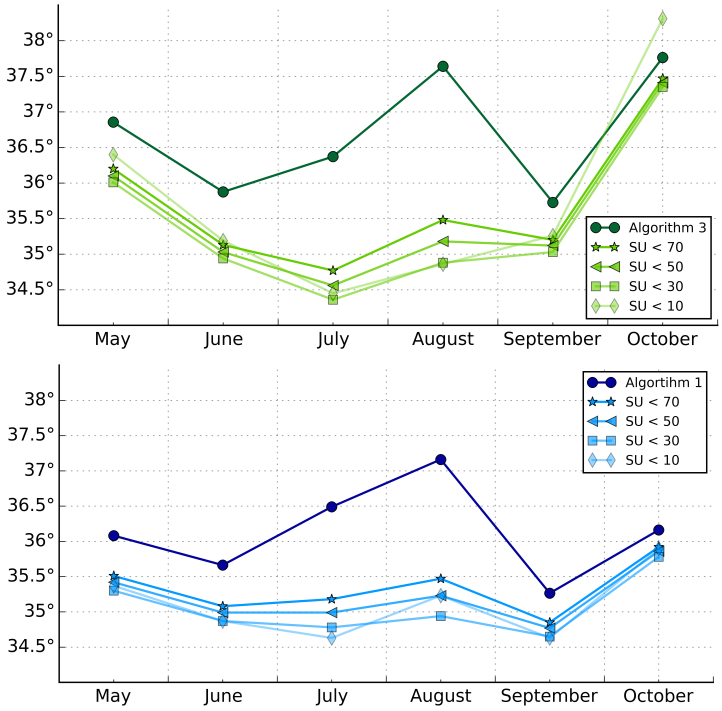


Fig. 4.8 Results of early experiments to improve home detection performance by filtering users on SU values. CSM values (in degrees) are shown when filtering users on different SU values (expressed in km). Experiments were done for the MA criterion (bottom) and the TC 19-9 criterion (top), but for a limited number of cell towers only (hence the difference in naming and CSM values compared to the other figures in the chapter).

Chapter 5

Sensitivities of Home Detection Performance

> Uncertainty is inevitable at the frontiers of knowledge.
>
> *Joel Achenbach*

Abstract

This chapter continues the investigation of home detection from CDR data. Building on the performance assessment introduced in chapter 4, the focus is on the sensitivity of home detection performance to the choice of criteria, parameter, time period, and duration of observation. Sensitivity is empirically tested by 9 Home Detection Algorithms (HDAs) with simple criteria, deployed on the French CDR dataset for 23 different time periods, and special attention is given to the spatial patterns of home detection. In total, around 3.7 billion home detections are investigated, sketching a better picture of the sensitivity of HDA performance. Findings are summarized by means of a discussion on the magnitudes of different sensitivities, and their implications for future research.

Related Publications and Acknowledgments

- The content of this chapter has been rewritten as an (accepted) conference paper for the BigSurv'18 conference: M Vanhoof, C Lee, Z Smoreda (2020) *Performance and sensitivities of home detection from mobile phone data*, that will appear in the conference proceedings planned to appear early 2020.
- Similar to chapter 4, acknowledgments go to the INSEE team for the construction of the ground truth dataset, and to Fernando Reis, Michail Skaliotis and Dr. Zbigniew Smoreda for their help in establishing the partnership between INSEE and Orange Labs France.

5.1 Investigating Sensitivities of Home Detection Performance

As was raised in the literature review in section 2.7 and in the discussion on the results of chapter 4, it is remarkable that few studies exist which discuss the sensitivity of home detection methods to user choices such as that of HDA, time period, or duration of observation. Concerning the choice of HDA, [27] found different HDAs to result in different attributed homes for 1% to 9% of users in a credit card transactions in Spain, and for 7% to 20% of users in a worldwide Flickr dataset. For the French CDR dataset, the SMC values calculated in chapter 4 revealed percentages between 4% and 40% of users (figure 4.5). In other words, the choice of HDA seems to have a significant effect on home detection, at least at user level. Consequently, it becomes important to investigate the sensitivity of home detection performance to this as well as other choices of the researchers.

The goal of this chapter is to empirically explore the sensitivities of nation-wide performance of home detection on the French CDR dataset. The focus in this chapter is on the research choice of HDA, time period choice, and duration of observation (although the latter can also be induced by data availability instead of researcher choice solely). A clearer insight into the combined effect of these choices on the quality of home detection is desirable, because, until such investigation is carried out, it will remain unclear what is influencing the quality of home detection methods, what kind of error might propagate in further analyses, and what decisions researcher can make to avoid quality or performance loss. In addition, investigation on the sensitivity of home detection performance could justify current underlying assumptions related to home detection on CDR data, or prove them wrong.

5.1.1 Defining 9 HDAs with Simple Criteria

In analogy with the methodology in previous chapter, HDAs with simple decision rules are created for the analysis in this chapter. The criteria used in the HDAs are the same as in chapter 4 but the space constraints (SC) and the time-space constraints (TC-SC) criteria are omitted because of consistent under-performance compared to the other criteria (tables 4.3, 4.4). The considered criteria, namely the maximum amount of actions (MA), the maximum distinct days (DD) and the time-constraints (TC) criteria, when combined with parameter choices discussed below, form the nine HDAs that will be used in this chapter.

Parameter Choice

The advantage of the MA and DD criteria is that they do not require any parameter choice. The TC criterion on the other hand demands a parameter choice on the time restriction. Despite this extra parameter choice, home detection algorithms that are based on the TC criterion are rather popular, especially when restricting observations to nighttime and/or weekend days (section 2.7.4). One reason for their popularity is that time restriction is intuitive and parameter

choices can sometimes be based on available time surveys, both of which lend justification to the parameter choice.

Nevertheless, there exists no studies on the sensitivity of the TC criterion's performance to such parameter choices, nor is it clear whether nighttime or weekend days in themselves are the best options to consider. Therefore, different parameter choices for the TC criterion will be investigated, incorporating either nighttimes, daytimes, week days, weekend days or a combination of them.

The nine HDAs that will be used, as well as the considered parameters, are defined and explained in table 5.1.

Criterion	Parameters	Acronym	'Home is cell tower where:'
Maximum Amount	/	MA	Most activities occurred
Distinct days	/	DD	the maximum active days were observed
Time constraints	19,9	TC-19-9	Most activities occurred between 19.00 and 9.00 (nighttime)
Time constraints	19,9,weekend	TC-19-9-WE	... between 19.00 and 9.00 (nighttime) and during weekend days
Time constraints	21,7	TC-21-7	... between 21.00 and 7.00 (nighttime)
Time constraints	21,7,weekend	TC-21-7-WE	... between 19.00 and 9.00 (nighttime) and during weekend days
Time constraints	9,19	TC-9-19	... between 9.00 and 19.00 (daytime)
Time constraints	9,19,week	TC-9-19-WK	... between 9.00 and 19.00 (daytime) but only during weekdays
Time constraints	weekend	TC-WE	... during weekend days only (Sat and Sun)

Table 5.1 Description of deployed HDAs with different criteria and parameters.

5.1.2 Defining 23 Time Periods

To investigate performance sensitivity to time period and duration of observations, the nine defined HDAs will be run on multiple time periods, starting at different dates and lasting different durations.

Duration of Observation Periods

The French dataset collects data from May 13 till October 15 2007. One straightforward duration of observation would be 154 days, or thus the entire duration of the dataset. Doing so, however, makes calculations computationally expensive and does not allow us to study sensitivity to time period. Investigation is therefore also done for the duration of discrete months similar to the analysis in chapter 4. Discrete months, in turn, bring about the problem that they have different number of days obscuring proper comparison. As such, durations of exactly 14 and 30 days are investigated too.

The different duration of observations considered are thus:

- Discrete months: May (18 days), June (30 days), July (31 days), August (31 days), September (30 days), October (15 days)
- 154 days: Entire time period
- 30 days: Moving window of 30 days
- 14 days: Moving window of 14 days

Time Periods

The actual time periods that are related to these durations are straightforward for the discrete months and the 154 days duration, but not necessarily for the 14 and 30 days. For the latter, time periods start the first day of the dataset (13 May) and take a moving window of respectively 14 and 30 days from that point in time. For the 14-day duration, this means that home detection is performed on 11 two-weeks periods, the first one being the period from 13 May till 26 May, the second one being from 27 May till 9 June, and so on. The same strategy is used for the 30-day duration, resulting in five periods of 30 days, the first one from 13 May till 11 June. Note that we omit the sixth and last 30-day period which ends up being only eight days (from 7 October till 15 October). Note also that, because of our dataset starting in mid-May, the 30-day periods are more or less complementary to the discrete months, offering interesting opportunities for comparison. All time periods for all different durations that will be used in the analysis are illustrated in figure 5.1. In total, they sum up to 23 different time periods.

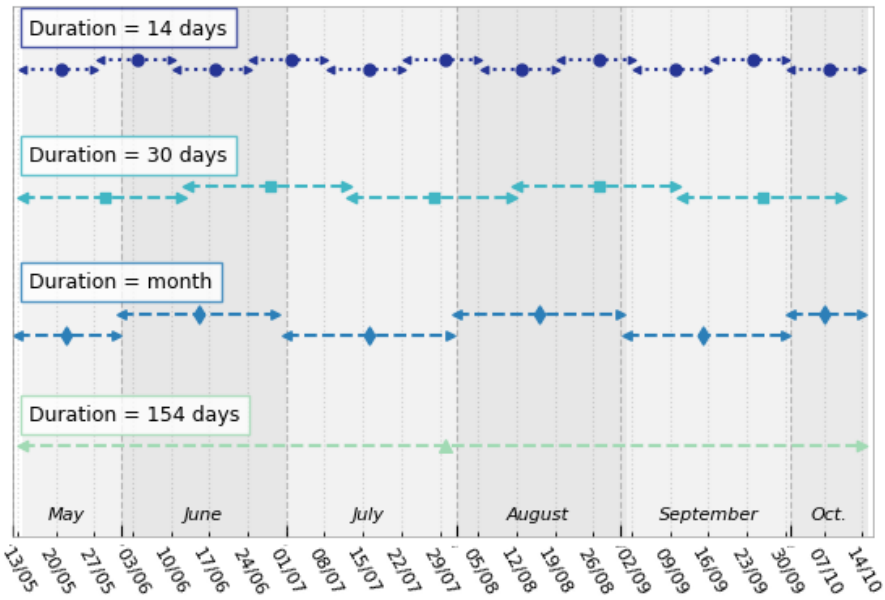


Fig. 5.1 Different durations and related time periods used in the analysis. Dots, squares, diamonds and triangles represent the middle of the time periods, dotted arrows note the durations of the time periods.

5.2 Assessing Performance and Sensitivity

5.2.1 Performance Measures at Nation-Scale

The performance of single HDAs at nation-scale is assessed by comparing the same ground truth data prepared by the French Official Statistics office in chapter 4 (section 4.2.1). Similarly, measures of nation-wide performance are expressed by the Pearson correlation coefficients (Pearson's R, see equation 4.1) or the Cosine Similarity Measure (CSM, see equation 4.2) that are calculated for the relation between user counts (based on a HDA) and population counts (based on the ground truth data), both of which are aggregated at cell tower level (section 4.2.1).

Performance measures are calculated for each of the 9 HDAs when applied to CDR data of all users for each of the 23 time periods, resulting in two performance measures (Pearson's R and CSM) for 207 (9×23) different combinations of HDA and time period.

5.2.2 Spatial Patterns of Home Detection Performance

While Pearson's R or CSM measures evaluate similarity between the vectors of user and population counts, they do not assess the differences between ground truth and HDA results for each cell tower. Having measures of similarity between user and population counts at cell tower level is useful as they enable the investigation of spatial patterns. One way to create such measures is by calculating the ratio between user and population counts for each cell tower i. Given a right long-tail in the distribution of ratios, a better way to describe them is by using the logarithm of the ratio, resulting in the definition of the *LogRatio* as:

$$LogRatio_i = log_e\left(\frac{x_i}{y_i}\right), \tag{5.1}$$

where i represents one cell tower,e is the natural log base, x_i is the user count at cell tower i estimated by one HDA, and y_i is the population attributed to cell tower i based on census data.

Note that, although local market shares of Orange can differ from cell tower to cell tower, LogRatios are distributed around -1.27 ($log_e(0.28)$), given the 0.28 overall market share of Orange (section 3.1.4. Large deviations from the expected 0.28 ratio can therefore be considered indicative for severe under- or overestimations. Because LogRatios are calculated at cell tower level, patterns of over- and underestimation can be mapped providing a spatial insight on the performance of HDAs.

5.2.3 Assessing Sensitivity

To assess sensitivity of HDA performance to the choice of criteria, parameter, time period and duration of observation choice, a systematic investigation of different combinations and their performance is carried out. Figure 5.2 provides an overview of the different investigations that will be performed in the following sections. User counts obtained by home detection are compared to the validation population count enabling nation-scale measures of performance to be calculated (A), and spatial pattern of over- and underestimation (LogRatios) to be mapped (B). Additionally, relations between HDA results and the validation dataset over time are investigated to assess the sensitivity of the user time periods (C). Correlations are compared between different durations of observations (**D**), or between different HDAs (E) in order to understand, respectively, sensitivity to duration of observations or to criteria and parameter choice .

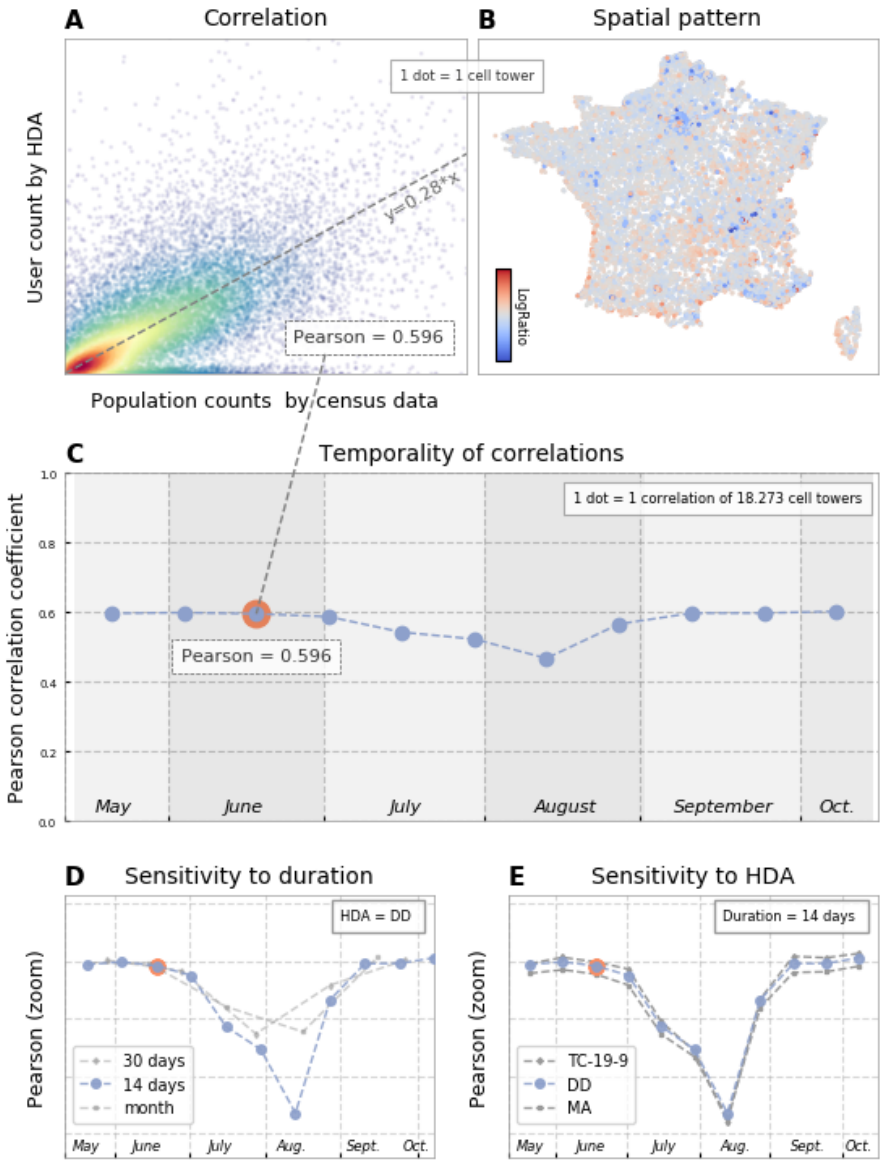


Fig. 5.2 Overview of the methodology to assess the sensitivity of home detection performance at nationwide scale to criteria, parameter, time period and duration of observation choice. (A) Correlation between user counts from one HDA for one time period and population counts from census data. Dots are colored by the density of dots in their surrounding in a gradient from red (high density) to blow (low density). The $y = 0.28 \times x$ is plotted as a guide to the eye and represents the expected relation based on the overall 28% market share of Orange. (B) Spatial patterns of the LogRatios at cell tower level. Colors correspond to LogRatios in a gradient from red (over-estimation) to blue (under-estimation). Grey areas correspond to the expected -1.27 ($log_e(0.28)$) LogRatio based on the overall 28% market share of Orange. (C) Temporal pattern of correlations between user counts from one HDA and population counts from census data for different time periods according to one duration of time period (here: 14 days). (D) Temporal patterns of correlations for one HDA over time periods with different durations. (E) Temporal patterns of correlations for different HDAs over time periods with one duration.

5.3 Results

5.3.1 Relations between User Counts and Population Counts

For all HDAs and time periods, the Pearson's R and CSM values are calculated. In line with the results in tables 4.3 and 4.4, performance ranges between 0.45 and 0.62 for Pearson's R and between 34° and 42° for CSM. As can be observed in figure 5.3, no clear one-to-one relation between Pearson's R and CSM values exists, but their correlation is close to linear. For this reason, hereafter, results will show Pearson's R values only, but analysis with CSM values render similar results.

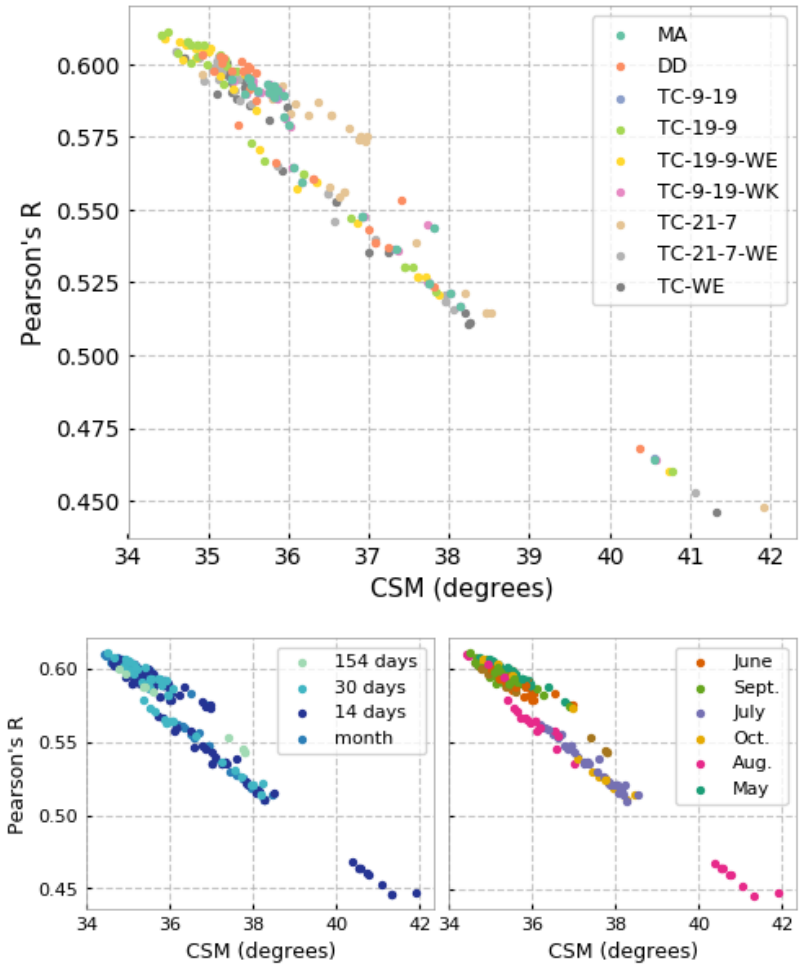


Fig. 5.3 Scatter of the Pearson's R and CSM values for all nine HDAs and 32 investigated time periods colored by HDA (top), duration of time period (bottom left) and month of time period (bottom right). Note that the months annotated in the bottom right subplot are based on the start date of their time period.

Also observable in figure 5.3 is that no clear HDA, time period or duration of observations can be distinguished as the best performers. Some algorithms, such as the TC 19-9-WE, TC 19-9 and DD depict better performance in most of the cases, but their worst performance is still far beneath the average performance of all algorithms. Regarding the worst performance (lower right parts of the plot where Pearson's R values are lowest and CSM values are highest), it is clear that 14 days periods in August seriously under-perform, regardless of the deployed HDAs.

In general, observations in May, June and September obtain better performance, but between these months no clear pattern is apparent. Home detection performance at nation-wide scale, in other words, is dependent on a complex combination of duration of observation, time period, criteria, and parameter choice.

5.3.2 Spatial Patterns of LogRatio

Spatial patterns of LogRatios values can be investigated to detect over- or underestimation of user counts compared to the validation population counts (section 5.2.2). Figure 5.4 shows the LogRatios at cell tower level for the HDA-time period combination that, in terms of Pearson's R, has the best performance (upper left), the worst performance (upper right) and two moderate performers (lower left and lower right).

One general observation in figure 5.4 is that all HDAs underestimate populations in major city centers and among major roads compared to the ground truth data. This observation is consistent for all other combinations of HDA and time periods investigated. One reason for this pattern is that, typically, city centers are highly competitive locations between operators (with better services offered by smaller, competitive operators), resulting in smaller market shares for all operators.

Overestimation, on the other hand, occurs mostly in rural areas and touristic areas, but these patterns vary depending on the deployed HDA and time period as will be discussed in more depth later. Overestimation in rural areas could potentially be explained by rural areas that have higher local market shares compared to the nation-wide market share because of historical (coverage) reasons and brand loyalty of small communities. Such potential causal relations between local market shares and home detection performance, however, need confirmation from proper investigation.

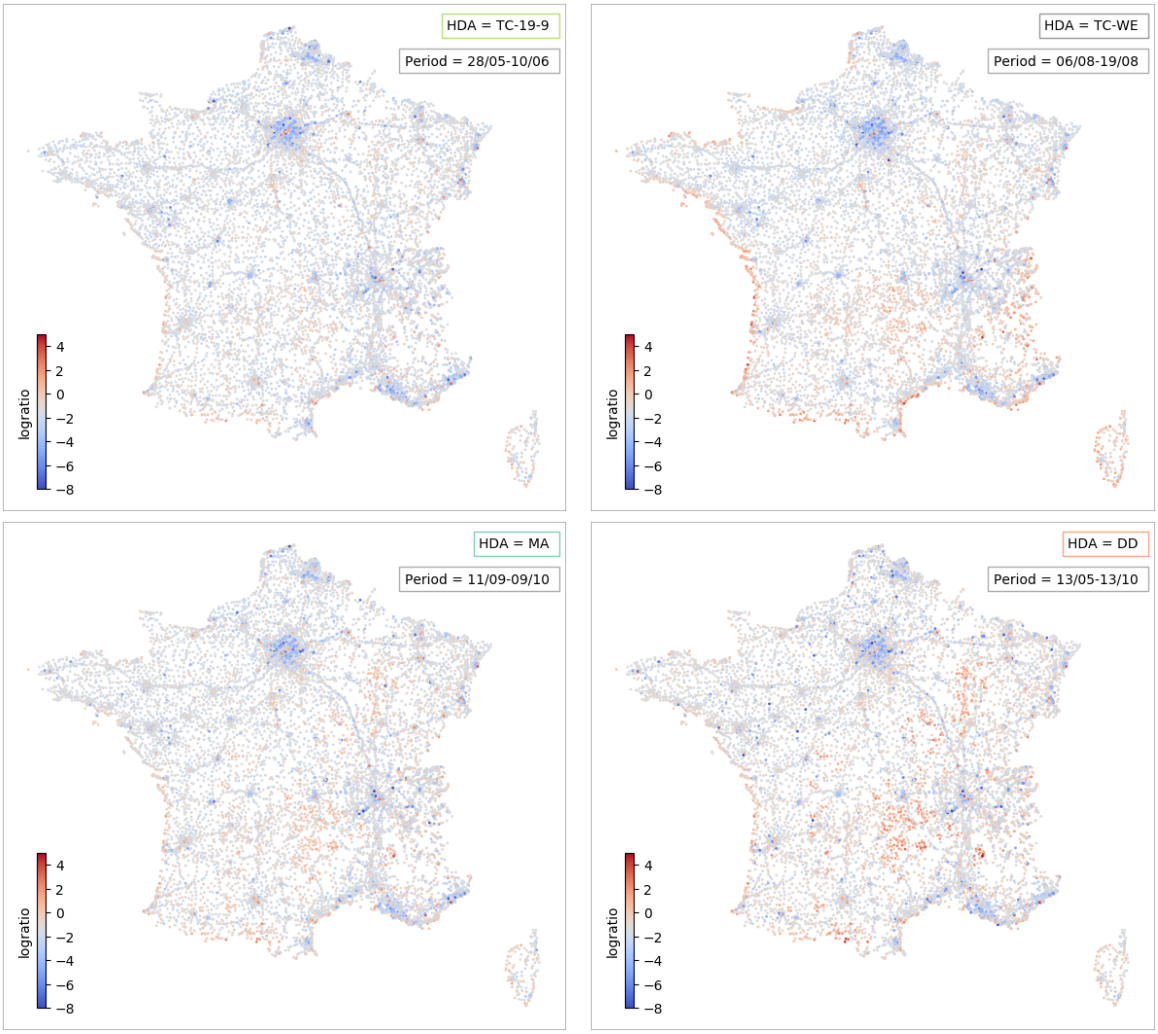


Fig. 5.4 Spatial patterns of LogRatios for different algorithms and time periods. Colors correspond to LogRatios in a gradient from red (over-estimation) to blue (under-estimation). Grey areas correspond to the expected -1.27 ($log_e(0.28)$) LogRatio based on the overall 28% market share of Orange. Subplots show the HDA-time period combinations with best performance (upper left), worst performance (upper right) and with moderate performance (lower left and lower right) in terms of Pearson's R.

5.3.3 Sensitivity of Performance to Time Period

As has been suggested by the results in chapter 4 and in figure 5.4 , a main influencer of home detection performance for the French CDR dataset is the time period, and more specifically the summer period. Investigation of the performance of HDAs over time revealed a clear drop in performance during July and August, regardless of the criteria or parameter used. In figure 5.5, this drop is illustrated for four different HDAs by plotting their performance for all periods with a duration of 14 days. The sensitivity to time period is the largest one in the analysis with Pearson's R values typically dropping around 0.10 to 0.15 for time periods in July and August compared to other time periods.

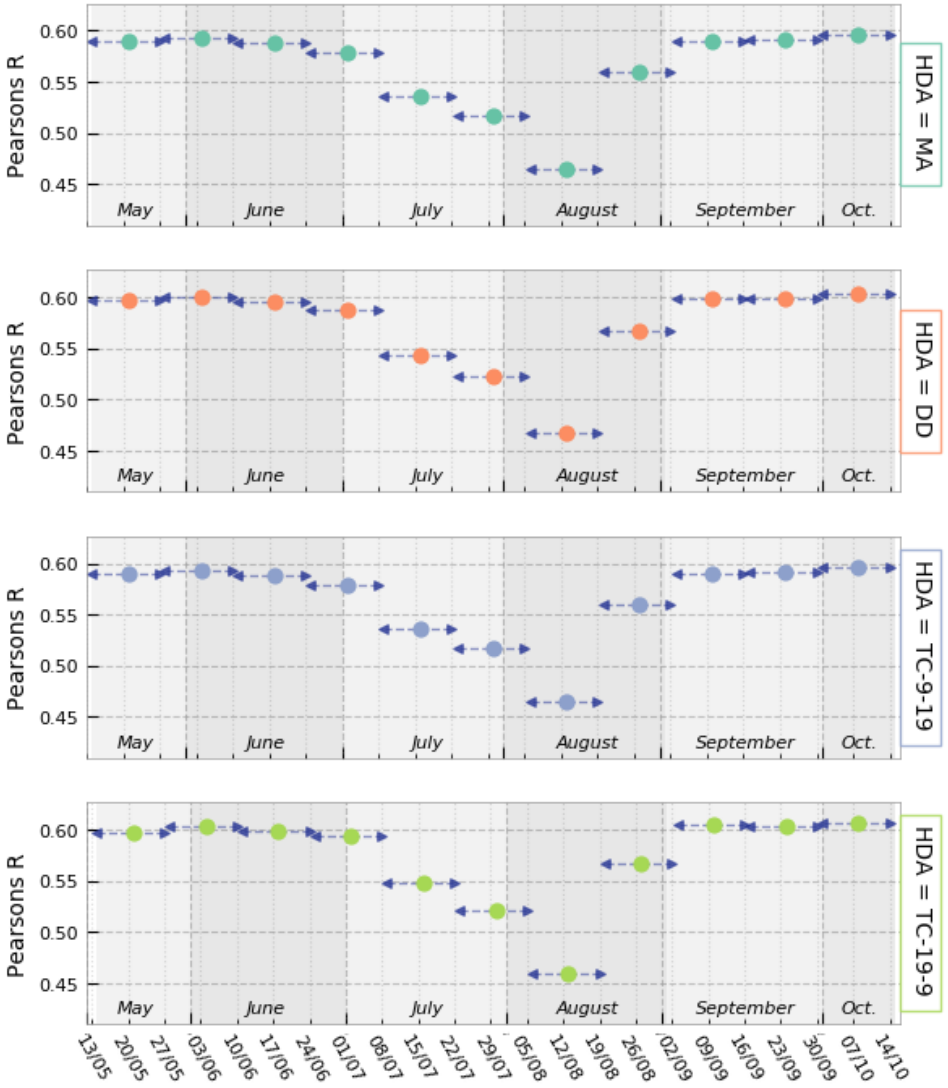


Fig. 5.5 Pearson R values for the relation between user counts and population counts from census data for different HDAs and time periods with a duration of 14 days. Pearson's R values are plotted in the middle of the deployed time periods, with blue dotted lines and blue arrows indicating the duration of observations. Other HDAs depict similar temporal patterns.

Influence of Holiday Mobility on Performance

A main reason for the performance drop during summer months was hypothesized to be the holiday movement of the French population during that period (section 4.4.1). This hypothesis can be confirmed when investigating the change in spatial patterns of user counts for an HDA between a time period outside and a time period during summer. In figure 5.6, the example of the MA algorithm shows a clear rise in user counts at seaside locations and, to a lesser degree, in mountainous areas (which can be linked to major touristic locations) for a 14-day period in August compared to a 14-day period in June. Similar patterns are observed for all HDAs, although the degree of change, or thus the magnitude of the performance drop, is dependent on the deployed HDA and the chosen duration of observation, which will also be discussed next.

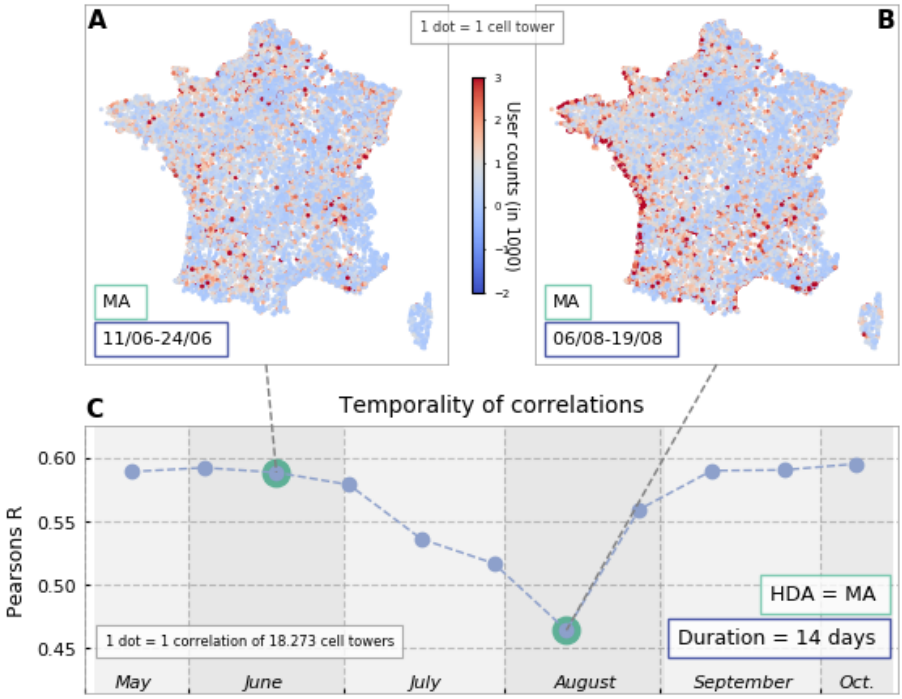


Fig. 5.6 Spatial patterns of the user counts obtained by the MA algorithm for a non-summer (A) and a summer (B) period of 14-day duration. (C) Temporal pattern of Pearson's R value for the correlation between user counts from the MA algorithm and population counts from census data.

5.3.4 Sensitivity of Performance to the Duration of Observation

One outstanding question is to which degree the duration of observation influences the HDAs' performance. For example, it seems only logical that when holiday mobility is influencing home detection, larger durations of observation periods could mitigate this effect. As can be observed in figure 5.7, which shows the sensitivity of HDA performance to different durations of time periods for 4 different HDAs, longer durations of observation periods indeed do mitigate the drop in performance during summer, although only to a certain extent. It forms an interesting realization that, actually, the sensitivity to the duration of observation is subordinate to the sensitivity to time period (in our case, to summer periods), in the sense that duration is influencing performance mainly by the proportion of the time period that occurs in August or July (although to a lesser degree for July).

More specifically, results in figure 5.7 show that shorter durations of observations (such as the 14-day duration) perform amongst the worst when observations are made during summer but amongst the best when the observations are made outside summer months. This is in contrast to long durations of observations, where performance is somewhere in between, depending on the proportion of the time period that occurred in July or August. This effect is also observable for the month and 30-day durations and is independent of the choice of HDA.

For the 154-day period an extra, interesting, observation can be made in figure 5.7. For some HDAs (such as TC-19-9 and TC-WE), given a long enough time period (such as 154 days), and despite including periods of large-scale mobility, it is possible to obtain the same (high) level of performance compared to time periods with shorter durations that are not characterized by large-scale mobility. This however is not true for all HDAs. Performance for the 154-duration period of the MA and TC-9-19-WK algorithms for example, are not even close to obtaining the performance levels of shorter duration periods that are not occurring in summer.

The consequence is that there is no clear single duration of observations that performs best, and that all is dependent on the combination between duration, phases of large-scale mobility, and HDA choice. One rule of thumb could be that, if no insights are available on periods of general mobility, performing home detection on longer durations might be the safest choice, but will probable lead to moderate performance compared to shorter periods of observations for which one can be sure that they are free of large-scale mobility.

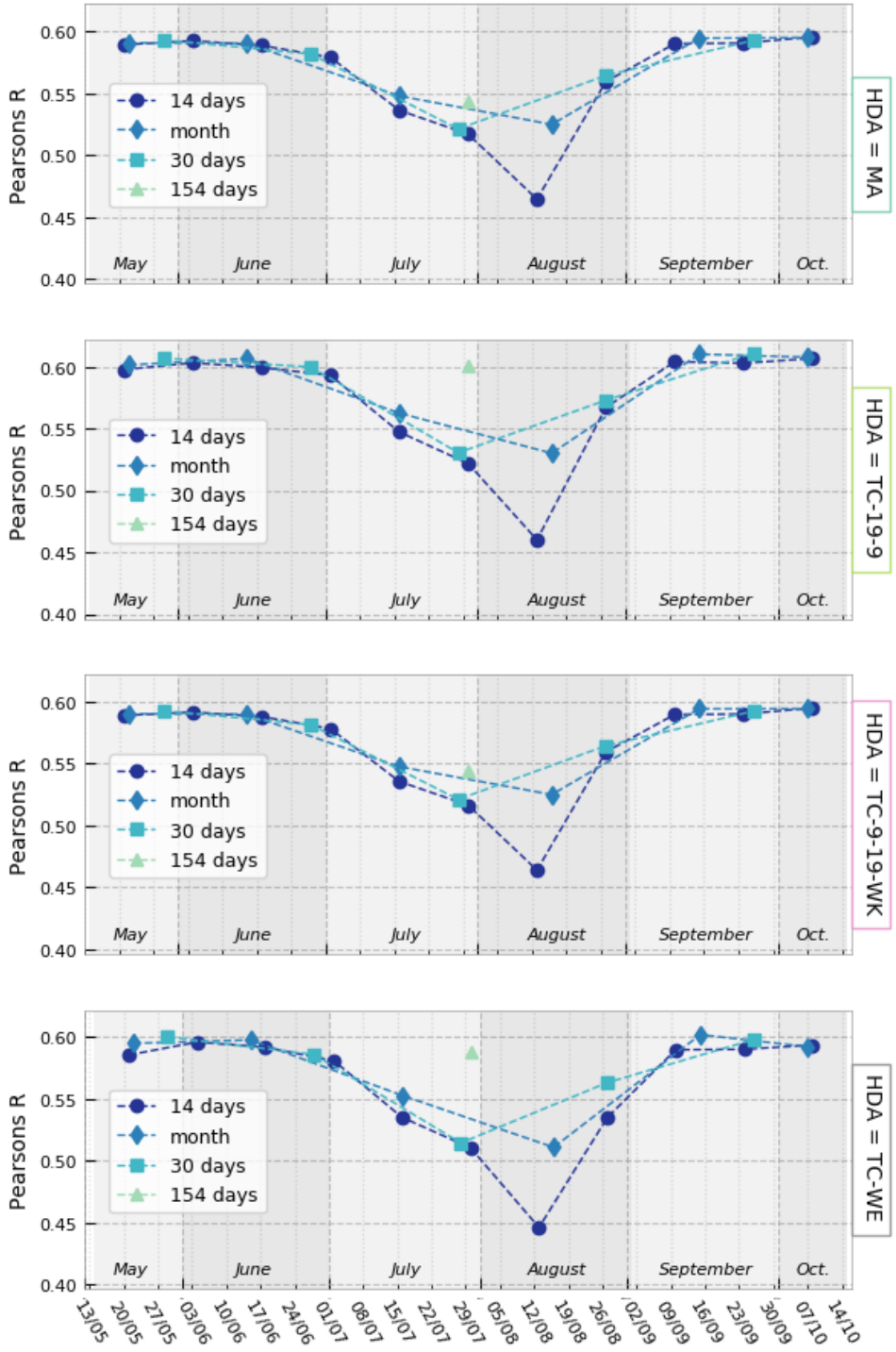


Fig. 5.7 Temporal pattern of Pearson's R values for the correlation between user counts and population counts from census data for different HDAs and different durations of observations.

5.3.5 Sensitivity of Performance to Criteria and Parameter Choice

Finally, investigation is on the sensitivity of HDA performance to the choices of criteria and/or parameter. It forms an interesting observation that they are less influential compared to time period or, sometimes, even duration of observation choice. For periods with a 14-day duration, for example, the effect of criteria choice is about a difference of 0.025 in Pearson's R, whereas the summer period effect is about 0.15, or thus an order higher (figure 5.8 B).

Still, some interesting observations can be made when comparing the performance of different HDAs. One observation is that the TC criterion outperforms the MA and DD criteria for some parameters (such as the 19-9) but not for all. In other words, parameter choice for the TC criterion does have an impact on (relative) performance. For the French CDR data, for instance, defining nighttime as 21-7 hours instead of 19-9 hours results in substantial performance loss for all 14-day periods investigated (figure 5.8 C). Even more remarkable is that the 21-7 parameter, at least for 14-day durations, is consistently outperformed by the 9-19 parameter, which is a daytime definition (figure 5.8 C). This finding drastically challenges the assumption that using nighttime would be better because people spend more time at home then. The performance of different TC parameters is also influenced by the time period. Using nighttime and weekends, for example, outperforms using weekdays during non-summer periods, but such performance is reversed for all 14-day duration periods in August (see figure 5.8 D).

The most remarkable finding regarding the sensitivity of performance to criteria and parameter choice is that it is strongly dependent on the duration of observation. While performances for 30-day and month durations are similar to the 14-day periods in figure 5.8 A,B,C,D, figures 5.8 E,F,G,H show how patterns of performance change drastically when considering the 154-day duration. Observe, for example, how similar the performance of the TC 9-19, TC 19-9, and TC 21-7 are for all 14-days periods (figure 5.8 C) and how different their performance is for the 154-day period (figure 5.8 G).

For the 154-day duration, the general logic seems to be that the TC criterion with parameters restricted to nighttime and weekend-days perform better than other HDAs (figure 5.8 E). From our experiments, it is not clear what exactly is driving this gain in performance during longer observation periods, nor is it clear from which duration onwards this gain comes to be (most likely between 30 and 154 days). One possible reason is that HDAs with time-constraints criteria need an observation period with long enough duration to operate properly according to their semantics, but this requires further inquiry.

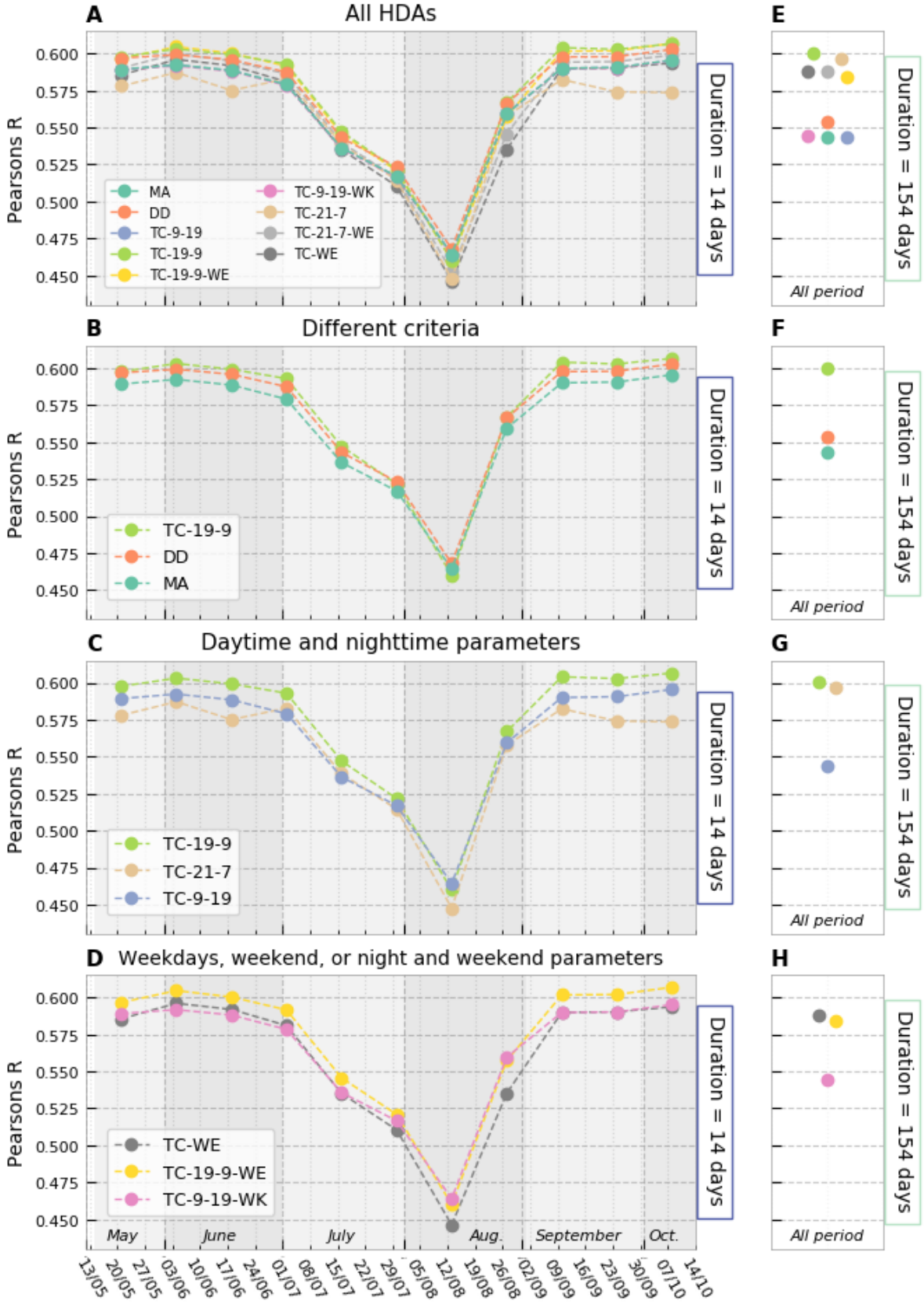


Fig. 5.8 Temporal pattern of Pearson's R values for the correlation between user counts and population counts from census data for all HDAs for the time periods with a 14-day (A,B,C,D) and 154-day (E,F,G,H) duration of observation.

5.4 Discussion

The presented results reveal the different effects of the choice of criteria, parameter, time period, and duration of observation on the performance of home detection on a French CDR dataset when performed by nine HDAs with simple decision rules. In line with the results in chapter 4 the performance of the nine HDAs is found to be moderate with correlations with validation data between 0.45 and 0.62 (figure 5.3). Regarding the spatial pattern of performance, results show consistent underestimation of population in city centers and along main roads for all investigated HDAs and time periods. Overestimation occurred mainly in rural and touristic areas, but is dependent on the deployed HDAs, investigated time period and duration of observation (figure 5.4). The definitions of under- and overestimation are, however, based on an assumed market share of the operator of 28%, which is correct at national level but can vary locally. Consequently, observed spatial patterns, especially when defined by the LogRatio, can be influenced by the pattern of variations in local market shares, although to an unknown degree. Unknown local market shares can potentially influence the upper bound of correlations with ground-truth too [142]. As such, future research is suggested to incorporate information on local market shares when investigating nation-wide performance of HDAs.

5.4.1 Magnitudes of Sensitivities

For France 2007, HDA performance is found to be sensitive, in descending order of magnitude, to the deployed time period, the duration of observation and, to a small degree only, to the criteria and parameter choices. The largest sensitivity is to the July and August period, in which performance drops significantly for all HDAs (figure 5.5). During July and August, HDAs consistently overestimate population counts in touristic regions, suggesting that they are influenced by large-scale holiday movement of the French population (figure 5.6). The effect of the duration of observation is found to be subordinate to this summer effect, with performance of the HDAs being directly related to the proportion of the time period occurring in July and August. Shorter periods which are outside the summer months outperform longer durations for which part of the observations were made in July and/or August (figure 5.7). The consequence is that there exists no clear advice on what duration of observation to use. When periods of large-scale population mobility are unknown to the researcher, longer durations might be the safest option. When such periods are known, however, avoiding them at all costs seems the best advice, even if this means giving up part of the dataset.

One most remarkable finding is that the criteria and parameter choices for HDAs seem to have little influence, especially when compared to the sensitivity to time period and duration of observation. Regarding the criteria we find that the distinct day-criterion slightly outperforms the amount of activities-criterion. The performance of the popular ‘time constraints’-criterion is highly dependent on the chosen parameters. It is remarkable, for example, that the 19-9 hours time constraint outperformed most algorithms, whereas the 21-7 time constraint performed the worst of all tested algorithms.

Additionally, the performance of HDAs based on the 'time constraints'-criterion is rather inconsistent, with intuitive (e.g. nighttime and weekends) and contra-intuitive (e.g. only working days) constraints outperforming each other at different time periods and for different durations of observations (figure 5.8). All of this makes one wonder why the 'time-contraints'-criterion, although popular in literature [142], should ever be opted for; if not for one observation that, for the 154-day period, time-constraint HDAs with intuitive parameters outperformed all other HDAs, with performance equaling the best performances of other HDAs for other periods and (shorter) durations in the analysis.

5.4.2 Contributions

The main contribution of the presented analysis is that it offers an insight in the combined effect of researcher's choices on the performance of home detection when deployed on the French CDR dataset. The results can help other practitioners to decide on suitable home detection algorithms and time periods of observation. It is however, strongly urged that other researchers reproduce similar analyses on their own datasets, mainly because performance in France was found to be most sensitive to summer periods (which likely corresponds with large-scale holiday movement), and partly because of the potential influence of unknown local market shares both of which are location and time dependent. Additionally, the presented results can inform the assessment of uncertainty and error related to home detection methods when performed on CDR data or, to a lesser degree, other large datasets of geo-located traces for individual users. One contribution is to the wider debate on integrating big data in official statistics in, at least, two ways. First, the results serve as a reminder that, despite showing large potential, many big data sources and related methods remain in need of quality assessment. Secondly, the presented work forms an example that collaboration between academia, private sector and official statistics offices can be extremely fruitful, even though the actual realization of such collaborations is, due to multiple practicalities, never easy.

Chapter 6

Correcting Mobility Entropy

> Explorers, who have a tendency to wander between a larger number of different locations.
>
> *Dr. Luca Pappalardo*

Abstract

In this chapter, focus is on one mobile phone indicator, namely the Mobility Entropy (ME). It is shown that the traditional way to calculate ME from CDR data is dependent on the density of cell towers. The consequence is that the ME is biased towards areas with higher cell tower densities, such as city centers, and that comparisons of ME values between regions with different cell tower densities are problematic. As a solution, the Corrected Mobility Entropy (CME) is proposed. Comparing CME and ME values for the French CDR data, CME is found to be less correlated with cell tower density (r = -0.17 instead of -0.59 for ME). Additionally, spatial patterns of ME and CME are shown to be different, with CME values revealing that mobility diversity is higher in sub-urban regions compared to their corresponding urban centers and that mobility diversity is decreasing with urban center size. Proceeding to an interpretation of the observed spatial pattens, relation with other indicators are investigated. Single linear regressions find that mobility diversity in France are related to indicators like income and employment rate. Multiple linear regressions enable a more nuanced interpretation, showing that employment rate, distance to cities, demographics, income, land use, and the role of cars form the main factors to predict mobility diversity in France.

Related Publications and Acknowledgments

- The structure and content of this chapter largely accords with [143], a publication that is authored by the PhD candidate and that builds on the master thesis of Willem Schoors: *Mobility entropy and space, A CDR-based study of entropy behavior in France*, which was presented at the department of Geography, KU Leuven in 2016 under co-supervision of the PhD candidate.
- Acknowledgments go to Willem Schoors for his tireless efforts in performing experiments and preparing the figures used in the paper. For this reason, figures that originate from the paper (and thus were made by Willem) will be cited in the caption.

6.1 A Flawed Mobility Entropy Indicator

6.1.1 Defining the Mobility Entropy Indicator

As discussed before in sections 2.6.2 and 3.3.2, the Mobility Entropy (ME) indicator, is one of the most popular indicators to describe individual movement patterns derived from CDR data. Given its definition, the ME indicator forms a quantification of the diversity of a user's movement pattern, reflecting how many different cell towers were visited by a user while, simultaneously, taking into account how evenly visits are distributed among cell towers. Several possibilities exist to calculate mobility entropy values (see section 2.6.2). Given the sparse temporal resolution of the users' activities in the French CDR dataset, the amount of visits is used to approximate the probability of a user being at a cell tower, resulting in the definition of Mobility Entropy (ME) as:

$$ME = -\sum_{i=1}^{n} p_i \log_{10}(p_i) \qquad and \qquad p_i = \frac{e_i}{\sum_{i=1}^{n} e_i}, \tag{6.1}$$

where ME is the mobility entropy of a user, i is a cell tower visited by the user, n is the total number of towers visited by the user, p_i is the probability of the user visiting tower i, and e_i is the number of mobile phone events by the user at tower i. The calculation of the ME from CDR data is illustrated in figure 6.1.

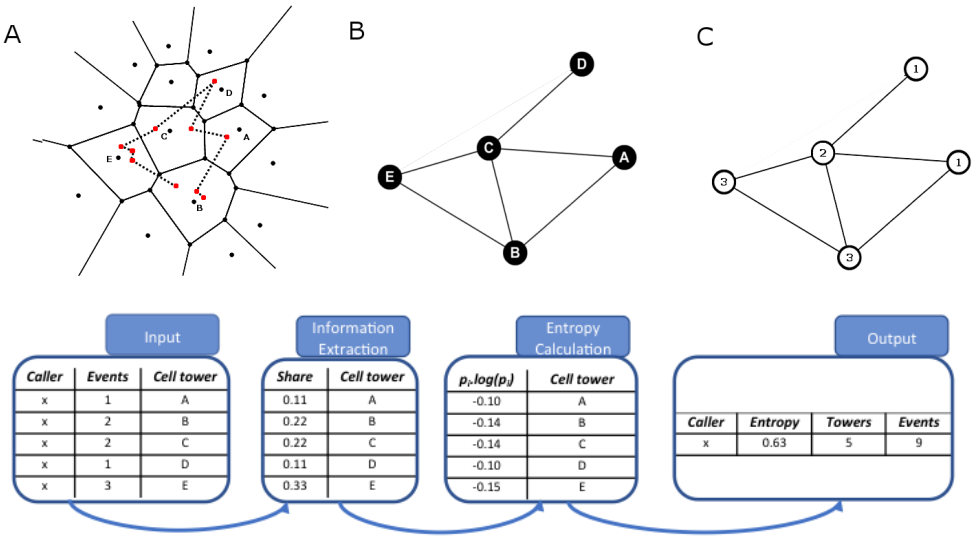
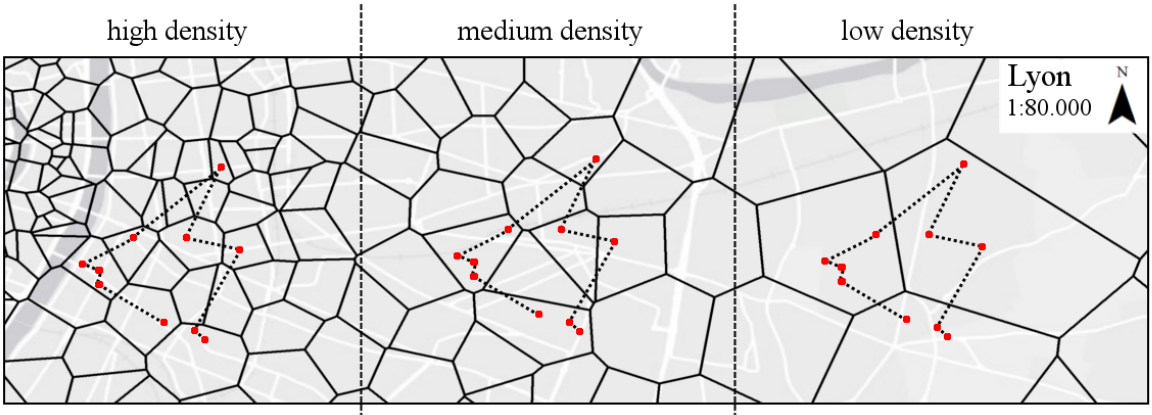


Fig. 6.1 Illustration of the calculation of the ME. (A) Movement patterns for single users are constructed (dotted lines) based on CDR data by means of events initiated (red dots) at cell towers (black dots). The Voronoi polygons (full lines) represent the borders of coverage areas for different cell towers. (B) The movement network based on the visited towers is constructed from the movement pattern. (C) Cell towers in the movement network are weighted by the number of visits to serve as a proxy for the probability that a user visits each cell tower. Shares of the numbers of visits for each cell towers are used to calculate the ME. (Bottom) The calculation of ME values by a database representation. Source: figure adapted from [143].

6.1.2 Bias of the Mobility Entropy Indicator

A major problem with the calculation of ME values is that they are dependent on the absolute number of cell towers a user visits while, in fact, they should not be. This bias becomes apparent when critically investigating equation 6.1: higher numbers of cell towers lead to higher values for n and lower values for p_i, both combined resulting in higher values for ME. Figure 6.2 illustrates the bias in detail, showing how the same movement pattern results in different ME values when calculated for three areas that are characterized by decreasing cell tower density. The obtained ME values differ significantly because the resolution of information derived from the cell tower network is different. The implication is that ME values are biased when cell tower densities are changing, meaning that ME values form a non-objective indicator for comparison between areas that have different cell tower densities. In the case of CDR data, this means that spatial distributions of ME values cannot be trusted especially when comparing areas with different cell tower densities.

The goal of this chapter is to unveil the bias of the ME and to elaborate a correction that reduces the dependency on cell tower density, enabling a more objective comparison of mobility diversity between regions. The Corrected Mobility Entropy (CME) then is calculated for all users in the French CDR dataset and its differences with the original ME are discussed. Ultimately, the relation with urban areas and other indicators are investigated for both the ME and the CME.



tower	events	p_i	$p_i \log p_i$
1	2	0.2	-0.14
2	1	0.1	-0.10
3	1	0.1	-0.10
4	1	0.1	-0.10
5	1	0.1	-0.10
6	3	0.3	-0.16
7	1	0.1	-0.10
total	10	1	**-0.80**

ME = 0.8

tower	events	p_i	$p_i \log p_i$
1	3	0.3	-0.16
2	1	0.1	-0.10
3	1	0.1	-0.10
4	1	0.1	-0.10
5	4	0.4	-0.16
total	10	1	**-0.63**

ME = 0.63

tower	events	p_i	$p_i \log p_i$
1	3	0.3	-0.16
2	3	0.3	-0.16
3	4	0.4	-0.16
total	10	1	**-0.47**

ME = 0.47

Fig. 6.2 Illustration of the bias to cell tower density when calculating ME values. ME values are calculated for the same movement pattern (red dots are cell phone events, black lines are Voronoi polygons of cell towers, dotted line is the presumed trajectory) in three different locations with varying cell tower density. The same pattern results in the calculation of different ME values depending on cell tower density. Source: [143].

6.2 Correcting Mobility Entropy

6.2.1 Defining the Corrected Mobility (CME) Indicator

To balance the influence of cell tower density on ME calculation, a normalization of the ME formula for cell tower density is proposed. Existing entropy normalization, like for example that in the Python toolbox *bandicoot.py*, consists of dividing the entropy by the logarithm of either the number of visited towers or the number of events [48]. Such approach might seem promising but in fact corrects for the amount of actions on cell towers, not for the cell tower density. For this reason, the proposed correction focuses on correcting each cell tower based on its surrounding cell tower density. The formula for Corrected Mobility Entropy (CME) then is:

$$CME = -\sum_{i=1}^{n} p_i \log_{10}(p_i) c_i \qquad and \qquad p_i = \frac{e_i}{\sum_{i=1}^{n} e_i}. \tag{6.2}$$

The formula for CME is the same as the formula for ME in equation 6.1 except for the correction factor c_i for each cell tower i which will be explained in the following paragraphs.

The idea behind the proposed correction factor c_i is straightforward. When cell tower density is high, visiting a new cell tower (and hereby increasing the mobility entropy) becomes easier and should therefore be assigned a lower weight compared to visiting a new cell tower in a low-density area, where the registration of a user on a new cell tower is more unique. The weights are assigned according to a simple correction factor that is directly related to the cell tower density.

A Proxy for Cell Tower Density

Before explaining how to derive the correction factors c_i from cell tower density, it is important to decide which proxy for cell tower density to use. Experiments were done on different proxies and, ultimately, it was opted to use the Voronoi polygons (which are calculated based on the locations of all cell towers) as a measure for cell tower density. This way, other proxies are avoided that would have user defined parameters like, for example, the radius of a circular area in which the density of cell towers could be counted.

Concerning the Voronoi polygons, both the area and the circumference were examined. Ultimately, the Voronoi circumference was selected because it was less influenced by irregular shapes. The intuitive meaning of using the Voronoi circumference is clear: high-density areas will have smaller Voronoi polygons, while the opposite happens in low-density areas. A possible disadvantage of this proxy is the limited density detection range, as the circumference of Voronoi polygons is only influenced by the towers directly surrounding the tower under consideration.

Constructing the Correction Factor

Converting the Voronoi circumference to correction factor c_i is done based on equation 6.3:

$$c_i = \log_{10}\left(\frac{(10^a - 10^b) \times (d_i - \min(d))}{\max(d) - \min(d)} + 10^b\right), \tag{6.3}$$

where c_i is the correction factor for cell tower i, d_i is the Voronoi circumference for cell tower i, (d) is the set of Voronoi circumferences for all cell towers, and a, b are the lower and the upper bound of the scaling range.

To construct the correction factor, logarithmic scaling is performed on the Voronoi circumferences and the resulting c_i values are bound to range within a and b. This means that the cell tower with the smallest Voronoi circumference (indicative for the highest cell tower density) will have a correction factor c_i that equals a. Similarly, the cell tower with the largest Voronoi circumference will have a correction factor c_i that equals b. Note that, in the case of France, it was opted to use a logarithmic scaling because of the disproportionately large number of cell towers with very small Voronoi circumferences in Paris. Note also that the scaling range was deliberately taken to be symmetrical around 1, implying that the correction factor increases when cell tower density deviates more from the mean cell tower density. The correction factor is thus bounded by the (a, b) parameter choice.

Parameterizing the Scaling Range

Sensitivity to the (a, b) parameter choice for the scaling range is tested for different (a, b) sets, after which the (0.7,1.3) range is chosen because it minimized the Pearson's correlation coefficient between CME and cell tower density to -0.17 (table 6.1). Note that for scaling range (0.6,1.4), the observed Pearson's R is even lower (0.07) but because of the change in direction of the correlation it is opted not to use this range. Sensitivity analysis at higher resolutions (e.g. 0.01 level) have not been performed because calculations of CME values for the entire CDR dataset are time-costly but could further optimize the parameter choice.

Scaling Range (*a*,*b*)	Pearson's R
0.9,1.1	-0.51
0.8,1.2	-0.38
0.7,1.3	-0.17
0.6,1.4	0.08
0.5,1.5	0.30

Table 6.1 Sensitivity to different scaling ranges (a, b) for the correlation between CME and cell tower density. Correlations are expressed by means of Pearson's R. Cell tower density is approximated by the Voronoi circumferences of the cell towers in the French CDR dataset.

Correction factors c_i are calculated for each cell tower in the CDR dataset based on their Voronoi circumference and the (0.7,1.3) scaling range and are mapped in figure 6.3. Unsurprisingly, correction factors with values <1 can be observed in all urban areas, for both large and small cities alike. Suburban areas for most cities display correction factors smaller, but very close to 1, as are the cases for coastal areas, border areas (which are quite densely populated in France), and some areas along major transport axes. Rural areas, on the other hand, have correction factors >1, with highest correction factors being attributed to mountainous areas (e.g. the French Alps) and nature reserves (e.g. the Landes de Gascogne natural park situated south of Bordeaux).

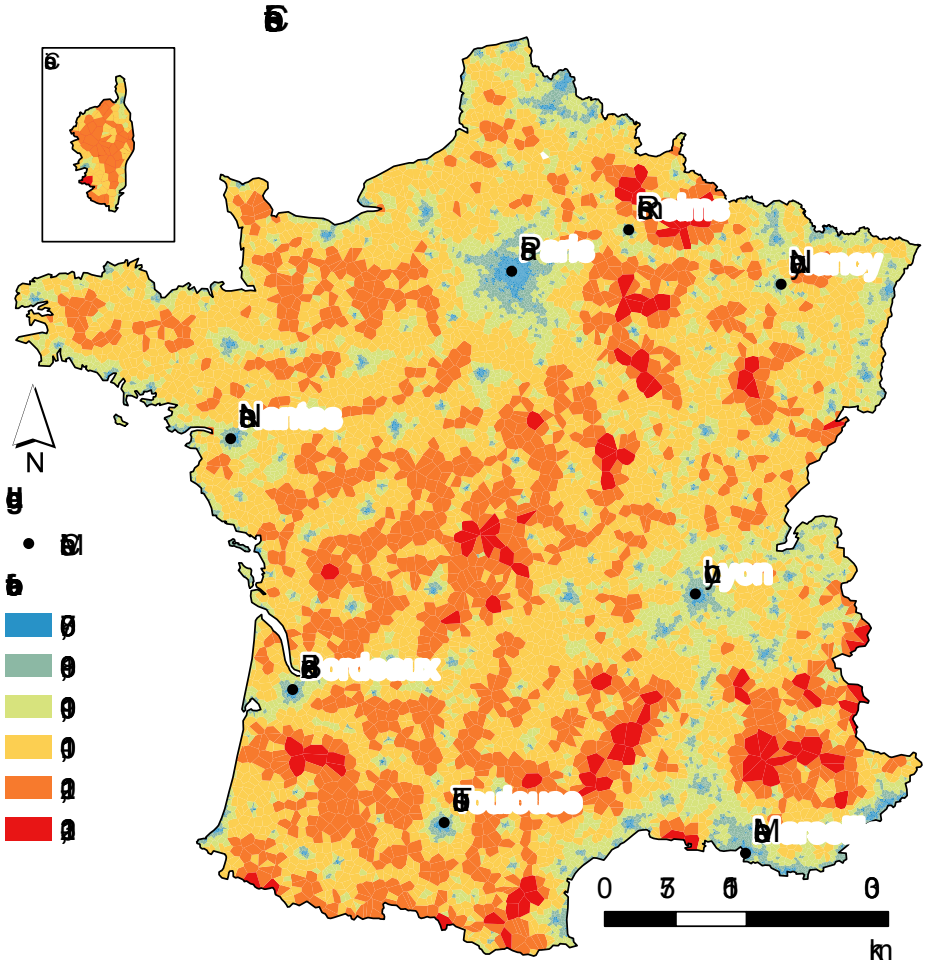


Fig. 6.3 Spatial pattern of the correction factors c_i calculated by using the Voronoi circumference as proxy for cell tower density and for a scaling range (0.7,1.3) for all cell towers in the French CDR dataset.

6.3 Relations between ME, CME, Urban Areas, and Other Indicators in France

To study the proposed correction, in a first step, obtained ME and CME values for France will be compared with each other, with the density of cell towers, and their spatial pattern will be shown (section 6.3.1). In this phase, values will be territorially aggregated at cell tower level by means of home detection. In a second step, distributions of ME and CME values for different urban areas in France are calculated and compared (section 6.3.2). In this phase, values will be aggregated on the level of the urban areas proposed by INSEE. Finally, values of CME and ME or put in relation with other indicators from census data (section 6.3.3). To do so, aggregation at the level of the French communes (more or less according to municipalities) is necessary to ensure comparability between both datasets.

6.3.1 Calculating CME from CDR data

The Mobility Entropy (ME) and the Corrected Mobility Entropy (CME) are calculated for all 18.5 million users in the French CDR dataset and aggregated at cell tower level based on the TC 19-9-WE algorithm defined in table 5.1. Indicators are calculated for the period from 1 September till 14 October 2017, that is, a 1.5 month period. This period was chosen because it was the longest time period the system could handle at the time of the analysis (see section 1.3.3 for more information on the limitations of the system). In addition, the time period is outside summer months in order to avoid unnecessary influence of holiday movement on the results (see section 5.3.3 for more information on the holiday movement in France during summer period).

To highlight spatial patterns, the Getis-Ord Gi* statistic [61] is used (e.g. in figure 6.7). This statistic reveals significant spatial clusters of high (low) values for an indicator, thus revealing hot-spots (cold-spots) of values. Within the statistic, each location and its wider environment (chosen to be 25 km in this analysis) are compared to values of all investigated locations. One remarkable observation when analyzing the Getis-Ord Gi* statistics of the ME and the CME is that multiple areas in France have statistically low values (cold-spots). Since most of these areas are remote or mountainous areas one could interpret human mobility to be different here. Another explanation, inspired by the extreme low numbers of average visited cell towers per user in these areas (figure 6.7 (top right)), is that individual movement is insufficiently captured here, resulting in abnormally low entropy values despite the proposed correction. In other words, the spatial resolution of the cell tower network in these areas might not be sufficient to study the diversity of human mobility. As a consequence, it was decided to filter out cell towers where the average number of visited towers over the 1.5-month period is smaller than 10. The exact spatial pattern of this filtering is indicated in green in figure 6.7 (bottom right). In total 814 out of 17,385[1] cell towers were omitted, which is about 4.6% of the cell towers in the analysis.

[1] Note that the number of total cell towers in this analysis is slightly lower compared to the number of cell towers in the previous chapter because of the limited time period (1.5 months against 6 months)

6.3.2 Distributions of ME and CME in Urban Areas

To investigate relations between mobility diversity and the French urban system, the ME and the CME are aggregated into the official urban area classification proposed by INSEE. Aggregation is done by merging the observations of all cell towers (and their related users) that are located in the same urban area. Differences in the distributions of ME and CME values within the same urban areas are tested by means of the pairwise Wilcoxon rank-sum test to understand the effect of the correction in different urban areas. Similarly, the differences in the distributions of CME values between different urban areas is tested by means of the Wilcoxon rank-sum test in order to investigate whether mobility diversity is significantly higher in different urban areas across France.

The urban area classification in France is based on the identification of employment centers and their area of influence through commuting data. As such it goes beyond the typical physical borders defined by the continuity of buildings often used in urban unit delineation, which facilitates studying city organization and development based on the dynamic interactions between locations [140]. As listed in table 6.2, the urban area classification in France consists of 9 classes, being distinguished mainly by the size of the employment center in the *central urban unit*. Major, medium and small poles are employment centers offering respectively >10,000, between 5,000 and 10,000, and between 1,500 and 5,000 jobs, respectively. The surroundings of a pole are made up of municipalities for which more than 40% of the working population commutes daily to this pole. Special cases are being recognized for municipalities that have several poles to commute to (multi-polarized municipalities) or municipalities that are not influenced by major, medium or small poles (isolated municipalities)[2]. In this analysis the 2010 urban area classification will be used, which is based on data collected between 2006 and 2010 in national census surveys and therefore accords best with the CDR dataset captured in 2007. The urban area classification for France anno 2010 is shown in figure 6.4.

Urban Area	Description	Municipalities (%)	Cell towers (%)
111	Major pole (>10,000 jobs)	9	54
112	Surroundings of a major pole	34	18
211	Medium pole (5,000 to 10,000 jobs)	1	3
212	Surroundings of a medium pole	2	1
221	Small pole (1,500 to 5,000 jobs)	2	3
222	Surroundings of a small pole	2	0.3
120	Multi-polarized, in a large urban area	11	5
300	Multi-polarized, other	19	6
400	Isolated, outside influence of poles	20	10

Table 6.2 Urban areas in France based on the official statistics office classification for 2010 and their related share of municipalities and Orange cell towers.

[2]For more information on the urban area classification in France visit: https://www.insee.fr/fr/statistiques/1281191.

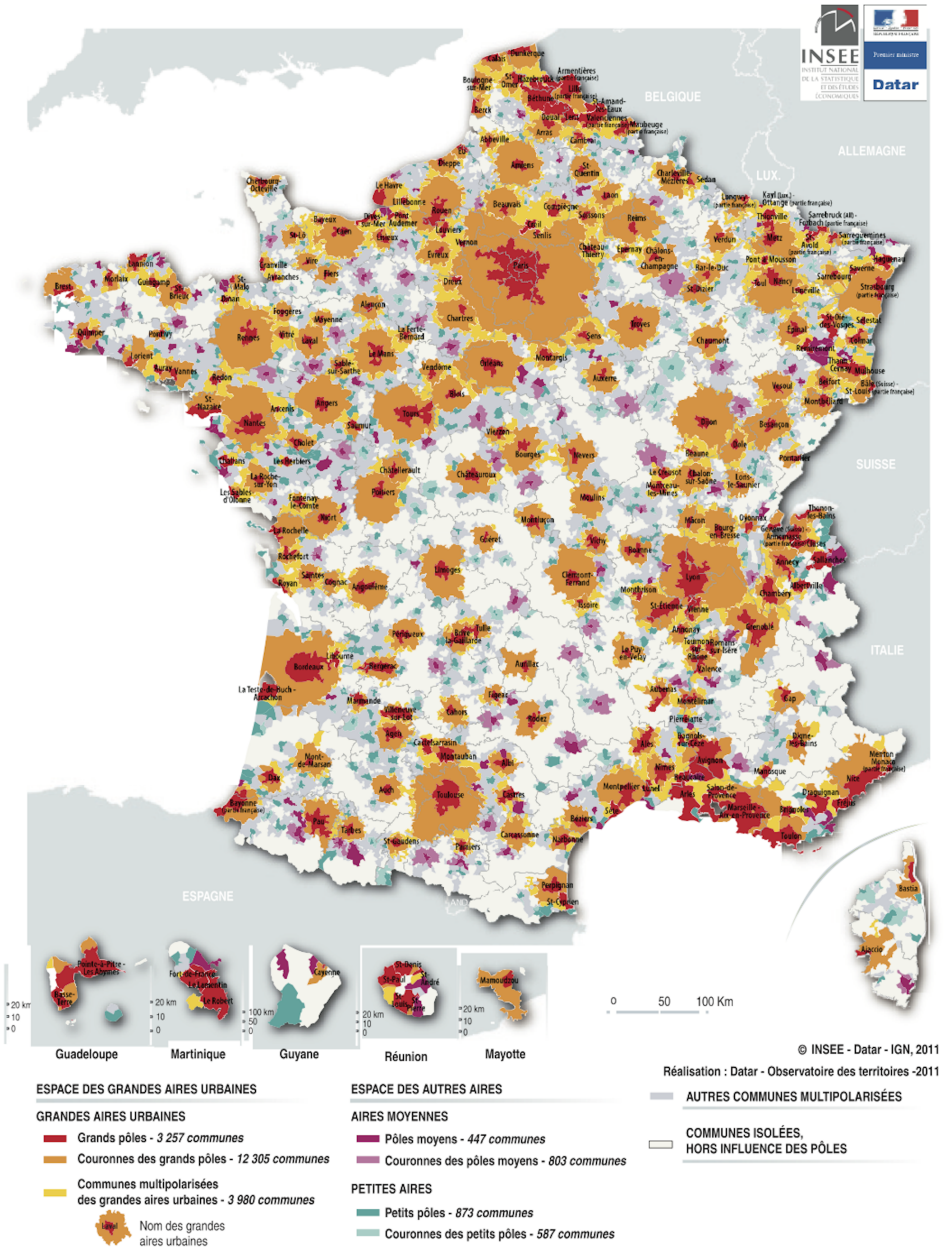


Fig. 6.4 Urban Areas Classification for 2010 in France. Source: Insee, zonage en aires urbaines 2010 (www.insee.fr/fr/statistiques/1281191).

6.3.3 Relations with Socio-Economic and Environmental Indicators

Ultimately, investigation is on the relation between the ME, the CME, and a set of socio-economic and environmental indicators that are listed in table 6.3. The question is whether such relations can help understand the observed patterns of mobility diversity in France.

Indicator	Description
Median Income[1]	Median income per household member (in euro)
Active population[1]	Share of municipality population between 15 and 65 years old (in %)
Working population[1]	Employed share of the active population (in %)
Commuting distance [2]	Median commuting distance for mobile phone users in a municipality (in km).
Car Share[1]	Share of trips made by car (in %)
Public transport share[1]	Share of trips made by public transport (in %)
Car ownership[1]	Average number of cars owned per household
Employment in municipality[1]	Share of people working in their municipality of residence (in %)
Mean elevation[3]	Mean elevation of the municipality (in m)
Artificial land use[4]	Share of artificial land use in a municipality (in %)
Agricultural land use[4]	Share of agricultural land use in a municipality (in %)
Natural land use[4]	Share of natural land use in a municipality (in %)
Road distance[5]	Average distance in the municipality to nearest regional road (in m)
City distance[5]	Average distance in the municipality to nearest city with > 20,000 inhabitants
Station distance[5]	Average distance in the municipality to nearest train station

Table 6.3 Socio-economic and environmental indicators considered, [1] based on census data, [2] based on CDR data, [3] based on SRTM data, [4] based on Corine land cover maps, and [5] based on GIS analysis

In a first step, pairwise linear regressions between indicators are performed. This provides a direct insight in the relation between the ME, the CME and each indicator but does not allow for a proper comparison between indicators regarding their share in predicting the ME or the CME. In the second step, multiple linear regressions are performed, using either the ME or the CME as the dependent variable, and the normalized socio-economic and environmental indicators as independent variables. All independent variables are normalized before applying the multiple regression models, allowing the comparison between estimated regression coefficients (and their statistical significance) for each indicator. In a second step, multivariate linear regressions with interaction terms are performed. The inclusion of interaction terms is done because of considerable multi-collinearity between the independent variable and offers insights in the degree to which the effect of one independent variable varies as a function of a second independent variable. They also help to better understand or, at least, hypothesize, the different relations and processes at play.

All socio-economic indicators in table 6.3 are obtained from INSEE for the year 2007. Information for the environmental variables is retrieved from Corine land cover maps based on satellite images of 2006. Both land cover maps and census data are freely available to the public. As census data is captured at the level of the administrative municipality, the spatial resolution of the investigation is changed to the French municipality level when the indicators are deployed. Aggregation of ME and CME values from cell tower to municipality level is done by attributing users with a presumed home at a certain cell tower to the municipality that cell tower is located in. Distributions at user level for all municipalities can therefore be calculated, but for the regression models average values per municipality are used.

6.4 Patterns of Mobility Diversity in France

6.4.1 Relations between ME, CME, and Cell Tower Density

Relations between ME, CME and cell tower density as approximated by the Voronoi circumferences of each cell tower, are given in figure 6.5. For the ME, the Pearson's R is found to be -0.59. Applying the proposed correction results in Pearson's R dropping to -0.17. Although still showing a small negative trend, most probably because of the influence of the highest tower density areas (lowest Voronoi circumferences), it is fair to say that calculating CME largely alleviates the effect of cell tower density, as was its initial goal. The result is promising as the CME values are almost completely liberated from the structural bias of cell tower density, rendering them more comparable between regions with differing cell tower densities.

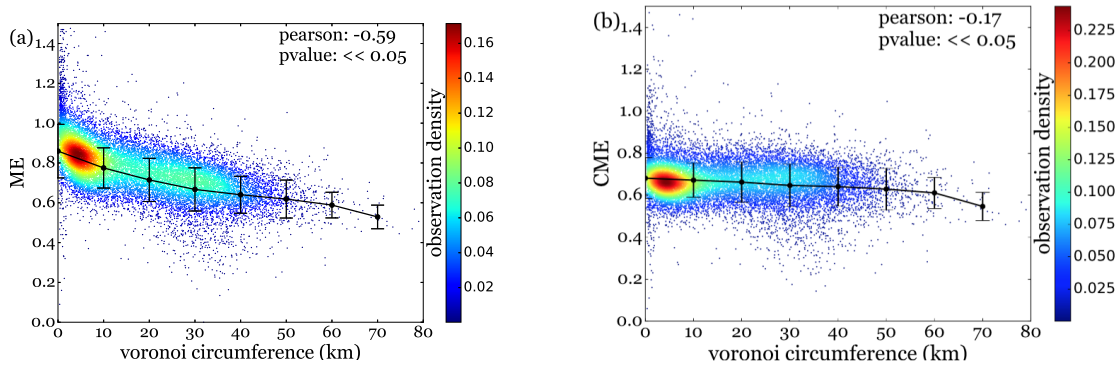


Fig. 6.5 The average ME (a) and CME (b) against cell tower density (approximated by means of the Voronoi circumference of each cell towers). Averages are calculated by means of aggregation at the home cell tower. Observations (one dot is one cell tower) are colored based on the local density of observations in the plot. Trends of the average values and standard deviation per deciles of the Voronoi circumference are shown by the black lines. Deciles 0-10, 10-20, and 20-30 are not shown because they obscure the red centers of observation in the left-hand side of the plots. Source: [143]

The relation between the ME and the CME for all cell towers is shown in figure 6.6 (a). Only a few cell towers that have equal average ME and CME values and are thus situated on the 1:1 line. Cell towers that diverge from the $y = x$ line are influenced by the correction, with most of them experiencing a small decrease from ME to CME. This is also obvious in figure 6.6 (b), which shows the distribution of the correction factors (c_i) as calculated by equation 6.3. The bimodal distribution indicates that most of the cell towers are attributed a correction factor <1, with a peak around 0.78. The spatial pattern of c_i in figure 6.3 showed this first peak of (strongest) corrections to be located in the city centers throughout the country, which is as expected given that city centers typically have higher densities of cell towers. Cell towers that have a correction factor >1 are fewer and, in general, corrected to a smaller degree as can be observed from the second peak around 1.03 in figure 6.6 (b). These cell towers are mostly located in rural and mountainous areas (figure 6.3).

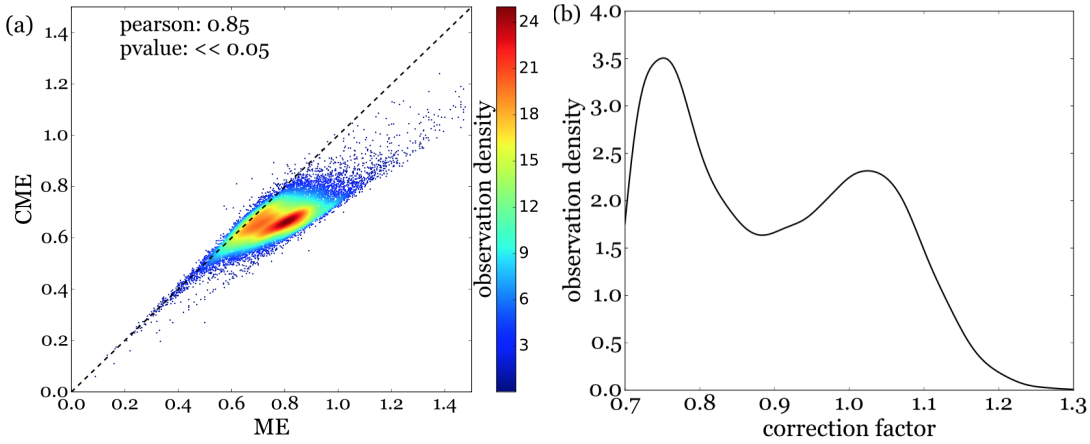


Fig. 6.6 (a) Relations between average ME and average CME values at cell tower level. Each dot represents a cell tower and is colored based on the local density of observations in the plot. The dotted line is the $y = x$ line. (b) Distribution of applied correction factors (c_i) to all cell towers in the French CDR dataset. The spatial pattern of this distribution can be observed in figure 6.3. Source: [143]

6.4.2 Spatial Patterns of (Corrected) Mobility Entropy

The bias of the ME with regard to cell tower density is well illustrated in figure 6.7 by the accordance between the spatial patterns of ME values and the average number of visited cell towers. The spatial pattern of ME shows a highly centered pattern around cities with extensions that relate to, respectively, areas with high population densities and roads. On the contrary, the spatial pattern of CME values renders a more homogeneous spatial pattern that, although still highlighting cities and main roads, is less dominated by them. Interestingly, it seems that the CME values, in comparison to the ME values, are capable of differentiating smaller geographical regions. The most obvious example of this is the wider Paris region where, despite having a rather constant density of cell towers, the spatial patterns of the CME values reveal differences at a higher resolution compared to the ME values.

Studying the hot- and cold-spots produced by the Getis-Ord Gi* statistic in figure 6.7, cold-spots are observed in border regions, near important passage roads, or in remote and mountainous areas that are close to the filtered-out areas (section 6.3.1). Mobility patterns are, possibly, different in these areas but it is more probably that these cold-spots form an artifact of the data collection process. Similar to the filtered-out areas, the mountainous and remote areas close to them might suffer from limited spatial resolution of cell towers resulting in insufficiently collected data to create representative mobility patterns. For border regions, this reasoning does not apply as numbers of visited towers in these regions are higher and thus the spatial resolution of cell towers can be deemed sufficient. Rather, it is highly plausible that movement patterns are collected only partially in border regions because CDR data are not collected outside the French territory.

Intriguingly, hot-spots of CME values are located just next to medium-sized and large-sized cities like Nantes, Reims, Toulouse, and Lyon, but not in their city centers. One hypothesis is that these hot-spots represent areas with high shares of commuters whom, besides commuting, also have a rather mobile lifestyle, as commuting solely would result in lower entropy values. The observation that such hot-spots are found in the vicinity of almost all medium-sizes cities in France forms a strong suggestion towards the nation-wide comparability of the CME values and an incentive to investigate their relation with urban areas, as will be discussed in the next section.

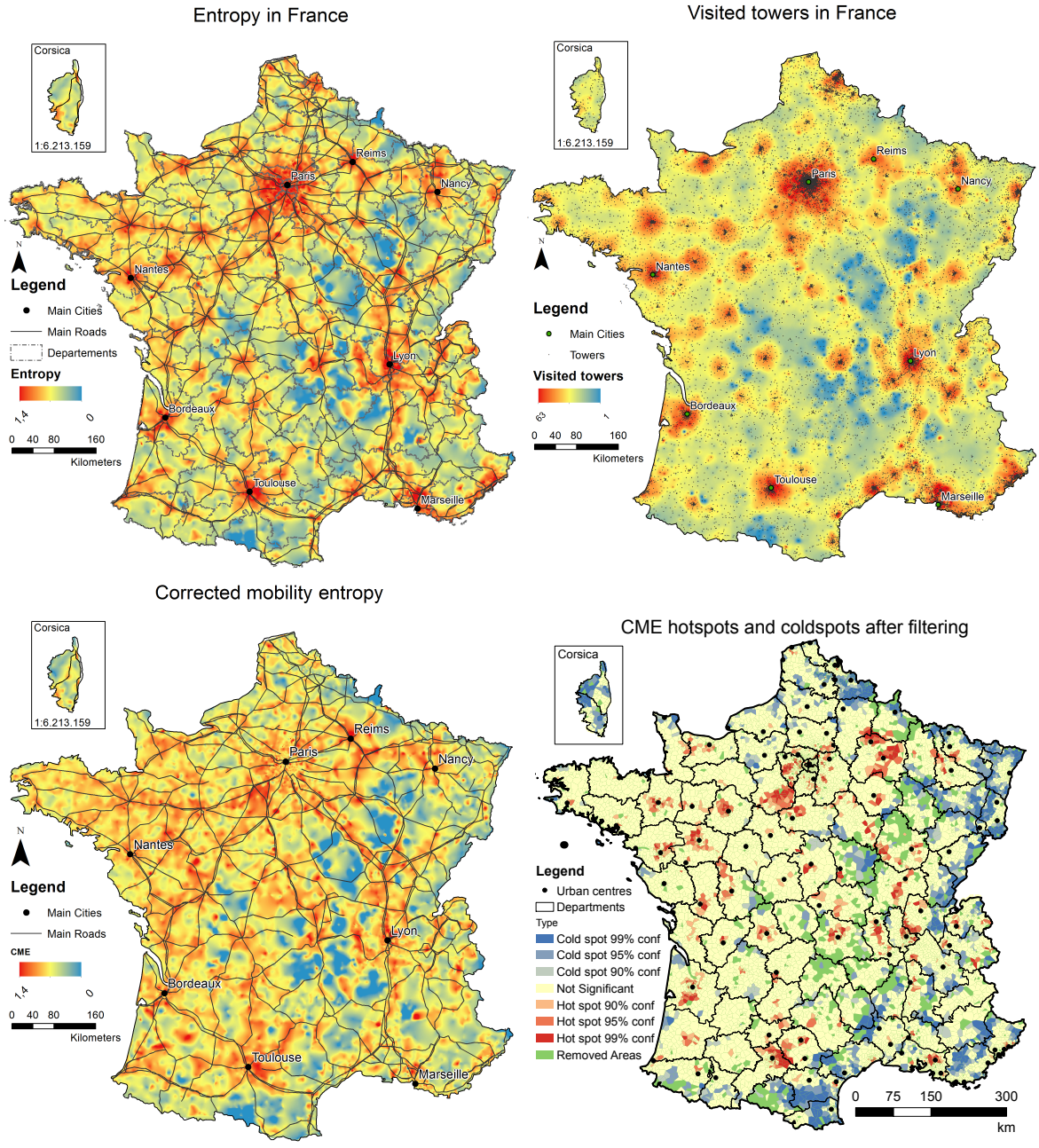


Fig. 6.7 Spatial patterns of the ME values (top left), number of visited cell towers (top right), the CME values (bottom left), and hot-spots (red) and cold-spots (blue) of the CME values at different significance levels as calculated by the Getis-Ord Gi*-statistic (bottom right). Green areas accord to filtered-out cell towers. Source: [143]

6.4.3 Mobility Diversity in Urban Areas

When tested by the Wilcoxon rank-sum test, the distributions of the ME and the CME values for each urban area are found to be statistically different except for the surroundings of small poles (table 6.4). The proposed correction, in other words, is significantly affecting measures for mobility diversity in all urban areas. The largest differences between ME and CME values are observed within the major poles urban area, where ME values are clearly higher than CME values. This is unsurprising given the spatial pattern of c_i, which are highest in city centers (figure 6.3).

Urban Area	ME (mean)	ME (stdv.)	CME (mean)	CME (stdv.)	Wilcoxon rank-sum test (p-value
111-Major pole	0.84	0.13	0.68	0.09	< 0.01
112-Surroundings of a major pole	0.76	0.09	0.69	0.08	< 0.01
211-Medium pole	0.68	0.09	0.62	0.07	< 0.01
212-Surroundings of a medium pole	0.67	0.10	0.66	0.10	< 0.01
221-Small pole	0.62	0.09	0.59	0.08	< 0.01
222-Surroundings of a small pole	0.63	0.09	0.63	0.09	0.69 (non-sign.)
120-Multi-polarized (in large urban area)	0.69	0.09	0.65	0.08	< 0.01
300-Multi-polarized (other)	0.67	0.10	0.66	0.09	< 0.01
400-Isolated	0.62	0.11	0.61	0.10	< 0.01

Table 6.4 Summary statistics of the ME and the CME values per urban area. Differences between ME and CME distribution per urban area are tested by means of the Wilcoxon rank-sum test. Most p-values are significant on 0.01 (***) level

Differences in CME Values between Urban Areas

Investigating the difference in the CME value distributions between urban areas in figure 6.8, shows interesting findings. First, urban areas around major poles depict higher mobility diversity compared to all other urban areas. Multi-polarized municipalities, too, have high CME values, especially when situated near major poles. Isolated municipalities outside the influence of poles have a rather high average CME value, but their wide distribution indicates that the range of different situations that can be encountered in this category. The contrary can be said about the medium poles (and to a smaller extent, the major poles), where a narrow distribution indicates similar observations of mobility diversity can be found in poles all over France.

Focusing on the relation between urban poles and their surroundings in figure 6.8, it becomes apparent that the surrounding areas of major and medium poles have higher CME values compared to their centers. Most remarkably, a clear decrease of CME values with size of the urban pole is found, both for the poles themselves and their surroundings. Larger poles imply a higher mobility diversity both in their surrounding areas and centers, suggesting that there is a potential urban scaling law to be uncovered between CME and urban size (at least when urban size expressed is by employment centers, as is the case for the urban area classification). When tested by the Wilcoxon rank-sum test, most differences in CME distributions between urban areas are found to be significant (table 6.5).

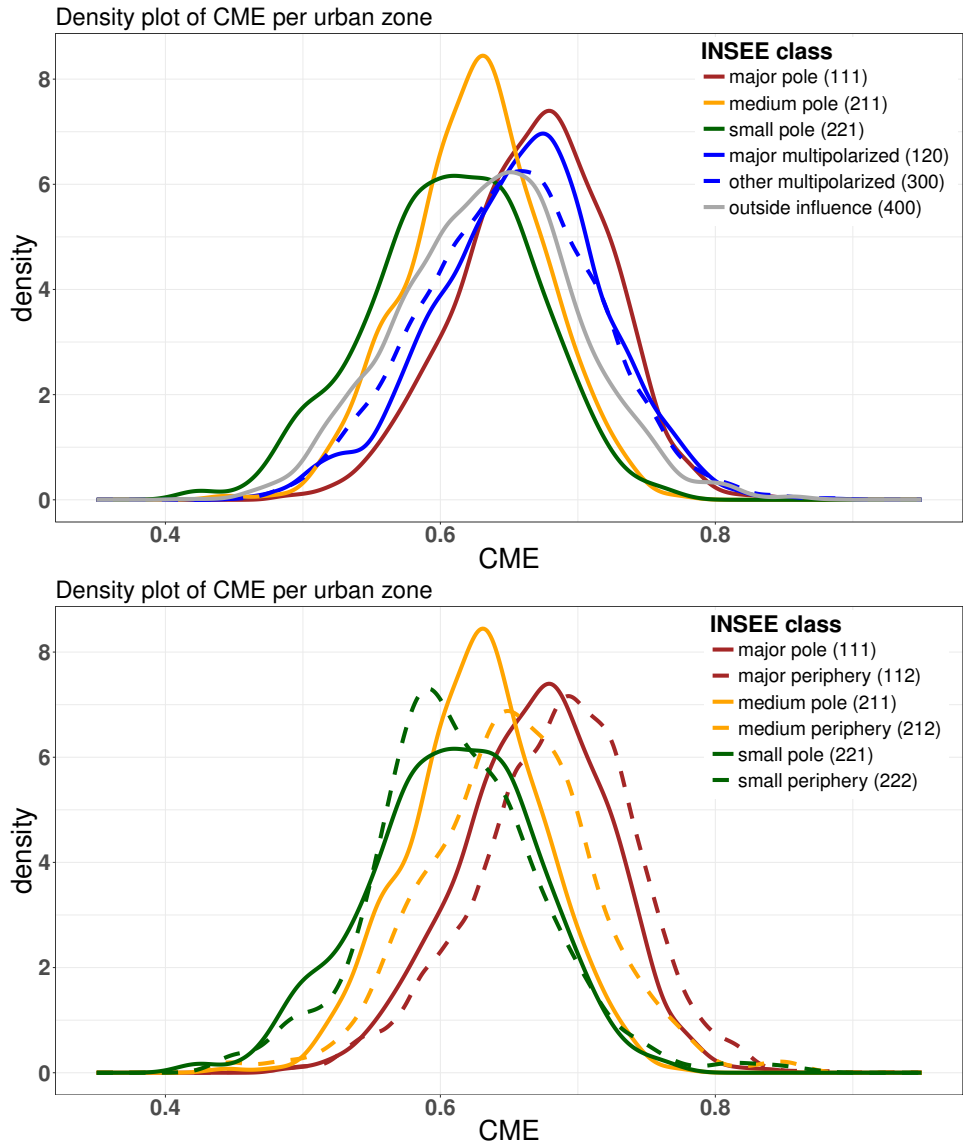


Fig. 6.8 Distributions of the CME values per urban area (a) for different urban areas and (b) for the three main urban poles and their sub-urban areas. Densities are calculated as the density function of average CME values for cell towers in each urban area class. Source: [143]

Urban Areas	111	112	211	212	221	222	120	300
111-Major pole		< 0.01	< 0.01	< 0.01	< 0.01	< 0.02	< 0.01	< 0.01
112-Surroundings of a major pole	< 0.01		< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
211-Medium pole	< 0.01	< 0.01		< 0.01	< 0.01	non-sign.	< 0.01	< 0.01
212-Surroundings of a medium pole	< 0.01	< 0.01	< 0.01		< 0.01	< 0.01	< 0.01	non-sign.
221-Small pole	< 0.01	< 0.01	< 0.01	< 0.01		< 0.01	< 0.01	< 0.01
222-Surroundings of a small pole	< 0.02	< 0.01	non-sign.	< 0.01	< 0.01		< 0.01	< 0.01
120-Multi-polarized (in large urban area)	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01		< 0.01
300-Multi-polarized (other)	< 0.01	< 0.01	< 0.01	non-sign.	< 0.01	< 0.01	< 0.01	
400-Isolated	< 0.01	< 0.01	non-sign.	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

Table 6.5 Significance tests for the pairwise differences in CME distributions between urban areas. Differences are tested by means of the Wilcoxon rank-sum test. The p-values are given on significance level < 0.01 (***), < 0.02 (**), and < 0.05(*).

6.4.4 Relations with Socio-Economic and Environmental Indicators

Single Linear Regressions

Figure 6.9 and table 6.6 show the results of single linear regressions between the ME and CME by a selection of the indicators listed in table 6.3. In general, single linear regressions have very little predictive power (low adjusted R^2-values) and using the CME instead of the ME values does not lead to radical shifts in the obtained regression coefficients. Relations do tend to be less pronounced when using CME, which is not really surprising given the smoothing effect of the correction factor.

What is remarkable though is that by using CME instead of ME, some relations shift significance (table 6.6). When correcting ME to CME, the share of artificial land use in a municipality becomes insignificant to explain mobility diversity. This is unsurprising as the share of artificial land use highly coincides with cell tower density, which was the focus of the correction. The opposite is observed for the share of agricultural land use and for the share of car as a transport mode. Both become significant when correcting to CME. These shifts in significance coincide well with previous findings on the spatial patterns of the ME and the CME (figure 6.7. They imply mobility diversity, when measured by CME, is higher in areas with higher agricultural shares (like rural areas, but also certain sub-urban areas, especially around smaller city centers). When based on ME values, mobility diversity is expected to be higher in areas with higher artificial land use (like larger city centers). With this pattern in mind, it is remarkable that the CME unveils the role of car use as transport mode as a possible explanatory factor for mobility diversity, whereas the ME does not.

	ME			CME		
Indicator	**Regr. coef.**	**Adjusted R^2**	**Signif.**	**Regr. coef.**	**Adjusted R^2**	**Signif.**
Median Income	0.000021	0.19	***	0.000006	0.030	***
Car share	-0.0073	3.5×10^{-5}	0.15	0.031	0.0022	***
Public transport share	0.54	0.12	***	0.068	0.0059	***
Employment in municipality	-0.17	0.076	***	-0.067	0.020	***
Artificial land use	0.2	0.10	***	0.0023	-7.7×10^{-7}	0.37
Agricultural land use	0.02	4.3×10^{-6}	0.28	0.05	0.044	***

Table 6.6 Results of single linear regression between ME, CME and a selection of other indicators. Regression coefficients, adjusted R^2 measures, and significance of the regression coefficient estimate based on p-values are given (< 0.01=***, < 0.02=**, < 0.05=*, or exact p-value).

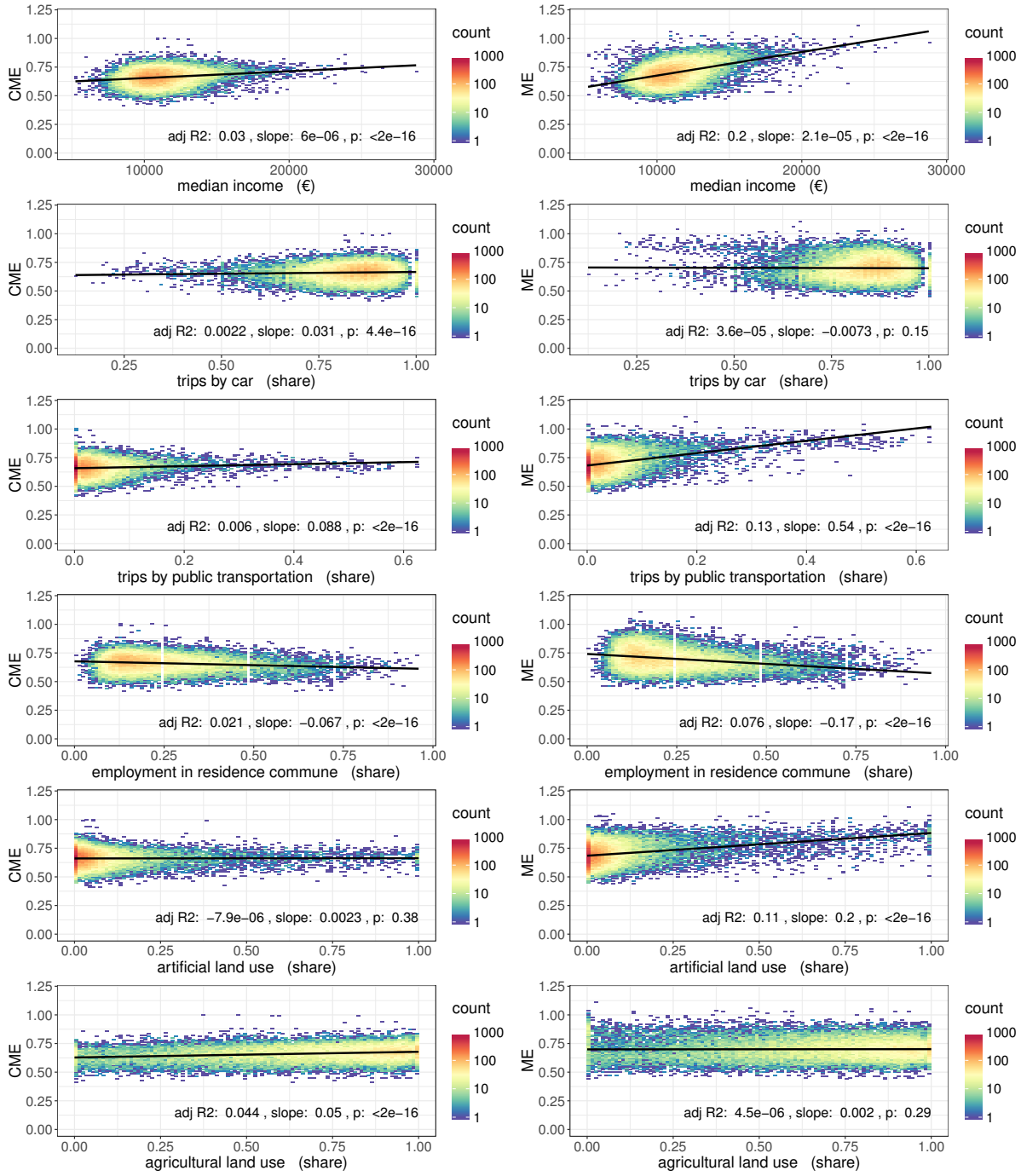


Fig. 6.9 Linear regressions of the ME (a) and the CME (b) by a selection of indicators at municipality level. Each dot represents one municipality in France. Densities of observations are color-coded based on the absolute numbers of observations in the surrounding plot area. Adjusted R^2 measures are given, as well as the p-values and estimates for intercept and slope of the regression. Source: [143]

Multiple Linear Regressions

The insights that can be obtained from simple regressions are rather limited, given their limited predictive power (adjusted R^2 in table 6.6). Therefore, two multiple regression models are explored incorporating all indicators in table 6.3 and their interaction terms as independent variables and either the ME or the CME as dependent variable.

Tables 6.7 and 6.8 show all statistically significant indicators, or interaction terms between indicators, with an estimated regression coefficient larger than 0.1 (or for negative coefficients smaller than -0.1). Remember that, before performing the multiple linear regression, all independent variables were normalized, meaning that the magnitude of the estimated regression coefficients can be compared between independent variables.

Investigating the multiple regression models, a first observation is that the predictive power of the ME model is higher than the CME model (R^2 of 0.40 versus 0.18), even though the same independent variables were used. Regarding the most significant explanatory indicators and their predictive power, the ME model is found to have a set of significant and strongly explanatory variables that, interpretatively, align well with the definitions of larger city centers such as distance to a large city, median income and share of public transport as transport mode (table 6.7). This is a finding that, given the nature of the ME calculations and the cell tower density bias, could be expected. Additionally, there is a second set of variables that are less explanatory and less significant but seem to deal with non-urban environments. It concerns locations with higher shares of forest or agricultural land, often in combination with an interaction term of remoteness (city distance).

The multiple regression model predicting CME gives rather different results (table 6.8). Although geographical elements (distance to large city, elevation) remain more or less similar compared to the ME model, in the CME model the median income factor shares its predictive force with the share of workforce employed in the municipality and the demographic indicator of active workers in the population gains importance in the model. More remarkable is that the share of public transport as travel mode has become insignificant and there is the appearance of the strong and significant factor of the share of artificial land use. The role of the artificial land use share is puzzling, as it has a large negative regression coefficient, which is not entirely expected given the non-existent relation in the simple linear regression discussed before. Its role, however, is partly balanced out by a strong positive interaction effect between artificial land use and the share of cars as transport mode. The interaction terms between share of cars and other available land-use classes also have a prominent role in the model, rendering the combination of car and land use one of the cornerstones for the interpretation of nation-wide CME patterns.

Reflecting on the lower predictive power of the model for CME ($R^2 = 0.18$) compared to the model for ME ($R^2 = 0.40$), one could argue that the latter's focus on typically (large) urban characteristics probably boosts its predictive power, especially since a large share of France municipalities are classified under the influence of major poles or within a large urbanized area (around 65%, see table 6.2). Supporting a more diverse set of municipalities and their related characteristics (e.g. the differences in use of public transport in different urban areas), the CME model is probably not capable of mining this quick win in predictive force. This, however, does not mean the CME model is less useful for research. After all, the model indicates that some general tendencies can be derived but that local situations are probably far more complex and in need of further in-depth study.

	ME ($R^2 = 0.404$, R^2-adj. $= 0.401$)		
Indicator or interaction term	**Coeff. est.**	**t-statistic**	**Pr(>|t|)**
Median Income	0.23	34.09	< 0.001(***)
Public transport share	0.20	20.08	< 0.001(***)
City distance	-0.20	-31.40	< 0.001(***)
Active population	0.17	29.86	< 0.001(***)
Mean elevation	-0.14	-15.36	< 0.001(***)
Natural land use	-0.11	-2.42	0.015(**)
Natural land use : city distance	0.11	2.33	0.020(**)
Agricultural land use : city distance	0.11	2.25	0.025(*)
Natural land use : Car share	0.11	2.06	0.040(*)

Table 6.7 Coefficients of most contributing indicators or interaction terms in a multiple linear regression between ME and a selection of indicators. Regression coefficients and related p-values (rounded at 3 decimals) based on the t-statistic are given with significance levels < 0.01 (***), < 0.02 (**), and < 0.05(*). Interaction terms between indicators *A* and *B* are notated as *A:B*. Indicators and interaction terms were selected by an estimated coefficient with absolute value higher than 0.1 and a p-value smaller than 0.05. Note that all indicators were normalized before performing the multiple regression. Indicators are sorted in descending order by significance level and absolute value of the coefficient estimate.

	CME ($R^2 = 0.189$, R^2-adj. $= 0.185$)		
Indicator or interaction term	**Coeff. est.**	**t-statistic**	**Pr(>|t|)**
Active population	0.20	30.09	< 0.001(***)
Artificial land use	-0.20	-5.98	< 0.001(***)
Mean elevation	-0.16	-14.55	< 0.001(***)
City distance	-0.13	-17.48	< 0.001(***)
Median income	0.12	15.53	< 0.001(***)
Employed in municipality	0.12	12.11	< 0.001(***)
Artificial land use : car share	0.12	2.97	0.003(***)
Natural land use : car share	0.16	2.49	0.013(**)
Agricultural land use : car share	0.15	2.21	0.027(*)

Table 6.8 Coefficients of most contributing indicators or interaction terms in a multiple linear regression between the CME and a selection of indicators. The setup of the table is equal to table 6.7.

6.5 Discussion

6.5.1 Plausibility of the CME Indicator

In this chapter, it is shown that the traditional mobility entropy indicator (ME), when calculated for CDR data, is not independent from cell tower density. Consequently, it is argued that the ME indicator from CDR data cannot be used for objective comparison between areas with different cell tower densities. As a solution, the Corrected Mobility Entropy (CME) has been proposed. One main point of discussion regarding the CME, and a point that can equally be raised for the ME or any other mobile phone indicator, is that it is impossible to validate its nation-wide patterns given the absence of individual ground truth data at such a large scale. Still, two main arguments can be put forward to support the credibility of the proposed correction .

First, for the case of CDR data in France, findings show the CME to be less correlated to cell tower density (Pearson's R of -0.17 in figure 6.5) compared to the ME (Pearson's R of -0.59 in figure 6.5), hence indicating the effectiveness of the correction. One important aspect in this regard is the effect of the scaling range (a,b) associated with the correction factor c_i. As shown in table 6.1, correlations between CME and cell tower density are sensitive to this parameter choice. Sensitivity tests also suggest that, potentially, even lower correlations could be obtained when scaling range parameters would be defined with a higher precision. To do so, however, a manifold of sensitivity tests are required, which becomes very time-costly rather quickly.

Second, when comparing between different regions, CME values are found to be rather interpretable. They form coherent spatial patterns (figure 6.7), have significant, and logical, relationships with other indicators derived from census data or land use maps (tables 6.6 , 6.8 and figure 6.9), and they depict significantly different distributions for different urban areas (table 6.5 and figure 6.8). All of this leads to an increased credibility of the CME, even if direct validation data is absent.

6.5.2 A Changing View on Mobility Diversity in France

Investigating mobility diversity based on the CME renders different insights compared to using the ME. One main difference is in the obtained spatial pattern, as is evident in figure 6.7, or in the differences in the distributions of the ME and CME values for different urban areas in table 6.4. Compared to the ME values, which imply mobility diversity to be highest in areas with high cell tower density such as city centers, the spatial pattern of the CME values appropriates sub-urban areas, especially from middle-sized cities (such as Rennes, Toulouse, or Reims), as locations with the highest diversity.

The relation between the CME and the French urban system becomes clearer when investigating the distributions of the CME values in different urban areas. Surprisingly, significantly different distributions of the CME are found between (almost) all urban areas (table 6.5). The overall pattern shown in figure 6.8 is that mobility diversity, when measured by the CME, is significantly higher in surrounding areas compared to their related urban poles and that mobility diversity, in both urban and sub-urban regions, increases considerably with urban center sizes, suggesting the existence of a scaling law for mobility diversity in France.

Interestingly, correcting the ME to the CME does not result in fundamentally changing relations with other indicators, at least when the relation is investigated by simple linear regressions (table 6.6 and figure 6.9). This is true for most indicators but not for the share of cars in trips, artificial land use and agricultural land use. With regard to the share of cars in trips, the change in the relation with mobility diversity when correcting the ME is interesting. While the relation between the share of cars in trips and ME is insignificant, the relation with CME becomes significant (table 6.6). Given the important role of cars in terms of accessibility and individual mobility, this is an important finding that seems to endorse the argument for CME.

Combining insights from the spatial patterns with insights from relations to other indicators seems to confirm that the CME appropriates sub-urban zones as the locations with the highest mobility diversity. This is in contrast with the ME, which indicates mobility diversity to be highest in city centers. The changing relation with artificial and agricultural land use between the ME and the CME in the single linear regressions in table 6.6 already suggested such an interpretation.

The results of both multiple regressions models in tables 6.7 and 6.8, add more nuance to the interpretation of ME and CME patterns. The multiple regression model with the ME as dependent variable can be interpreted as operating in an urban-remote areas dichotomy, with the most important contributing indicators (or interaction terms) enabling one to distinguish between city centers and remote areas either by economic (median income, share of active population, and share of public transport as transport mode) or geographic factors (city distance, elevation, land use) (table 6.7). In contrast, the multiple regression model that tries to predict CME values, although less explanatory in its whole, does not focus on this urban-remote dichotomy. It sketches a more complex process that encompasses interactions between employment centers, distances to cities, demographics, incomes, land use, and the role of cars as transport mode as its main explanatory dimensions (table 6.8). According to the model, higher CME values can be expected in areas that are economically active (higher active population, higher median income, higher employment within the municipality), where the share of car use for trips is higher and independently from which land use is predominant (although interaction terms strongly suggest there to be different regimes for different land use mixes), and that are situated closer to large cities, typically on lower elevations, and with a smaller share of artificial land use. It is reassuring, that such a description matches well with what can be expected from sub-urban areas, especially around medium and large cities, which are also the locations of hot-spots for CME values observed in figure 6.7.

6.5.3 Wider Relevance

The main finding of this chapter is that the ME indicator from CDR data cannot be used for objective comparison between areas that have different densities of cell towers but the relevance of this finding goes further than CDR data only. Other datasets that collect data on human movement for large areas might be equally prone to this bias if they have observation points that are unevenly distributed over a territory. Check-in data from location based social networks or services form an example, as do credit card datasets.

In this regard, one important question is if the proposed correction for CDR data can be translated to other datasets. The underlying assumption when constructing the correction factor for CDR data is that mobile phone use (calling/texting) is independent from the (passive) location recording. This assumption does not hold for location-based social media or online social networks where location recording is an active act embedded in the mediated use of the service. As such, for these applications the heterogeneous possibility of being detected in a location is not only given by a spatial, infrastructural element (the density of possible check-in locations in an area) but also, and possibly even more so, by the mediation of these locations within the application (the 'attractivity' of sharing this location within the context of the application). This extra layer of heterogeneity most probably needs to be accounted for when constructing a correction factor for such datasets.

On a more general note, the presented study forms a good reminder that even though current data coverage might be large-scale, this does not automatically imply that analytics are objective at such a scale. Biases occur in different ways and at different scales, and might be due to, amongst others, changing context and user groups, non-objective data collection methods, or unthoughtful methodology. As a consequence, critical evaluation remains crucial when performing empirical studies on large-scale data.

One problem that remains is the absence of proper validation data that, even if only to a small degree, could match the extent and diversity of the newly gathered data sources. Obviously, this absence strongly complicates the validation of methods but it also discourages efforts to challenge existing practices given that neither the claimer nor the challenger has solid ground to prove its *trustworthiness*. The situation of this study is similar. Through publications (e.g.[126, 95]) and implementation in software packages like bandicoot.py [48], the use of the biased mobility entropy indicator is slowly becoming institutionalized.

Any attempt to challenge this development, such as the argument that correction of ME is necessary (whether or not in the form of CME), will face critique that it cannot provide validation to prove that it is outperforming the established indicator which in itself has not been validated but is merely supported 'by literature'. Nevertheless, findings suggest that the proposed corrected mobility entropy (CME) results in a more reliable indicator when it comes to comparing the diversity of human movement between large-scale regions based on CDR data. Using CME enables a better description, understanding, and delineation of regions, or urban areas, with respect to individual mobility. It is therefore relevant for planning and policy (especially from the perspective of urban development) and has clear applications in official statistics, urban planning, and mobility research.

Chapter 7

Scaling Relations of Mobile Phone Indicators

From the landscape: a sense of scale.
From the dead: a sense of scale.

Richard Siken

Abstract

This chapter studies urban scaling laws of mobile phone indicators. This way, it investigates whether mobile phone indicators show consistent patterns across different cities in France and to which degree they might be influenced by magnitude of populations. Secondly, it investigates whether the found empirical scaling laws are sensitive to the used city definition by simulating 4914 city definitions based on different combinations of a density of population threshold, a threshold on the percentage of commuters and a population threshold. Lastly, this chapter investigates whether correlations between mobile phone indicators and census data on income, such as the ones presented in chapter 3 (figure 3.11) are influenced by city definitions. The presented results are the first to systematically investigate the influence of city definition on urban scaling laws of mobile phone indicators and on the correlation between mobile phone indicators and income indicators. Regarding the latter, the found sensitivity to city definitions challenges existing work in literature that use predictive models of mobile phone indicators based on one city definition only.

Related publications and Acknowledgments

- The structure and content of this chapter is based on the paper in preparation: C Cottineau, M Vanhoof, E Arcaute. *Urban scaling laws of mobile phone indicators and their relation to census data*. For this publication, the research design, the preparation of the mobile phone data, the analysis of the scaling results, the analysis of the correlation results, the creation of the figures, the discussion of the results and large parts of the writing were all done by the PhD-candidate in discussion with Dr. Clementine Cottineau, Dr. Clement Lee and Prof. Dr. Elsa Arcaute.

- Acknowledgments go to Dr. Clementine Cottineau who performed the preparation of the Gini and Segregation indices based on [41] as well as the simulation of city definitions based on [42] and the subsequent calculations of scaling laws and correlations.

7.1 Outstanding Knowledge Gaps for Mobile Phone Indicators

While the spatial knowledge gap with regard to mobile phone indicators can be illustrated by studies that reveal spatial uncertainties (chapter 4) or spatial bias (chapter 6), it can also be illustrated by what has not yet been studied or understood. This chapter explores three spatial aspects of mobile phone indicators that have not yet been investigated: their urban scaling laws, the sensitivity of scaling laws to city definitions, and the sensitivity of the relations between mobile phone indicators and income indicators to city definitions.

7.1.1 Urban Scaling Laws

A first investigation is on the urban scaling laws of mobile phone indicators. As discussed in section 2.6.4, urban scaling laws summarize the variation of an indicator with city size and can be estimated by:

$$Indicator_i = \alpha \times Pop_i^{\gamma} + \varepsilon, \tag{7.1}$$

where $Indicator_i$ is an indicator value for a city i in a set of cities I, α is a normalization constant, γ is the scaling exponent, Pop_i is measure of size of a city i, often expressed in population or population density, and ε is an error term. Remark that the notation Pop_i, instead of e.g. $Size_i$, is used to avoid confusion with city definitions later on in the chapter.

When the considered indicator is an absolute quantity such as, for example, the number of calls, the obtained scaling exponent γ is considered in relation to 1. When $\gamma \approx 1$, the growth of indicator values scales proportional to city population. When $\gamma < 1$, the scaling law reveals a sublinear regime, meaning that there is a non-linear growth of the indicator with city size, in this case the indicator values grow slower than the city population. When $\gamma > 1$ a superlinear regime is observed, again indication non-linear growth of an indicator with city size, only this time with the indicator values growing faster than the city population.

When the considered indicator is not an absolute quantity but an index, or a per-capita indicator, similar interpretations for the scaling exponent γ apply but this time with reference to 0. After all, when an indicator Y scales with city size $Y \sim cPop^{\gamma}$ than the per capita indicator $\frac{Y}{Pop} \equiv y$ should scale by $y \sim cPop^{\gamma-1}$ [118]. Simply defining $\gamma - 1$ by β for the case of per-capita indicators means that, $\beta \approx 0$ corresponds to the absence of significant size effects or thus a linear regime. $\beta < 0$ indicates sublinear regimes characterized by decreasing per-capita indicators with city size, and $\beta > 0$ indicates superlinear regimes that are characterized by increasing per-capita indicators with city size.

Estimates of β are easy to calculate as scaling laws can by transformed to a single regression model:

$$log(Indicator_i) = \beta \times log(Pop_i) + \delta + \varepsilon, \tag{7.2}$$

with δ being an normalization constant and ε an error term. The statistical fitness of the regression model can be calculated by means of the R^2 values which, in the case of single linear regressions, are equal to the square of the Pearson correlation coefficient between the indicator values and the population values of all different cities i in a set of cities I. Note, in this perspective, that literature observes scaling laws of per-capita indicators to produce much lower R^2 values compared to absolute quantities [118, 41].

Following equation 7.2, the interpretation of β values can be derived. Doubling city populations will lead to an increase of 2^β in the per-capita indicator value. For example, imagine that the number of hospitals per capita are found to have a scaling indicator β of 0.05. This means that, when doubling city populations, the number of hospitals per capita is expected to increase by factor $2^{0.05} = 1.035$ or thus by 3.5%. Table 7.1 displays the increase factors when doubling population size for a common range of β values.

Scaling indicator (β)	Increase factor	Increase (in%)	Scaling regime
-0.1	0.933	-6.7	Sublinear
-0.05	0.966	-3.4	Sublinear
-0.01	0.993	-0.7	(Sub)Linear
0	1	0	Linear
0.01	1.007	0.7	(Super)Linear
0.05	1.035	3.5	Superlinear
0.1	1.072	7.2	Superlinear

Table 7.1 Increase factors when doubling population size for per-capita indicators given a range of β values.

Although scaling laws for census indicators have been widely studied [103, 77, 22, 10], little work has yet evaluated the scaling laws of mobile phone indicators. Studying scaling laws of mobile phone indicators aims at answering the question whether consistent patterns of human behavior can be captured by CDR data across different cities and for different city sizes. To answer such questions for France, scaling laws are calculated for the same mobile phone indicators used in chapter 3. Their definitions can be found in table 3.5 in section 3.2.1. Indicators are calculated for all users in the French CDR dataset for September and territorial aggregation is performed by the distinct days algorithm defined in chapter 4 (section 4.1.1). Testing the effect of time period choice against the month June (the other, entirely captured, non-summer month), and testing the HDA choice against the maximum amount of activities algorithm, no substantial differences between results are found.

To define a measure of city size for the urban scaling law, one could use the total population count provided by census data or the Orange user count as resulting from the distinct days HDA. Both were tested but no substantial difference was found, probably because the Orange user count scales linearly with total population (indicating a bias in the share of Orange clients by city size). The following analyses will use the total population from census data given that it is a more generic reference in urban scaling literature.

7.1.2 Urban Scaling Laws and City Definitions

Building upon recent empirical work that revealed the influence of city definition when assessing scaling laws [10, 42], a second outstanding question is to which degree do city definitions influence empirical scaling laws from mobile phone indicators. Supposing that mobile phone indicators depict different types of spatial variation compared to other indicators the sensitivity of their scaling laws to city definitions might be fundamentally different and thus worth considering. One could imagine, for example, that calling patterns might be less influenced by infrastructural elements and built-up environment and therefore be more homogeneously spread across the country than, for instance, the number of hospitals which, in France, are found to be sublinear (see figure 2.12 in section 2.6.4).

Simulating City Definitions

To investigate the effect of city definition on empirical scaling laws, a clear way to simulate city definitions is needed. Following previous work on simulating city definitions in France [10, 42], different definitions are set by combining three parameters: a density threshold to define a center of the city (the density threshold: d), a minimum percentage of workers in a commune commuting to this city center (the flow threshold: f) and a minimum population within the defined city (the population threshold: p). More practically, French communes (municipalities) are aggregated into a city i following a city definition I, representing the set of all cities i that can be defined by the parameter combination of the density threshold d, the flow threshold f and the population threshold p. Note that depending on the threshold combinations used for the city definition, the amount of cities in the French territory and their constitution in terms of municipalities will differ, as can be observed in figure 7.1.

Equal to [10, 42], the city definitions that will be simulated are all combinations of the thresholds, which are parametrized between:

- 1,000 and 20,000 inhabitants/ha in steps of 5,000 for d
- 0 and 100 percent of the active population in steps of 5 for f
- 0 and 50,000 inhabitants in steps of 10,000 for p

In total 4914 city definitions are simulated (39 density thresholds $\times$ 21 flow thresholds $\times$ 6 population thresholds). A selection of city definitions are shown in figure 7.1. Note that in this figure the population threshold (p) is fixed to 0.

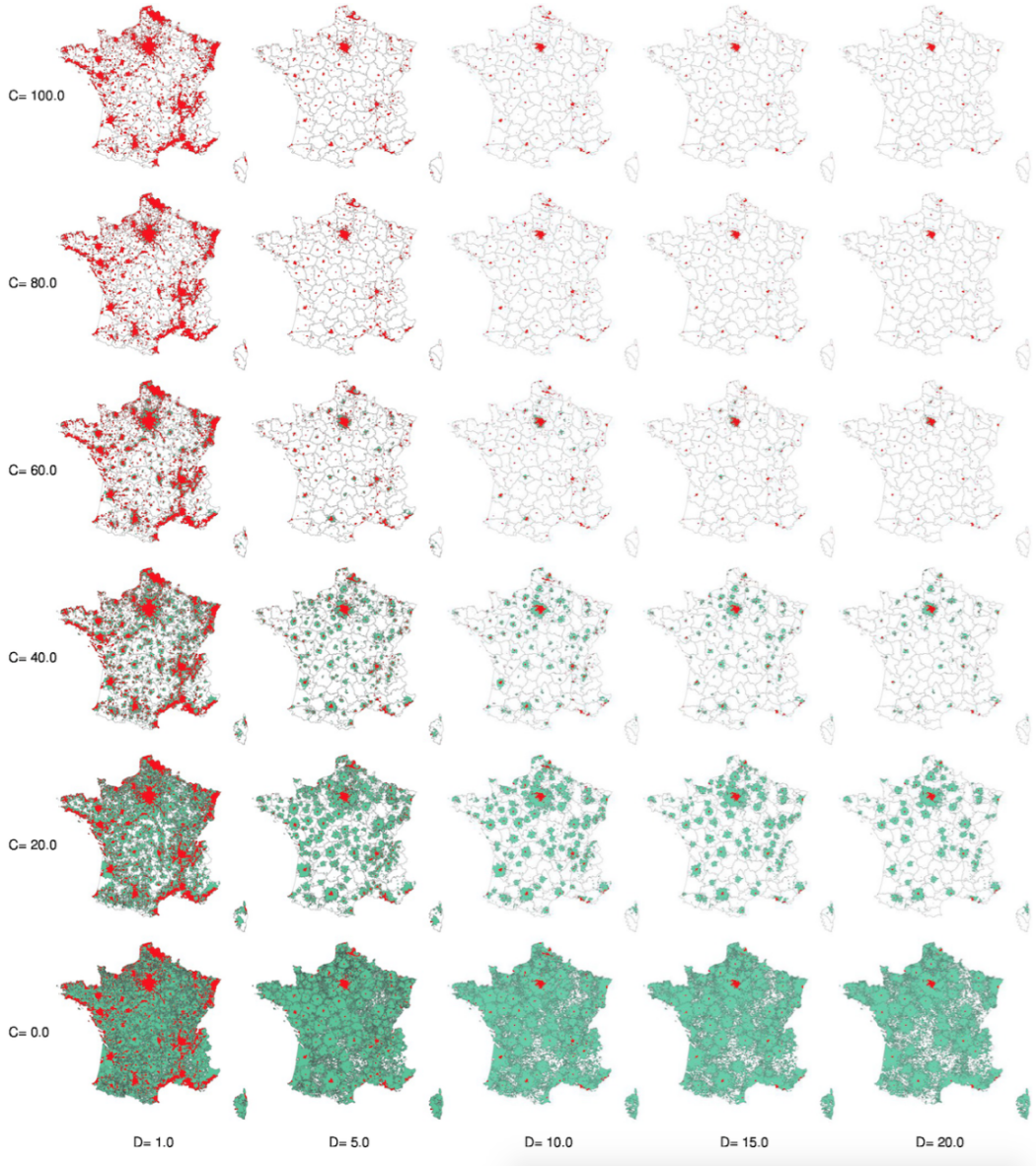


Fig. 7.1 A range of city definitions for different density and flow thresholds. D is the minimum density of residents per hectare that defines the urban centers (in red). C is the share of commuters (in %) living in the periphery and working in the density-based urban clusters (in green). P is the population minimum, and is set to 0. Note that D,C and P accord to, respectively, d, f and p in our analyses but differ merely in notation. Aggregation is performed using the 2013 GeoFla geometry of communes. Source: [42].

Calculating Scaling Laws per City Definition

For each of the 4914 city definitions I, scaling laws of mobile phone indicators can be estimated by the linear regression with population counts following equation 7.2:

$$log(Indicator_i) = \beta \times log(Pop_i) + \delta_I + \varepsilon \qquad with\, i \in I_{d,f,p}, \tag{7.3}$$

with i being one city in a city definition I, δ_I a normalization constant and ε an error term. Note that the parameter β is interpreted with reference to 0 because aggregation of mobile phone indicators from commune level to a specific city definition I is done by taking the weighted average of commune values with respect to their populations. As such, the aggregated average indicators at city level can be considered as a per-capita quantity. Aggregation of population counts from commune level to a city definition is done by summing population counts.

7.1.3 Relations between Indicators and Sensitivity to City Definitions

Related to the second question is an outstanding question that is even more relevant: to which degree are relations between mobile phone indicators and census data dependent upon city definitions. As discussed before (e.g. sections 2.6.4, 3.4, or 6.4.4), multiple relations between mobile phone indicators and census data have been revealed, but all of them do so based on one aggregation level, one city definition only. If, however, relations between indicators are sensitive to city definitions, then this would have considerable implications for their validity and interpretation. One implication is that it would force research to focus on the spatiality of found relations, reaching for an explanation why relations shift with city definitions and an understanding of the real-life processes that produce these relations. The analysis in this chapter will limit itself to a methodological discussion only, meaning that found scaling laws, relations, and their sensitivity to city definition will be merely revealed, not interpreted. Nevertheless, future research would undoubtedly profit from a more interpretative analysis.

Income Inequality, Segregation, and Deprivation

To investigate the sensitivity of relations between mobile phone and other indicators, three indicators based on census data are considered, each describing a different aspect of income: inequality, segregation and deprivation. The indicators are iteratively calculated for each city under the different city definitions.

The first two indicators, the inequality and segregation of incomes, are based on data from the public database CLAP[1], which provides total wages and and number of employees of firms in France aggregated at municipality level for the year 2008. Following [41], the Gini index [60] of the wages in the CLAP database is used as the inequality indicator:

$$G_i = \frac{\sum_{a=1}^{n_i} \sum_{b=1}^{n_i} |x_a - x_b|}{2n \sum_{a=1}^{n_i} x_a}, \tag{7.4}$$

where G_i is the Gini index of a city i, x_a is the wage of one employee, x_b is the wage of one employee, and n_i are the number of employees in a city i according to the CLAP database.

A segregation indicator can also be derived from the CLAP database. Again following [41], the ordinal variation ratio index (R^o) is calculated, expressing the proportion of the total variation in wages for a city that is due to differences in the composition of different wage categories between different municipalities of that city. Following [109], the ordinal variation ratio index can be defined as:

$$R_i^0 = \sum_{m=1}^{M_i} \frac{t_m}{T_v}(v - v_m) \qquad \text{and} \qquad v = \frac{1}{K-1} \sum_{j=1}^{K-1} 4c_j(1 - c_j), \tag{7.5}$$

where R_i^o is the ordinal variation index of the wages in a city i, m indexes the set of different municipalities M in a single city i, t_m is the population count in municipality m, v is a measure of variation of the wages, T_v is is a measure of variation in the wages in the entire population in a city, v_m is is a measure of variation of the wages in municipality m, j indexes the different categories in K the wages can belong to, and c_j is the cumulative proportion of the sample with wages in category j, expressing the distribution of wages in that category. Note that in this definition, the variation is maximum when half the population has wages in category $k = 1$ and half the population has wages in category $k = K$). Variation is at its minimum, when all wages are in category c with $c \in (1, 2, ...K)$. Measuring ordinal variation then amounts to measuring how close the observed distribution is to the minimum and maximum variation states [109].

A third indicator is the European Deprivation Index (EDI) [102], which was also used in chapter 3 (section 3.4). The EDI is an individual deprivation indicator constructed from an European survey specifically designed to study deprivation [102, p. 982]. It is created as a composite measure incorporating information on both subjective and objective poverty and the attribution of the weights for different contributing factors is done specifically for France [102]. As such, the EDI forms a general measure for deprivation, which is interesting for comparison with mobile phone indicators as, for example, has been done by [98]. The lower the EDI measure, the less deprived a city is and thus the better the socio-economic situation.

[1] https://www.insee.fr/fr/metadonnees/definition/c1232

Sensitivity of Relations between Mobile Phone Indicators and Income measures to City Definitions

Similar to the investigation on the sensitivity of scaling laws to city definition (section 7.1.2), the relations between mobile phone indicators and income indicators are calculated for each of the 4914 city definitions. Relations between indicators are expressed by means of the Pearson Correlation Coefficient (Pearson's R):

$$Pearson's\,R(\vec{x},\vec{y}) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}} \qquad with\, i \in I_{d,f,p}, \tag{7.6}$$

where $\vec{x}$ and $\vec{y}$ are vectors with, respectively, mobile phone indicator values and income indicator values for all cities i defined by a city definition I. Note that, depending on the city definition I, the amount of defined cities i varies.

7.2 Scaling Laws of Mobile Phone Indicators in France

7.2.1 Significance of Scaling Laws

Investigating the scaling laws of mobile phone indicators for all 4914 city definitions, a first observation is that the majority of mobile phone indicators do not depict significant scaling laws. The R^2 found never exceed 60%, regardless of whether scaling to be sublinear, linear, or superlinear. Low R^2 values for per-capita indicators are not uncommon in literature [118, 41]. For example the Gini index that will be used in further analysis has R^2 values that do not exceed 40% [41]. Still, R^2 values lower than 20% depict very little significance and were found for most mobile phone indicators. Figure 7.2 exemplifies the limited significance for two indicators: the percentage of nocturnal calls and the median duration of calls. While the former shows a superlinear regime for most city definitions, the latter shows a linear regime for most city definitions, but both have extremely low R^2 values for the obtained scaling laws.

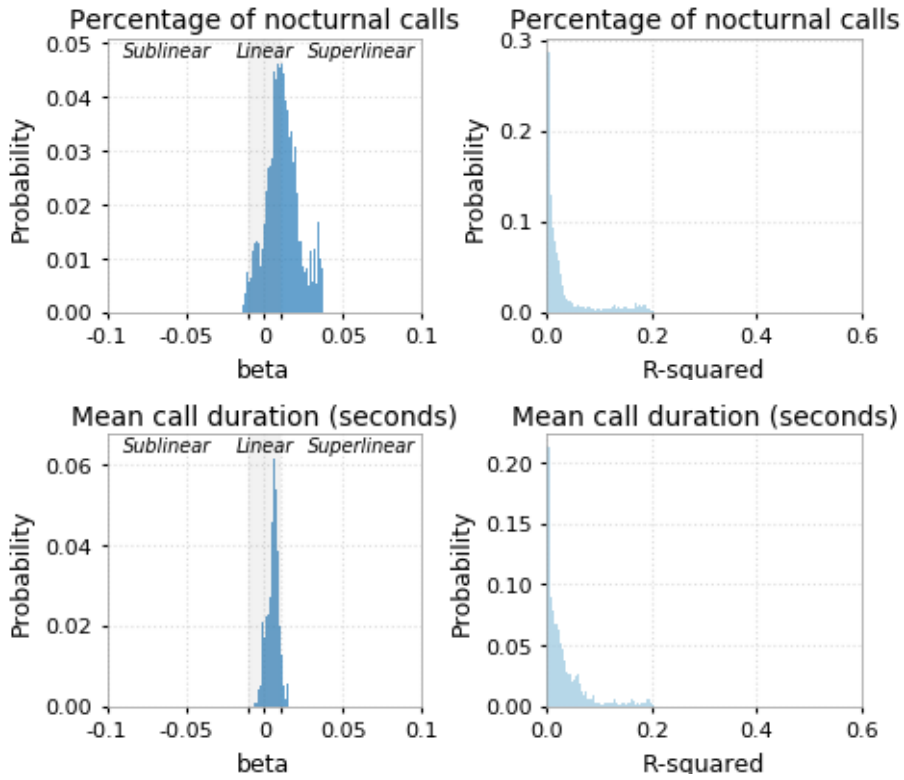


Fig. 7.2 Distributions of β (left) and R^2 (right) for the urban scaling laws of the percentage of nocturnal calls (top) and the mean duration of calls (bottom). The distributions are represented by histograms of 100 bins. The maximum value of the x-axis is defined by the highest observed R^2 value for all mobile phone indicators.

7.2.2 Superlinear Scaling Regimes

Three mobile phone indicators are found to be superlinear with reasonable significance: the average entropy of places, the average number of places visited and, although to a lesser extent, the average number of contacts. Figure 7.3 shows the β and R^2 values for scaling laws of all city definitions. Superlinear regimes suggest that the number of visited cell towers, the diversity of mobility, and the number of contacts increases with city size, across different cities in France, and independent from city definitions.

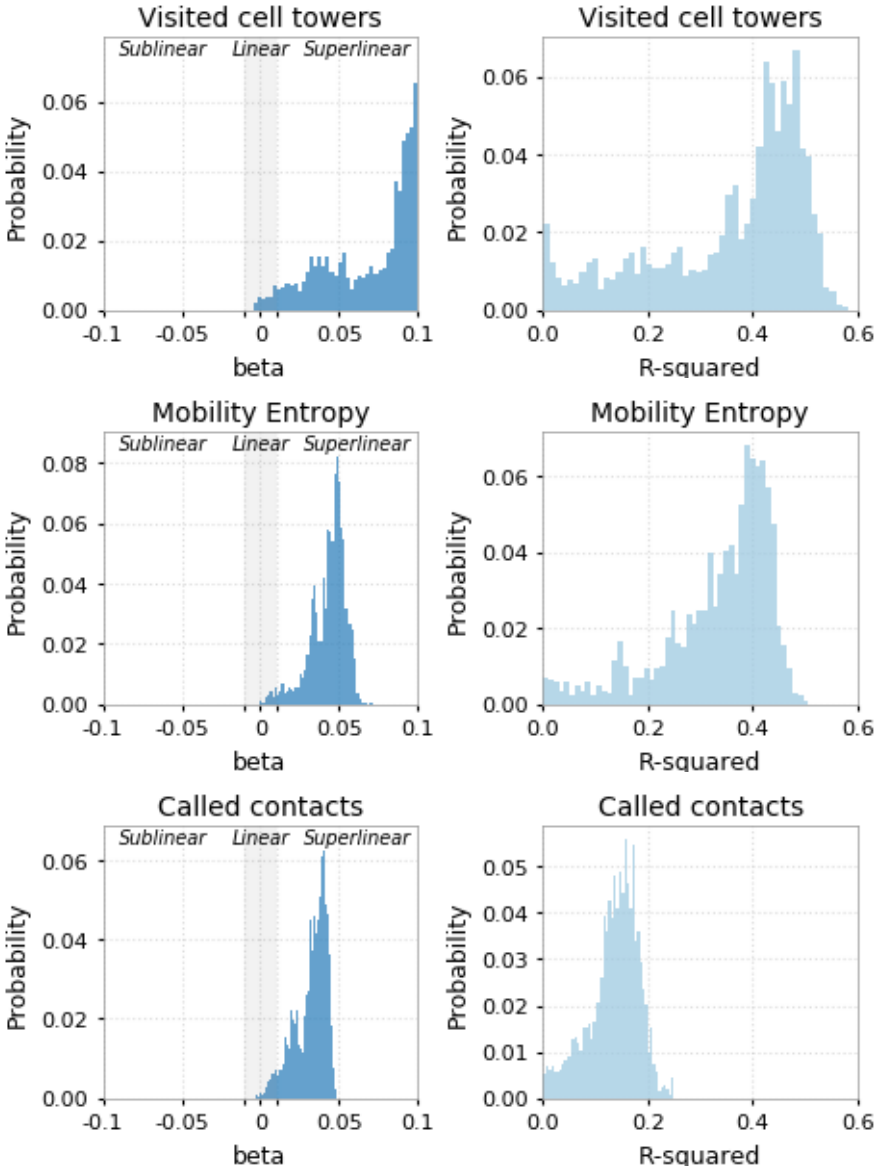


Fig. 7.3 Distributions of β (left) and R^2 (right) for the urban scaling laws of the number of visited cell towers (top), the mobility entropy (middle) and the number of contacts (bottom). Figure setup is equal to figure 7.2.

Superlinear scaling for the number of visited cell towers and mobility entropy in figure 7.3 is not surprising when remembering, respectively, the spatial distribution of cell towers in France (figure 3.1) and the bias of the mobility entropy to cell tower density as discussed in chapter 6 (figure 6.7). The superlinear regime of the number of contacts is an interesting finding as it suggests that users have more contacts in bigger cities. This relation accords well with the spatial distribution of the number of contacts indicator in figure 3.9. Additionally, a similar finding has been revealed in Portugal where the number of calls is found to be superlinear with high significance for three city definitions[114]. In the French case, significance for most of the 4914 city definitions is found to be moderate, at best, with R^2 values reaching a maximum of 0.2.

7.2.3 Sublinear Scaling Regimes

Three mobile phone indicators are found to have sublinear scaling regimes with reasonable significance: the mean, median and standard deviation of the time between mobile phone events. Results in figure 7.4 suggest that the time periods between mobile phone events become shorter with increasing city size or thus that the frequency of calls increases with city size. This seems a reasonable finding for mobile phone usage in 2007 when mobile phones were adapted but probably to a different degrees in urbanized regions.

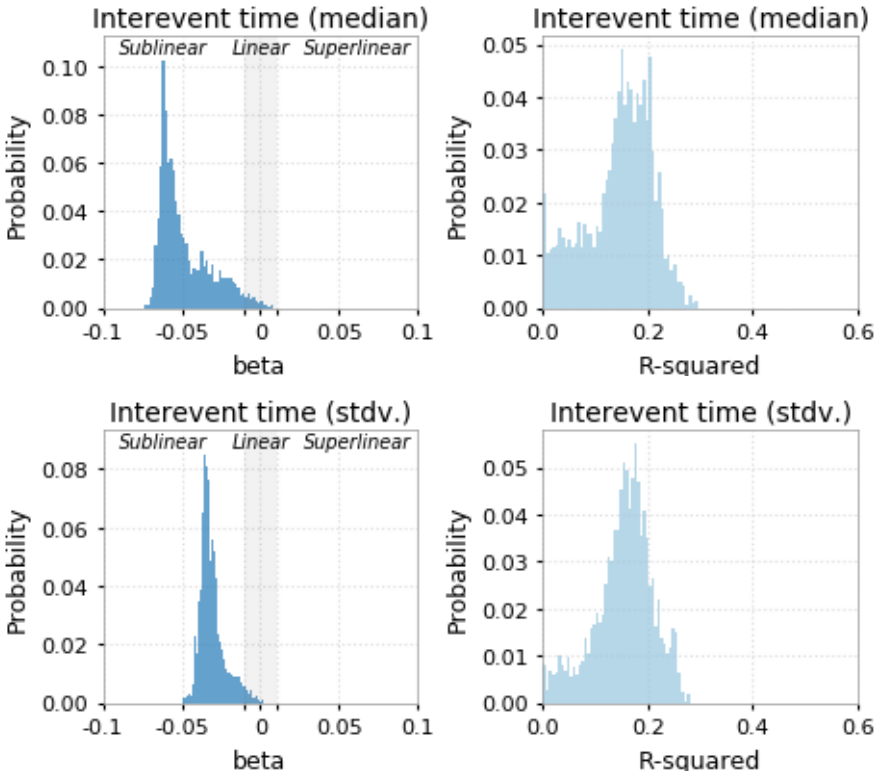


Fig. 7.4 Distributions of β (left) and R^2 (right) for the urban scaling laws of the median (top) and standard deviation (bottom) of the duration between mobile phone events. Figure setup is equal to figure 7.2.

7.2.4 Changing Scaling Regimes

As can be observed in figures 7.2, 7.3, and 7.4, city definitions do not substantially affect scaling laws for most mobile phone indicators. In other words, for most mobile phone indicators, the found scaling regimes are similar for most, if not all, of the 4914 city definitions. This is not true for indicators related to home detection, such as the spatial uncertainty at the L1 location, or the distance between L1 and L2 locations. As shown in figure 7.5, scaling laws for these indicators are sublinear when cities are defined compactly (low cut-offs on the population (p) and population density (d) thresholds), and superlinear when cities are defined more loosely. Although the scaling laws in figure 7.5 are insignificant, the change of regimes with different city definitions speaks to the idea that uncertainties on home detection depict more complex patterns than a simple relation to city size.

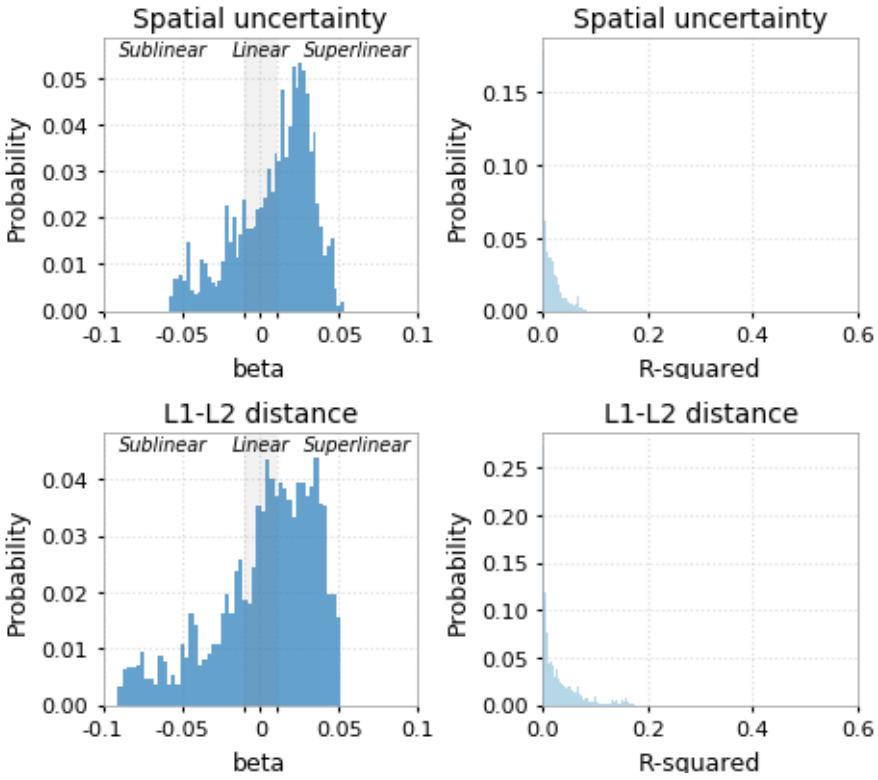


Fig. 7.5 Distributions of β (left) and R^2 (right) for the urban scaling laws of spatial uncertainty of the L1 home detection (top) and the distance between L1 and L2 for the home detection (bottom). Figure setup is equal to figure 7.2.

7.3 Relations between Mobile Phone Indicators and Income Measures

One outstanding question is how stable relations between mobile phone indicators and (the three) income indicators would be with regard to different city definitions? Are relations independent from city definition, as is implicitly assumed in analyses such as the ones in chapter 3 (figure 3.11)? Or are they more unpredictable and less reliable than previously thought?

7.3.1 Relations for All City Definitions

EDI vs. Mobile Phone Indicators

The answer to the previous question is that it depends on the combination of mobile phone indicator and census variable. Figure 7.6 shows the relation between EDI and a selection of mobile phone indicators by plotting the histogram of the resulting correlation coefficient for each of the 4914 city definitions.

The relation between EDI and some indicators, such as the interactions per contact or the amount of calls at home, remains similar for all city definitions (positive in these cases), although a rather wide range of significance is observed. For most relations between EDI and mobile phone indicators, however, directions of the relation can change with city definitions. In these cases, correlations are thus either positive or negative, depending on the used city definition.

One prominent example is the relation between EDI and the radius of gyration. Here, a similar amount of city definitions lead to either positive or negative relations, some of them with rather high correlation coefficients. In this case, observed correlations are thus extremely sensitive to city definition rendering them rather unreliable to use even though the correlations themselves might be high.

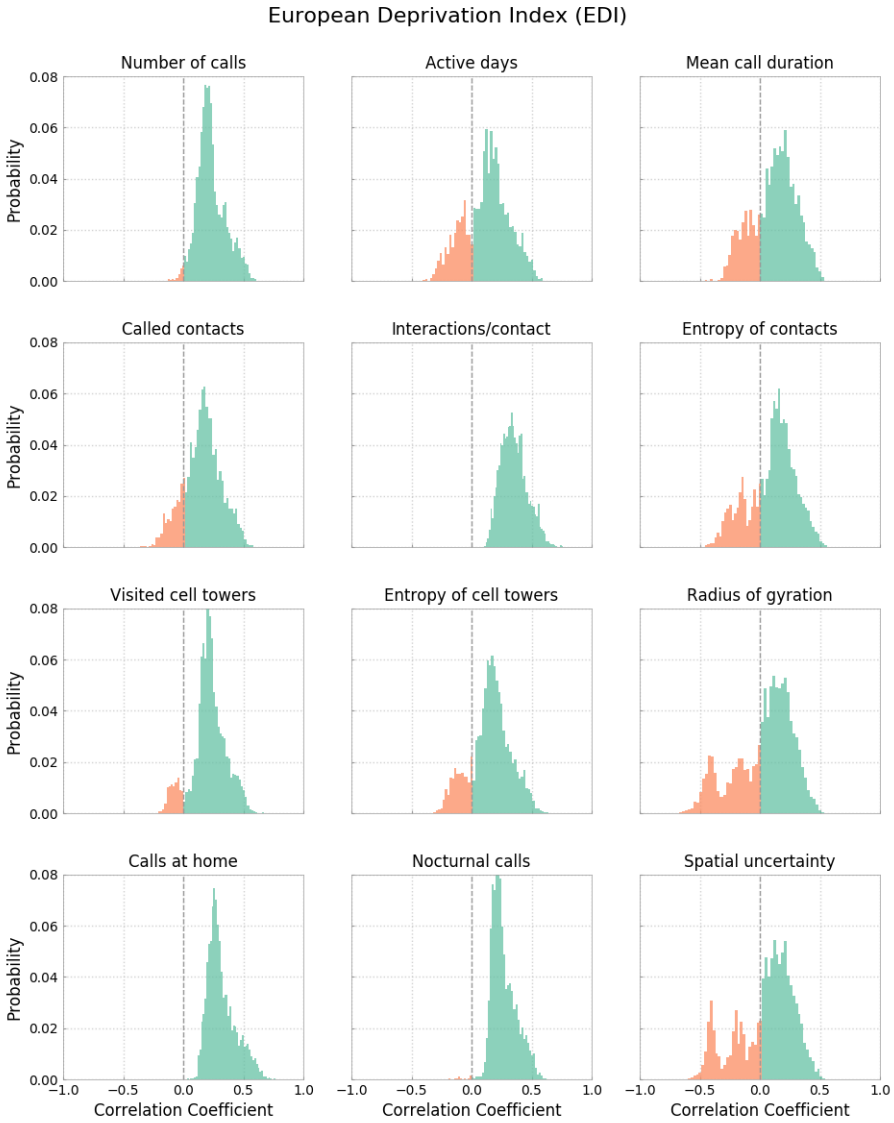


Fig. 7.6 Histogram of the Pearson correlation coefficient for the relation between EDI and a selection of mobile phone indicators calculated for all 4914 city definitions. The histogram is colored green when correlation coefficient is positive and orange when negative.

Gini Index of Wages vs. Mobile Phone Indicators

Regarding the relations between the Gini index of wages and mobile phone indicators, figure 7.7 shows relations to be more stable, meaning that less direction changes are observed compared to the relations between EDI and mobile phone indicators. Exceptions are the relations with the mean call duration, the amount of calls at home, and the amount of nocturnal calls, but the coefficients of these correlations (either positive are negative) is very low, rendering them less trustworthy anyhow.

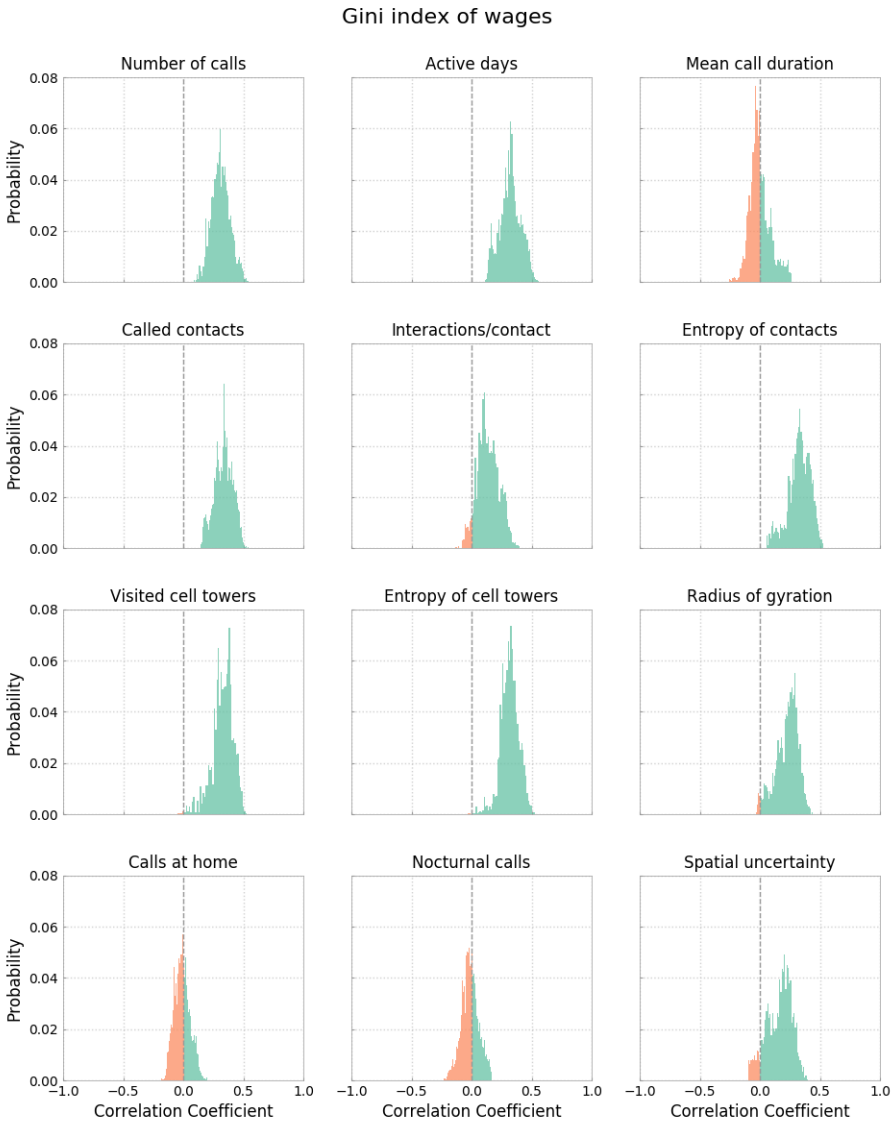


Fig. 7.7 Histogram of the Pearson correlation coefficient for the relation between the Gini index of wages and a selection of mobile phone indicators calculated for all 4914 city definitions. Figure setup is equal to figure 7.6.

Segregation Index of Wages vs. Mobile Phone Indicators

The relations between the segregation index of wages and mobile phone indicators in figure 7.8 reveal mostly negative correlation coefficients, although for most indicators a small amount of city definitions lead to positive correlations. For some indicators such as the interactions per contact or the nocturnal calls, the amount of city definitions that lead to either positive or negative correlation coefficients is balanced, raising the questions which relation can be deemed more trustworthy, and what city definitions would lead towards what relations uncovered.

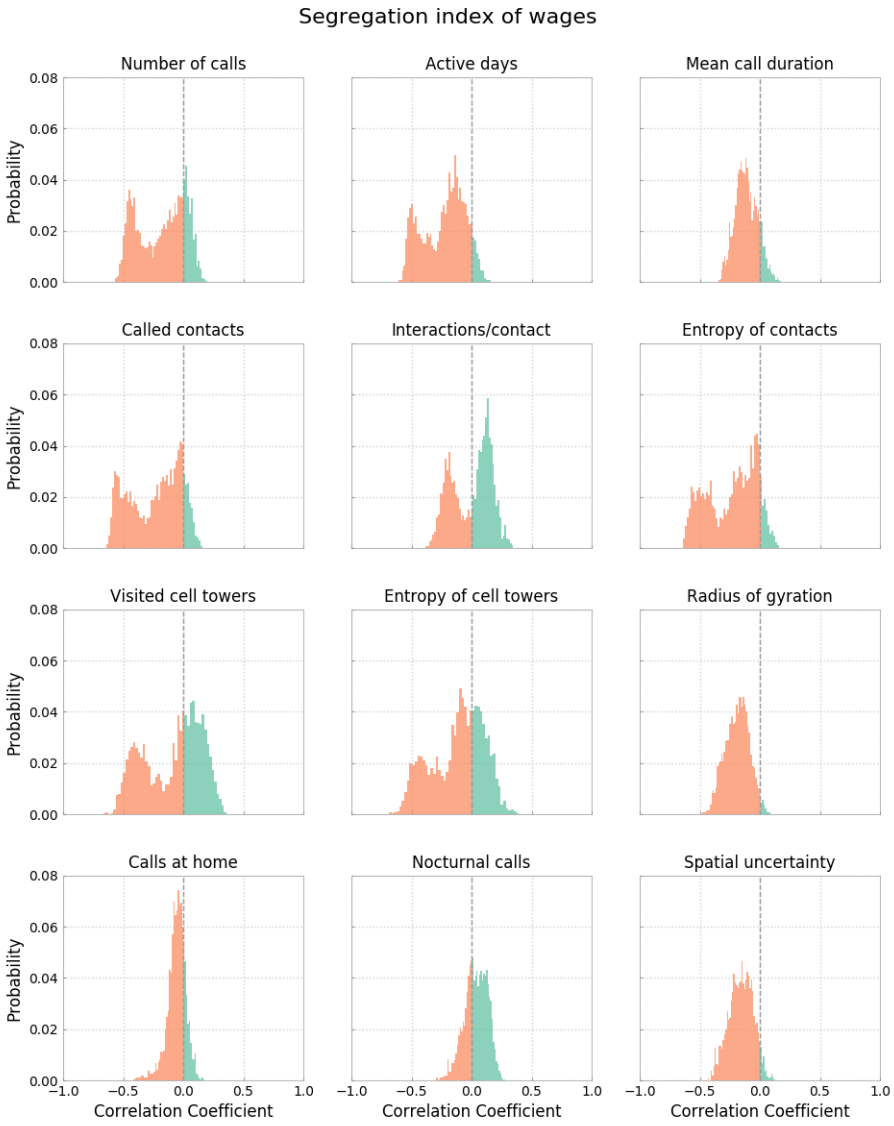


Fig. 7.8 Histogram of the Pearson correlation coefficient for the relation between the segregation index of wages and a selection of mobile phone indicators calculated for all 4914 city definitions.. Figure setup is equal to figure 7.6.

7.3.2 Sensitivity to City Definitions Parameters

Given the sensitivity of relations to city definition observed in figures 7.6, 7.7, and 7.8, the question is what city definitions lead to what relations. In other words, how do combinations of the p, f, and d influence the observed correlations?

Combinations of Three Thresholds

The influence of city definition on obtained correlations becomes clear when mapping the obtained correlations to the parameter-space used for the city definition. The heatmaps in figures 7.9 and 7.10 do so for the relation between, respectively, the entropy of contacts and the segregation index of wages, and the radius of gyration and the EDI. Clearly, different combinations of thresholds influence the observed correlations, their strength, and their significance.

The results in figure 7.9, for example, show that the entropy of mobile phone contacts is inversely related to income segregation. This suggests that less diverse mobile calling occurs together with higher segregation measures, but such relations only gain significance when cities are defined by low flow thresholds (f), so for rather loose definitions of suburban regions.

In figure 7.10 the influence of city definition is shown to be even more complex for the relation between the radius of gyration (a measure for the volume of mobility) and the EDI (a measure for deprivation). When city centers are defined without population threshold ($p = 0$, upper row in the figure), the relation between mobility volume and deprivation is negative and quite significant. However, when restricting the population threshold, the relation gradually becomes positive until moderate significance is found for positive correlations at higher flow thresholds. The found ambivalence, as well as the combined influence of different city definition parameters, adds to the uncertainty of such relations and their interpretations.

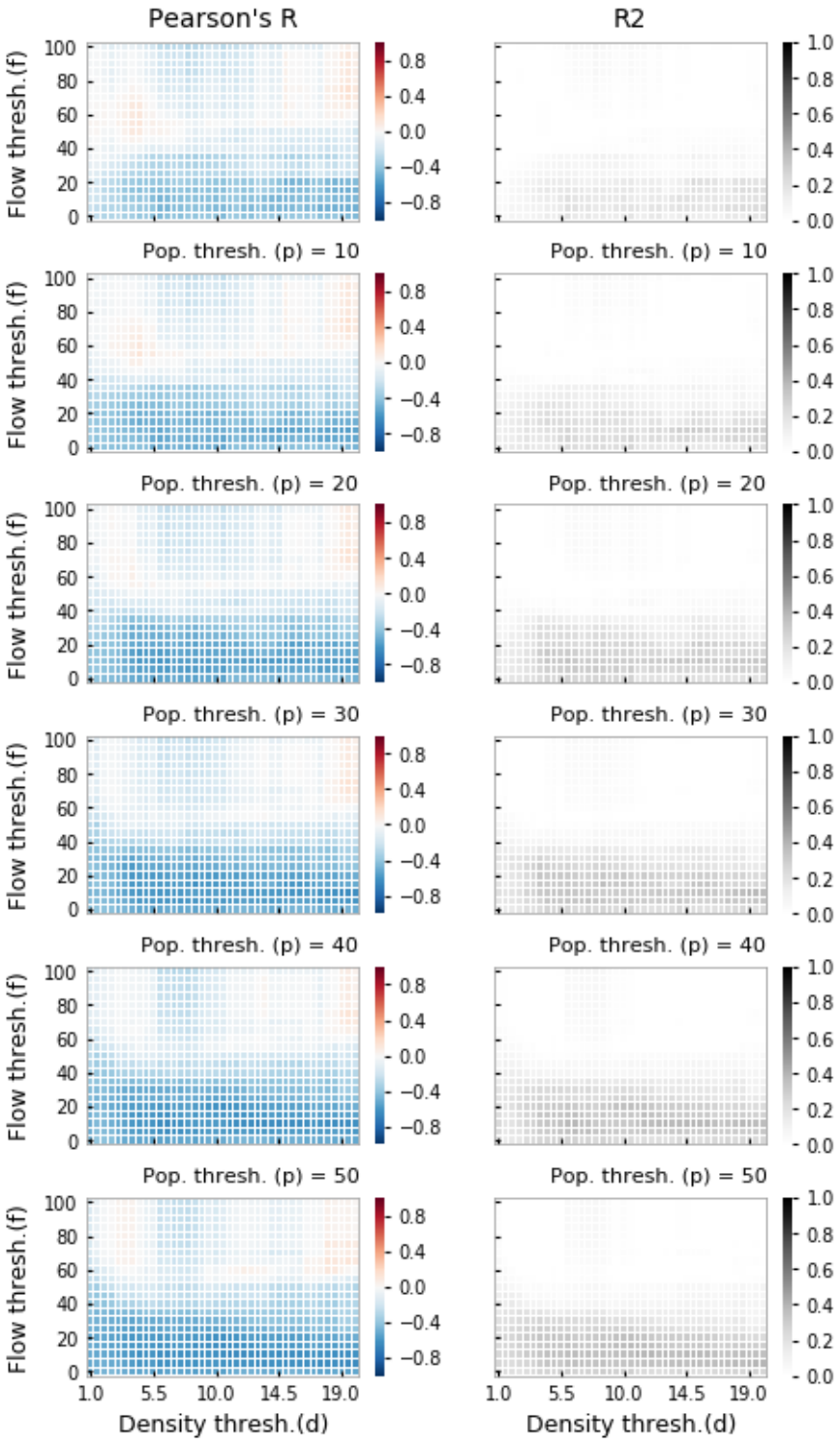


Fig. 7.9 Heatmap of the Pearsons's R (left) and R^2 values (right) for the relation between the entropy of contacts and the segregation index of wages for all city definitions represented in their according parameter-space. Each box represents one of the 4914 city definitions. Density thresholds (d) are for the city centers and in thousands inhabitants/hectare, flow thresholds (f) are in percentage of population commuting to the city center, and population thresholds (p) are in thousands inhabitants in the wider city. As can be deduced, the top row plots have a population threshold (p) of 0.

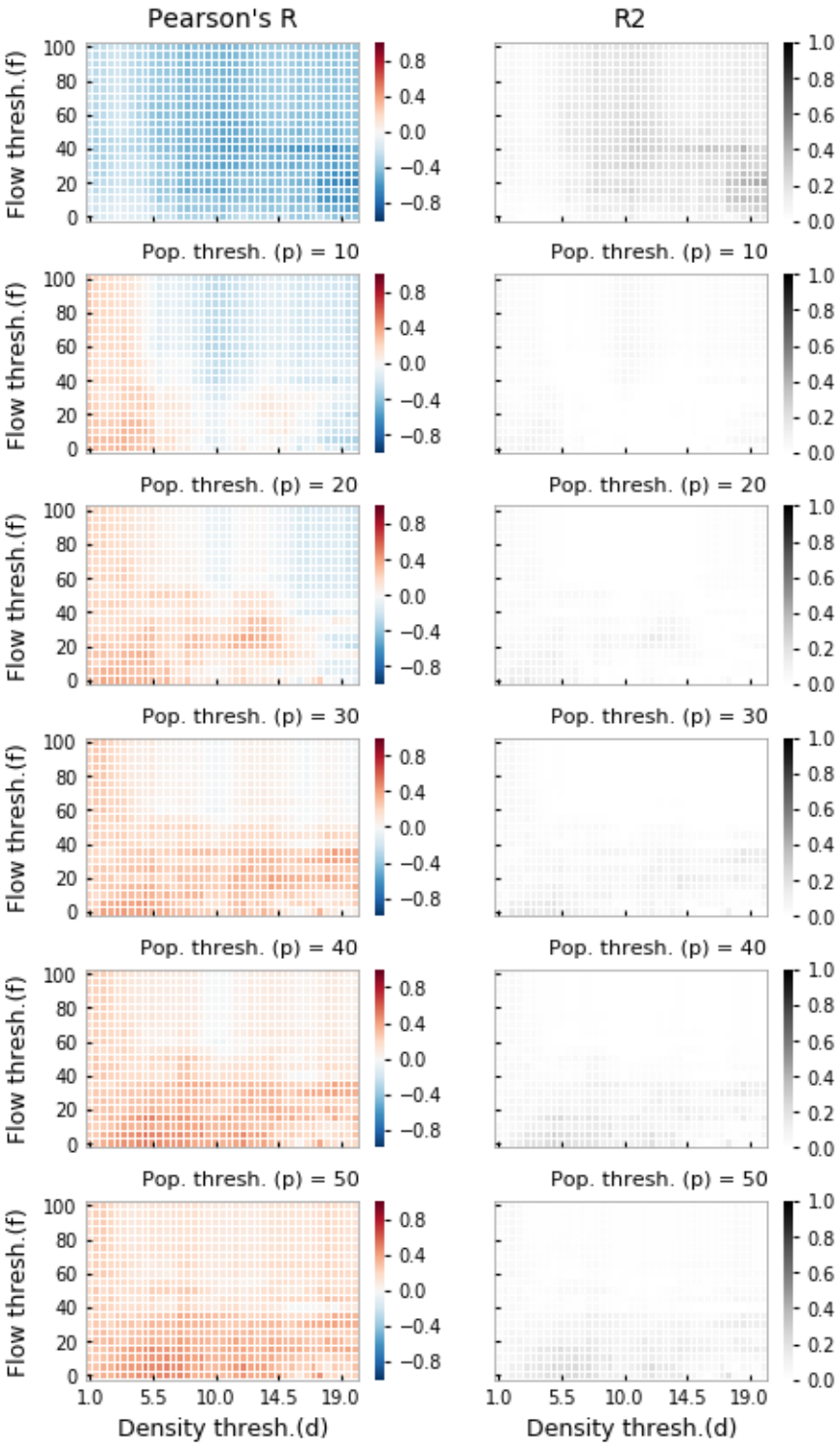


Fig. 7.10 Heatmap of the Pearsons's R (left) and R^2 values (right) for the relation between the radius of gyration and the EDI for all city definitions in their parameter-space. Figure setup is equal to figure 7.9.

Individual Thresholds

To assess the influence of individual thresholds, obtained correlations for all city definitions can be grouped by one individual threshold, resulting in a distribution of correlations for each parameter of the considered threshold. In figure 7.11, correlations between the segregation index and a selection of mobile phone indicators are grouped by the different parameters of the commuting threshold, revealing their influence on observed correlations. For the relation between the segregation index and the number of active days, for example, the influence of commuting thresholds seems limited, as all correlations obtained for the different thresholds remain negative. For the relation between segregation index and the number of interactions per contact, the influence of the commuting thresholds seems clearer: lower commuting thresholds in the city definition lead to negative correlations, while higher thresholds lead to positive correlations, with a tilting point situated at the 40-45% threshold.

Similar to figure 7.11, figure 7.12 investigates the sensitivity of relations between EDI and mobile phone indicators only this time with respect to the different density thresholds. Here, most correlations are found to be positive, and the effect of the density thresholds seems limited, especially with regard to changes in the direction of the relation. However, there seems to be an effect of the density threshold on the variation of the distributions, with higher density thresholds leading towards larger interquartile ranges, more extreme outliers, and correlation coefficients that can either be positive or negative for a single threshold. In other words, for the relations between EDI and mobile phone indicators, city definitions with higher density thresholds are found to imply a higher uncertainty on the observed correlations.

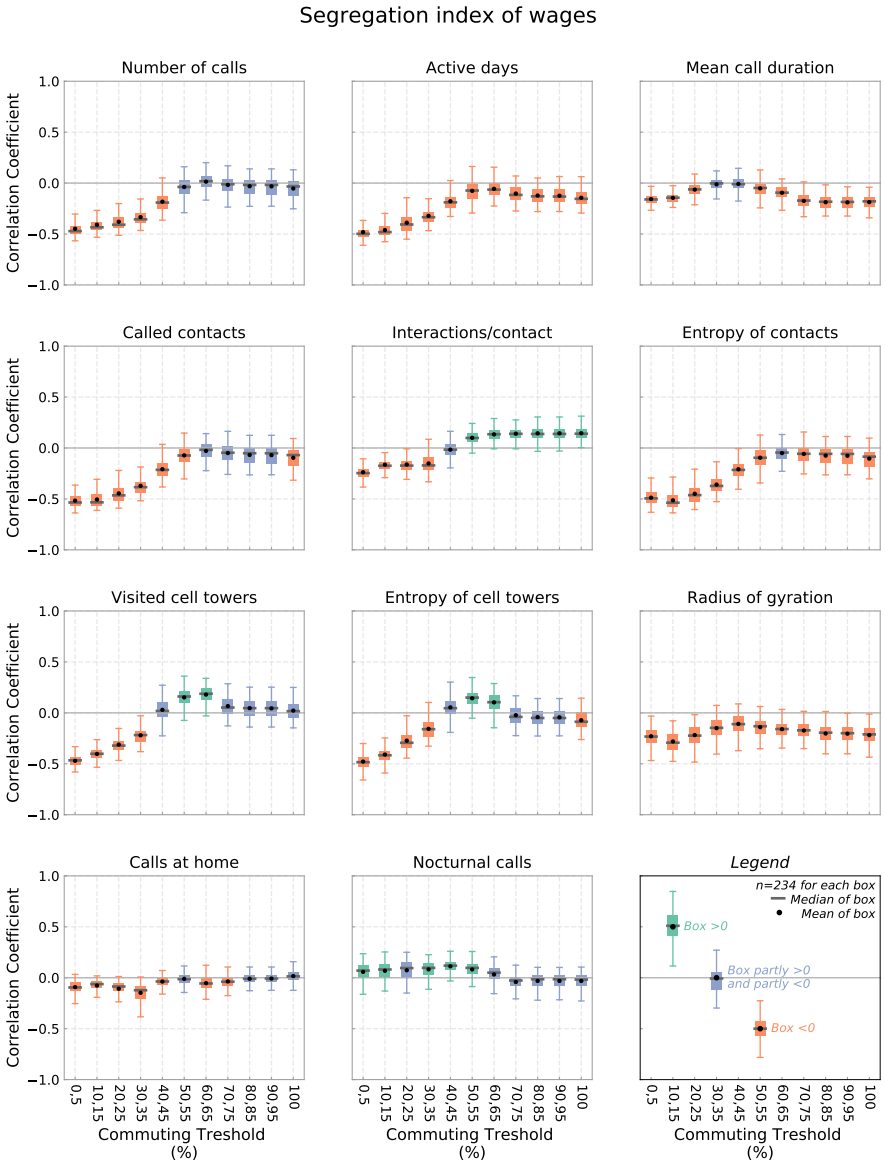


Fig. 7.11 Boxplots of the distributions of Pearson's R for the relations between the segregation index of wages and a selection of mobile phone indicators given different commuting thresholds. Distributions are obtained by grouping results of all city definitions by different commuting thresholds (in % of residents). Boxplots are colored based on the values in the interquartile range (the box of the boxplot, see legend).

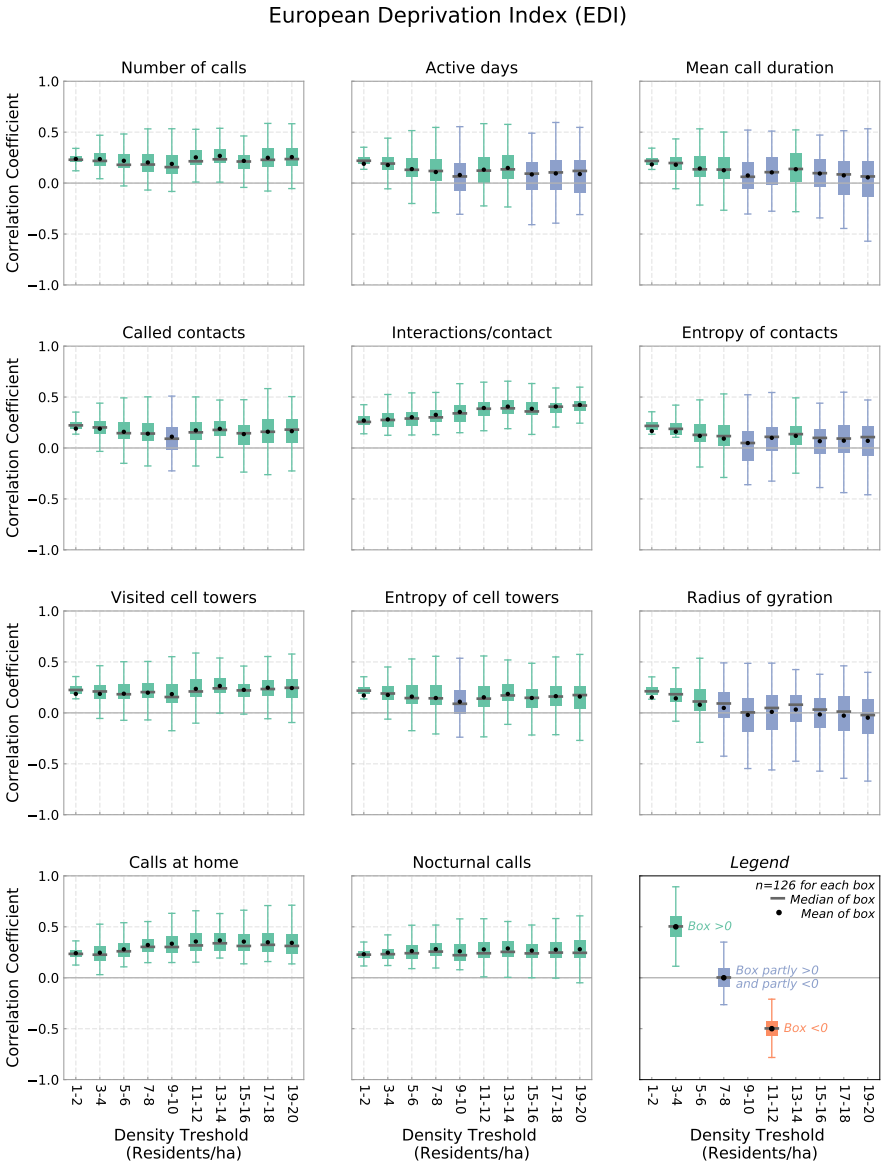


Fig. 7.12 Boxplots of the distribution of correlation coefficients for the relation between the EDI and a selection of mobile phone indicators given different density thresholds. Distributions are obtained by grouping results of all city definitions by different density thresholds (in residents/ha). Figure setup is equal to figure 7.11.

7.4 Discussion

7.4.1 Urban Scaling Regimes of Mobile Phone Indicators

A main finding of this chapter is that the majority of mobile phone indicators do not depict significant scaling laws, in the sense that the statistical fit of the relation with population size is (too) low as can, for example, be observed for two indicators in figure 7.2. Another finding is that the influence of city definition on the revealed scaling regimes is limited. Both findings are interesting, because they contrast with what has been found for indicators from census data in France before [42]. Such census indicators often have (significant) scaling regimes that tend to change with different city definitions as can be observed also in figure 2.12 in section 2.6.4.

Technical Limitations

One reason for this difference could be in the way the mobile phone indicators are treated. Because of technical limitations, mobile phone indicators are aggregated at cell tower level, leading towards weighted averages per city, meaning that they have to be considered as per-capita indicators (hence the β that is interpreted in reference to 0.). Scaling laws on per-capita indicators are known to depict lower significance [118, 41] compared to scaling laws of absolute quantities, and the reduction of information by using summary measures could result in lower significance of findings. Ideally, absolute quantities of mobile phone indicators should be investigated in the future. For example, the number of calls, which is now treated as the average number of calls in a city, could in the future be summed up resulting in an absolute quantities of the number of calls per city. Such summing up, however, is technically challenging because it requires a coupling between the geo-computational iteration of city definitions and individual user in the Orange big data system, something which has not been done (yet).

An in-between solution could also be imagined in which summing up is done based on the distributions of mobile phone indicators per cell tower which are available by summary measures (see for example figure 3.3 or 3.4 in chapter 3). Based on such distributions, aggregation into a city can be done by means of a weighted sum, which would form an estimate of the actual absolute quantities of an indicator. Using absolute quantities instead of per-capita indicators could lead towards higher significance of scaling laws, but it still remains to be seen whether this would also influence the observed scaling regimes and their sensitivity to city definitions. However, given the large variations in mobile phone indicators at cell tower level observed in chapter 3 (figures 3.6, 3.7 and 3.8) it is reasonable to expect an effect on observed scaling regimes and their sensitivity to city definitions too.

Captured Behavior

Another potential reason for the difference in scaling laws is that mobile phone indicators, compared to census data, capture human behavior which is less dependent to structural elements such as built-up environment or population densities, therefore more homogeneous in space, and subsequently less prone to scaling laws. This interpretation is merely a hypothesis and deserves further investigation but it introduces an intriguing tension between mobile phone indicators and census data when it comes to space. While census data tend to capture more contextual information on elements that are more spatially anchored (or at least that change slower over time), mobile phone data, and in extension other new big data sources, tap into behavioral aspects of society which have a different relation with space (or at least with a higher variability over space and time).

An example of hospitals can illustrate this train of thought. While census data would have information on the number of hospitals in a country, an aspect of society that is rather fixed in space and changes over time only slowly, a typical type of information captured by new (big) data sources could be the way people use mobile phones in hospital environments, an indicator that is far less *fixed* in space and time. Logically, methods or techniques that are deployed on the former, traditional type of data such as the investigation of scaling laws, might not always form the best ways to investigate the latter types of data.

Reasonable Significance and Home Detection

Nevertheless, some mobile phone indicators depict scaling regimes with reasonable significance, that is with R^2 values of maximum 60 (figure 7.3, and 7.4). The number of contacts, for example, depicts a superlinear regime, whereas the average time between mobile phone events of users depict a sublinear regime. This finding is interesting in two ways. First, it emphasizes cities as a facilitator for human interaction, pointing out the role of urban social networks in, for example, the diffusion of ideas or information.

Secondly, this finding is interesting with regard to home detection practices. For the 2007 dataset in France, sublinear scaling of inter-event times indicate that diminishing frequency of calls, and thus observations in the CDR dataset, comes with diminishing city size. The relation between frequency of observation and successful home detection remains poorly understood, but it is not unreasonable to imagine a better performance of home detection with more frequent observations. Following the observed scaling laws, the latter thus occurs more in larger cities.

Continuing on home detection, an interesting finding was observed in figure 7.5, which revealed high sensitivity to city definition of the scaling laws for mobile phone indicators related to home detection. The spatial uncertainty of the distinct days algorithm, for example, can be found to be either sublinear, linear or superlinear depending on the city definition. This observations adds to the complexity of the home detection problem discussed already in chapters 4 and 5. The uncertainties of home detection, in other words, depict more complex patterns than a relation to city size for different city definitions.

7.4.2 Sensitivity of Relations with Mobile Phone Indicators to City Definition

Another main finding of this chapter is the high sensitivity to city definitions of the relations between mobile phone indicators and three measures on income capturing inequality, segregation, and deprivation (figures 7.6, 7.7, and 7.8). For many relations, the found Pearson's R's differ substantially with city definitions, and for some relations even change direction. Especially in the latter case, this finding challenges the use of relations between mobile phone indicators and census data without performing proper sensitivity testing to city definitions and, in extension, without reflection on aspects of spatial analysis such as optimal spatial scale choice or the well-known Modifiable Area Unit Problem (MAUP) [56, 154].

Testing sensitivity to city definitions is one way to reveal the uncertainty that is related to interpreting mobile phone indicators, in the absence of proper validation data. An entire different strand of research is to try and interpret or theorize the observed relations and their sensitivity to city definitions. This chapter has withdrawn from such interpretations as they would require much more investigation outside the scope of this chapter. Nevertheless, the analyses of the found correlations with respect to the parameter space of the city definitions, as presented in figures 7.9, 7.10, 7.11, and 7.12, offer one way to start such investigations since many questions still remain to be answered. How come, for example, that the relations between the segregation index of wages and multiple mobile phone indicators such as the number of calls, the number of visited call tower, the interactions per contact and the mobility entropy all depict a tilting point around the commuting thresholds of 40 to 50% (a threshold which, by the way, is also used in the official Urban Area zoning of INSEE [140])? What is the consequence of the growing variability in the observed correlations between EDI and mobile phone indicators for city definitions with growing density thresholds? Does this mean that such city definitions capture urban areas that are too large for the more local spatial variability of mobile phone indicators? All questions that deserve further investigation.

Chapter 8

Discussion, Relevance, and Conclusion

> I take the position that I'm always to some degree wrong, and the aspiration is to be less wrong
>
> *Elon Musk*

Abstract

This chapter discusses the main results presented in the thesis, the relevance of the results, and their relation with potential future work. It concludes by revisiting the main research questions presented in chapter 1.

8.1 Discussion

8.1.1 Spatial Knowledge Gaps of Mobile Phone Indicators

Indicators from Mobile Phone Data

Starting from about 2006, mobile phone data research has been expanding into multiple applications and across various research domains. Adhering to the emerging Computational Social Science Paradigm (CSSP), the quantitative analysis of human interactions and movement patterns as captured by Call Detail Records (CDR) data has led to unprecedented insights on human behavior at scales that were previously hard to imagine. The observation of holiday movement performed by a large share of the French population and high temporal and spatial resolution in figures 4.4 and 5.4 is only one example of information that is now available in digital data sources that would have been (almost) impossible to obtain by traditional data collection methods.

It is no wonder that, currently, network operators, who possess the majority of mobile phone data, are actively bringing *their* data to the market while public institutions such as official statistics offices or ministries are heavily investing in exploring the potential of such datasets for their own workings. For both cases, a dominant way in which mobile phone data are used or conceived is in the form of indicators. Although it would form an interesting discussion whether this *indicator-format* is the best format available to simultaneously capture the available information in mobile phone data and the end users' needs, indicators do have the advantage that they summarize abundant and messy source data. Additionally, indicators align well with existing practices, know-hows, or protocols used for more traditional datasets such as census or survey data.

Main Argument

The main argument of this thesis is that for CDR data, and in extension for all large-scale, geo-located, digital datasets capturing human behavior, the translation from raw data to indicators cannot, and should not, be deemed trivial. There are multiple reasons that support this argument and all are, one way or another, related to each other. One reason is that CDR data are simply not collected to serve as indicators. The problem of unevenly distributed cell towers that is influencing the construction of the Mobility Entropy (ME) indicator, as is revealed in chapter 6, forms a clear example of this. The unknown local market shares that are obscuring the validation of home detection practices discussed in chapter 4 form another example. A second reason, and one that speaks to the (academic) contribution of this thesis, is that few investigations have yet been performed on the spatial patterns of mobile phone indicators, the methods that lead to them, and the spatial errors or uncertainties that surround both.

This thesis elaborates a critical investigation of the different steps presented in chapter 3 that lie between raw CDR data and the deployment of mobile phone indicators. Taking on an explicitly spatial perspective on these steps, the different chapters reveals several knowledge gaps related to the use of mobile phone data. Note that a critical investigation is not to be understood in the epistemological, radical sense of the word here, especially because the social atom problem, a main critique on the CSSP that is briefly discussed in chapter 2, holds true for many of the used methods in this thesis. Rather, the presented investigations are deemed critical because they raise questions on the spatial aspects of existing research that have either been neglected or gone unnoticed.

8.1.2 Uncertainty on the Performance of Home Detection Methods

Performance at Nation Level

A first spatial knowledge gap is on the uncertainty on the performance of home detection methods for CDR data, which is an example of non-continuous location traces, just like check-in data for credit cards or location-based services and online social networks. Despite forming a cornerstone of mobile phone data research, home detection in the literature has been deployed rather ad-hoc with few works performing validation or discussing user decision such as the home criteria choice or the optimal duration of observation. Chapters 4 and 5 present a first systematic investigation of home detection methods performed on a French CDR dataset. Results are far from encouraging and reveal that a large uncertainty on the validation of home detection still persists.

For example, when it comes to the validation of home detection methods at nation-wide scale, comparison with census data in the form of population counts or commuting figures form a popular way of validation but they reveal little information on how, where, or for which users home detection performs well (or not). For the French case, such high-level validation resulted in moderate performance of home detection at best, with correlations between user counts from home detection and population count from census data ranging between 0.42 and 0.62 (in Pearson's R, table 4.3 and figure 5.3) for different Home Detection Algorithms (HDAs), using different home criteria and deployed on different time periods. The consequence is that there exists a structural uncertainty with regard to home detection at nation scale that has yet to be understood as it can be made up by different elements such as unknown local market shares, differing definitions of home between HDAs and census data, or differences in HDA performance between users that are induced by their mobile phone usage.

Individual Level Validation

The latter is an interesting point as it reveals that the relations between mobile phone usage and home detection performance are ill understood. The reality is even worse as relations between mobile phone usage (or thus the characteristics of the digital traces a user leaves in CDR datasets) and home detection are not understood at all for the simple reason that validation data at user level is unavailable/unaccessible by law. As a consequence, no validation of home detection at individual user level is possible, blocking several advancements that could or should be made in order to render home detection, and in extension mobile phone indicators or other types of analysis that are supporting on home detection, more trustworthy. One advancement that could have been pursued by now (in the sense that it has become technically feasible over the past few years), is the development of learning HDAs that would optimize general performance by incorporating information on the characteristics of a user's CDR traces, whether or not expressed by indicators, in order to choose the most suitable HDA parameters for the job. Such a learning approach, however, would require individual validation data for, at least, a large enough test set of users, which in seldom available.

Spatial Uncertainty Instead of Spatial Error

In this perspective, an interesting contribution of chapter 4 is the construction of the Spatial Uncertainty (SU) measure. In the absence of validation data, and thus in the absence of the possibility to derive spatial error, the SU forms a data-driven estimation of spatial uncertainty that comes with a home detection decision at user level. Median SU values of all users in the French dataset, as can be observed in figure 4.6, are found to be low, indicating that home detection is often narrowed down to a local decision. Variations in SU values between users, however, are extremely high, indicating that a wrongful home decision can result in a large spatial error for many users. Estimators of the uncertainty on home detection, such as the SU, clearly cannot determine whether home detection is correct or wrong, but they can play a role as surrogates for proper validation measures for example by showcasing the relevance of user segmentation for home detection, by revealing the necessity to investigate spatial patterns of home detection error/uncertainty, by establishing the relation between user-level and nation-level validation, or by kick-starting the development of learning HDAs. Some of these roles are briefly touched upon in chapter 4, but all could from the subject of future research.

Individual Level vs. Nation Level Performance

As stated before, estimators of uncertainty cannot replace validation and so the effect of researcher decisions, such as the choice of HDAs criteria or of the duration of observations used, cannot be evaluated at user level. At the same time, evaluation of performance sensitivity to researcher choices can be pursued, as shown in chapter 5, but the nation-wide level validation measures might not be subtle enough to capture the large heterogeneity between users and areas.

This problem is illustrated in chapter 4 where pairwise comparison between different HDAs reveal that 4% to 40% of users get accorded a different home when using different HDAs (figure 4.5). In absolute terms, these percentages represent a huge amount of users, but nation-wide performance measures, such as correlations between user counts from HDAs and population counts from census data, do not always pick up such differences. This is partly because differences in detected homes of a local nature (remember the rather limited SU values), and mainly because nation-wide performance measures are too general. This limitation of nation-wide performance measures is equally illustrated when investigating the spatial patterns of home detection in summer as, for example, shown in figure 5.6. Here, a small deviation in the nation-wide performance measure is linked to a rather substantial shift in the amount of detected homes from city centers to touristic areas, representing the holiday movement of (hundreds of) thousands of mobile phone users.

Sensitivities of Performance to Researcher Choices

When investigating the sensitivity of nation-wide performance to researcher choices on HDA criteria, HDA parameters, the chosen time period and the chosen period of observation in chapter 5, it is found that, in France, performance is most sensitive to the previously mentioned holiday movement of users in July and, especially, August (figures 5.7 and 5.8). Performance, in other words, is most of all sensitive to the chosen time period, with the effect on the period of observation being subordinate to the degree for which time periods are in July and August. Concerning criteria and performance choice, there exists an effect on performance, but it is dependent to the combination with the duration of observation and, to a lesser degree, time period. Popular time-constraints algorithms that impose a restriction on parts of the days or nights to be considered for home detection, for example, seem to live up to their semantic logic only when (very) long duration of observations are considered (figure 5.8). The consequence is that, based on nation-wide performance measures, no clear suggestions can be made on which criteria or parameters best to use. For France, the most important effects on performance are dictated by the characteristics of the collected data itself which, by definition, is location and time dependent (the holiday movement period might, for example, not exist in other countries or CDR datasets). Consequently, a good a-priori knowledge of the investigated dataset in combination with sensitivity testing is crucial to support researchers' decisions on how to apply HDAs.

8.1.3 Spatial Patterns of Mobile Phone Indicators

Small Differences between Areas, Large Variations within Areas

A second knowledge gap is on the spatial patterns of mobile phone indicators. Once calculated and territorially aggregated, distributions of mobile phone indicators can be investigated, as for example done at cell tower level in chapter 3. Summary measures of indicator distributions do depict clear spatial patterns (figures 3.9, and 3.10), but variation within distributions is large (figures 3.4, 3.4, 3.6, 3.7), which is an expected property of indicators on human behavior captured for large sample sizes. In other words, it seems only logical that big data on human behavior can depict differences between areas even though variations within areas are much larger. Although expected, this property complicates the interpretation of differences between areas in the sense that small differences between overlapping distributions of two areas will probably end up statistically significant because of the large sample sizes, even though large parts of the distribution in one area could, with a high probability, have been drawn from the distributions of the other area. This makes it difficult to point out what exactly makes up the difference between areas, hence complicating interpretation. One good example of this complication is in the distributions of the Mobility Entropy (ME) indicator shown in figure 3.12. Aggregation here is not in areas but based on different deciles of EDI, but the problem remains the same. Clear summary differences exist between distributions, but the distributions themselves overlap to such a large degree that it is difficult to assess the real value of this result.

Regional Differences of Indicators

Nevertheless, spatial patterns of mobile phone indicators often depict clear regional differences, offering a welcome help in the interpretation of observed variations. Regarding indicators that are related to the volume of mobile phone activity, for example, the spatial patterns in figure 3.9 reveal that mobile phone activity is typically higher in urban areas, lower in rural areas, and that there exists a small gradient from north to south in France, probably expressing economic activity and cultural differences respectively. Other spatial patterns are more difficult to interpret, indicating how little yet is known on the processes that generate mobile phone indicators and, in extension, the information gathered in CDR datasets. The case of the number of calls at home forms a perfect illustration of this last point. As shown in figure 3.10, the spatial pattern of the number of calls at home is extremely pronounced, but it remains unclear as to why this patterns exists. In a way, this is problematic as the number of calls at home might influence both the construction of movement patterns, as well as home detection, from CDR data. Clear regional differences in the number of calls at home thus imply regional differences in the applicability of certain analyses and, in absence of proper validation data, the trustworthiness of the dataset, or at least of the obtained results from the dataset.

Spatial Bias of the Mobility Entropy

The main example of why the investigation and interpretation (as limited as it might be sometimes) of spatial patterns from mobile phone indicators can be crucial is presented in chapter 6. Investigating the spatial patterns of the Mobility Entropy (ME) indicator, a popular measure to express the diversity of users' movement patterns, reveals an extremely high accordance with the spatial pattern of the number of visited cell towers and, in turn, with the cell tower density (figures 6.6 and 6.7). In a following investigation, the traditional ME formula is found to be biased for cell tower density, which severely challenges its trustworthiness for comparisons between regions that have different cell tower densities. To mitigate the spatial bias, a correction in the form of the Corrected Mobility Entropy (CME) is elaborated. This solution is not generic, meaning that it cannot directly be applied to other CDR dataset, mainly due to a parameter choice for the scaling range (a, b) that needs testing. The real merits of the CME are that i) it clearly points out the spatial bias of the traditional ME and ii) that it forms a perfect illustration on how small adaptations to mobile phone indicators, in this case a correction for cell tower density, can result in substantially different insights with regard to the resulting spatial patterns (figure 6.7), the relations with urban areas (table 6.4), or the relations with information from other sources such as census data or satellite imagery (table 6.6 and figure 6.9). In the case of the CME, the mobility diversity of mobile phone users is assessed to be highest in sub-urban areas (figure 6.8), and relations with share of car use in trips and land use are put forward as explaining factors for the observed patterns (table 6.8), providing different insights compared to the traditional calculation of mobility entropy. Critical investigation and interpretations of spatial patterns can thus be crucial to assess the veracity of mobile phone indicators.

8.1.4 Urban Scaling Laws of Mobile Phone Indicators

Insignificant Scaling Laws with Stable Regimes

A third and last spatial knowledge gap with regard to mobile phone indicators is their relation with city size, as often expressed by means of urban scaling laws. Compared to indicators from census data, chapter 7 finds that most mobile phone indicators do not depict significant scaling laws (figure 7.2), with the exception of four indicators (figures 7.3 and 7.4): the number of visited cell towers, the ME, the number of called contacts, and the frequency of calls. The former two are logical given the demand-driven positioning of cell towers in France (fig 3.1) and the spatial bias of the ME towards them (chapter 6). The latter two form an interesting finding as they accentuate the effect of cities on interpersonal contact and their role in, for example, the spreading of ideas or innovation. Results in chapter 7 also find scaling regimes of mobile phone indicators to be largely independent from city definitions, although the scaling exponent (β) does vary with city definition. For this observation, the exception on the rule are indicators related to home detection, such as the SU or the distance between L1 and L2, for which regimes can change according to city definition (figure 7.5), indicating that the complexity of the home detection problem cannot be reduced to a simple relation with city size.

Relations with other Indicators and Sensitivity to City Definitions

Most probably, the limited significance of urban scaling laws for mobile phone indicators is related to the large variation of indicator distributions within areas opposed to the small differences in distributions between areas, as discussed before. The implication is that there is a large base of users with similar indicator values that are independent from urban characteristics. Although they might not depict significant scaling laws, the pronounced spatial patterns and the large variations in local area distributions makes one wonder how mobile phone indicators would correlate with other indicators such as income measures, and whether such correlations are sensitive to city definitions or not? As shown in chapter 7, it turns out such correlations are highly sensitive to city definitions (figures 7.6, 7.7, and 7.8), and in a different way to the parameters that make up these definitions such as density, population size or commuting thresholds (figures 7.9, 7.10, 7.11, and 7.12). Because the goal of the investigations in chapter 7 is merely to quantify this variability and not to elaborate on its interpretation, reasons of existence, or consequences for other work, a wide range of future work on this topic is still open.

One consequence, at least, is that claims in literature on the possibility to *nowcast* socio-economic states of territories based on mobile phone indicators will have to be revised if they are only based on one city definition. The deployment of mobile phone indicators in predictive models, or any other application for that matter, should ideally be backed by an advanced understanding of their spatial pattern, by a good insight on the (spatial) error related to user-territory mapping, and by sensitivity testing of the chosen spatial delineation ($\approx$ city definition); by a thorough assessment of their geographical veracity, in other words.

8.2 Transferability and Relevance of Findings

8.2.1 Transferability

Data Accessibility

The use of the French CDR dataset comes with a situatedness of the presented research that cannot go undiscussed as it limits the reproducibility and transferability of findings in several ways. Concerning reproducibility, the most important aspect is undoubtedly the limited access to the dataset as is discussed in chapter 1 too. Researchers that want to reproduce the results in this thesis will have to come to an agreement with the Orange Labs France in order to obtain access to their systems and the therein stored data. This extra barrier impedes quality control. In a wider perspective, limited access to CDR datasets hinders the further, equal, development of the mobile phone data research domain and forms the basis on which partnerships between operators and (top-)universities can maintain a hegemony on research profits and directions.

Technical and Technological Context

One important limiting factor to transferability of findings is related to the technical and technological context in which they are produced. A clear example of this is given in chapter 3 where it is shown that some event types cannot be used to construct movement patterns or contact networks due to technical issues. This limitation might not exist for other datasets although, undoubtedly, those will have their own technical limitations. Mismatches in the types of information and the form in which information is collected, hold or treated between datasets thus introduce a limitation on transferability that might not seem large at first but that can potentially lead to substantial differences in applied pre-processing, in used methods or in obtained results. Similarly, transferability is limited by technological context, which is rapidly changing when it comes to big data technologies. The discrepancy between the depth of investigation between chapters 4 and 5 forms a good illustration of this point. First investigations on home detection were started in 2014, but were limited to discrete months only because of the time cost of analysis (chapter 4). Two years later, by mid-2016, system capacity and technology were advanced to such a degree that they allowed sensitivity testing of the same analysis for many more durations, parameters and periods (chapter 5). Being a major player in the world market of telecommunication, Orange invests a fair amount in keeping their systems up-to-date with the latest technologies, but is indisputable that many other CDR datasets are not being governed by similar advanced systems, limiting transferability of certain methods. It is, for example, reasonable to assume that big data systems in developing countries are not even that advanced as the Orange system was in 2014, making sensitivity testing not feasible, or very time costly.

Historical and Geographical Context

Other limitations on transferability of findings are the historical and geographical context of the collected data. As discussed in length in the literature review of chapter 2, CDR data captures snippets of human behavior that are mediated by mobile phone usage but that can be used to study movement patterns for example. However, both human behavior and mobile phone usage are ever changing, rendering each empirical analysis of CDR data time-bound and location-bound. At this moment, the effect of changing mobile phone usage on results are ill understood, mainly because data are available for limited durations only (remember that the 154-day duration of the French CDR dataset is one of the largest in literature). This adds a lot of uncertainty to the transferability of findings. For example, the influence of holiday movement on home detection performance, as observed for France in 2007 (chapter 5), might very well change if the French population changes its holiday behavior over time. Similarly, changing mobile phone usage could, for example, increase the temporal resolutions of traces in CDR data compared to the CDR data from 2007, potentially leading towards a different appreciation of different home detection criteria. Such effects cannot be foreseen, and their influence on results remains unknown until longer periods of CDR data become available and careful comparisons between regions are carried out. The (short) investigation on the spatial patterns of mobile phone indicators in chapter 3 reveals at least one way to start doing the latter.

Users Composition

A final limitation to transferability stems from the unknown composition of the Orange users populating the CDR dataset. Due to legal restrictions, the combination of user information, as for example stored in Customer Relationship Management (CRM) databases, and CDR data is not allowed. Consequently, it is impossible to know the exact composition of the Orange users, and thus their representativeness for the French population, or for populations in other CDR datasets. This problem is augmented by the unknown local market shares of Orange that impede higher level insights on population representations of the dataset in the same way as they are obscuring validation efforts for home detection as is discussed in chapter 4. Knowledge on the composition of mobile phone users is relevant in many ways. One of them to understand patterns of mobile phone usage, as different subpopulations might have different ways of using their phones. More important though, it is relevant to questions regarding exclusion and discrimination. If mobile phone data are going to be used to inform decision making, whether by private companies or public offices, or, in an even further step, if they are going to be a part of the official statistic production chain, then it forms an absolute priority to have a clear understanding of who is represented in the data and to which degree, especially when it comes to minorities or other vulnerable subpopulations.

Data Sharing and Open Algorithm Systems

Interestingly, operators and public institutions in Europe are, nowadays, actively working on solutions that might lower several transferability barriers. One of such projects is the OPen ALgorithm (OPAL[1]) project. This project aims at constructing a big data system to which different operators can contribute their (source) CDR data and on which, in a next step, partners can run algorithms obtaining aggregated results based on all data. Having a common format in which operators contribute the data, as well as a system with the same technological capabilities, already eliminates large parts of the technical and technological transferability barriers that exist today. Additionally, when supported by all operators in one country, such systems could alleviate the local market share problem, as well as part of the user composition problem discussed before. Ultimately, one can imagine such systems to establish a link between operators and academic institutes. Equally, one could imagine a similar system to be governed by public institutions such as the European Commission, opening up at least two potential advantages. First, it might open the way to a controlled enrichment of mobile phone data with information for, for example, census data. And secondly, it might enable legislation to directly enforce regulations by parameterizing access to different datasets and algorithms for different users in the system directly, cutting the long and cumbersome road of implementation of, for example, privacy policy encountered today.

Commercial vs. Research Purpose

Clear implementation of regulations are necessary, especially in the light of the very recently (May 2018) introduced General Data Protection Regulation (GDPR). Although sending a clear message on the importance of *data privacy*, one thing the current version of GDPR excels in is the very unclear directives as to what can or cannot be done, and for which purposes. Especially the latter point is important for academic work, as there exists a fundamental difference between using data for commercial purposes and for research. In the current state of affairs, this distinction is being blurred by unspecific directions. In the case of mobile phone data, there exists a perfect example of this point. In anticipation of the GDPR, the French data protection agency (CNIL) has been restricting the use of individual CDR data to 24-hour periods only, after which user identifiers in the dataset need to be irreversibly decoded. In terms of commercial purposes this makes sense as little applications actually need information at that resolution. For research, however, this restriction is a severe blockage for the understanding of CDR datasets, their potential, and their deficits. Such restrictions imply that no analysis on individual mobile phone usage for longer periods can be made, that the construction of call networks and movement patterns are extremely restricted, that the potential to perform correct home detection is reduced to an unknown degree, and that long term patterns such as holiday movement become increasingly difficult to observe, just to name a few. As there is no clear directive on the distinction between commercial and research purpose, it is understandable that the safest restrictions are implemented but clearly this does hinder scientific advancements.

[1] https://www.opalproject.org/about-opal/

Academic Brain Drain

In addition, there is the question to which degree the current GDPR will impact academic research and specifically, academic access to (big) data in the long-term. Severing privacy regulations, although for the best of the public interest, let that be clear, might very well lead to a conservative reflex of operators and other data-holding private companies. After all, sharing data with partners has become increasingly difficult and each collaboration entails extras risk on privacy infringement both of which form extra costs to cover for. Although perhaps indirectly, this stimulates expertise to be brought (or should one say: bought) in-house. This is a development that can already be observed in other cutting-edge technological research domains such as Artificial Intelligence (AI), where it is draining expertise from academia hereby limiting the dissemination of knowledge to and the occurrence of debated in the public domain. Putting in place data-sharing systems such as OPAL while providing educational institutions with access might counteract such developments but probably will have to happen rather sooner than later.

8.2.2 Relevance

Relevance of the Main Argument

Limited transferability does not equal limited relevance. The main relevance of this thesis is that it demonstrates the usability of spatial thinking to assess mobile phone indicators and related methods in terms of error, uncertainty, sensitivity to researcher decision, and limitation of interpretation. Space, in this perspective, serves a role as mediator for the critical investigation of empirical practices. In the light of the tremendous growth of empirical analyses related to big data, this forms a welcome counteract to the quickly (re-)emerging quantitative research paradigms such as the Social Physics that clearly have their shortcomings, as is discussed in chapter 2. The relevance of the main argument of this thesis therefore goes further than mobile phone data research alone. It can be extended to all empirical analyses of geo-located datasets that take on an explicit quantitative perspective without incorporating notions of space to confront their own findings.

Relevance of Mobile Phone Indicators

The creation of mobile phone indicators from CDR data as discussed in chapter 3, together with the discussion on the statistical and spatial variations of the indicators, is relevant in multiple ways. First, it is relevant for official statistics that has a long tradition in the creation of nation-wide indicators but that has been actively investigating the possibility to integrate big data in their production process. Obviously, there exists many more ways to integrate mobile phone data in official statistics, but the creation of mobile phone indicators definitely bears some relevance in accelerating these developments and stimulating discussion.

Secondly, indicators are relevant for all scholars that want to study regional differences in human behavior. The discussion of the CME indicator and its relations to urban areas in France in chapter 6 is only one example of how nation-wide indicators from mobile phone data could be used by, geographers, transport planners, urban planners, sociologists, regional economists, policy makers, etc. Thirdly, the observation of large local variations in mobile phone indicators forms a strong incentive for (spatial) statisticians and spatial analysts to (continue to) take on the challenges that stand in the way of a correct and unambiguous interpretation of regional patterns of human behavior as captured by big data sources.

Relevance of the Home Detection Problem

The problems raised with regard to home detection from mobile phone data are extremely relevant to the wider mobile phone research domain and in extension all domains that use non-continuous location data. In short, the analyses presented in chapters 4 and 5 form a clear illustration of the total lack of validation and error assessment of home detection, which is a crucial step in many mobile phone data research. Although the presented analyses are limited in many ways, such as the fact that only single-step HDAs based on one criteria only are used, uncovering this shortcoming is relevant for future research. One direct contribution to the latter is in the first systematic investigation of the effect of researcher decisions on nation-wide performance of home detection performed in chapter 5. The finding of highest relevance, however, is that the absence of individual level validation data is severely impeding correct validation and error assessment of current home detection practices. This finding, hopefully, contributes to form an incentive to all parties involved in mobile phone data research to tackle this outstanding problem.

Relevance of Mobility Entropy Correction

The discussion on the bias of the ME indicator in chapter 6 is relevant for all mobility studies that use this indicator while having (spatially) heterogeneous observation points (cell towers for CDR data) resulting in different observation point densities. Note that for other data sources that, compared to CDR data, are not based on passive location recording but on an active and mediated recording, such as location-based social media or online social networks where location recording is an active act embedded in the mediated use of the service, heterogeneity is not only given by a spatial, infrastructural element (the density of possible check-in locations in an area) but also, and possibly even more so, by the mediation of these locations within the application (the attractiveness of sharing this location within the context of the application). The relevance of the proposed solution in chapter 6 is confined in the sense that it is not an elegant, generic solution. Its main contribution lies in the revealing of the bias and the effect it has on interpretations at nation-scale.

Relevance of Urban Scaling Laws for Mobile Phone Indicators

The study of urban scaling laws for mobile phone indicators presented in chapter 7 forms one of the first explicit connections between mobile phone data and urban scaling literature, but its scope was limited in two ways. First, the study was limited by scaling laws of mobile phone indicators that, in general, depicted too low of significance (figure 7.2). As discussed, this result might be due to specific technical limitations that do not allow to calculate scaling laws for absolute quantities of mobile phone indicators but that should, theoretically, be solvable in the future. The relation between per-capita indicators, as used in chapter 7, and absolute quantities with regard to the significance and interpretation of urban scaling laws remains largely unexplored in literature and forms, together with the outlined technical challenge, an interesting line for future research. Secondly, the study of urban scaling laws for mobile phone indicators is confine in its interpretation, mostly because such interpretations would require much more in-depth investigation than what could be presented in this thesis. Undoubtedly, a clear relevance exists in the interpretation of scaling laws from mobile phone indicators for studying urban systems, social processes and movement patterns but this largely remains future work.

Relevance of the Sensitivity to City Definition

The revealed sensitivity to city definitions of relations between mobile phone indicators and indicators on income (figures 7.6, 7.7, and 7.8) is of high relevance for the applicability of mobile phone indicators. Although it might very well be that such sensitivities exist between census indicators themselves too, the argument that high local variation of mobile phone indicators imply nation-wide level relations with them to be sensitivity to the used spatial delineation should be taken into account every time mobile phone indicators are deployed in this fashion. Simplification of locally rich datasets, such as CDR, into high-level indicators and their correlations with other measures should be challenged until such correlations are proven to be insensitive to the used spatial delineation or this sensitivity is clearly documented, regardless whether their application is in policy, development studies, official statistics, or economics, just to name a few. In terms of relevance for future work, the analyses in chapter 7 offer a wide range of possibilities for more interpretative studies on the relations between mobile phone indicators, that typically capture behavioral aspects, other information sources such as census, that typically capture contextual information, and urban systems. As was the case with the urban scaling laws, here too, discussions in chapter 7 restrain from interpretation of findings but, clearly, the form an avenue for future research. For example, exampling prospects exist on the further investigation of relations between, for example, income segregation and calling behavior or deprivation and movement patterns, and this for different city definitions.

8.3 Conclusion

8.3.1 Answering the Research Questions

By way of conclusion, some short answers can be provided to the main research questions introduced in chapter 1.

What is the state of the art of mobile phone data research, especially with regard to the creation of indicators from mobile phone data?

The state of the art shows that it is possible to use mobile phone data to study different dimensions of human behavior, most notably human interactions and movement patterns. Such investigations are mostly embedded in an emerging Computational Social Science Paradigm and have a tendency to treat mobile phone data from a network perspective. Apart from studying the larger structures, such as contact networks of millions of people, both technology and methods nowadays allow for the efficient calculation of measures at individual user level, such as the amount of contacts, or the number of visited cell towers.

Being calculated for large sample sizes, these measures have been used to describe statistical properties of human behavior, and are increasingly being deployed in relation to other sources of informations such as census data, leading towards their use as *indicators*. Uncovering relations between mobile phone indicators and other indicators such as socio-economic indicators, has led to claims that mobile phone indicators could be used to *nowcast* characteristics territories, or to support economic development, subsequently stirring interests from official statistics and public institutions. The latter claims, however, have seldom been investigated from a spatial perspective, meaning that relations are not contextualized as to where and why they occur.

Is it possible to create mobile phone indicators from a mobile phone dataset available for France?

Experiments show that it is possible, and thanks to rapid technological developments increasingly so, to calculate mobile phone indicators from a CDR dataset of approximately 18.5 millions users spanning 154 days of observations.

To what extent can we interpret mobile phone indicators? What do they describe? How are they distributed in space? Can we validate the spatial patterns of mobile phone indicators?

Although most mobile phone indicators are rather straightforward in their definition, their interpretation is not. The reason is that CDR data possesses very limited information on the context in which it has been captured, rendering it difficult to evoke an interpretation that goes further than purely descriptive measures such as than an user has made 25 calls. Additionally, very little validation data exists at the individual user level that could support the completeness of information collected by mobile phone data. In terms of movement patterns, for example, CDR data is restricted in observations by the temporal resolution of a user's phone calls. As such movement patterns are always collected partially, but it is unsure to which degree and how this influences the validity of results.

Nevertheless, at nation-scale, most indicators do depict clear spatial patterns, allowing for interpretations related to spatial occurrence but, then again, little interpretation can be made on the underlying processes without enriching the available data with information from other data sources such as census data or satellite imagery. One clear example of this is the spatial pattern of the amount of calls at the most used cell tower. In France, clear regional differences are revealed with Northern regions, such as Nord-Pas-de-Calais, Champagne, Alsace, Bretagne, and Normandy, showing much higher values than southern regions or Paris, but the reasons for the existence of this pattern remain unknown unless combination with other data sources is pursued. Enrichment of mobile phone indicators with other information can be helpful to interpret observed patterns, as was shown for the case of the diversity of mobility; but here again the true validity of such interpretation would be dependent on external confirmations such as individual user validation data, theory development or simulation models.

Can we investigate the relation between mobile phone indicators and indicators from census data? What problems arise? What relations can be uncovered and how are these relations shaped geographically?

Relations between mobile phone indicators and indicators from other data can be investigated, but typically require a form of spatial aggregation from the mobile phone indicators to the level in which other data is available. Apart from the uncertainty that comes with user-territory mapping, in many cases this aggregation provokes a problem as it is influencing found relations. Different aggregations, in other words, result in different relations, and it remains a difficult exercise to understand which finding results from actual processes at play, and which form an artifact of the methodology. One element that is disadvantageous for mobile phone indicators in this perspective is that most of them depict large local variations and only limited variations between different areas, making them more susceptible for the aggregation problem.

What are the uncertainties that come with mobile phone indicators? Can we quantify what errors relate to them?

The uncertainties that comes with mobile phone indicators are rather high but difficult to quantify. One main contributor to uncertainty is in the user-territory mapping phase that is typically based on home detection methods. Such home detection methods suffer from limited validation possibilities, making it difficult to properly assess the related error, or even the uncertainty that comes with the home detection decisions. The main problem in this perspective is that, ideally, validation should be performed at individual user level but, given the size of CDR data, such validation level is practically impossible to correct. High-level validation practices such as comparison with census data do exist but have severe limitations when it comes to validating home detection practices. The consequence is that validation of home detection practices need much more effort and investigation than what has been carried out so far, making proper, systematic assessments of the errors and uncertainties related to this step in the analysis non-existent.

References

[1] Aguiléra, V., Allio, S., Transportation, C. M. P. o. t. t., and undefined 2014 (2014). Territory analysis using cell-phone data. In *Transport Research Arena*, pages 1–10, Paris.

[2] Ahas, R., Aasa, A., Mark, Ü., Pae, T., and Kull, A. (2007a). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, 28(3):898–910.

[3] Ahas, R., Aasa, A., Roose, A., Mark, Ü., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

[4] Ahas, R., Aasa, A., Silm, S., and Tiru, M. (2007b). Mobile Positioning Data in Tourism Studies and Monitoring: Case Study in Tartu, Estonia. In *Information and Communication Technologies in Tourism 2007*, pages 119–128. Springer Vienna, Vienna.

[5] Ahas, R., Mark, Ü., Järv, O., and Nuga, M. (2006). Mobile positioning in sustainability studies: the social positioning method in studying commuter's activity spaces in Tallinn. In *The Sustainable City IV: Urban Regeneration and Sustainability*, volume 1 of *WIT Transactions on Ecology and the Environment, Vol 93*, pages 127–135, Southampton, UK. WIT Press.

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

[7] Alessandretti, L., Sapiezynski, P., Lehmann, S., and Baronchelli, A. (2017). Multi-scale spatio-temporal analysis of human mobility. *PLOS ONE*, 12(2):e0171686.

[8] Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., and Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*.

[9] Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240–250.

[10] Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., and Batty, M. (2015). Constructing cities, deconstructing scaling laws. *Journal of the Royal Society, Interface*, 12(102):20140745.

[11] ARCEP (Autorité de régulation des communications électroniques et des postes) (2008). Le Suivi des Indicateurs Mobiles - les chiffres au 31 décembre 2007.

[12] Ashbrook, D. and Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286.

[13] Bajardi, P., Delfino, M., Panisson, A., Petri, G., and Tizzoni, M. (2015). Unveiling patterns of international communities in a global city using mobile phone data. *EPJ Data Science*, 4(3):1–17.

[14] Baldacci, E., Buono, D., Kapetanios, G., and Krische, S. (2016). Big Data and Macroeconomic Nowcasting: from data access to modelling. Technical report, Eurostat, Luxembourg.

[15] Bandini, S., Manzoni, S., Vizzari, G., Bandini, S., Manzoni, S., and Vizzari, G. (2009). Agent Based Modeling and Simulation: An Informatics Perspective. *Journal of Artificial Societies and Social Simulation*, 12(4).

[16] Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.

[17] Barabasi, A.-L. and Bonabeau, E. (2003). Scale-Free Networks. *Scientific American*, 288:60–69.

[18] Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., and Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, 734:1–74.

[19] Barnes, T. J. and Wilson, M. W. (2014). Big Data, social physics, and spatial analysis: The early years. *Big Data & Society*, 1(1):205395171453536.

[20] Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3):274–279.

[21] Beckers, J., Vanhoof, M., and Verhetsel, A. (2017). Returning the particular: Understanding hierarchies in the Belgian logistics system. *Journal of Transport Geography*, In Press.

[22] Bettencourt, L. M., Lobo, J., and Strumsky, D. (2007). Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Policy*, 36(1):107–120.

[23] Bharti, N., Lu, X., Bengtsson, L., Wetter, E., and Tatem, A. J. (2015). Remotely measuring populations during a crisis by overlaying two data sources. *International Health*, 7(2):90–98.

[24] Blondel, V. D., Decuyper, A., and Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10.

[25] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[26] Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. *Information Technology for Development*, 18(2):107–125.

[27] Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., and Ratti, C. (2015). Choosing the Right Home Location Definition Method for the Given Dataset. In Liu, T.-Y., Scollon, C. N., and Zhu, W., editors, *7th International Conference on Social Informatics (SocInfo)*, pages 194–208, Beijing. Springer.

[28] Boyd, D. and Crawford, K. (2012). Critical Question For Big Data: Provociations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

[29] Buchanan, M. (2007). *The social atom : why the rich get richer, cheaters get caught, and your neighbor usually looks like you*. Bloomsbury USA.

[30] Bulger, M., Taylor, G., and Schroeder, R. (2014). Data-driven business models: challenges and opportunities of big data.

[31] Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2011). Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. *EEE Pervasive Computing*, 10(4):36–44.

[32] Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301–313.

[33] Calabrese, F., Ferrari, L., and Blondel, V. D. (2014). Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Computing Surveys*, 47(2):1–20.

[34] Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

[35] Carrasco, J. A., Hogan, B., Wellman, B., and Miller, E. J. (2008). Collecting Social Network Data to Study Social Activity-Travel Behavior: An Egocentric Approach. *Environment and Planning B: Planning and Design*, 35(6):961–980.

[36] Chang, R. M., Kauffman, R. J., and Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63:67–80.

[37] Chen, C., Bian, L., and Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, 46:326–337.

[38] Chen, C., Ma, J., Susilo, Y., Liu, Y., and Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68:285–299.

[39] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, pages 1082–1090, New York, New York, USA. ACM Press.

[40] Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J. P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., and Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346.

[41] Cottineau, C., Finance, O., Hatna, E., Arcaute, E., and Batty, M. (2018). Defining urban clusters to detect agglomeration economies. *Environment and Planning B: Urban Analytics and City Science*, page 239980831875514.

[42] Cottineau, C., Hatna, E., Arcaute, E., and Batty, M. (2017). Diverse cities or the systematic paradox of Urban Scaling Laws. *Computers, Environment and Urban Systems*, 63:80–94.

[43] Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., and Zook, M. (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2):130–139.

[44] Cranshaw, J., Toch, E., Hong, J., Kittur, A., and Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing - Ubicomp '10*, pages 119–128, New York, New York, USA. ACM Press.

[45] Csáji, B. C., Browet, A., Traag, V., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., and Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473.

[46] Daas, P. J., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, 31(2):249–262.

[47] Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., and Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology Advances in database technology - EDBT '08*, pages 668–677, New York, New York, USA. ACM Press.

[48] De Montjoye, Y.-A., Rocher, L., and Pentland, A. S. (2016). Bandicoot: a python toolbox for mobile phone metadata. *The Journal of Machine Learning Research*, 17(1):6100–6104.

[49] Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J.-M., Krings, G., Gutierrez, T., Blondel, V. D., and Luengo-Oroz, M. A. (2014). Estimating food consumption and poverty indices with mobile phone data.

[50] den Hoed, W. and Russo, A. P. (2017). Professional travellers and tourist practices. *Annals of Tourism Research*, 63:60–72.

[51] Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45):15888–93.

[52] Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A.-L., and Wang, D. (2016). Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 113(26):7047–7052.

[53] Dong, Z.-B., Song, G.-J., Xie, K.-Q., and Wang, J.-Y. (2009). An experimental study of large-scale mobile social network. In *Proceedings of the 18th international conference on World wide web - WWW '09*, page 1175, New York, New York, USA. ACM Press.

[54] Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981):1029–1031.

[55] Expert, P., Evans, T. S., Blondel, V. D., and Lambiotte, R. (2011). Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19):7663–7668.

[56] Fotheringham, A. S. and Wong, D. (1991). The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning A*, 23(7):1025–1044.

[57] Frias-martinez, V., Soto, V., Virseda, J., and Frias-martinez, E. (2013). Can cell phone traces measure social development ? In Vincent Blondel, Adeline Decuyper, Pierre, D., Yves-Alexandre, De Montjoye Jameson, T., Vincent, T., and Dashun, W., editors, *Third Conference on the Analysis of Mobile Phone datasets, NetMob*, pages 62–65, Boston.

[58] Frias-Martinez, V. and Virseda, J. (2012). On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development - ICTD '12*, pages 76–84, New York, New York, USA. ACM Press.

[59] Frias-Martinez, V., Virseda, J., Rubio, A., and Frias-Martinez, E. (2010). Towards large scale technology impact analyses. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development - ICTD '10*, pages 1–10, New York, New York, USA. ACM Press.

[60] Fuller, M. (1979). The estimation of Gini coefficients from grouped data: Upper and Lower Bounds. *Economics Letters*, 3(2):187–192.

[61] Getis, A. and Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3):189–206.

[62] Giannotti, F., Pedreschi, D., Pentland, A., Lukowicz, P., Kossmann, D., Crowley, J., and Helbing, D. (2012). A planetary nervous system for social mining and collective awareness. *European Physical Journal: Special Topics*, 214(1):49–75.

[63] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

[64] Gong, H., Chen, C., Bialostozky, E., and Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 36(2):131–139.

[65] González, M. C., Hidalgo, C. A., and Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

[66] Graham, M. and Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, 3(3):255–261.

[67] Grauwin, S., Szell, M., Sobolevsky, S., Hövel, P., Simini, F., Vanhoof, M., Smoreda, Z., Barabási, A.-L., and Ratti, C. (2017). Identifying and modeling the structural discontinuities of human interactions. *Scientific Reports*, 7:46677.

[68] Hightower, J., Consolvo, S., LaMarca, A., Smith, I., and Hughes, J. (2005). Learning and Recognizing the Places We Go. In Beigl, M., Intille, J., Rekimoto, J., and Tokuda, H., editors, *Ubicomp 2005: Ubiquitous Computing*, pages 159–176. Springer, Berlin, Heidelberg.

[69] Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., and Smoreda, Z. (2013). Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies. In *Geographic Information Science at the Heart of Europe*, pages 247–265. Springer, Cham.

[70] Iqbal, M. S., Choudhury, C. F., Wang, P., and González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74.

[71] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In Lyons, K., Hightower, J., and Huang, E., editors, *Pervasive 2011: Pervasive Computing*, pages 133–151. Springer, Berlin, Heidelberg.

[72] Janzen, M., Vanhoof, M., Smoreda, Z., and Axhausen, K. W. (2018). Closer to the total? Long-distance travel of French mobile phone users. *Travel Behaviour and Society*, 11:31–42.

[73] Järv, O., Ahas, R., and Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38:122–135.

[74] Kang, C., Ma, X., Tong, D., and Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391(4):1702–1717.

[75] Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1–12.

[76] Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. (2009). Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, page L07003.

[77] Kühnert, C. and West, G. B. (2006). Scaling laws in urban supply networks. *Physica A: Statistical Mechanics and its Applications*, 363(1):96–103.

[78] Kung, K. S., Greco, K., Sobolevsky, S., and Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE*, 9(6):e96180.

[79] Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325.

[80] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Social science. Computational social science. *Science (New York, N.Y.)*, 323(5915):721–723.

[81] Lima, A., De Domenico, M., Pejovic, V., and Musolesi, M. (2015). Disease Containment Strategies based on Mobility and Information Dissemination. *Scientific Reports*, 5(1):10650.

[82] Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G., and Cools, M. (2014). Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications*, 41(14):6174–6189.

[83] Louail, T., Lenormand, M., Cantu Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J. J., and Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. *Scientific Reports*, 4(1):5276.

[84] Lu, X., Bengtsson, L., and Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11576–81.

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

[86] Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., and Gabrielli, L. (2015). Small Area Model-Based Estimators Using Big Data Sources. *Journal of Official Statistics*, 31(2):263–281.

[87] Matamalas, J. T., De Domenico, M., and Arenas, A. (2016). Assessing reliable human mobility patterns from higher order memory in mobile communications. *Journal of the Royal Society, Interface*, 13(121):20160203.

[88] Menezes, T. and Roth, C. (2017). Natural Scales in Geographical Patterns. *Scientific Reports*, 7:45823.

[89] Nanavati, A. A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., and Joshi, A. (2006). On the structural properties of massive telecom call graphs. In *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*, pages 435–444, New York, New York, USA. ACM Press.

[90] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. (2012). A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE*, 7(5):e37027.

[91] Nurmi, P. and Bhattacharya, S. (2008). Identifying Meaningful Places: The Non-parametric Way. In Indulska, J., Patterson, D., Rodden, J., and Ott, M., editors, *Pervasive 2008: Pervasive Computing*, pages 111–127. Springer, Berlin, Heidelberg.

[92] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., de Menezes, M. A., Kaski, K., Barabási, A.-L., and Kertész, J. (2007a). Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179–179.

[93] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007b). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7332–7336.

[94] Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667.

[95] Pappalardo, L., Pedreschi, D., Smoreda, Z., and Giannotti, F. (2015a). Using big data to study the link between human mobility and socio-economic development. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 871–878. IEEE.

[96] Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., and Giannotti, F. (2013). Understanding the patterns of car travel. *The European Physical Journal Special Topics*, 215(1):61–73.

[97] Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A.-L. (2015b). Returners and explorers dichotomy in human mobility. *Nature Communications*, 6(1):8166.

[98] Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., and Giannotti, F. (2016). An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1-2):75–92.

[99] Pentland, A. (2015). *Social physics : how social networks can make us smarter*. Penguin.

[100] Phithakkitnukoon, S., Smoreda, Z., and Olivier, P. (2012). Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. *PLoS ONE*, 7(6):e39253.

[101] Poletto, C., Tizzoni, M., and Colizza, V. (2013). Human mobility and time spent at destination: Impact on spatial epidemic spreading. *Journal of Theoretical Biology*, 338:41–58.

[102] Pornet, C., Delpierre, C., Dejardin, O., Grosclaude, P., Launay, L., Guittet, L., Lang, T., and Launoy, G. (2012). Construction of an adaptable European transnational ecological deprivation index: The French version. *Journal of Epidemiology and Community Health*, 66(11):982–989.

[103] Pumain, D., Paulus, F., Vacchiani-Marcuzzo, C., and Lobo, J. (2006). An evolutionary theory for interpreting urban scaling laws. *Cybergeo*, (343).

[104] Ranjan, G., Zang, H., Zhang, Z.-L., and Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44.

[105] Ratti, C., Frenchman, D., Pulselli, R. M., and Williams, S. (2006). Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748.

[106] Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., and Strogatz, S. H. (2010). Redrawing the Map of Great Britain from a Network of Human Interactions. *PLoS ONE*, 5(12):e14248.

[107] Raun, J., Ahas, R., and Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management*, 57:202–212.

[108] Reades, J., Calabrese, F., and Ratti, C. (2009). Eigenplaces: analysing cities using the space – time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836.

[109] Reardon, S. F. (2009). Measures of ordinal segregation. In Flückiger, Y., Reardon, S. F., and Silber, J., editors, *Occupational and Residential Segregation*, pages 129–155. Emeral Group Publishing Limited, researcg o edition.

[110] Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M. C., and Toroczkai, Z. (2014). Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature Communications*, 5:5347.

[111] Ricciato, F., Widhalm, P., Craglia, M., and Pantisano, F. (2015). Estimating population density distribution from network-based mobile phone data. Technical report, Joint Research Center European Commission, Luxembourg.

[112] Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–1123.

[113] Rubrichi, S., Smoreda, Z., and Musolesi, M. (2018). A comparison of spatial-based targeted disease mitigation strategies using mobile phone data. *EPJ Data Science*, 7:1–17.

[114] Schläpfer, M., Bettencourt, L. M. A., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G. B., and Ratti, C. (2014). The scaling of human interactions with city size. *Journal of the Royal Society, Interface*, 11(98):20130789.

[115] Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data & Society*, 1(2):205395171456319.

[116] Schwanen, T. (2017). Geographies of transport II. *Progress in Human Geography*, 41(3):355–364.

[117] Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., and Leskove, J. (2008). Mobile call graphs: beyond power-law and lognormal distributions. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, pages 596–604, New York, New York, USA. ACM Press.

[118] Shalizi, C. R. (2011). Scaling and Hierarchy in Urban Economies.

[119] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

[120] Shen, L. and Stopher, P. R. (2014). Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews*, 34(3):316–334.

[121] Silm, S. and Ahas, R. (2010). The Seasonal Variability of Population in Estonian Municipalities. *Environment and Planning A*, 42(10):2527–2546.

[122] Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

[123] Sobolevsky, S., Campari, R., Belyi, A., and Ratti, C. (2014). General optimization technique for high-quality community detection in complex networks. *Physical Review E*, 90(1):012811.

[124] Sobolevsky, S., Szell, M., Campari, R., Couronné, T., Smoreda, Z., and Ratti, C. (2013). Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLoS ONE*, 8(12):e81707.

[125] Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010a). Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823.

[126] Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010b). Limits of predictability in human mobility. *Science (New York, N.Y.)*, 327(5968):1018–21.

[127] Soto, V. and Frías-Martínez, E. (2011). Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch - HotPlanet '11*, pages 17–22, Bethesda, Maryland, USA. ACM Press.

[128] Sridharan, A. and Bolot, J. (2013). Location patterns of mobile users: A large-scale tudy. In *2013 Proceedings IEEE INFOCOM*, pages 1007–1015. IEEE.

[129] Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J., and Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society, Interface*, 14(127):20160690.

[130] Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P., and Scholten, H. (2013). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, 78(2):223–243.

[131] Szell, M., Sinatra, R., Petri, G., Thurner, S., and Latora, V. (2012). Understanding mobility in a social petri dish. *Scientific Reports*, 2(1):457.

[132] Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, 4:170004.

[133] Tibély, G., Kovanen, L., Karsai, M., Kaski, K., Kertész, J., and Saramäki, J. (2011). Communities and beyond: Mesoscopic analysis of a large social network with complementary methods. *Physical Review E*, 83(5):056125.

[134] Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., Smoreda, Z., González, M. C., and Colizza, V. (2014). On the Use of Human Mobility Proxies for Modeling Epidemics. *PLoS Computational Biology*, 10(7):e1003716.

[135] Toole, J. L., Herrera-Yaqüe, C., Schneider, C. M., and González, M. C. (2015). Coupling human mobility and social ties. *Journal of the Royal Society, Interface*, 12(105):20141128.

[136] Toole, J. L., Ulm, M., González, M. C., and Bauer, D. (2012). Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12*, pages 1–18, Beijing, China. ACM Press.

[137] Trasarti, R., Olteanu-Raimond, A.-M., Nanni, M., Couronné, T., Furletti, B., Giannotti, F., Smoreda, Z., and Ziemlicki, C. (2015). Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy*, 39(3-4):347–362.

[138] Travers, J. and Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443.

[139] Urry, J. (2004). Small Worlds and the New 'Social Physics'. *Global Networks*, 4(2):109–130.

[140] Vanhoof, M., Combes, S., and De Bellefon, M.-P. (2017a). Mining mobile phone data to detect urban areas. In Petrucci, A. and Verde, R., editors, *Statistics and Data Science: New challenges, new generations, SIS 2017*, pages 1005–1012, Firenze. Firenze University Press.

[141] Vanhoof, M., Hendrickx, L., Puussaar, A., Verstraeten, G., Ploetz, T., and Smoreda, Z. (2017b). Exploring the use of mobile phones during domestic tourism trips. *Netcom*, 31(3/4):335–372.

[142] Vanhoof, M., Reis, F., Ploetz, T., and Smoreda, Z. (2018a). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, page In Press.

[143] Vanhoof, M., Schoors, W., Van Rompaey, A., Ploetz, T., and Smoreda, Z. (2018b). Comparing Regional Patterns of Individual Movement Using Corrected Mobility Entropy. *Journal of Urban Technology*, pages 1–35.

[144] Venturini, T., Jensen, P., and Latour, B. (2015). Fill in the Gap: A New Alliance for Social and Natural Sciences. *Journal of Artificial Societies and Social Simulation*, 18(2).

[145] Wang, X.-W., Han, X.-P., and Wang, B.-H. (2014). Correlations and Scaling Laws in Human Mobility. *PLoS ONE*, 9(1):e84954.

[146] Wang, Z., He, S. Y., and Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 11:141–155.

[147] Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., and Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences of the United States of America*, 115(14):3529–3537.

[148] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[149] Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science (New York, N.Y.)*, 338(6104):267–70.

[150] Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., and Dobra, A. (2015). Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. *PLOS ONE*, 10(7):e0133630.

[151] Wilson, A. (1967). A statistical theory of spatial distribution models. *Transportation Research*, 1(3):253–269.

[152] Wilson, A. (2010). Entropy in Urban and Regional Modelling: Retrospect and Prospect. *Geographical Analysis*, 42(4):364–394.

[153] Wolf, J., Guensler, R., and Bachman, W. (2001). Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record: Journal of the Transportation Research Board*, 1768:125–134.

[154] Wong, D. (2009). The Modifiable Areal Unit Problem (MAUP). In Fotheringham, A. S. and Rogerson, P., editors, *The SAGE handbook of spatial analysis*, chapter 7, pages 105–123. SAGE Publications, London.

[155] Yan, X.-Y., Zhao, C., Fan, Y., Di, Z., and Wang, W.-X. (2014). Universal predictability of mobility patterns in cities. *Journal of the Royal Society, Interface*, 11(100):1–7.

[156] Yuan, Y. and Raubal, M. (2012). Correlating mobile phone usage and travel behavior – A case study of Harbin, China. *Computers, Environment and Urban Systems*, 36(2):118–130.

[157] Zipf, G. K. (1946). Some Determinants of the Circulation of Information. *The American Journal of Psychology*, 59(3):401–421.

The deployment of indicators from mobile phone data should, ideally, be backed by a thorough assessment of their geographical veracity.

Maarten Vanhoof