European Commission

JRC TECHNICAL REPORT

Estimating Population Density Distribution from Network-based Mobile Phone Data

Fabio Ricciato, Peter Widhalm,
Massimo Craglia and Francesco Pantisano

2015

Joint Research Centre

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY

European Commission
Joint Research Centre
Institute for Environment and Sustainability

Contact information
Dr. Massimo Craglia
Address: Joint Research Centre,
Via Enrico Fermi 2749, TP 262,
21027 Ispra (VA), Italy
E-mail: massimo.craglia@jrc.ec.europa.eu
Tel.: +39 0332 78 6269

JRC Science Hub: https://ec.europa.eu/jrc

Legal Notice
This publication is a Technical Report by the Joint Research Centre, the European Commission's in-house science service. It aims to provide evidence-based scientific support to the European policy-making process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

JRC96568

© European Union, 2015

Reproduction is authorised provided the source is acknowledged.

Abstract

In this study we address the problem of leveraging mobile phone network-based data for the task of estimating population density distribution at pan-European level. The primary goal is to develop a methodological framework for the collection and processing of network-based data that can be plausibly applied across multiple Mobile Network Operators (MNOs). The proposed method exploits more extensive network topology information than is considered in most state-of-the-art literature, i.e., (approximate) knowledge of cell coverage areas is assumed instead of merely cell tower locations. A distinguishing feature of the proposed methodology is the capability of taking as input a combination of cell-level and Location Area-level data, thus enabling the integration of data from Call Detail Records (CDR) with other network-based data sources, e.g., Visitor Location Register (VLR). Different scenarios are considered in terms of input data availability at individual MNOs (CDR only, VLR only, combinations of CDR and VLR) and for multi-MNO data fusion, and the relevant tradeoff dimensions are discussed. At the core of the proposed method lies a novel formulation of the population distribution estimation as a Maximum Likelihood estimation problem. The proposed estimation method is validated for consistency with artificially- generated data in a simplified simulation scenario. Final considerations are provided as input for a future pilot study validating the proposed methodology on real-world data.

Extraction of population density distribution from network-based mobile phone data

Fabio Ricciato[1], Pete Widhalm[2], Massimo Craglia[3] and Francesco Pantisano[3].

July 22, 2015

[1]Fabio Ricciato is with the University of Ljublijana, Faculty of Computer and Information Science, Ljublijana, Slovenia, and with the Austrian Institute of Technology (AIT), Mobility Department, Vienna, Austria. Email: fabio.ricciato@fri.uni-lj.si

[2]Peter Widhalm is with the Austrian Institute of Technology (AIT), Mobility Department, Vienna. Email peter.widhalm@ait.ac.at

[3]Massimo Craglia and Francesco Pantisano are with the Institute for Environment and Sustainability of the Joint Research Centre (JRC), European Commission, Ispra, Italy. Email: {massimo.craglia, francesco.pantisano}@jrc.ec.europa.eu

Executive Summary

The vast majority of people nowadays carries (at least) a mobile phone, and every mobile phone is logically "attached" to the network infrastructure of a Mobile Network Operator (MNO). The MNO infrastructure is composed of multiple radio "cells" of different size — ranging from tens of meters up to several kilometers — and at any time the phone is logically "camped" to one cell. Upon certain events — e.g., when initiating or receiving a phone call or SMS — the mobile phone reveals its current cell location to the network, and the latter stores this information (permanently) in the so-called Call Detail Record (CDR) database for billing purposes. Moreover, radio cells are hierarchically organised into larger spatial entities called Location Areas (LAs): whenever the phone moves from one LA to another, it informs the network, and the latter stores this information (temporarily) in the so-called Visitor Location Register (VLR) as a routine network operation. Therefore, both types of network-based data, CDR and VLR, embed information about the location of every mobile phone at the level of radio cells and/or LAs. Several research work in the last decade has shown that, in principle, it is possible to leverage network-based data from MNO to infer human mobility patterns (e.g., periodic commutes, favorite locations, average speed). The majority of this work has focused exclusively on CDR data, and was based on sample dataset from a single MNO.

In this study we address the problem of leveraging network-based data (CDR and/or VLR) for the task of estimating population density distribution at pan-European level. The primary goal of the study was to develop a methodological framework for the collection and processing of network-based data that can be plausibly applied *across multiple MNOs*. The main challenge of this task is to design a methodology that achieves *general applicability* in a highly heterogenous scenario, where several technical details of network configuration and data organisation remain highly MNO-specific. To this aim, we pursue the design of an "resilient" methodological framework, whereas the core set of functions does not rely on any non-standard MNO-specific configuration — hence, it can be implemented by any MNO — and, at the same time, it is flexible enough to *optionally* leverage additional MNO-specific network and/or data characteristics so as to improve the fidelity of the final results to the "ground truth". Owing to such flexibility, the proposed methodology lends itself to be extended and further refined, by taking advantage of the future evolutions of mobile network infrastructures (e.g., availability of additional data sources).

The main outcome of this study is a proposal for a *systematic methodological framework for population density estimation based on mobile network data*. In our intention, this shall represent an initial reference for future discussion with and between experts from MNOs and public institutions, with the goal of ultimately consolidating a realistic implementation plan. Along the process, it is likely that the methodology proposed in this document will undergo extensions and refinements, and in general shall benefit from technical inputs from MNO expert.

The methodology developed in this study yields several important novelties with respect to the current state-of-the-art work in this field. In particular, we highlight the following:

- Use of extended network topology data: the proposed methodology takes in input (an approximation of) the whole coverage area of the generic radio cell, not only the antenna tower location. Based on such data, a novel tessellation scheme is proposed that yields more accurate results than the the classic Voronoi tessellation method.
- Beyond CDR-only data: the proposed method can be casted in different implementation scenarios with different combinations of cell-level and LA-level location data, from both CDR and/or VLR databases (or other proprietary systems). In this way, it supports the CDR-only scenario — that is likely the preferred option by most MNOs — but at the same time enables (and motivates) initial experimentation with combined CDR/VLR data fusion.
- Multi-MNO: the proposed method is designed upfront for application across different MNOs, and for the fusion of data from multiple MNOs serving the same spatial region (e.g., same country).

In order to facilitate the reading for non-technical experts, the present report contains an initial introductory section about mobile networks. In this sense, the report is self-contained and does not require frequent reference to external specialised technical sources. The proposed estimation method is validated for consistency with artificially generated data in a simplified simulation scenario. A set of final considerations are provided as input for the process of preparing a future inter-MNO pilot study for the proof-of-concept validation on real-world data.

Contents

Foreword

There is an increasing recognition that good policy should be grounded on solid scientific evidence that is traceable, open, and participated. This is the rationale of the many open data initiatives across the world, including the open government partnership[1] launched in 2011 to promote more open and accountable governance, and the Research Data Alliance[2] supporting open research data. The European Union is at the forefront of these initiatives and INSPIRE[3] is the legal framework adopted in 2007 to make existing environmental and spatial data more visible, interoperable, and shared among public authorities to support environmental policy and policies that affect the environment.

The Joint Research Centre (JRC) of the European Commission, as overall technical coordinator of INSPIRE, is supporting the European Member States in the implementation of this key policy. It is also assessing the interoperability between INSPIRE and the increased heterogeneity of data sources that can support public policy, such as data from space, commercial transactions, sensor networks, the Internet, and the public, including social media. The Big Data revolution is creating many opportunities but also posing new challenges to public authorities, including issues of data access, analytical methodologies, ethics and trust. The increasing shift in knowledge about society from the public to the private sector requires new partnerships to ensure that sound policy is still based on relevant and timely data. For example, many environmental and social policies need to have a good understanding about population distribution to prepare strategies and assess impacts. Natural disasters, like floods and earthquakes, are obvious cases but urban and regional planning, environmental impact assessment, and the effects of environmental exposure on health are equally important areas where using census and administrative data about the resident population at night may considerably misrepresent reality at different times of day and night. In this respect, one potential source of much more timely and accurate data about the population distribution could come from mobile network operators, and the scientific literature shows many cases in which this data was successfully exploited. Several European National Statistical Institutes are exploring this data source to complement their own data but access to data is often difficult and only successful on the basis of individual ad-hoc arrangements. This is potentially creating inequalities in the knowledge base on which to develop and assess European policy.

To address this challenge and support the activities of the European Statistical System Big Data Task Force, the JRC commissioned this study to the Austrian Institute of Technology on a general methodology enabling mobile network operator to process and integrate different types of network data in their possess (e.g., anonymised Call Detail Records, Visitor Location Register data) with

[1] http://www.opengovpartnership.org/
[2] https://rd-alliance.org/
[3] http://inspire.ec.europa.eu/index.cfm

the aim of estimating population density, for public policy purposes. The methodology described in this report has been designed to be flexible and scalable, mindful of commercial sensitivity, as well as the need to protect personal privacy and confidentiality. The proposed methodology has been tested with a sample of synthetic data and, the next steps following publication of the report and gathering of feedback from interested parties, will be to test it with partner mobile network operators. In this way feasibility and costs can be properly assessed and become the basis for a dialogue with all willing operators in Europe with a view to define a common framework for data access and use to support public policy.

Chapter 1

Essentials of mobile phone networks and network-based data

1.1 Mobile Communication Technologies

A mobile cellular network is a large-scale communication network that provides wireless connectivity over a large area in which Mobile Stations (MS), e.g., mobile phones, are deployed. It consists of multiple Public Land Mobile Networks (PLMN), each one spanning a country's territory and typically being operated by a single Mobile Network Operator (MNO). Hereafter we will use the term "MNO" to refer both to the technical/administrative entity (the "network operator") and to the associated infrastructure (the operated network, i.e., the PLMN).

For the past 30 years, mobile communication technology has been progressively evolving, under different international standards which have not always been compatible across different countries. While the first generation of cellular networks was developed in the 80s within national systems (notably in Japan and the USA) with consequent cross-country compatibility issues, mobile communications became a worldwide mass market during the 90s with the Global System for Mobile Communications (GSM) system developed by the European Telecommunications Standards Institute (ETSI). GSM networks represent the"second generation" (2G) of cellular systems, and were designed for the transition from analog to digital transmission, which ultimately enabled voice and data traffic coexistence (e.g., Short Message Services (SMS)). In a successive evolution and in light of the rise of data traffic demand, it was later upgraded (with the introduction of GPRS and EDGE) to enhance packet-switched data communication. The universality of the technology standards is, therefore, a relatively recent achievement, pioneered at European level with the Global System for Mobile Communications (GSM) and followed by worldwide standard Universal Mobile Telecommunications System (UMTS). UMTS – the "third-generation" (3G) of mobile communication systems – was launched in 2004 for supporting Internet multimedia services (e.g., web browsing, video streaming). Similarly to GSM, UMTS was later upgraded to higher quality of service standards with the introduction of High Speed Packet Access (HSPA), and UMTS penetration and coverage are now pretty advanced throughout Europe. The "fourth-generation" (4G) system, called LTE (Long Term Evolution), has been rolled out in Europe in 2011 and it promises to meet the requirements of upcoming communication network concepts, including the Internet-of-things (IoT), smart cities, smart grid, and vehicular networks.

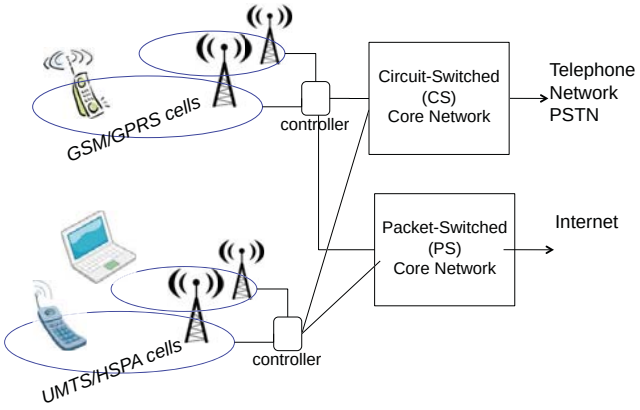


Figure 1.1: High-level view of a combined 2G/3G network.

The methodology proposed in this document is based on GSM and UMTS standards and network architecture, although, with opportune modifications, it can be adapted to other mobile communication standards, such as LTE. Hence, throughout the document, we will purposefully omit technical details (e.g. additional components of the network architecture), under the assumption that the method developed here can also be adapted to 4G network architectures.

Hereafter we will use the term “2G” to refer to the ”GSM” access and “3G” for UMTS/HSPA access. Most operators maintain both a 2G and 3G network infrastructure, and therefore we will refer to a single “2G/3G” infrastructure, like the one depicted in Fig. 1.1.

1.2 Mobile Network Architecture

The network architecture is composed of two main parts: the Radio Access Network (RAN) and the Core Network (CN). The RAN includes all the “peripheral” components, i.e. the base stations [1] that transmit / receive on the radio link from / to the MSs, and their respective controllers — called Base Station Controller (BSC) in GSM and Radio Network Controller (RNC) in UMTS. The CN includes “back-end” equipments, whose physical location is normally concentrated at a few sites.

It should be noted that there are actually two distinct CNs domains: the Circuit-Switched (CS), mainly for voice calls, and the Packet-Switched (PS) for data calls. The resulting high-level architecture is sketched in Fig. 1.1. The network element that connect the CN to the RAN is the Mobile Switching Center (MSC) in the CS domain, and the SGSN in the PS domain. At any given time, a generic MS can be logically “attached” to the CS domain, to the PS domain, or both. Since our primary focus is on 2G/3G MSs that support voice services (as this are more likely associated to persons, as discussed later in Section 1.6) hereafter we will restrict our attention to the CS domain, unless differently specified[2].

[1] The term “base station” is used hereafter to refer to jointly to the Base Transceiver Station (BTS) in GSM and to the Node-B in UMTS.

[2] The distinction between CS and PS domains is slowly vanishing, with the progressive introduction of integrated MSC/SGSN equipments. However, for the purpose of this study it is useful to keep in mind the *logical* separation

In modern networks, 2G and 3G systems coexist over the same infrastructure, as they operate on different portions of the frequency spectrum, i.e., different bands. Every MNO is assigned a different sub-band (or set thereof) for each system. Therefore, a generic point in space is generally serviced by different radio access technologies (2G and 3G) and by multiple MNOs. However, each MS can be "attached" only to one MNO and one access technology at any given time[3].

1.3 Cells

We now introduce the notion of "radio cell", or simply "cell". In cellular networks, geographical radio coverage is provided by a multitude of base stations distributed across the serviced area. Each base station services one "cell"[4]. Each base station services a limited portion of space, called "cell coverage area", or simply a "cell". In turn, only MS terminals within a cell can connect to the associated base station.

The transmissions from each base station are optimised according to a set of modulation parameters (e.g., carrier frequency in 2G, spreading code in 3G, antenna settings, transmit power) that ultimately affect the shape of the cell. Also, in order to avoid interference, each cell operates on a preassigned frequency band, which is different from that of the adjacent cells. Such a frequency band allocation pattern, which is regularly repeated all over the network, can be described as a chromatic range. Therefore, adjacent cells within a cluster can be denoted with different"colours", indicating the operating frequency band. Finally, every point in space may be "covered" by multiple cells of different colours.

Moreover, due to the different transmission settings, cells may have different shapes and sizes. The largest cells are found in 2G, with diameter in the order of a few tens of kilometers. In urban and suburban areas, cells areas tend to cover distances between hundreds of meters (micro-cells) and a few kilometers. Smaller cells (pico-cell and femto-cells) can be deployed at specific high-density points, both outdoor and indoor, such as in shopping malls, train stations, airports. Generally speaking, within each technology (2G, 3G and 4G) the cell density determines the local network capacity, i.e., the maximum amount of data traffic that a radio network can deliver. The latter depends on the spatial density of people, and on the intensity of their individual traffic (e.g., frequency and duration of phone calls and data connections). For this reason, areas with high population density (urban areas, especially business districts) will be typically covered by many small cells (possibly in addition to few large "umbrella cells") while sparsely populated areas (e.g., countryside, forests) will be covered by few large macro-cells (see Fig. 1.2). Hence, in such a scenario, the spatial granularity of cellular coverage varies from tens of meters (in hot-spots) to hundreds of meters (in urban areas) up to tens of kilometers[5] (in the countryside), depending primarily on the *density* of people, and secondarily on their traffic intensity. Since density and intensity are time-varying — following the typical daily and weekly cycles of human activity — the

between CS and PS domains.

[3]The MS refers to the combination of one Subscriber Identity Module (SIM) and one Mobile Equipment (ME). A mobile phone with dual-sim is therefore considered as two distinct MSs.

[4]Sometimes the term "sector" is used to refer to an individual cell, especially in GSM jargon. Throughout this paper, we use the terms sector and cell interchangeably. Also, for the sake of a simplified terminology, a single 3-sector BTS will be considered as a bundle of 3 co-located base stations.

[5]The maximum distance between the base station and a generic MS attached to it is 35 km.

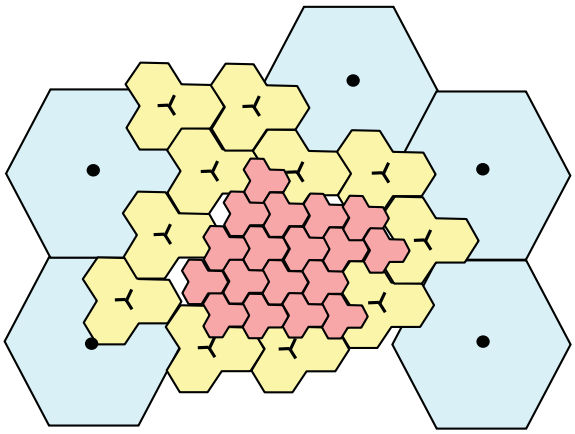


Figure 1.2: Example of multi-layer cell coverage, with increasing cell sizes (and decreasing cell density) from inner towards outer city areas.

network coverage tends to be designed based on their peak values.

Due to the heterogeneous factors discussed above, real cellular networks do not exhibit a regular pattern, hence cell coverage areas can be estimated only approximately. However, for the purposes of this work, it is sufficient to assume that every MNO knows, at least approximately, the expected coverage area for each cell. This information, for example, can be obtained from field measurements and/or from simulations conducted as part of the radio planning and optimization processes. In the worst case, a coarse estimation of the cell coverage area can be derived from antenna configuration parameters (e.g., antenna height, beam-width, tilt).

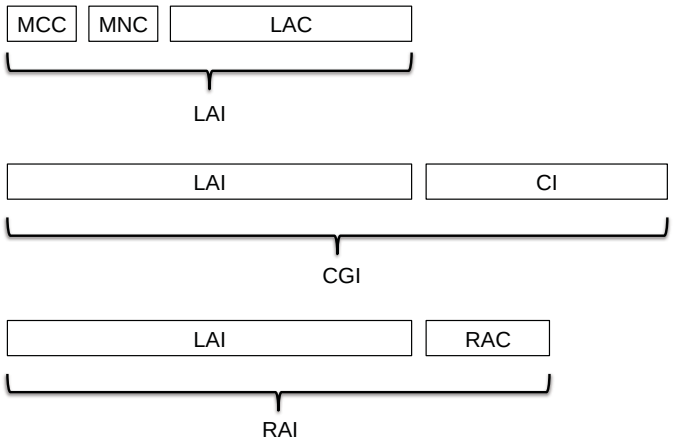
Every cell has an associated unique identifier, the Cell Global Identification (CGI), that is broadcast by the base station in the whole cell area. As shown in Fig. 1.3, the CGI has a prefix structure that allows the MS to immediately identify the country, the MNO and the Location Area (introduced below) to which the cell belongs.

The cell area, as defined insofar, should be distinguished from the position of the antenna tower. The tower position can be either outside or inside the cell coverage area, as depicted in Fig. 1.4 for cells A and C . Notably, more base stations can share the same tower, meaning that cells with different areas (possibly but not necessarily overlapping) might be associated to the same tower position, as seen in Fig. 1.4 for cells A and B.

Upon occurrence of certain events (e.g., starting a phone call), the network learns the cell-level location of a generic MS, and stores the corresponding cell identifier — namely, the CGI— in some internal database, as discussed below in §1.5. In other words, cell-level locations are encoded in the form of CGI values.

1.4 Location Areas (LAs)

Neighboring cells from the same MNO are logically grouped into so-called Location Areas (LA). Every LA is identified by a unique Location Area Code (LAC) that, together with the MNO identifier, forms the Location Area Identity (LAI) as sketched in Fig. 1.3. The grouping of cells into LAs is



- MCC = Mobile Country Code (3 decimal digits)
- MNC = Mobile Network Code (2-3 decimal digits)
- LAC = Location Area Code (16 bit)
- CI = Cell Identifier (16 bit)
- RAC = Routing Area Code (8 bit)
- LAI = Location Area Identity
- CGI = Cell Global Identifier
- RAI = Routing Area Identifier

Figure 1.3: Structure of unique identifiers for Location/Routing Areas and Cells.

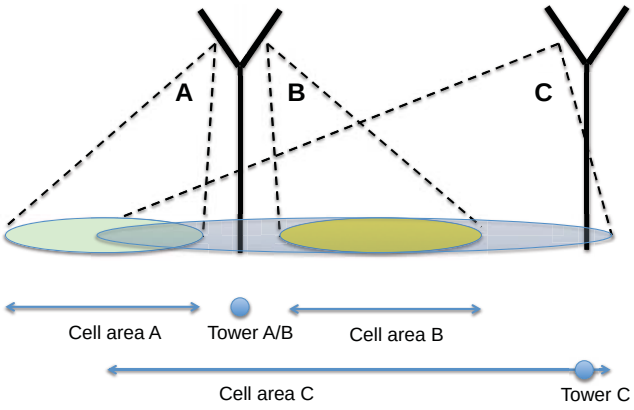


Figure 1.4: Examples of cell areas and tower positions.

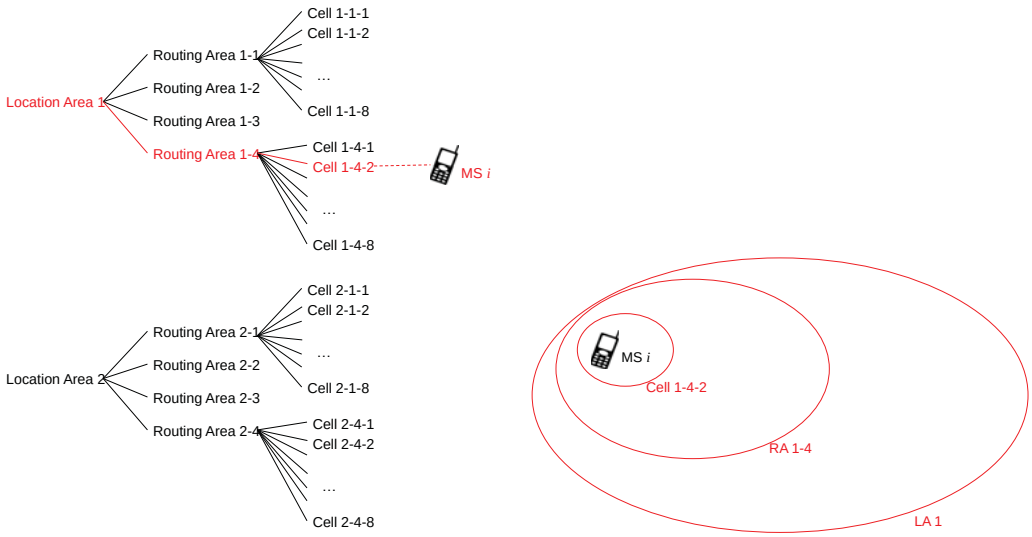


Figure 1.5: Hierarchical relation between LA, RA and individual cells.

decided by the MNO and is completely independent from the "colour" of each cells. The union of all cell areas belonging to the same LA (equivalently: sharing the same LAC prefix) defines the geographical "footprint" of the LA, i.e., the LA-level location. In practice, the cell-to-LA assignment is accomplished implicitly when configuring the CGI: all cells belonging to the same LA (and only those) are assigned CGI with the same common LAC prefix. Thanks to this prefix structure, a moving MS can easily recognise whether a cell change involved a LA changes, by simply comparing the LAC prefix of the new and old cells. When moving to a new LA, the MS must report this event to the network that stores the new LAI in an internal database (the VLR, introduced below). In other words, the LA-level locations are encoded in the form of LAI values.

The notion of LA was introduced in 2G. With the deployment of 3G cellular systems, the additional notion of "Routing Area" (RA) has been introduced. Accordingly, in the PS domain, every LA may be further divided into smaller sub-groups (up to 8) called "Routing Areas" (RA)[6]. The hierarchical relation between a cell, the outer LA and the (intermediate) RA is depicted in Fig. 1.5. Furthermore, the new term "Tracking Area" (TA) has been introduced in 4G. To keep the discussion simple, we will refer hereafter only to LAs, with the understanding that the more spatially accurate RA (or TA) information could be used instead of LA whenever available.

The typical geographical size of LAs varies across MNOs and between urban and rural areas. Our analysis of data samples from the `OpenCellID` database [2] reveals that the median LA radius is around 10 km in big cities, while non suburban and rural areas the median LA diameter is found in the order of 20-25 km, with values up to 40 km (see Appendix B).

For a thorough understanding of the role of LAs, we need to introduce (a simplified view of) MS states. In a nutshell, every MS can be found at any given time in one of two different states: active

[6] Some operators maintain a 1:1 mapping between RAs and LAs, and the two terms can be used interchangeably. If instead LA are split into smaller RAs, some MSs will be tracked at LA level while others at the (finer) RA level, depending on whether MSs are "attached" or not to the PS domain. It should be noted that, for MSs that are attached to both PS and CS, the RA information could be included (optionally) in the CS VLR associated to the MSC, in addition to the mandatory LA information, due to direct communication between the SGSN and MSC.

or idle. The MS spend most of its time in the "idle" state. It switch to "active" during voice calls and when engaged in the exchange of data packets with the network[7]. It switches to active state also when exchanging signalling messages, without any trigger by the data or voice applications. It is important to remark that, *at any given time, only a small minority of all MSs are found in active state, the vast majority being in "idle" mode [13]*.

There are fundamental differences in the "behaviour" of MS during idle and active states, that translate into different levels of temporal and spatial accuracy when it comes to estimate their location from network-side data, as explained below.

- **MS in idle state**. The MS is logically "attached" to one network[8] but is not assigned any radio resource. The MS "listens" (the broadcast channel of) one cell, but does not transmit. In idle states, decisions are taken autonomously by the MT: which cell to listen, and whether and when to "jump" towards another cell (cell change), is determined autonomously by the MS internal logic, not by the network. The MS decision logic depends on the device vendor and is takes into account local measurements as well as past history.

  By definition, MS in idle mode are passive receivers (i.e., they are not transmitting) therefore the network has no way of detecting a cell change unless the MS decides to report this event explicitly. The MS reports the cell change only when it enters a new LAs, while cell changes inside the same LAs are not reported. In this way, *the network can track the position of idle MSs only at the LA level, not at the cell level*.

- **MS in active state**. The MS is assigned radio resources and is engaged in traffic exchange (voice, data or signalling) to and from the network. In active state, all decisions involving radio resources are taken by network: this includes the determination of channel and cell, as well as whether and when to "jump" (handover) to another channel or cell. In this way, *the network tracks the position of active MSs at the cell level*.

From the above discussion, it should be clear that the network can "observe" the cell-level location of each MS only at some specific times, and with a finite spatial resolution. In other words, given the "real" trajectory of a generic MS, continuous in time and space, the cellular network acts like a sensor that applies some form of *sampling in time* and *quantisation in space*.

1.5 Network-side data

There are several elements and subsystems within the network that maintain information about the MS. Hereafter, we will discuss the ones more relevant for our study.

[7] Having a "data connection" (i.e., a PDP-context in 3G terminology) open does not imply that the MS is in "active" state. In fact, the MS can maintain the connection (logically) open for a long time without (physical) sending or receiving data packets, in which case it would be persist in idle state. Generally speaking, the transition from "active" to "idle" is triggered by a short timeout (typically between 2 and 5 seconds) that is reset upon transmission or reception of new data packets).

[8] Preferably their home MNO, if available, otherwise it will be "roaming" to another MNO

1.5.1 Billing data: Call Detail Records (CDR)

For each voice and data connection (or part of it) the network elements generate "tickets" that are sent to the billing system for charging purposes. The billing system stores these data in large databases, normally in the MNO warehouse. The term "Call Detail Records", and especially its acronym "CDR"[9], is commonly used nowadays to indicate generically all billing records, including those originated from data connections.

The format of CDR is not standardised [3, 15] and there is a great deal of variability across different implementations regarding the type of data contained in every CDR, as well as other details of the CDR generation process (e.g., whether long calls are chunked into multiple CDRs). It is safe to assume that mobile CDR data contain at least the following information:

- International Mobile Subscriber Identifier (IMSI) (possibly encrypted).
- Starting time and duration of the call or connection.
- Type of call or connection (e.g. voice, SMS, data).
- Cell Global Identifier (CGI) of the *starting* cell, where the call or connection was initiated[10].

Additional data might be optionally available for specific CDR implementations. For example, in case of handovers, CDR might include the identifiers of the subsequent visited cells, after the starting cell. This is particularly relevant for long-lasting connections (e.g. always-on data connections for mobile phones). Other additional data include the IMEI, APN (for data connections) etc.

Historically, the CDR data were the first data source used in mobile phone data research, and still the overwhelming majority of studies and research project rely exclusively on CDR (see e.g. the recent survey [14].) This is mainly due to the fact that extracting CDR data for off-line processing is *technically* simple, given the non-volatile nature of such data, as discussed below.

1.5.2 Visitor Location Register (VLR)

The Visitor Location Register (VLR) and the Home Location Register (HLR) are database for subscriber data. The HLR stores the "permanent" subscriber parameter that are logically associated to the Subscriber Identity Module (SIM), like e.g. the IMSI. The HLR is a central module serving the whole MNO network, but is not very relevant for this study.

Basic VLR data

Logically speaking, each Mobile Switching Center (MSC) has its own associated VLR. The VLR contains the "temporary" subscriber data for the MS currently "visiting" this MSC area. The most relevant VLR data for this study are the following *mandatory* fields:

[9] The terms "Call Data Records" and "Charging Data Records" are occasionally found in the literature in association to their common acronym "CDR".

[10] Strictly speaking, this is not a mandatory field [3] but we expect that most if not all MNOs actually include this information in their CDR.

- Location Area Identity (LAI)
- Temporary IMSI (T-IMSI).

These data, and especially the LAI, are used by the basic VLR-based method described later in §3.3.2. In addition to the mandatory fields above, some proprietary VLR implementations support the option of storing additional details, e.g., the time and CGI of the last message received by the MS. In case that such optional data are available, they can be used to considerably improve the spatial accuracy of the VLR method, as discussed later in §3.3.5.

Besides the MSCs, every Serving GPRS Support Node (SGSN) has also an associated VLR. The main difference between the VLR of circuit switching (CS) domain (traditionally associated to voice traffic, at the mobile switching center (MSC)) and those of the packed switching (PS) domain (associated to data traffic at SGSN) is that the latter contain the Routing Area Identity (RAI) field instead of the LAI. A generic MS that is attached to both the CS and PS domains will logically appear in two VLR, one for CS and one for PS. However, the distinction between CS VLR and PS VLR might not be important in practice, since the MSC and its neighbouring SGSN might share a single combined VLR — especially if the MSC and SGSN are themselves combined in a single physical equipment. However, since our focus is on voice-enabled MSs, hereafter we will refer exclusively to the VLR serving the CS domain — or both CS and PS, in case of combined VLR.

The set of all VLR pertaining to all MSC in the MNO network collectively form a distributed database. Therefore, hereafter we will use the singular term “VLR” to refer to the entire set of VLR data across all MSCs.

Augmented VLR data

The standard Mobility Management procedures for 2G and 3G systems foresee the involvement of the MSC and/or SGSN whenever the MS engages in a new data connection, voice call or SMS and in general whenever the MS interacts with the network. During the message exchange between the MS and the MSC/SGSN the latter learns the current MS cell location. Although it is not mandatory for the VLR to record the cell nor the timestamp associated to such message exchange, it is reasonable to expect that certain MNOs might decide to configure their VLR to retain these (optional) data in addition to the mandatory LAI/T-IMSI fields[11].

In this case, the VLR data is enriched with the identifier of the last “observed” cell within the current LA along with the associated timestamp, for every generic MS. Such “augmented” VLR would therefore merge together the two types of data that we have previously encountered, separately, in the basic VLR-only and CDR-only methods: cell-level and LA-level locations. Furthermore, augmented VLR data could provide cell-level location also for MS that did not engage in SMS/voice/data connections, provided that they performed some kind of signalling procedure, e.g. Location Area Update (LAU). In other words, they bear the potential to “observe” the cell-level location of a larger fraction of MS than what is possible with CDR data. The estimation method described later in Chapter 3 is designed to cope with the data heterogeneity deriving from a combination of cell-level and LA-level records.

[11] In fact, the marginal cost of storing this information in the VLR is in general small, and augmented VLR data can be exploited to implement supplementary (non standard) functions and/or certain forms of MNO-specific optimisations.

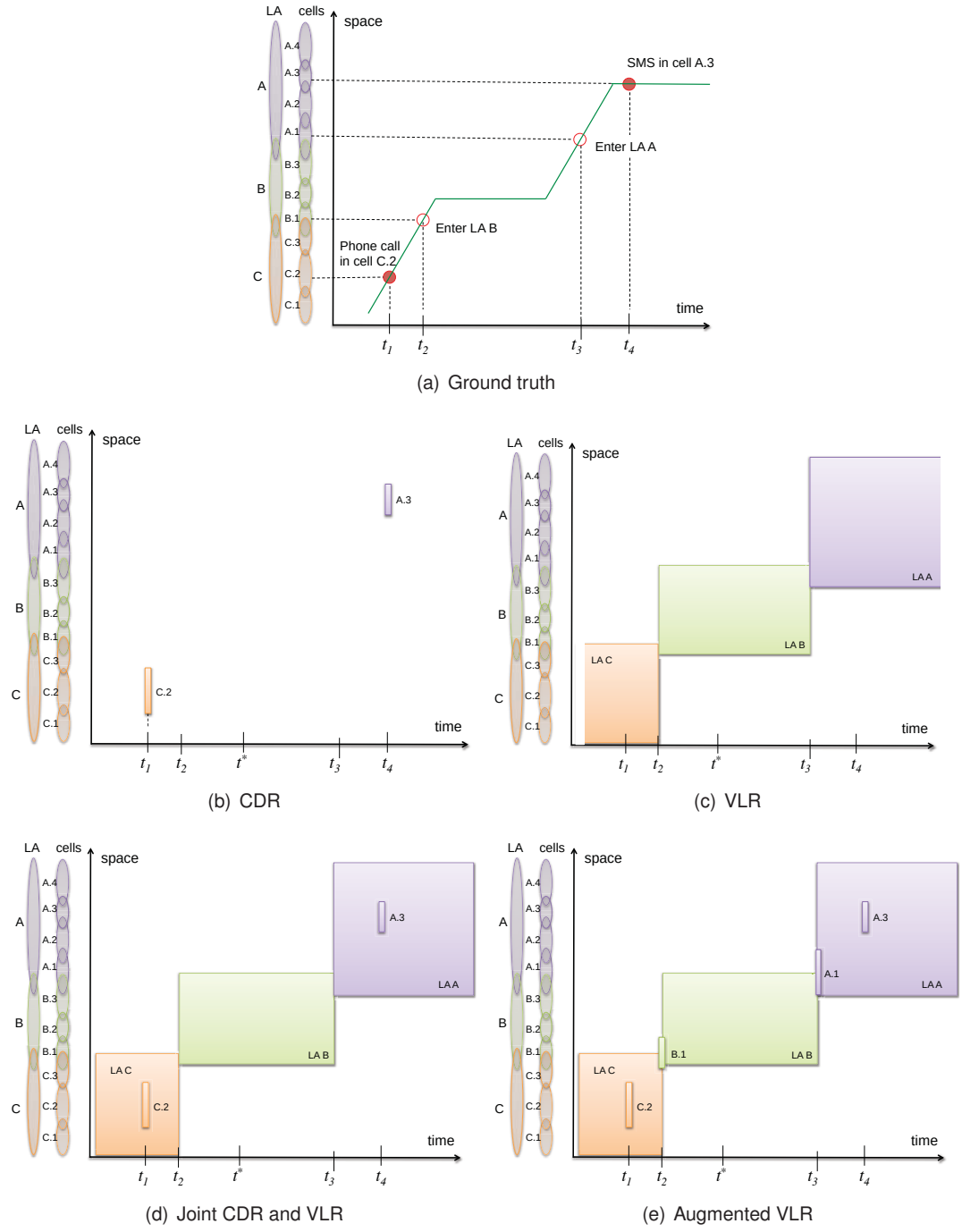


Figure 1.6: Schematic representation of observed trajectory for different network-based data.

1.5.3 Other systems

For the sake of completeness we mention below additional systems that contain network-side data but are not in the focus of this study.

- **Customer Database.** Every MNO maintains a data warehouse with private customer data. These are necessary e.g. for administrative, accounting and contractual purposes. The customer database is not to be confused with the HLR: the latter contains data associated to the SIM (e.g., IMSI) while the former contains information directly referred to the individual persons (identities, residential address, bank account coordinates, etc.).
- **Lawful Interception.** Every MNO is obliged to maintain a lawful interception system and store certain data about the position and activity of its customers, to be made available to law enforcement staff upon order by a judge. We assume that it is not possible to use such systems, and the data therein, for any other purpose than legal interception and without judge order, therefore we leave this system out of consideration.
- **Location-Based Servers (LBS).** Some operators deploy in their network commercial solutions to deliver so-called Location-based Services to part of their customers. These systems often involve one or more LBS servers connected to the network elements. These solutions are based on proprietary vendor technology, and their capabilities (in terms of share of population coverage and spatial accuracy) are highly dependent on the specific network configuration.
- **Passive Monitoring systems.** Some operators implement additional passive monitoring system in support of network operation and troubleshooting (e.g. [9, 10]). These systems observe the whole signalling and traffic exchange between the network and the MSs and can be used to infer the location of every MS with the highest possible spatial and temporal accuracy allowed by network-based data [13]. As these systems are proprietary and not available at all operators, they are left out of the focus of this study. Note that however that the location data obtained from such systems are conceptually similar to the "augmented VLR" data discussed earlier in §1.5.2, i.e., a combination of cell-level and LA-level data, therefore the methodology presented in Chapter 3 can be naturally applied to data obtained with such systems, if available.

1.6 Mobile Stations $\neq$ Persons

Strictly speaking, the cellular network "observes" MSs, not people, and the association between individual persons and MS is not always 1:1 (ref. Fig. 1.7). This represents a source of error when leveraging the mobile network to estimate density of "people". More in detail, the following cases are possible:

- **1:1** — the ideal case (for the purpose of this study) is a single person carrying a single mobile device.
- **1:many** — Individuals that carry multiple devices: it is becoming more and more popular to carry more than one phone (e.g., one for private communications and another for work) and other mobile devices like, e.g., tablets and laptop with 2G/3G/4G radio interface.

- **1:0** — some persons do not carry any mobile phone.
- **0:1** — MS that are not associated to any person: these MS are associated to "things", not individual persons, and use the mobile network for machine-to-machine (M2M) communications.

The 1:many and 0:1 cases introduce positive errors (overcounting), while 1:0 introduces negative error (undercounting). We expect that the frequency of 1:many and 1:0 cases varies across demographic groups, i.e., that correlations exist between the number of personal devices and certain demographic attributes (age and profession above all). For this reason, 1:0 and 1:many cases are likely to introduce a bias, with certain age/professional groups under- or over-represented.

In order to mitigate (yet, not completely eliminate) the over-counting errors "0:1" and "1:many", a possible approach is to restrict the analysis to data from the CS domain. This will automatically exclude those data-only devices that are designed to attach only to the PS domain. For VLR data, this implies restricting to MSC data, and to exclude SGSN data.

Besides this initial filtering, it is possible to further mitigate the over-counting error by adopting more sophisticated (i.e., implicit or explicit) filtering strategies. For instance, one approach is to identify and filter out MSs that are not enabled for voice calls. This can be done by accounting for the Type Allocation Code (TAC) code included in the International Mobile Station Equipment Identity (IMEI) – if available in the CDR/VLR, or by integration with other data sources– from the APN, or heuristically by simply picking MS that never engaged in a voice call during a reasonably long observation period (e.g. over 24 hours). All the above methods tend to rely on data fields that are optional and/or additional data sources, and their cost of implementation and effectiveness are highly dependent on the particular network setting. In other words, it is not possible to define a single mitigation approach that fits for all MNOs, but this heterogeneity should not discourage a MNO to put in place additional processing function, based on MNO-specific configuration, aimed at removing or anyway reducing some of the known sources of error (e.g. filtering of M2M terminals).

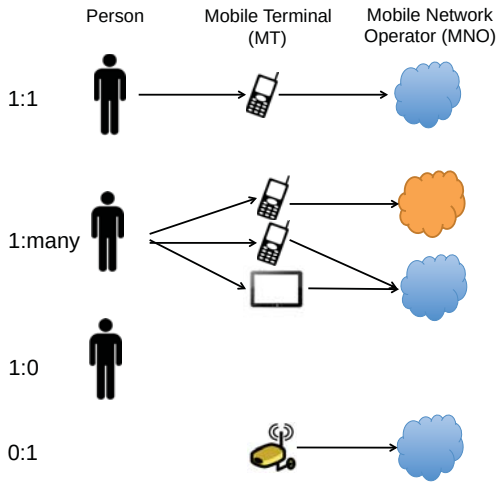


Figure 1.7: Possible association schemes between Mobile Stations and persons.

Chapter 2

Measuring population density distribution in support of public policy: requirements and definitions

2.1 Overview of the general approach

The vast literature on mobile phone data insofar is constituted by studies conducted for a specific purpose on datasets from a single MNO (see [14] for a recent survey). In rare cases datasets from different MNOs were *compared* (e.g. [8]). One distinctive goal of this study is to develop a methodology that allows data from different MNOs to be *fused*. The union of data from MNOs across different countries would allow to produce a pan-European view of population density. Furthermore, the proper fusion of multi-MNO data from the same country bears the potential of improving the accuracy of the estimation *within the same country* along different directions, namely: (i) increase the population coverage; (ii) mitigate the potential bias caused by MNO-specific network configurations and (iii) improve the spatial accuracy (this point is discussed later at the end of §3.5).

In order to be applicable to multiple MNOs, the proposed methodology must *rely on data that are commonly available at every MNO* — as needed for the operation of the network and associated mobile services — and that *can be extracted at reasonable cost*. Moreover, particular attention must be paid to avoid jeopardisation of *business confidentiality* and *user privacy*.

We envision the data and computation flow depicted in Fig. 2.1, consisting of two stages. The first stage algorithm, termed “local processing”, is run independently *within each MNO*: it takes in input a set of “micro-data” and returns in output a set of highly aggregated intermediate data.

The input data are termed “micro” because every record (from CDR and/or VLR databases) is referred to individual MS. The local processing module will take in input also network topology data about position and coverage area (footprint) of every cell, and optionally additional data sources available within the MNO that might help to identify and filter out MS not associated to human users (e.g., M2M terminals).

It is important to remark that with the proposed method *micro-data do not leave the MNO domain*. For every MNO, the output of the local processing module is a set of vector data that collectively

represent the "view" of MS distribution by this specific MNO. Such data are highly aggregate over hundreds or even thousands of MSs: it is not possible to infer from there any information about individual MSs (location, trajectory, identity, calling patterns, etc.) and therefore such data are free from any user-privacy criticality[1]. In order to preserve business confidentiality, the per-MNO vector data must be constructed in a way to avoid leaking business sensitive information — e.g., about the structure and load of the MNO infrastructure, or the characteristics of his customer basis — *beyond what is already available in the public space or anyway deducible from public sources*[2]. However, we envision a conservative scenario where vector data from each MNO is acquired and processed under strict non-disclosure conditions by a trusted public entity (e.g., the JRC or Eurostat) or some private organisation with an established trust basis with the MNOs (e.g., the GSM Association[3]). The central trusted entity is in charge of combining the individual vector data from multiple MNOs and produce a single global density map. In order to ease the combination of multiple MNO data, vector data need to adhere to a common format.

From the discussion in Chapter 1 it should be clear that the problem of inferring the spatial people distribution from the set of available MNO data does not have a unique solution. Starting from a reference resolution method, such as the one described later in Chapter 3, based on a minimum common set of data records available across all MNOs, it is possible to introduce additional MNO-specific refinements (e.g. filtering functions for M2M terminals), leveraging additional MNO-internal data sources (e.g., terminal type databases) in order to reduce some sources of error. Such potential refinements are MNO-specific and cannot be applied in the same way to all MNOs — otherwise they could be included in the "basic" version of the processing procedure, common for all MNOs — and it is desirable that the overall methodology be sufficiently versatile to take advantage of MNO-specific refinements, if available. In other words, the proposed methodology should be designed according to the principle of pursuing the "best possible accuracy" given the specific configuration of each MNO infrastructure, accepting that the actual level of accuracy might differ across MNOs, instead of levelling down all MNOs output towards the worst-case level.

This vision fits well with the two-stage model sketched in Fig. 2.1: it is possible to tailor part of the local processing stage to the specific MNO conditions, by including more advanced "optional" functions that exploit the additional data that might be available at the specific MNO (but not necessarily other MNOs). In other words, the local processing stage should be sufficiently "elastic" to adapt to the heterogeneity of MNO setting, so as to exploit the potential for more accurate estimation than the basic version whenever possible.

2.2 Definitions of "density"

The term "density" (of people) might take on different meanings. This is especially true when we want to "measure" density, since in general the definition of "what" is measured is intimately tied to "how" it is measured. In this section we discuss this point and define unambiguously the notion of "density" adopted in the remainder of this document.

[1] Occasional records with very low value can be set to an arbitrary common minimum threshold to prevent personal identifiability in areas with very low population density.

[2] In this regard, it is important to remark that a certain amount of information about the radio coverage of every MNOs is already publicly available, e.g. from crowdsourcing databases like `OpenCellID` [2].

[3] www.gsma.com.

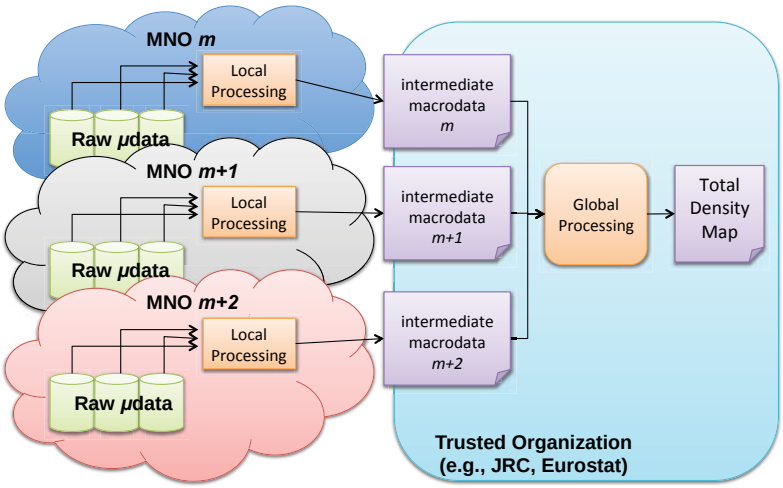


Figure 2.1: General scheme of data and computation flow. Micro-data do not leave the respective MNO domains. Only (intermediate) macro-data are exported by MNOs to the central organisation for multi-MNO data fusion.

Spatial Density. Consider an ideal oracle that tracks the exact geographical point position $y_i(t)$ of every individual $i \in \mathcal{I}$ at any time t. We can take a snapshot $\boldsymbol{y}(t^*) \stackrel{\text{def}}{=} \{y_i(t^*),\ \forall i\}$ of all individual positions at a particular reference time t^*. One possible way to reduce these data is to divide the geographic surface into a tessellation of countable units. We shall consider here a grid of fixed-size squares, called "tiles" hereinafter, without gaps or overlapping areas between adjacent tiles. We shall indicate by a the tile area: e.g. if tiles have $200\,m \times 200\,m$ then $a = 0.04\,km^2$. The tile size should be smaller than the typical cell footprint in order to avoid introducing too much spatial approximation error during the process of mapping cell coverage areas to the reference grid. However, since the spatial granularity of the final estimated density depends primarily on the (distribution of) cell and LA sizes rather than the tile size, reducing the tile size below a certain level does not bring any gain in accuracy, while causing unnecessary additional burden on the computation procedure.

Assume that we have an ideal measurement tool that is able to track the *exact position* $y_i(t)$ of every individual *at any time* t. Denote by $n_k(t^*)$ the number of individuals falling in the generic tile k at time t^*. With these positions, we can define the *spatial density* in tile k as

$$\Delta_k(t^*) \stackrel{\text{def}}{=} \frac{n_k(t^*)}{a}. \tag{2.1}$$

The above definition is unambiguous, and the term "density" in the sense of (2.1) is defined exclusively in the spatial domain.

Probabilistic Density. Now consider a less ideal measurement tool, that is able to track individual positions only approximately. Assume that for every MS i and time t, it returns a bounded region $\nu_i(t)$ that is guaranteed to contain the actual (unknown) point position $y_i(t)$. Hereafter we use the term "location" to refer to the region $\nu_i(t)$. In other words, we do not know exactly the point position $y_i(t)$, but we know that it falls within the location $\nu_i(t)$, formally $y_i(t) \in \nu_i(t)$. In

practice, the location will represent (an approximation of) of the coverage area of a cell or LA, hereafter referred to as "cell-level locations" and "LA-level locations" respectively.

For the sake of simplicity, consider a quantised geographical space where every location $\nu_i(t)$ maps to a set of tiles on the regular reference grid. Let $|\nu_i(t)|$ denote the (integer) number of tiles enclosed by $\nu_i(t)$. Without any further information, we must assume that a MS i can be found equally likely at every point within $\nu_i(t)$. This means that the MS i is present (i-th uniform probability $\frac{1}{|\nu_i(t)|}$) in each tile within the associated location (and with zero probability outside). We now introduce the binary indicator function $\delta_{k\in u_i(t)}$ to indicate whether the generic tile k is included in location $\nu_i(t)$, formally: $\delta_{k\in\nu_i(t)} = 1 \Leftrightarrow k \in \nu_i(t)$. From such data, we can still define the "density" in the generic tile k as:

$$\Delta_k(t^*) \stackrel{\text{def}}{=} \frac{1}{a} \cdot \sum_{i=1}^{I} \frac{\delta_{k\in\nu_i(t)}}{|\nu_i(t^*)|} \tag{2.2}$$

wherein I denotes the total number of MS. Definition (2.2) has a different interpretation than (2.1) as it embeds a probabilistic dimension in addition to the spatial one. In fact, the value of $\Delta_k(t^*)$ defined in (2.2) represents the *average* MS density in tile k in a scenario where the actual position of every MS i is a random variable uniformly distributed within the associated location. The meaning of "density" embodied by (2.2) is similar to the one adopted in this study.

Temporal Density. Strictly speaking, the individual point position $y_i(t^*)$ and the associated location $\nu_i(t^*)$ are defined unambiguously only if the time instant t^* is univocally specified. If we consider an extended time *interval* $[t_1, t_2]$ of duration $T \stackrel{\text{def}}{=} t_2 - t_1 > 0$ we must take into account the possibility that a moving MS i visits multiple locations in this interval. To illustrate, assume that during said interval the MS i has visited three adjacent tiles, namely k_1, k_2 and k_3. In principle, we could "distribute the presence" of individual i to these tiles proportionally to the dwell time, i.e., we could assign to each tile k a fractional weight proportional to the share of interval T that i spent in k. By summing the weights over the index i, we would obtain a new "density" that embeds also the temporal dimension. This approach is viable only if we have full knowledge of the exact trajectory of i during the whole interval of interest, i.e., *if we can observe exactly the point position* $\{y_i(t),\ t \in [t_1, t_2]\}$ *continuously over time*. Unfortunately, this is never the case with MNO data: recall from the discussion in §1.5 (see also Fig. 2.2) that the information available from the network about the actual MS trajectory is *coarse spatially* (LA-level for VLR, cell-level for CDR), and furthermore cell-level location data are *incomplete temporally* — since cell-level locations are available at given sample times, upon occurrence of certain events (e.g. phone call or SMS for CDR). Because of that, the temporal ambiguity intermingles with the spatial ambiguity in a way that complicates the task of "distributing the presence" of moving individuals in a clear manner. In this context, aiming at capturing the *temporal* dimensions of "density" — in addition to the intrinsic spatial and probabilistic dimensions in the sense of equation (2.2) — would represent a major complication. Motivated by this argument, for ease of simplicity we shall seek to exclude the temporal dimension from our definition of "density". In other words, we aim at imposing a "static" definition of MS position — even for MSs that are actually in motion.

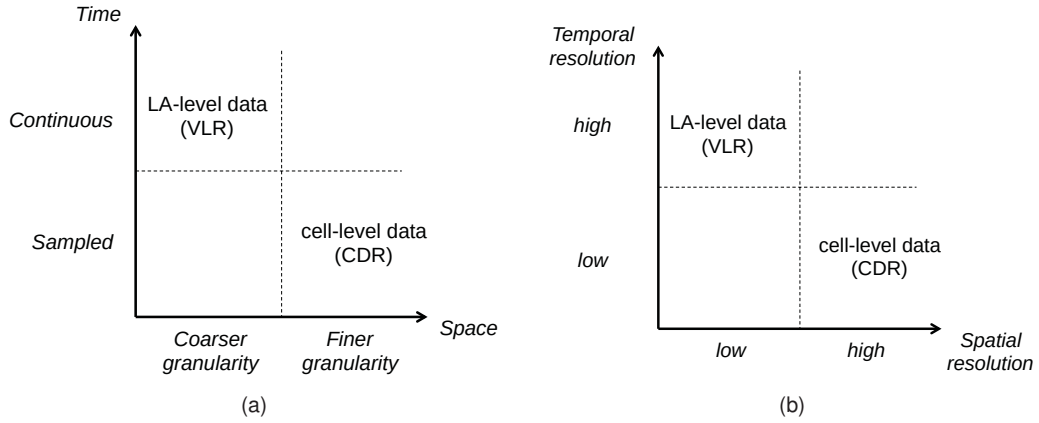


Figure 2.2: High-level comparison between the spatial and temporal dimensions of cell-level and LA-level data respectively in CDR and VLR.

2.3 Dealing with MS movements

Assume we aim at measuring the population density at a reference time t^*. If we were able to "sample" the position of *all* MSs at the same reference time t^*, than we would simply ignore whether each MS is moving or not at this time, and the problem of temporal ambiguity would simply not arise. In our context, this is possible only with LA-level locations obtained from VLR: recall that the MS must communicate to the network every change of LA (via so-called Location Area Update procedure), therefore the LA-level location is monitored *continuously in time*.

With cell-level locations instead (from CDR or augmented VLR), the number of MS that can be "observed" at a generic time t^* is only a small fraction of the whole MS population, also at peak hour. This is due to the fact that cell-level locations are revealed to the network only upon occurrence of specific events (starting a phone call or SMS, engaging in a data connection, initiating a signalling procedure etc.), therefore are observed only at specific "sampling times".

The duality between cell-level and LA-level data in terms of temporal continuity and spatial granularity is summarised in Fig. 2.2(a).

When cell-level locations are considered (e.g., from CDR) we need to consider records along an interval of reasonably long duration, say one or a few hours, in order to "observe" (the cell-level locations of) a sufficiently large number of MSs. But then the problem arises: which location to pick as representative of the position of MS i during an interval of non-null duration? We propose to pick the *observed location nearest in time to the reference time* t^*, i.e., the cell location with the closest timestamp to t^*, subject to minimum and maximum temporal limits. Formally: consider a generic MS i that was observed at the set of locations $\{\nu_i(t_1), \nu_i(t_2), ...\}$ respectively at the set of observation times $\mathcal{T} \stackrel{\text{def}}{=} \{t_1, t_2...\}$; denote by $t^* \notin \mathcal{T}$ the reference time and by $\mathcal{W} \stackrel{\text{def}}{=} [t^* - \theta_l, t^* + \theta_u]$ an observation window of duration $W = \theta_l + \theta_u$ around the reference time; we define the "proxy" location $\hat{\nu}_i(t^*)$ of MS i at time t^* as the location observed at the nearest observation time $\hat{t}^*$, i.e., $\hat{\nu}_i(t^*) \stackrel{\text{def}}{=} \nu_i(\hat{t}^*)$ with:

$$\hat{t}^* \stackrel{\text{def}}{=} \underset{t \in \mathcal{T} \cap \mathcal{W}}{\operatorname{argmin}} \{|t - t^*|\} \tag{2.3}$$

To illustrate, consider the sample trajectory depicted in Fig. 2.3(a) that is represented in the CDR dataset as depicted in Fig. 2.3(b). In this example, CDR data do not contain the cell location at the reference time t_A (cell B.2), hence the observed position at closest observation time t_1, namely cell C.2, would be used as a proxy[4]. If VLR data are available, and we are satisfied with LA-level locations, we can simply pick the actual LA location (ref. 2.3(c)). When both cell-level and LA-level data are available, as with joint CDR/VLR and Augmented VLR data (ref. Fig. 2.3(d) and Fig. 2.3(e)) it is possible to choose between the actual LA-level location and the proxy cell-level location — the choice can be based, for example, on the basis of the time delay between the reference time and the cell location timestamp, i.e. $t_A - t_1$ and $t_A - t_2$ respectively for Fig. 2.3(d) and Fig. 2.3(e). Similar considerations apply for the other case depicted in Fig. 2.3 when the reference time fall in t_B.

It should be noted that, while it is certainly possible that the actual (unknown) cell location of MS i at the exact time t^* does not coincide with its proxy value, i.e., $\nu_i(t^*) \neq \hat{\nu}_i(t^*)$, nevertheless our approach guarantees that i was present at this position at some time within the observation window $\mathcal{W}$. In other words, we can interpret the error on the cell location $|\hat{\nu}_i(t^*) - \nu_i(t^*)| = |\nu_i(\hat{t}^*) - \nu_i(t^*)|$ as a purely temporal (rather than spatial) error. This leads to an interesting interpretation of the *choice between the (proxy) cell-level location and the (actual) LA-level location as a matter of tradeoff between temporal and spatial resolution*, ad sketched in Fig. 2.2(b).

[4] Fig. 2.3(a) could suggest the possibility of resorting to some kind of interpolation method, where an intermediate position between the observed positions at times t_1 and t_4 is taken as proxy value for $\nu_i(t^*)$. However, when one takes into account the various sources of spatio/temporal uncertainty — spatial quantisation in the bi-dimensional space; unknown start and stop time of trips; unknown speed and mode of transport — and the spatial constraints due to the underlying transportation network(s) — it becomes evident that any such "interpolation" heuristic bears a serious risk of increasing, rather than reducing, the potential final error.

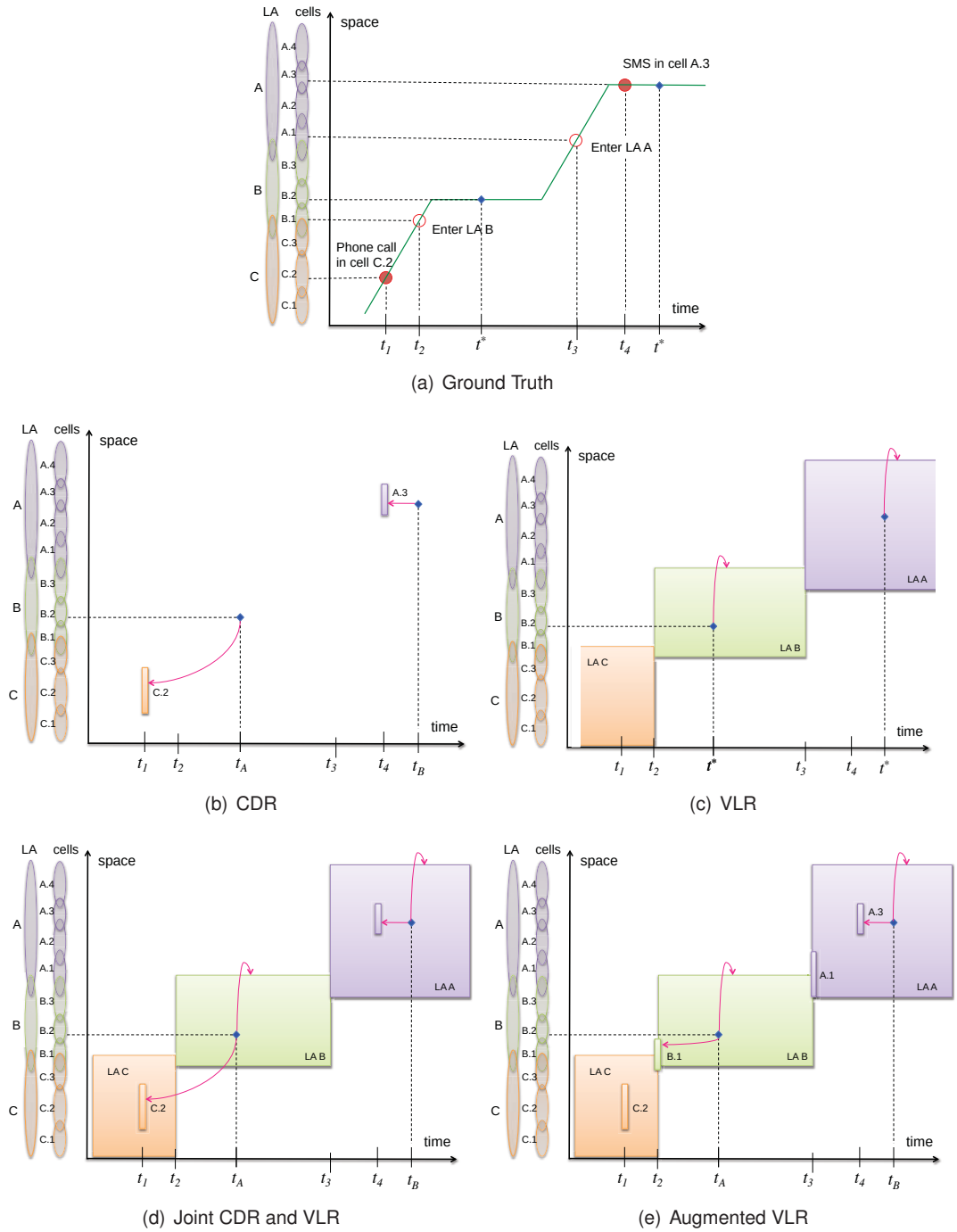


Figure 2.3: Examples of "proxy" locations for the MS trajectory of Fig. 1.6 for two sample reference times t_A and t_B, for different network-based data. In both cases the cell-level location is not observed at the exact reference time, therefore the MS position can be mapped to the (actual) LA location or to the nearest-in-time observed cell location.

Chapter 3

Measurement Methodology

In this Chapter we describe the proposed methodological framework for the task of estimating population density from multi-MNO data. We aim at providing a framework that is general enough to be implemented by any European MNOs — hence, does not rely on MNO-specific aspects like network configuration, data organisation etc. — but at the same time is flexible enough to take advantage (optionally) from potential MNO-specific improvements (e.g., availability of more accurate location data).

The proposed methodology can be applied to one-time analyses as well as to the periodical (offline) analyses, e.g., based on daily or monthly activity. In addition, the proposed approach is suitable to continuous online analyses, although such an option requires considerably more engineering efforts, especially at network modeling level, in order to ensure consistency of network topology data accounting for changes and upgrades. As the engineering aspects remain outside the scope of this study, hereafter we assume a static (known) network topology.

3.1 Overview of the measurement methodology

The proposed methodology relies on two distinct types of data:

- **Network Topology data** about the geographical location and coverage areas of radio cells.
- **MS Counters** of the number of MS observed (at the reference time) on every cell and LAs.

Two main contributions of this work are:

- We consider extended topology data and assume (approximate) knowledge of the whole cell coverage area, instead of merely the (exact) tower location.
- Our method can combine MS counters at different spatial granularity, i.e., at cell-level and LA-level, obtained from CDR and/or VLR databases, rather than exclusively cell-level data from CDR.

The proposed measurement method can be described as a chain of intermediate data processing stages. A high-level view of the data workflow is sketched in Fig. 3.1. Each processing stage is detailed in the following sections of this chapter.

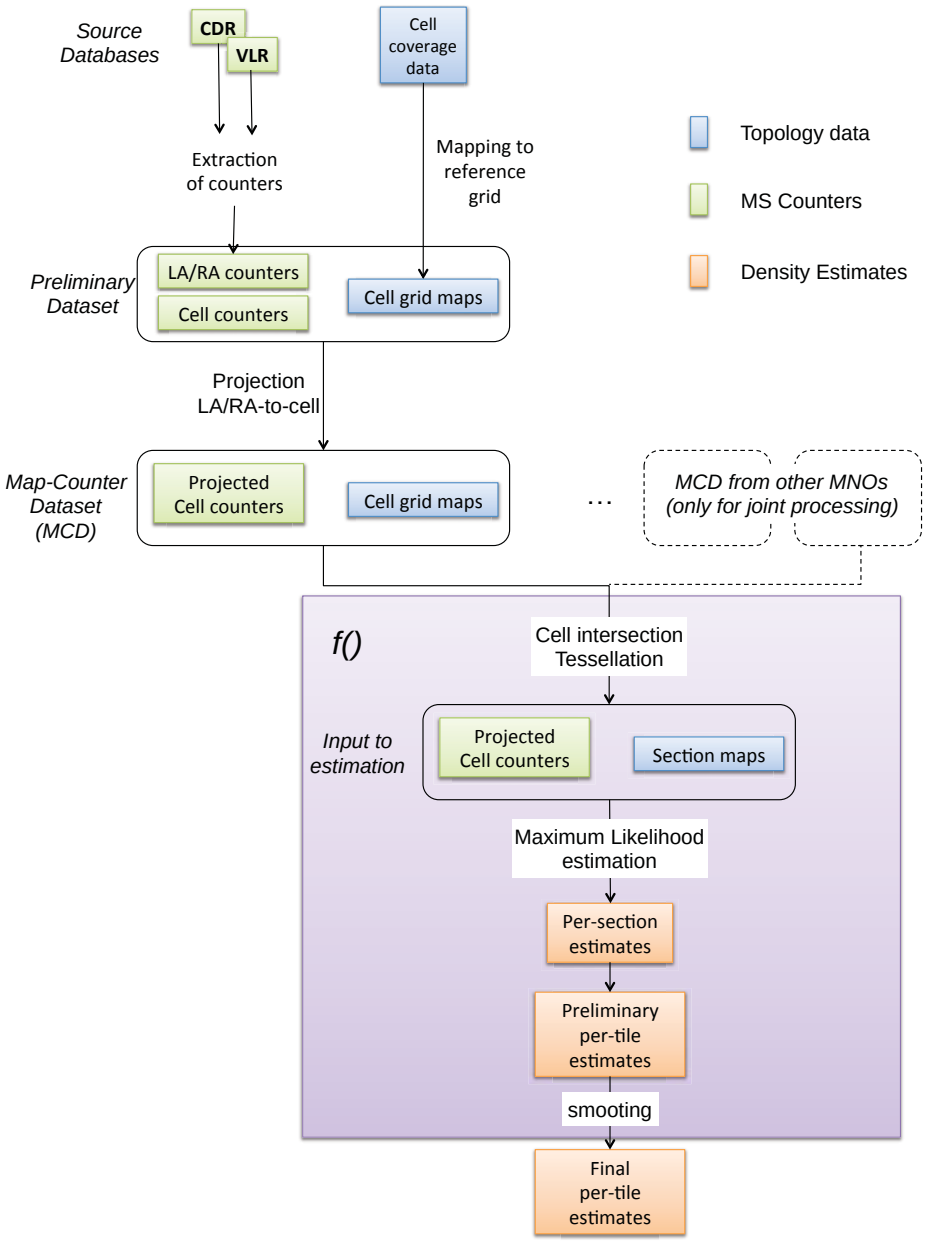


Figure 3.1: Overview of the data processing workflow. The processing method $f()$ can be applied to a single MCD from one MNO, or to a combined multi-MNO MCD.

The network topology data (i.e., cell maps) for each MNO are mapped to a common reference grid in order to facilitate the fusion of data from different MNO. We recommend to adopt the INSPIRE reference grid specified in [12] for this purpose. In fact, the INSPIRE specification provides a common framework for harmonized and interoperable geographic localization of different types of spatial objects and quantities, and it is specifically intended for statistical reporting purposes. It appears to be perfectly suited for the purpose of fusing aggregated data from different European MNOs. Furthermore, it greatly facilitates the prospective integration of multi-MNO data with other sources of spatial data and services. Hereafter we will adopt the term “tile" to refer to a generic spatial unit in the reference grid[1].

At some point during the workflow, the generic MNO m generates a set of “map-counter" records $(\boldsymbol{b}_j, c_j)$, each record referring to a different radio cell j in its network. In a nutshell, $\boldsymbol{b}_j$ denotes the map of cell j on the reference grid, while c_j denotes the number of MS “observed" in cell j according to the available CDR/VLR data — both elements are formally introduced in Fig. 3.5. The whole set of map-counter records from a generic MNO m constitutes the the “Map-Counter Dataset" (MCD for short) and will be denoted by $\mathcal{S}_m$ (ref. Fig. 3.2(a)). MCD is an important intermediate data along the data processing flow.

We can envision two possible options with respect to the subsequent processing of MCD data from different MNOs. In the first option, depicted in Fig. 3.2(c), all MNOs would agree to pass their MCD datasets to a central trusted entity (e.g., Eurostat or JRC). The latter would then estimate the total density map $\boldsymbol{D}_T$ by jointly processing the union of individual MCDs from all MNOs, i.e.:

$$\boldsymbol{D}_J = f\left(\mathcal{S}_1, \mathcal{S}_2, \ldots\right) = f\left(\bigcup_m \mathcal{S}_m\right) \tag{3.1}$$

where $f()$ denotes the data processing method that is detailed later through sections §3.5-§3.7.

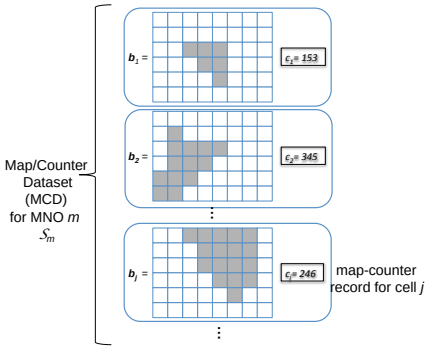
The advantage of this option is that the final density estimation can leverage in the best possible way data *diversity* — in terms of spatial coverage and population coverage — across different MNOs. Note that no privacy-critical information would be disclosed in this way, since map-counter records are aggregate data, not micro-data. However, this approach requires every MNOs to export information that might be regarded as critical from a business perspective (e.g., detailed size, location and traffic load of individual cells). Although the recipient of such data would be anyway a trusted entity, bound to non-disclosure legal constraints, it is not clear whether such model would be accepted by MNOs.

This motivates the definition of an alternative, more conservative scenario, where the MCD processing is split into two stages as sketched in Fig. 3.2(b) (see also Fig. 2.1). In the first stage, each MNO computes a “partial" density map $\boldsymbol{D}_m$ from its local MCD data, independently from other MNOs. In the second stage, the central entity simply combines the density maps from different MNOs into the final “global" density map $\boldsymbol{D}_\Sigma$. In other words, the function $f()$ of equation (3.1) is run by every MNO based exclusively on local data, and the (local) outputs are then exported to the central entity for final (weighted) summation, formally:

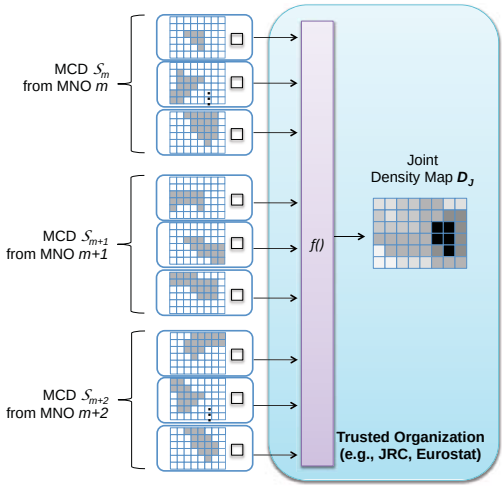
$$\boldsymbol{D}_m = f\left(\mathcal{S}_m\right), \quad \forall m. \tag{3.2}$$

$$\boldsymbol{D}_\Sigma = \sum_m w_m \boldsymbol{D}_m \tag{3.3}$$

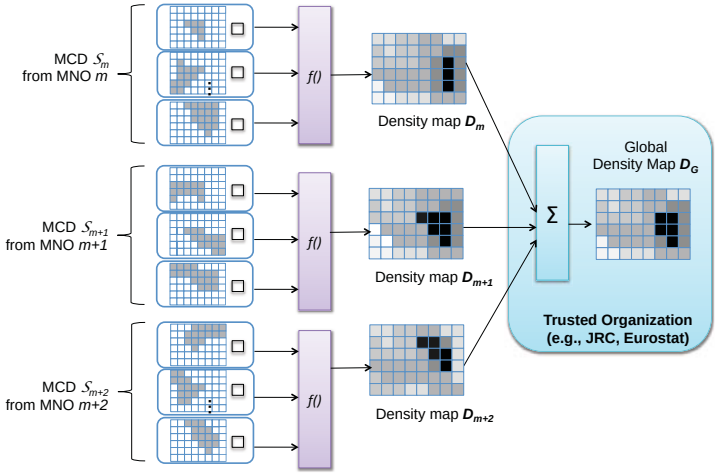
[1]Note that in [12] the term “cell" is used to refer to the spatial grid units. In the context of the present work, this collides with the usage of the term “cell" to denote radio coverage areas for the mobile network.



(a) MCD dataset from individual MNO

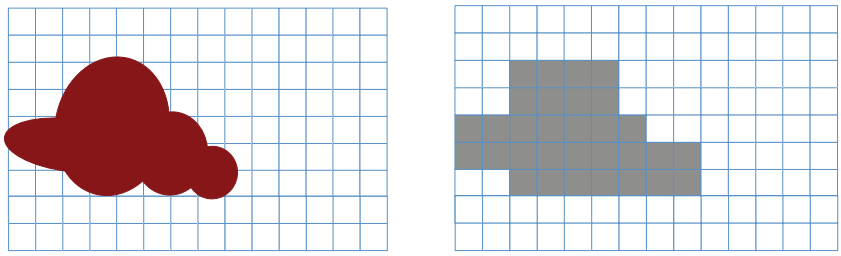


(b) Joint MCD processing from multiple MNOs



(c) Local MCD processing within each MNO

Figure 3.2: Schematic representation of multi-MNO data processing. The function $f()$ denotes the data processing method detailed through sections §3.5-§3.7. It can be applied for the joint processing of all MCDs (b) as well as for the separate processing of each individual MCD (c).



(a) Footprint of cell j (b) Grid map $\boldsymbol{b}_j$

Figure 3.3: Cell coverage area and corresponding map on the reference grid.

wherein the weights w_m are derived from the (normalised) MNO penetration rates.

In this model every MNO m must communicate only a density map $\boldsymbol{D}_m$, not the full MCD $\mathcal{S}_m$. We conjecture that the final estimate $\boldsymbol{D}_\Sigma$ obtained in this way might be somewhat less spatially accurate than the one that can be computed from the joint processing of the map-counter records from all MNOs, namely $\boldsymbol{D}_J$ (we will motivate this claim later at the end of §3.5, after introducing the notion of "section tessellation"). A comparison between these two strategies, i.e. the quantitative assessment of the fidelity of $\boldsymbol{D}_J$ and $\boldsymbol{D}_\Sigma$ versus the ground truth is an interesting direction for future research.

3.2 Construction of cell maps

We assume that every MNO knows — at least approximately — the geographical coverage area of every cell, i.e. the "cell footprint". This information can be embodied in different formats across different MNOs, and can be derived from different sources, for example "best server" maps produced during the planning process (via simulations) and/or from field measurements. In the worst case, a coarse estimation of the cell footprint can be derived directly from antenna configuration parameters (height, tilt, beam-width) in combination with cell tower location. Therefore, for every cell the MNO is able to produce the associated "grid map" (refer to Fig. 3.3) by projecting its footprint to the INSPIRE reference grid that we consider in this work [12]. Considering the typical differences in cell/LA size between urban, sub-urban and countryside areas, it might make sense to vary the Resolution Level of the reference grid between different types of regions. A possible choice is to adopt Resolution Level 11 (tile size 100 meters) in urban areas, and Resolution Level 10 (250 meters) or 9 (500 meters) in sub-urban and countryside areas (refer to [12] for further details).

3.3 Extraction of initial counters from CDR and/or VLR database

The proposed method ultimately relies on the possibility to infer the approximate location (cell or LA) of every MS from the network databases available at the serving MNO. More specifically, given an observation window $\mathcal{W} = [t^* - \theta_l, t^* + \theta_u]$ around the reference time t^*, the generic MS i served by MNO m during the said observation is mapped to the smallest spatial unit ν_i that can

be inferred from the available network database(s): the cell (identified by the CGI) if available, otherwise the RA or (in the worst case) the LA. In this work we are not interested in individual MS positions, but only in spatial densities, therefore such data can be immediately aggregated: for every cell j and LA ℓ denote by c_j and c_ℓ the total counts of MS observed in said location. In other words, the data flows embeds two distinct stages:

- **MS mapping**: MS $i \rightarrow$ individual MS location ν_i.
- **Aggregation**: set of MS locations $\{\nu_i\} \rightarrow$ set of cell/LA counters $\{c_j, c_\ell\}$.

In the remaining of this section we discuss the possible options for the initial MS mapping.

Two potential data sources are relevant for our study: Call Detail Records (CDR) and Visiting Location Register (VLR). Both CDR and VLR can be regarded as databases and, in principle, can be queried by the MNO staff.

The implementation of CDR and VLR databases varies greatly across MNOs. It is possible to identify a minimum set of mandatory fields that are necessarily present in every CDR / VLR implementation, since they are needed to perform standard procedures (mobility management, billing). This basic set of mandatory fields represent a sort of "minimum common denominator" across the CDR/VLR of different MNOs. However, when one considers the technical details of how such basic fields are encoded, and how such data can be retrieved, important differences between different MNOs emerge. For instance, it is not uncommon that MNOs configure their CDR/VLR systems to store additional (optional) data fields besides the minimum common set of mandatory fields. It is important to remark that several MNO-specific technical details about *what* information is stored (on top of mandatory data) and *how* it is encoded determine also the *feasibility and cost* of (i) extracting the data and (ii) implementing additional processing and data correlation functions aimed at improving the quality of the final output.

To allow for flexibility, hereafter we will present a palette of different data acquisition methods, based on different assumptions about the availability of certain data dimensions, that enable varying degrees of estimation "quality" in terms of population coverage, spatial / temporal resolution and risk of bias. In fact, we envision a flexible scenario where each MNO can contribute with the "best" possible data[2] that can be extracted at reasonable cost given the specific configuration of its CDR/VLR databases.

3.3.1 Basic CDR-only method

In the simplest scenario, the MNO relies exclusively on CDR data, i.e., VLR data are not considered. Given an observation window $\mathcal{W} \stackrel{\text{def}}{=} [t^* - \theta_l, t^* + \theta_u]$ of duration $W = \theta_l + \theta_u$ around the reference time t^*, only MS that have been somehow active in $\mathcal{W}$ (e.g., received or started a phone call or SMS) will be "observed" with this method along with their (proxy) location at the cell level. The main advantage of this scheme is the *high spatial resolution*, since CDR embed only cell-level MS location (typically, the call start CGI). The disadvantages of this scheme are:

[2] When a trade-off is in place between different quality criteria — e.g., spatial vs. temporal resolution, or spatial resolution vs. risk of bias, as discussed below in §3.3.3 — the operational definition of "best" data should be adapted to the particular application (use-case) for which the population density map is intended.

- Incomplete (possibly small) population coverage and *low temporal resolution*. The fraction of MS observed with CDR data depends on the duration W of the observation window and on the activity behaviour of the MS population, and the latter varies with the time-of-day. The population coverage could be very small during night time, even with observation window of several hours. The need to increase population coverage drives towards the choice of long observation windows (several hours) with consequent reduction of *temporal resolution*. The combination of these aspects will probably hinder the viability of certain types of analysis, e.g., time-of-day variability of population density.

- Bias due to calling habit. Generally speaking, the probability that a generic mobile phone user starts a call, SMS or data connection depends on the type of activity (s)he is currently engaged (working, leisure, traveling, etc.) which, in turn, depends on time and position. Therefore, the MS call activity, hence the probability of the MS being "observed" by the CDR method in the given temporal window, is correlated with the MS position. This introduces a certain degree of statistical bias, i.e., under- or over-representation of particular locations. Furthermore, as user activity patterns change in time, the structure of the bias error may vary in time.

The above disadvantages might be mitigated by integrating CDR data with VLR data, as explained in the following subsections.

3.3.2 Basic VLR-only method

In this alternative extreme scenario, the MNO relies exclusively on VLR data, i.e., CDR data are not considered. We assume here that the VLR database includes only the mandatory fields, namely the LAI[3] and the T-IMSI. The main advantages of this method are:

- Complete coverage: all MSs served by the MNO network are represented in the VLR database, for any choice of the reference observation time t^*.

- Reduced bias: the LA-level locations encoded in the VLR does not depend on the user calling habit, therefore the risk of under- or over-representation of particular locations (LA/RA in this case) is dramatically reduced.

- Perfect temporal resolution: at any generic instant t the VLR records the current LAI for every MS. Therefore, the location of the MS can be referred exactly to the reference time t^*, rather than to a reference observation interval of duration W. In other words, VLR have perfect *temporal resolution* (ref. Fig. 2.2).

The main disadvantage of this method is the low spatial resolution, since only LAI locations are encoded in the basic VLR, with spatial resolution in the order of kilometres (in urban areas) or even tens of kilometres in sub-urban areas (ref. to Appendix B).

[3] And possibly also the RAI, if the VLR is shared between the CS and PS domain, see discussion in §1.4.

3.3.3 Comparison between basic schemes: CDR-only vs. VLR-only

The mere comparison between the advantages and disadvantages listed above for the CDR and VLR methods clearly show that these two schemes are somewhat antipodal with respect to the tradeoff between spatial and temporal resolution: the basic CDR-only method yields the highest possible spatial resolution (cell level information) but with very low temporal resolution (due to limited coverage), while conversely the basic VLR-only method combines and excellent time resolution with the worst spatial resolution. This trade-off was depicted earlier in Fig. 2.2. Also, CDR data (and in general cell-level locations) suffer from a considerable risk of spatial bias, which might lead to non-negligible distortion of the final estimate. These simple considerations tell that, if the choice between the two is given, one method might be preferable over the other depending on the specific use-case and type of analysis required, but neither of them can be considered "superior" to the other in the general case. In other words, neither method dominates completely the other along the whole spectrum of performance dimensions.

However, more on the operational side, it is important to highlight one key difference between the VLR and CDR in terms of data extraction. The CDR database is "static", in the sense that new CDR records are added continuously, but past records are not modified. Therefore, they can be read off-line during pre-planned periods of minimum network load, typically during night time. In this way, it can be easily guaranteed that the extraction of CDR data will not interfere with the network operation. In contrast, the VLR is a "dynamic" database, as its role is to support the network operation by serving as a sort of temporary "cache" for volatile data that are continuously updated. Moreover, for given storage capabilities (as well as per the MNO data storage policy), an operator might consider not to store any of the VLR location data, or to save only the data relative to the last known location. Therefore, if one wishes to extract a snapshot of VLR locations for the reference time t^*, the VLR query must be actually accomplished on-line at the same time t^*: while relatively short delays can be tolerated, the VLR query can not be deferred indefinitely — as is typically done with CDR data. Considering that VLR is accessed continuously by the operational network equipment (mainly MSC and SGSN), particular care must be taken to avoid that the resources consumed by the VLR query/extraction process interfere with the normal network operation.

The main differences between the two basic methods are summarised in Table 3.1.

3.3.4 Augmented VLR data

Some MNOs might configure their VLR to maintain additional (optional) data fields besides the current LA-level location (LAI), for instance (i) the cell-level location (CGI) and (ii) timestamp of the last interaction with the MS. Other MNOs might collect similar data by means of other proprietary monitoring systems (e.g., [9, 10]). We shall refer to such data as "Augmented VLR" data, as they represent an augmentation the basic VLR data (i.e., LA-level locations for all MSs) with additional finer-grained data (cell-level locations, but for a subset of MS).

If such data are available, for every MS the more accurate cell-level location can be used in place of the LA-level location whenever the associated timestamp falls within the reference observation window $\mathcal{W}$. This approach merges the advantages of the VLR-only and CDR-only schemes in the sense that it yields the best possible combination of coverage, spatial resolution and temporal resolution allowed by network-side data. However, the risk of bias is not eliminated, because cell-

	Basic CDR data	**Basic VLR data**	**Notes**
Spatial resolution	high (cell level)	low (LA level)	the spatial resolution (for both cell-level and LA-level locations) varies between urban and sub-urban areas
Temporal resolution	low	very high	ref. Fig. 2.2
MS coverage	possibly low	very high	CDR coverage possibly very low (e.g., at night). VLR coverage virtually complete: all MS "attached" to the MNO network are always tracked at LA level.
Risk of bias	high	low	Cell-level location are intrinsically correlated to MS activity. Bias in CDR is due to call habit.
Data type	static	dynamic	VLR data are volatile, i.e., old data are continuously overwritten by new data. In CDR new data are appended to (not replaced by) past data.
Off-line data extraction	possible	not possible	CDR data query can be deferred arbitrarily. VLR data must be queried at the reference time t^*, as VLR fields are updated (overwritten) continuously. For VLR, attention must be paid to avoid interfering with network operation (especially critical at peak-hour).

Table 3.1: Summary comparison between basic CDR-only and VLR-only schemes.

level locations remain conditioned to the occurrence of certain events: the difference with CDR is that the set of event types is larger, since certain signalling procedures that would be "missed" by CDR are instead "observed" by VLR (e.g., Location Area Update (LAU), Attach Request, etc.) Therefore, while the bias due to calling habit is somewhat reduced in comparison with CDR, in principle the cell-level information contained in the augmented VLR data might be affected by additional sources of spatial bias. For example, LAU procedures are likely to occur at the LA borders, hence cells located at the boundaries between different LA would tend to be over-represented. In other words, the risk of bias associated to cell-level records is not due to the adoption of a particular type of data source (CDR or VLR), but is rather intrinsic to the functional dynamic of the mobile phone network, where the detection of cell location by the network is always conditioned to some particular type of MS action (starting a phone call or performing a signalling procedure) that, in general, is not completely independent from the MS location.

Similarly to the basic VLR method, also the augmented VLR method requires on-line data extraction, hence caution is needed to avoid interference with the network operation, especially at times of peak load.

3.3.5 Joint VLR and CDR

Even without augmented VLR data, it is still possible to "merge" CDR and basic VLR data that have been acquired independently. In general, it might not be possible to match the same MS identifier between the two datasets: for example, the same MS might be identified with the T-IMSI

in the VLR, and with the encrypted IMSI in the CDR. However, it is not necessary to perform a detailed MS-by-MS matching between the two datasets: in order to avoid double counting of the same MS between the two dataset, it is just sufficient to reduce the counter c_ℓ for every LA ℓ in the VLR data by an amount equal to the sum of MS observed in the corresponding cells in the CDR dataset. In this way, it is possible to build a single "combined" CDR+VLR dataset from two dataset acquired independenty.

3.3.6 Practical considerations on the practical adoption of CDR-only vs. other methods

In practice, we expect CDR data will be available at all MNOs, owing to the simplicity of extracting static data off-line (ref. Table 3.1). Additionally, a few MNOs might be willing to pioneer the extraction and processing of VLR data, possibly with "augmented" fields, and some of them might decide to complement (or even replace) CDR with more accurate data extracted with other (proprietary) monitoring systems (e.g., [9, 10]). In other words, the CDR-only case should be regarded as the most common "minimal" scenario, not the unique one.

The methodological framework presented in the remainder of this Chapter provides a basis for the combination of LA-level and cell-level location data, and for the experimental comparison between the CDR-only and other approaches (combined CDR/VLR, augmented VLR) in terms of spatial/temporal accuracy, bias, etc. Should such an experimental demonstrate a substantial gain of complementing CDR data with VLR data (or any other combination of cell-level and LA-level location data), the proposed methodological framework provides a reference evolutionary platform for the incremental addition of additional data by a larger number of MNOs.

3.4 Projection of LA counters to cell counters

In a first pre-processing step the MS counter value for each LA (as obtained from VLR data) is distributed to its cell counters. Consider a generic cell j included in zone ℓ (i.e., $i \in \mathcal{A}_\ell$). Denote by c_j and c_ℓ their respective counters before projection, and by d_j the new cell counter after projection. Recall that $\beta_j \stackrel{\text{def}}{=} \sum_k b_{kj}$ denotes the size of of cell j on the reference grid.

There are two extreme options for projecting the value of c_ℓ across its component cells:

- Proportionally to the cell counter c_j
- Proportionally to the cell area β_j.

In general, we can follow an hybrid approach where a share $\gamma \in [0, 1]$ of the LA counter c_ℓ is assigned proportionally to the cell counter, and the remaining share $1 - \gamma$ is assigned proportionally to the cell area, i.e.:

$$d_j = c_j + \gamma \cdot c_\ell \frac{c_j}{\sum_{h \in \mathcal{A}_\ell} c_h} + (1 - \gamma) \cdot c_\ell \frac{\beta_j}{\sum_{h \in \mathcal{A}_\ell} \beta_h} \tag{3.4}$$

In this way, the total set of cell and LA counters is transformed into a set of (projected) cell counters:

$$\{c_\ell, c_j, \quad j = 1, \ldots, J; \quad \ell = 1, \ldots, L\} \quad \rightarrow \quad \{d_j,\ j = 1, \ldots, J\}$$

The value of γ can be seen as a "tuning knob" in the trade-off between spatial accuracy vs. risk of bias due to call activity that is in place between LA data and cell-level data (ref. Table 3.1). At one extreme, for $\gamma = 1$ the (potential) bias affecting cell-level data (e.g., from CDR) is entirely projected on the whole LA data (from VLR). At the opposite extreme, for $\gamma = 0$ the LA data remain unbiased but at the cost of a major loss of spatial resolution. In practice, the more convenient setting for γ will depend on the relative impact of bias vs. spatial resolution for the specific application at hand.

3.5 Cell intersection tessellation and the notion of "section"

Hereafter, we shall use the term "section" to indicate a group of adjacent tiles covered by the same set of cells. Equivalently, each section represents the intersection area of a specific set of cells, different sections referring to different cell sets. An illustrative example is given in Fig. 3.4, in which 4 neighboring cells originate 11 sections (out of $2^4 = 16$ theoretically possible combinations).

Since sections do not overlap by definition, the division into sections constitutes a (irregular) tessellation of the area of interest. Such tessellation is different from the one resulting from a Voronoi tessellation technique [4], which, instead, is built by considering a single reference point for each cell (i.e., the tower location, or the centroid of the coverage area). In fact, the section tessellation, considered in this work, takes into account the entire cell footprint and overlapping areas with other cells, which avoids double-counting of users. Note that, in general, the number of sections is greater than the number of cells — consequently, the section tessellation is denser than in the Voronoi case — but still much smaller than the number of tiles (11 vs. $9 \times 14 = 126$ in the toy example of Fig. 3.4).

Moreover, we introduce the notion of section motivated by the fact that it is more appropriate to formulate the estimation problem in terms of per-section variables, rather than per-tile variables. In fact, it can be easily seen that the MCD dataset embeds information about the distribution of MS density *across different sections*, but does not tell anything about the distribution *within individual sections*. In other words, from the perspective of the available measurement data, tiles within the same section are identical, and there is no information therein that allows to discriminate the intra-section density differences. Such loss of detail is intrinsic to the spatial aggregation (or quantization) introduced by the network-based observation process.

The formulation of the estimation problem in terms of per-section variables, instead of per-tile variables, is also convenient from a computation perspective. First, it brings a considerable reduction of the search space dimension, by more than one order of magnitude (e.g., the simulation scenario introduced later in Section 4.3 consists of about 400 sections, versus 10,000 tiles). Second, it prevents the numerical solution to introduce artificial density gradients *within* individual sections, e.g. resulting from incorrect matrix conditioning and/or numerical instabilities. On the negative side, this approach introduces fictitious discontinuities at the border *between* adjacent sections. However, the latter can be easily counteracted in post-processing by means of a simple smoothing filter, as discussed later in §3.7.

Generally speaking, a denser and finer tessellation (i.e., a higher number of sections of smaller size) will lead to better spatial resolution. The former depends on the *number of radio cells* but also on the *topological diversity of cell footprints*: both these factors increase when combining

Map/Counter Datasets (MCD) from different MNOs. Based on this simple argument, it can be expected that the approach of fusing intermediate MCD data from different MNOs (ref. Fig. 3.2(c)) bears the potential of achieving a more accurate estimation than the mere (weighted) summation of density maps obtained from individual MCD (ref. Fig. 3.2(b)).

3.6 Maximum Likelihood Estimation of per-section densities

Let the variable r_n denote the probability that a generic MS is located in section n, and recall from Fig. 3.5 that q_{nj} represents the (conditional) probability that a generic MS located in section n is assigned to cell j. By the law of total probability it follows that the probability ι_j that a generic MS is assigned to cell j is given by:

$$\iota_j = \sum_{n=1}^{N} q_{nj} r_n \tag{3.5}$$

Recall that d_j denotes the total number of MS assigned to each cell j, and $\boldsymbol{d} \stackrel{\text{def}}{=} [d_1, \ldots, d_{J_c}]^T$ the total vector of per-cell counters. Considering that the assignment process is independent across MSs, the vector $\boldsymbol{d}$ has a multinomial distribution:

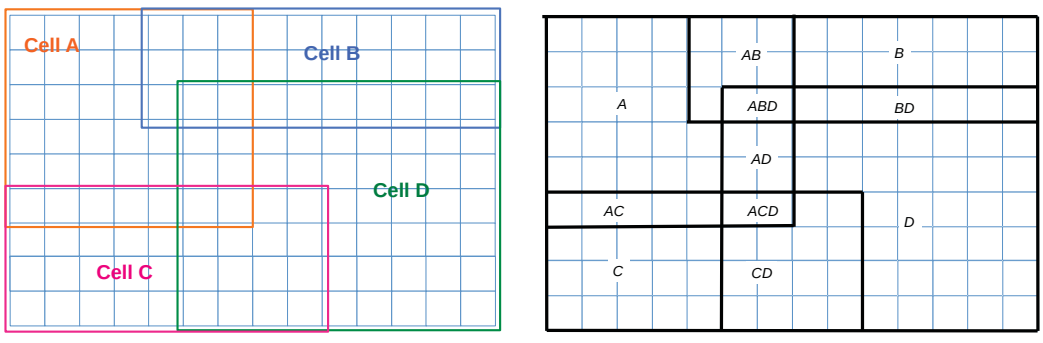
$$Prob\{\boldsymbol{d}|\boldsymbol{r}\} = \frac{D!}{d_1!d_2!\ldots d_J!} \prod_{j=1}^{J} (\iota_j)^{d_j} \tag{3.6}$$

wherein $D \stackrel{\text{def}}{=} \sum_j d_j$ is the total number of MS in the dataset. The corresponding likelihood function is therefore (omitting the irrelevant multiplicative factor):

$$\mathcal{L}(\boldsymbol{r}|\mathbf{d}) = \prod_{j=1}^{J_c} (\iota_j)^{d_j} \tag{3.7}$$

and the corresponding log-likelihood:

$$\log \mathcal{L}(\boldsymbol{r}|\mathbf{d}) = \sum_{j=1}^{J} d_j \log \iota_j = \sum_{j=1}^{J} d_j \log \sum_{n=1}^{N} q_{nj} r_n. \tag{3.8}$$



(a) Cells (b) Sections

Figure 3.4: Example of section tessellation: the different intersections of 4 cells (left) produce a tessellation of 11 non-overlapping “sections”.

In the considered setting, we have four types of spatial entities: tiles, cells, LAs and sections. Hererby, we distinguish between indices for each type of spatial entity, and ways to encode the associations (mapping) between different types. We shall use vectorial notation to encode the cell-to-tile mapping, and set notation for the other mappings.

The symbols K, J, N, L denote the total number of tiles, cells, sections and LA, respectively. We shall use a distinct index for every type of object:

- $k = 1, \ldots, K$ the tile index.
- $j = 1, \ldots, J$ the cell index.
- $n = 1, \ldots, N$ the section index
- $\ell = 1, \ldots, L$ the LA index.

Mappings and associated quantities:

- $b_{kj} \in \{0, 1\}$ boolean variable indicating whether tile k is included in cell j footprint.
- $\boldsymbol{b}_j \stackrel{\text{def}}{=} [b_{1j}, \ldots, b_{Kj}]^T$ the boolean vector representing the map of cell j.
- $\beta_j \stackrel{\text{def}}{=} \sum_k b_{kj}$ the size of cell j on the reference grid, i.e., the number of tiles spanned by cell j footprint.
- $\mathcal{A}_\ell$ the set of cells included in LA ℓ.
- $\mathcal{V}_n$ the set of tiles included in section n and $v_n \stackrel{\text{def}}{=} |\mathcal{V}_n|$ the size of section n.
- $\mathcal{Z}_n$ the set of cells defining section n and $z_n \stackrel{\text{def}}{=} |\mathcal{Z}_n|$ the number thereof.
- $q_{nj} \stackrel{\text{def}}{=} \begin{cases} z_n^{-1} & \text{if} \quad j \in \mathcal{Z}_n, \\ 0 & \text{if} \quad j \notin \mathcal{Z}_n. \end{cases}$ a set of model parameters derived from the cell coverage pattern. More in detail, $q_{nj} \in [0, 1]$ represents the probability that a generic MS in section n is assigned to cell j in the generative model described in Appendix A.

Variables and parameters:

- c_j and c_ℓ the number of MS observed in cell j and LA ℓ, respectively.
- d_j the MS counter for cell j after projection of LA counters (ref §3.4)
- $D \stackrel{\text{def}}{=} \sum_j d_j$ the total number of MS observed in the whole network.
- $\gamma \in [0, 1]$ a tunable parameter in the LA projection procedure described in §3.4.
- x_k the (unknown) number of MS in tile k and $\hat{x}_k$ the final estimated value obtained with the procedure described in §3.6.

Figure 3.5: Notation used in the presentation of the estimation method.

Therefore the Maximum Likelihood (ML) estimate $\hat{\boldsymbol{r}}$ given the data $\boldsymbol{d}$ and the model $\{q_{nj}\}$ is given by the solution of the following constrained optimization:

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{J} d_j \log \sum_{n=1}^{N} q_{nj} r_n \\ \text{subject to} \quad & \sum_{n=1}^{N} r_n = 1, \\ & r_n \geq 0, \quad \forall n, \end{aligned} \tag{3.9}$$

or, equivalently, to find:

$$\hat{\boldsymbol{r}} = \underset{\substack{\boldsymbol{r} \geq \boldsymbol{0} \\ \|\boldsymbol{r}\|_1 = 1}}{\arg\max} \sum_{j=1}^{J} d_j \log \sum_{n=1}^{N} q_{nj} r_n. \tag{3.10}$$

3.7 Deriving per-tile estimates

The solution $\hat{\boldsymbol{r}}$ to (3.10) represent the estimate of (normalised) per-section counters. For every tile $k \in \mathcal{V}_n$ in section n we derive a preliminary per-tile estimate by simply distributing of the per-section value uniformly across the component tiles, and rescaling by D, formally:

$$\hat{u}_k = \frac{\hat{r}_n D}{v_n}, \quad \forall k \in \mathcal{V}_n, \quad \forall n \tag{3.11}$$

wherein v_n denotes the size (in number of tiles) of section n. Finally, a simple 2D smoothing filter (e.g. circular gaussian) is run on the values of $\hat{u}_k$ in order to soften the artefactual discontinuities introduced by the hard-boundary tessellation, formally:

$$\hat{\boldsymbol{x}} = \boldsymbol{S}\hat{\boldsymbol{u}} \tag{3.12}$$

wherein $\boldsymbol{S}$ denotes the smoothing matrix.

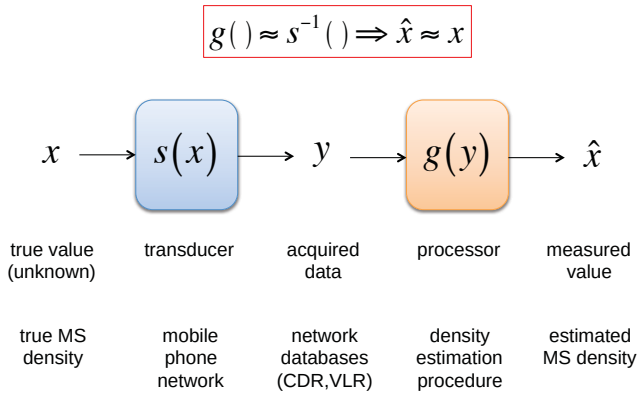


Figure 3.6: Abstract view of a generic measurement process.

3.8 Considerations on possible sources of error

Any measurement process or involves two logically distinct stages: data acquisition and processing. In the first stage, a "sensor" element (e.g., the retina or the camera) transforms some physical quantity x related to the object or phenomenon under measurement into a "signal" (data) y, leveraging some physical phenomenon that relates y to x through a transduction function $y = s(x)$. In the subsequent processing stage, an "intelligent" element (e.g., the brain of the computer) applies a processing procedure $g()$ to the acquired data and computes the final measured value $\hat{x} = g(y)$. The goal of the processing stage is to invert the transduction function and then reconstruct the original quantity with the highest possible fidelity level, i.e.

$$g() \approx s^{-1}() \Rightarrow \hat{x} = g(y) = g\left(s(x)\right) \approx x$$

This general process is depicted in Fig. 3.6. Generally speaking, two distinct types of error impede the exact reconstruction of the target quantity x:

- transduction function $s()$ being **not perfectly invertible**, e.g., due to quantisation or aggregation.
- transduction function $s()$ being **not perfectly known**, e.g., due to noise, incomplete knowledge of parameters, or any other unknown effect (deterministic or stochastic) taking place in the sensor.

It is important to remark that the loss of information due to quantisation/aggregation of the transduction function cannot be recovered by the subsequent processing stage. In our case, where the "transducer" role is played by the mobile phone network, this is accounted to the unavoidable loss of spatial detail due to the fact that MS positions can be "sensed" (at best) at the level of individual radio cells. In this respect, note that we can still infer density gradients within individual cells by leveraging the partial overlaps between adjacent cells — this is indeed captured by the formulation of the estimation process in terms of per-section variables (ref. §3.5). However, no information can be extracted about density gradients within sections and we refer to this source of error as *spatial quantisation error*.

In more concrete terms, the amount of information loss due to spatial quantisation depends (among other factors) on the particular network configuration, and particularly on the radio coverage patterns, i.e., location and size of radio cells. For this reason, different MNO networks might be "sensing" the same population with different levels of accuracy, and the network of the same MNO might yield different accuracy in different areas.

From the perspective of the population density estimation process, the "transducer" (i.e., the mobile phone network infrastructure) is given and cannot be changed. In other words, the spatial quantisation error represents an irreducible error floor for any network-based estimation method.

Besides that, our knowledge of the transduction function is not perfect, and this results in an additional source of error during the data processing stage which we call *estimation error*. This is due to several factors, most prominently: (i) a certain number of simplifying assumptions in the modelling of the network dynamics, hence in the "model" of the transduction function to be inverted; (ii) coarsely approximated knowledge of the *real* cell footprints (i.e., the area effectively serviced by a cell site); (iii) stochastic fluctuations (e.g., due to the wireless channel randomness) and (iv) spurious correlations between the transduction process and the phenomenon under observation. The latter is particularly insidious as it introduces a systematic distortion (or bias) in the final estimate.

Hereafter we provide a list of the main sources of errors that affect the estimation method presented in this Chapter.

- Inaccurate knowledge of cell coverage area: it is reasonable to expect that only a very coarse approximation of the cell footprint is available to the MNO, due to the intrinsic complexity and variability of the radio propagation channel.
- M2M devices: as discussed earlier in §1.6, the presence of MS for machine-to-machine (M2M) communications may inflate the MS counters and therefore lead to an over-estimation of population density. The problem will become more serious in the future, due to the expected growth of M2M devices served by mobile networks ("Internet of Things" paradigm). The problem can be counteract by applying more sophisticated M2M identification and prefiltering routines already in the data collection stage, but unavoidably the implementation of these routines will be highly MNO-specific.
- Biased cell-level location data: as discussed earlier in §3.3, the generation of cell-level locations data is conditioned to the occurrence of certain events (phone call, SMS, data connection, signalling procedure) related to the MS activity pattern. Since MS activity is not independent from time and space, the probability that the MS location is "observed" (sampled) at cell-level is correlated with the location itself (and with time). This in general leads to possible distortions in the final estimation, i.e., over- or under-representation of certain locations at certain times. Note that (at least part of) such correlations are systematic across different MNOs, and therefore can be perhaps mitigated but not completely eliminated by fusing data from diverse MNOs.
- MNO-specific customer base: the network infrastructure of a generic MNO can observe only part of the total population, and specifically (i) the customers of the MNO itself and (ii) customers of other MNO roaming into this network. In other words, every MNO "samples" part of the population, i.e., it can observe only a subset of all population members.

Note that the same person can be observed by two (or more) distinct MNOs if (s)he carries multiple subscriptions (e.g., a personal phone and a company phone). This leads to over-estimation of certain user groups (e.g., professionals), which demands systematic corrections. Besides duplications, attention must be paid to the fact that the customer base composition in general differs across MNOs. This introduces distortion (bias) representation of groups, among the subscribers of each individual MNO[4]. However such a bias can be reduced by jointly analysing data from different — possibly *all* — MNOs within the same market (country).

In principle, one can seek to reduce the estimation error by developing more sophisticated estimation algorithms based on more accurate (and complex) models of MS-to-network interactions and/or by leveraging external information from other systems. The *quantitative* assessment of the actual magnitude of these errors in real-world data remains a central direction for future research.

[4]Consider for example two MNOs m_1 and m_2 that are preferred, respectively, by low-income and high-income professionals. This "market specialisation" will cause luxury residential areas to be under-represented in m_1's data and over-represented in m_2's data.

Chapter 4

Exemplary Results with Synthetic data

This section provides numerical results for a simplified synthetic scenario. The main goal is to validate the correctness of the proposed method, and specifically the consistency of the maximum likelihood (ML) estimation method described in §3.6, and at the same time illustrate visually the type of outcome that can be expected. The simulation results presented hereafter represent the starting point for a future in-depth analysis of the performances, complexity and, most prominently, sensitivity of the proposed method to several real-world situations and potential sources of errors encountered in practice. In this sense, we do not aim here at reproducing each and every aspect of a "realistic" real-world scenario — a task that we leave for future work — but merely to illustrate the correctness of our approach in a simplified, reasonably well-behaved synthetic scenario.

4.1 Description of simulation scenario

We consider a network consisting of a square grid of 100×100 tiles hosting a total of $650,000$ MS. The MS are distributed randomly according to a bivariate distribution consisting of a mixture of three distinct Gaussian clusters, as shown in the "ground truth" map of Fig. 4.2(a).

We consider three types of cells with different footprint size and shape: (i) small sectors with 120° beam, (ii) medium-size circular cells and (iii) large circular cells. Note that three neighbouring 120°-sectors share the same cell tower. Cells are placed randomly according to an arbitrary design process that favour placement of more and smaller cells in most densely populated areas. This reflects the characteristic of real-world deployments, where the spatial distribution of radio capacity (i.e., more and smaller cells) tend to follows the *maximum traffic intensity* (peak-hour traffic), and therefore correlates positively with the (maximum) population density.

We consider two distinct scenarios (Scenario #1 and Scenario #2) with two different cell coverage patterns extracted randomly from the same process. For Scenario #2 we manually modified the cell placement in order to introduce a strong local mismatching in a particular region, as explained in detail later in §4.3.2.

The number of simulated cells is 56 for Scenario #1 and 117 for Scenario #2. In each scenario, cells are arbitrarily grouped into 5 simulated LAs of different size. An excerpt of the cell/LA footprints in Scenario #1 is depicted in Fig. 4.1.

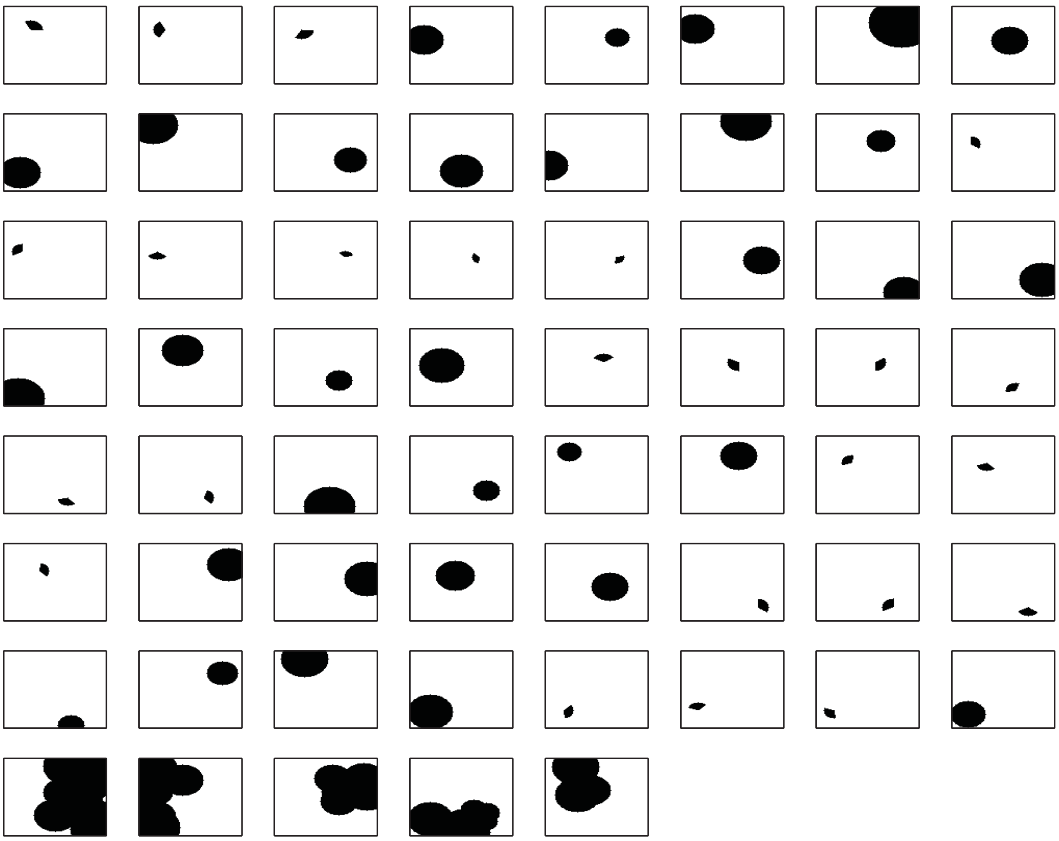


Figure 4.1: Examples of cell footprints (top seven rows) and LA footprints (bottom row) used in Scenario #1. Every square depicts the entire area of interest (toy world) with a single footprint in black. Triplets of 120°-sectors sharing the same cell tower are plotted in consecutive squares (see e.g. three leftmost squares in the top row).

A synthetic assignment process emulates the extraction of cell and LA counters from combined CDR+VLR data. The initial set of cell and LA counters is generated according to the probabilistic model described in Appendix A: in summary, a generic MS covered by z cells attaches to a randomly selected cell, all z cells being equally likely to be selected, and is assigned with probability ρ and $1-\rho$ respectively to the cell or to the corresponding LA. In our simulations we have set $\rho = 0.22$, which is a good approximation for most networks. By considering a constant value of the "activity probability" ρ we obtain a synthetic dataset free from spatial bias. This motivates the setting $\gamma = 1$ in the stage of LA counter projection (ref. §3.4).

4.2 Reference method: CDR with Voronoi tessellation

For the sake of completeness, we compare the proposed method with an alternative approach based on Voronoi tessellation that reflects the current state-of-the-art in the research literature.

Given a set of V points called "seeds", the Voronoi tessellation (or Voronoi diagram) assigns every point in the area of interest to the nearest seed in terms of euclidean distance [4]. The locus of all points assigned to one seed defines a Voronoi "region" with the shape of a (irregular) polygon. Generally speaking, the size of a generic region scales inversely with the local seed density.

The key components of the "basic Voronoi" method adopted by most previous literature (including the recent work by Deville *et al.* [8]) are:

- Only cell-level locations from CDR data are considered: LA counters (that could be extracted from VLR) are not available.
- The only spatial information associated to the cell is the location of the cell tower: no cell footprint nor cell size data are available.

Therefore, with the basic Voronoi method all cell counters are mapped to the Voronoi region corresponding to the cell tower, and local density is obtained by dividing this value by the size of the region. Note that in our toy-world (as well as any real-world network) the number of cell towers is smaller than the number of cells, since one tower can serve multiple cells (e.g., three adjacent 120°-sectors). Generally speaking, the basic Voronoi method uses *less information* than our method, and therefore it can be easily expected that it will lead to a less accurate final estimate — the interesting question is whether the accuracy gain of our method is substantial or not.

Recall that in our toy-world the call activity ρ does not vary in space, hence the initial set of cell-level counters is free from bias, and consequently the loss of information due to disregarding LA counters has a negligible impact on the estimation of the relative spatial density. Instead, the lack of cell footprint information represents a serious disadvantage of the basic Voronoi method compared to our approach. For this reason, we consider also an "improved Voronoi" scheme that takes in input the same cell footprint data as our method, but handles it in a different way. In the improved Voronoi version, the centroid of every cell constitutes an independent seed, hence the number of Voronoi regions equals the number of cells, not towers. Furthermore, the cell counter is adjusted to account for the actual size of the cell footprint. A similar method was adopted in [5].

It can be easily expected that augmenting the Voronoi method with accurate cell footprint information will improve the fidelity of the final result with respect to the basic Voronoi scheme. Here we

are interested to compare the improved Voronoi approach with the proposed estimation method that uses the same information — cell footprints and counters — but in a different way.

4.3 Numerical results

4.3.1 Scenario #1: a well-behaved case

The color map in Fig. 4.2(a) shows the "ground truth" distribution generated for Scenario #1, aggregated at the tile level (K =10,000 total tiles). The three clusters A, B and C are evident.

Identification of reference bound

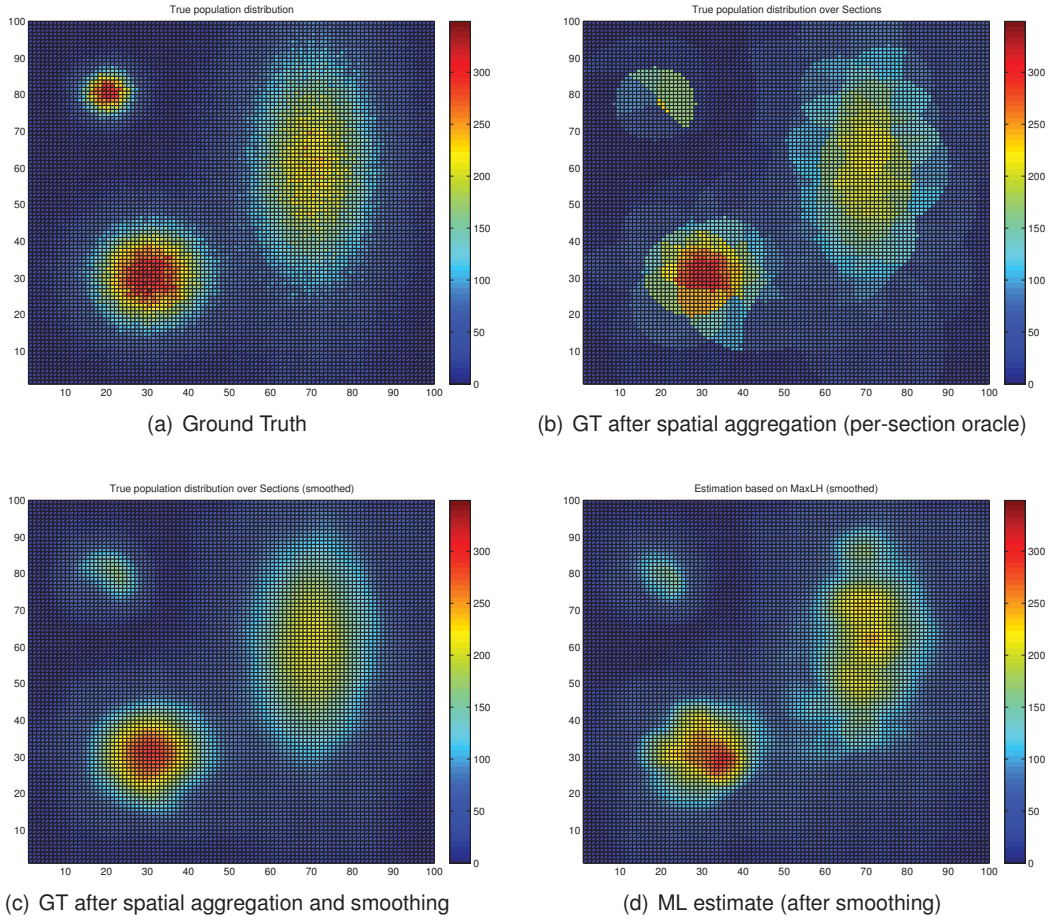
Fig. 4.2(b) shows the ground truth distribution aggregated at section level — recall that every section represents the intersection area of a specific subset of cells. After passing the latter through a smoothing filter we obtain the map in Fig. 4.2(b), which represents the output (after smoothing) of an ideal "oracle" that knows without error the ground truth distribution *at the level of individual sections*. In other words, moving from the per-tile ground truth of Fig. 4.2(a) to the map in Fig. 4.2(c) has introduced exclusively a *spatial aggregation error* but no *estimation error*. It is important to realise that the spatial aggregation error (at per-section level) is intrinsic to the usage of the mobile phone network, and specifically of network-based data, for the detection of MS locations: the unavoidable loss of spatial detail is due to the fact that MS locations can be "sensed" (at best) at the level of individual radio cell. For this reason, given a radio network coverage pattern (i.e., the given set of cell footprints) and without any further external information, the "oracle" map in Fig. 4.2(c) represents the ideal reference bound against which any density estimation method based on network data must be compared.

Output of the proposed method

In Fig. 4.2(d) we report the density map obtained by the ML estimation procedure described in §3.6. Moving from Fig. 4.2(c) to (4.2(d)) a certain *estimation error* has been introduced. The comparison between the two maps shows that the quality of the final estimate is rather good: all three clusters are clearly distinguishable. Note that while cluster C (upper left) has been slightly faded out, the cluster B (lower left) has been resolved very accurately. Such differences are due to the local coverage pattern in the cluster region: the more redundant the local coverage (higher number and smaller size of cells) the better the estimation accuracy.

Comparison with Voronoi schemes

In Fig. 4.3 we plot the results obtained by the two Voronoi schemes. As expected, the injection of cell footprint information improves somewhat the performance of the improved Voronoi approach compared with the basic Voronoi (compare Fig. 4.3(c) against Fig. 4.3(b)), but in both cases the result is considerably less accurate than the proposed method. This is further confirmed by the distribution of the absolute errors plotted in Fig. 4.4.



(a) Ground Truth

(b) GT after spatial aggregation (per-section oracle)

(c) GT after spatial aggregation and smoothing

(d) ML estimate (after smoothing)

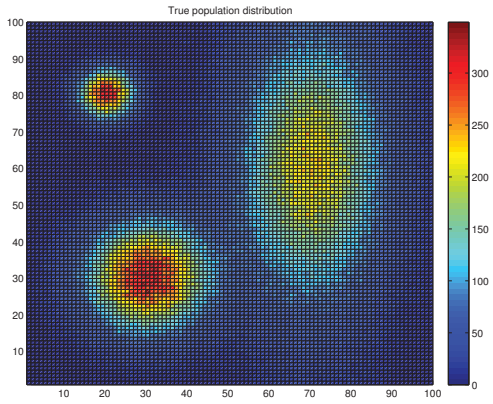
Figure 4.2: Spatial distributions for Scenario #1. Three density clusters are clearly visible, respectively, on the right side (cluster A), on the bottom left corner (cluster B) and on the top left corner (cluster C).

Recall from the previous discussion that the "improved Voronoi" scheme is fed with the same topological data as the proposed scheme (full cell footprint), but it uses these data in a considerably less effective manner. In other words, as with any estimation task, the quality of the solution is not only a matter of *what* information is used, but also *how* it is used.

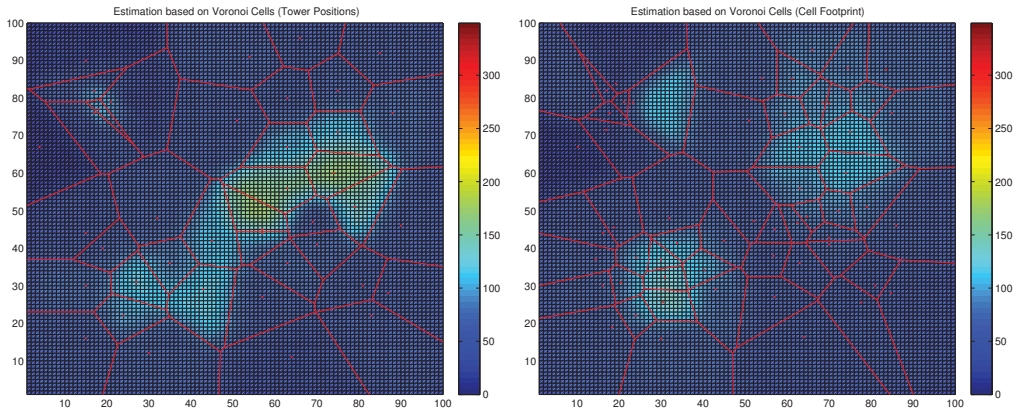
4.3.2 Scenario #2: a stressed scenario

Motivation

The goal of this second set of simulations is to illustrate one possible limitation of the general approach of estimating people density from mobile phone network data. Recall the discussion in §3.8 about the distinction between "spatial quantisation errors" and "estimation errors", and unrecoverable loss of information that, in principle, might be caused by the former. The previous



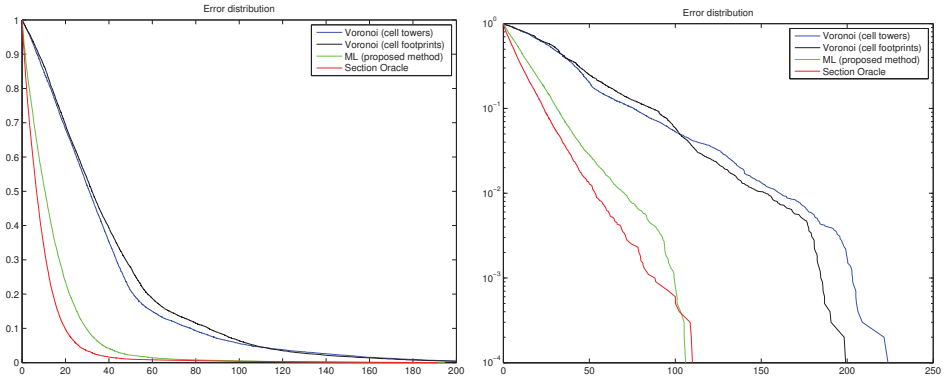
(a) Ground Truth



(b) Basic Voronoi method (Cell Towers)

(c) Improved Voronoi method (Cell Footprints)

Figure 4.3: Estimated distributions with Voronoi method for Scenario #1 (compare with Fig. 4.2).



(a) Per-tile error

(b) Per-section error

Figure 4.4: Estimation error distributions (CCDF) for Scenario #1.

Scenario #1 has shown a case where the sum of both errors is somewhat acceptable, in the sense that all three population clusters could be properly "sensed", though with different levels of accuracy. In this second Scenario #2 we provide a negative example, where a particular configuration of the radio network coverage would cause one of the clusters to be missed.

Differences with previous scenario

In Scenario #2 we consider the same population distribution of the previous scenario but a different radio coverage pattern. We introduce the following two modifications with respect to the previous scenario:

- The number of cells is higher, roughly doubled from 56 in Scenario #1 to 117 in Scenario #2.
- We have manually repositioned some cells away from the area around cluster C in order to create a strong local "density mismatching" between the MS density (high) and the cell density (very low) in this specific area.

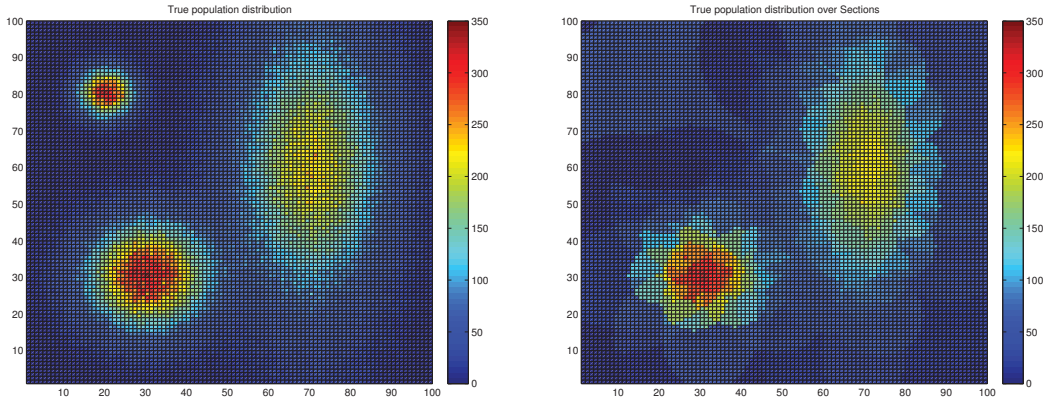
Roughly speaking, the first modification brings a potential advantage for all estimation methods (the proposed schemes as well as the Voronoi methods), while the second one represent a serious disadvantage, as we show in the following.

Interpretation of the results

The new maps are shown in Fig. 4.5 (ground truth and proposed method) and Fig. 4.6 (Voronoi). The fidelity of the basic Voronoi scheme remains pretty poor. It appears that the increase of cell number (hence cell density) benefits especially the "improved Voronoi" method, particularly in the region of cluster B that now becomes clearly visible. However, a closer comparison of Fig. 4.6(c) with the ground truth map of 4.6(a) reveals that cluster B is being seriously *overestimated* by Voronoi.

Note from Fig. 4.5(d) that cluster C has been missed by al estimation methods, including the ML estimation approach. This is exactly the sort of "information loss" that we intended to reproduce by purposely introducing a marked local mismatching. In fact, in Scenario #2 cluster C is covered only by the edges of a couple of large cells, and for this reason the corresponding MS observations are "diluted" over a large area spanning the whole upper left quadrant of the toy-world area. A close look at the "oracle" maps in Fig. 4.5(b) and Fig. 4.5(c) reveals that the disappearance of cluster C is to be accounted to the spatial aggregation error that is intrinsic to the usage of a mobile phone network, rather than to the subsequent data processing stage. In other words, the problem is not due to the ML estimation algorithm failing to detect cluster C in the input dataset, but rather to the fact that cluster C has not been captured by the network-based dataset in the first place, due to the extreme sparsity of (local) radio coverage. However, we expect that similar cases of strong local mismatching to be very rare (though certainly not impossible) in real-worlds deployments.

Finally note that also in Scenario #2 the accuracy of the Voronoi method fall well behind ML estimation (ref. also the error distributions in Fig. 4.7).



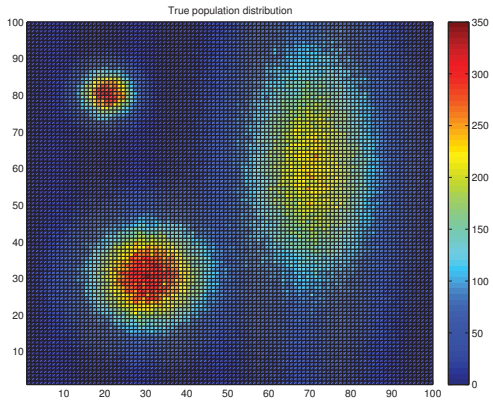
(a) Ground Truth (GT)

(b) GT after spatial aggregation (per-section oracle)

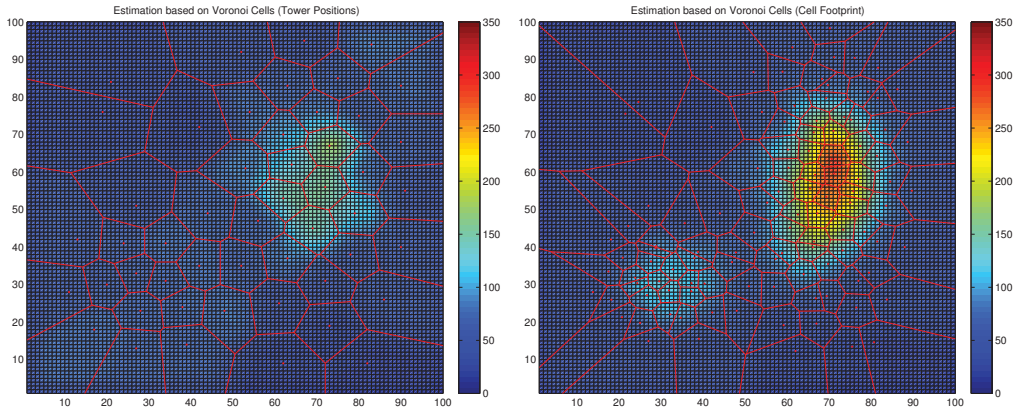
(c) GT after spatial aggregation and smoothing

(d) ML estimate (after smoothing)

Figure 4.5: Spatial density maps for Scenario #2. Note that Cluster C is missing already in the "oracle" map due to the particularly "low" degree of radio coverage in that area. Consequently, Cluster C is missed also by the final ML estimate)



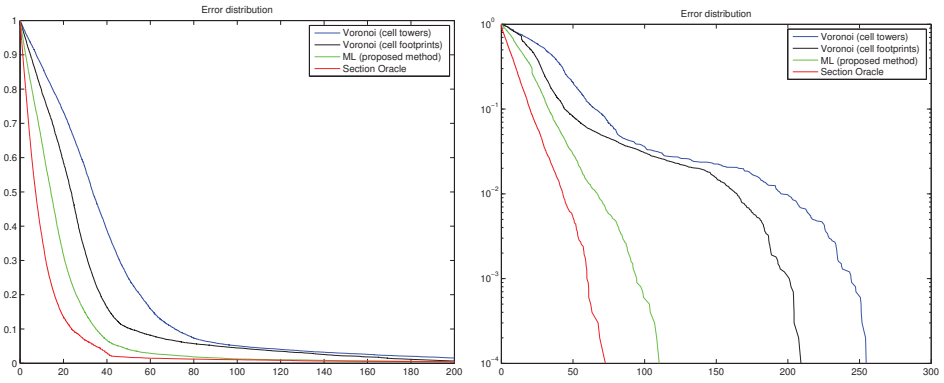
(a) Ground Truth



(b) Basic Voronoi method (Cell Towers)

(c) Improved Voronoi method (Cell Footprints)

Figure 4.6: Estimated density maps with Voronoi method for Scenario #2 (compare with Fig. 4.5).



(a) Per-tile error

(b) Per-section error

Figure 4.7: Estimation error distributions (CCDF) for Scenario #2.

s

4.3.3 Considerations about the representativeness of simulations for real-world scenarios

Recall that the above simulation results were obtained on synthetic data generated according to a simplified toy-model. While they cannot be taken as a final "proof" of performance with real-world data, they are nevertheless informative and provide an initial indication of what might be expected in a practical deployment. Here below we summarise the most important points learned from the above results, along with some considerations and conjectures based on our expert knowledge, with the high-level goal of motivating further experimentation with samples data from (possibly multiple) real-world MNOs.

First of all, we remark that the good result obtained in Scenario #1 validate the correctness of the ML estimation method formulated in §3.6. Recall that the synthetic data were produced according to the same generative model underlying the ML estimation model (ref. Appendix A). Further work is needed (i) to assess the sensitivity of the estimation process to various possible sources of model mismatching encountered in real-world data (e.g., unequal cell selection probability) and from there (ii) to develop robust estimation processes. Of particular importance is to gain a better understanding of additional bias due to (possibly time-varying) correlations between MS location and MS activity (for both CDR and augmented VLR data) and develop ways to counteract it. The research along these lines must be based on real-world data, possibly from different MNOs.

Second, the "disappearance" of Cluster C (from the oracle maps in Fig. 4.5(b) and Fig. 4.5(b)) in Scenario #2 should be taken as an instructive warning of the type of information loss that in principle might occur in areas where the radio coverage is particularly "thin" (i.e., with only few large cells). However, we expect that such cases will occur rarely in practice and will represent "anomalous" patterns rather than typical behaviour. In fact, real cellular networks are engineered and regularly (re)optimised to "match" the radio capacity to the "normal" traffic density observed locally in the peak-hour period. Occasional mismatching might be generated when a very high number of people gather in a country area that normally yields very low traffic density. Furthermore, even in these cases one might conjecture that the impact of spatial mismatching is somehow mitigated by an increased call activity of the people involved in that event — think for example to real-world cases like a big one-time concert in a remote area, or a severe road congestion in the countryside.

Chapter 5

Summary of main findings and points for further study

In this chapter, we summarise the main results of this study and we point out promising directions for future work.

Importance of better network topology data

The simulation results in Chapter 4 indicate that major gains in estimation accuracy can be obtained by integrating mobile operator data with additional topology data. For example, the density estimation procedure (and likely any other spatial analysis based on mobile phone data) would greatly benefit from the use of accurate cell coverage maps. Now, while cell coverage maps are measured or estimated at great computational costs, coarse approximations are in general available. In addition, any coarse approximation of cell footprint (e.g. obtained by static antenna configuration parameters) improves the data location resolution of mere (exact) tower location, and the simulation results presented in this study support this claim. As MNO typically possess this information, we propose to implement internal processes so as to prepare cell coverage data for their use in combination with CDR/VLR data for improved spatial analysis. Note also that, when cell coverage maps are made available, the (inter-)section tessellation defined in this study can greatly improve the results obtained by assuming a Voronoi tessellation method.

Understanding and quantifying the risk of spatial bias

In this work, we have often commented on cell-level location data being exposed to the risk of spatial estimation uncertainty (or bias). In fact, due to the functional dynamics of the mobile phone network, the estimation of a cell location is always conditioned to the event of a subscriber starting a phone call or sending an SMS, whose probability of occurrence typically depends on the MS's context and location. Furthermore, the correlation structure between MS location and MS activity might be varying with time. This introduces a certain risk of under- or over-representing certain specific locations in cell-level data, leading to distorted view of the population distribution in space and/or time.

Notably, the problem lies exclusively in cell-level location data (both from CDR and augmented VLR), i.e., it is contained at a small scale. For this reason, we conjecture the existence of a fundamental trade-off between spatial accuracy and risk of bias — a phenomenon that is somewhat reminiscent of the bias-variance trade-off in statistics and machine learning [11].

Hence, additional work is required for a better understanding of the various sources of spatial/temporal bias in real MNO dataset, and to quantify the resulting distortion in the final density estimation. In this respect, the complexity of this task is aggravated by the fact that reference “ground truth"data might not be available in practice, and that it might be necessary to resort to comparative studies across different MNO, with different network configurations and customer population characteristics. Nevertheless, the integration of dataset from different operators (e.g., on the basis of joint pilot studies, or projects) is deemed as a promising strategy for reducing uncertainty and obtain accurate estimations.

Counteracting the risk of bias

Another important challenge is to develop effective approaches to counteract the spatial/temporal bias that is possibly present in cell-level data. The adoption of adjustable parameters, such as γ in §3.4 should be considered as a very simple initial attempt to address such a problem. Alternative approaches might consider calibration strategies based on reference data (e.g., census data [8]) or leveraging external data (e.g. land use), which, however, require additional countermeasures to prevent error propagation across datasets.

Quantifying the cost and benefit of VLR data

Numerous case-studies investigating CDR applications demonstrate that the effort required for the extraction and preparation of such data is affordable for many MNOs. Unfortunately, there is no indication about the feasibility (and costs) of large-scale extraction of VLR data, nor about the achievable gains (e.g., in terms of population coverage, reduced bias, temporal resolution) that VLR data can bring to the task of population density estimation. As a result, further experimental work is required to quantify the cost and the potential benefits of complementing CDR with VLR data. The intention of this study was to provide a unified methodological basis for the joint processing of cell-level and LA-level data, hence for the fusion of CDR and VLR, and to shed light on the opportunity of network data exploration besides traditional CDR data sets.

Towards a multi-MNO pilot study

A number of research directions identified during this study would involve the fusion of, or at least the comparison between, network-based data extracted by different MNOs. We do not refer here to “raw" CDR/VLR data nor any other type of micro-data — that in our data processing model never leave the MNO domain — but to highly-aggregated intermediate data: preferably Map/Counter Dataset (MCD), or at least density maps (ref. §3.1). In order to pioneer the joint processing of multi-MNO dataset we envision the launch of pilot projects involving different MNOs for the coordinated extraction of sample datasets to be further processed and analysed by a trusted entity (e.g. JRC or Eurostat). It would be highly desirable to involve in the pilot study at least

two or three MNOs competing on the same national market. This would allow the investigation of the relative differences in the individual MNO's "views" (due to different network configurations and customer basis) as well as the quantitative assessment of the relative gain — in terms of spatial accuracy and/or bias mitigation — achieved by the two multi-MNO data fusion strategies presented in §3.1, namely MCD fusion vs. individual map fusion.

We expect that several European MNOs will favourably consider the perspective of engaging in a common multi-MNO pilot study, simply by considering it as an opportunity cost. Hereby, the efforts for the preparation of a sample dataset (e.g., CDR plus cell topology data) to serve as input for the pilot study will probably not exceed a few person-months, considering that such data are anyway available inside MNOs — still they need to be properly prepared, curated and pre-processed. Among expected benefits, there is a growing consensus among MNOs on the commercial value of the data in their possess, and the federation of multi-MNO data — at least within the limited scope of a pilot project — bears the potential to stimulate new applications, attract new customers that are not at reach of individual MNOs, as well as European institutions supporting public policies. We hope this study will contribute concretely to move some steps in this direction.

Appendix A

Reference generative model

In this appendix we detail the simple generative model underlying the model parameters in the ML estimation procedure developed in §3.6. The same model was used to generate the synthetic data in Chapter 4.

Consider a generic MS i placed in section n and attached to the network of the mth MNO at the reference time t^*. Recall that $\mathcal{Z}_n$ represents the set of cells covering (each tile of) section n, and z_n the number thereof. Every MS can be camped only in a single cell at any given time, and in case that multiple cells are available ($z_n > 1$) we assume that every cell has the same probability of being selected. Hence:

$$p_{nj} = Prob\{\text{MS } i \text{ camped in cell } j \mid \text{MS } i \text{ located in section } n\} = \begin{cases} 0 & \text{if} \quad j \notin \mathcal{Z}_n, \\ z_n^{-1} & \text{if} \quad j \in \mathcal{Z}_n. \end{cases}$$

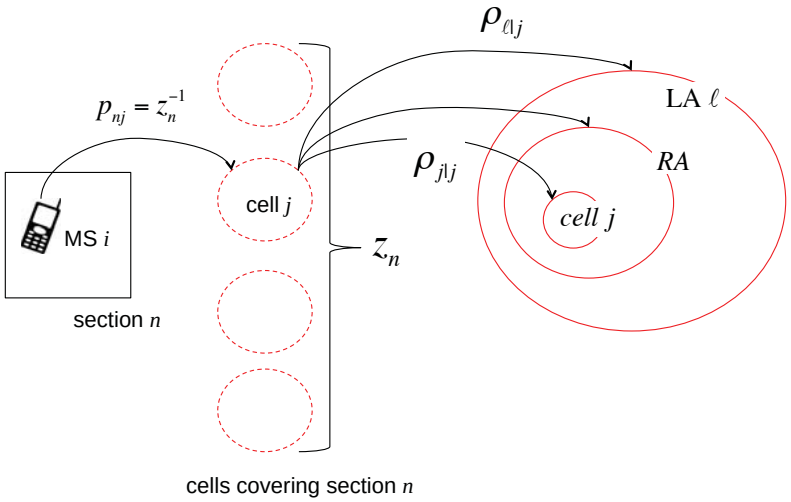


Figure A.1: Scheme of the simple probabilistic model for MS-to-location assignment underlying the estimating method described in §3.6.

The fact that MS i was camping to cell j is a necessary but not sufficient condition for i to be "observed" in cell j by the network measurement process. More precisely, MS i can be assigned to three different locations: the cell j itself (in the best case), the associated RA or the greater LA (in the worst case), as depicted in Fig. A.1. In practice, several factors concur to determine the mapping area for MS i: (i) which source databases are considered by the measurement process (CDR and/or VLR); (ii) the configuration of the MS (e.g., whether it is attached exclusively to the CS domain or to the PS domain too) and (iii) the recent activity pattern of the MS (e.g., whether it has performed voice calls in the reference observation interval). As discussed earlier in §3.3, the MS activity is not independent from the current MS location, and this introduce a certain risk of bias, due to possible over- or under-representation of specific locations. In the simple generative model we disregard these types of correlations.

Formally, consider a cell j included in LA ℓ, i.e., $j \in \mathcal{A}_\ell$. Denote by $\rho_n \leq 1$ the activity coefficient in section n, i.e.

$$\rho_n \stackrel{\text{def}}{=} Prob\{\text{MS } i \text{ active} \mid \text{MS } i \text{ located in section } n\}.$$

By assuming that every MS is observed either in the respective cell of LA (for the sake of simplicity we do not consider RA here), it holds that:

$$\begin{aligned} q'_{nj} \stackrel{\text{def}}{=} & Prob\{\text{MS } i \text{ observed in cell } j \mid \text{MS } i \text{ located in section } n\} \\ = & Prob\{\text{MS } i \text{ observed in cell } j \mid \text{MS } i \text{ camped in cell } j, \text{ located in section } n\} \\ & \cdot Prob\{\text{MS } i \text{ camped in cell } j \mid \text{MS } i \text{ located in section } n\} \\ = & \rho_n p_{nj}. \end{aligned}$$

In the simple generative model we assume an uniform activity coefficient, i.e:

$$\rho_n = \rho, \quad \forall n. \tag{A.1}$$

With this simplifying assumption, the optimal value of the projection factor is $\gamma = 1$ (ref. §3.4), and therefore the probability that a generic MS located in section n is mapped to the (projected) counter d_j reduces to $q_{nj} = p_{nj}$.

Appendix B

Preliminary analysis of LA sizes from OpenCellID database

The mobile network can track the position of all MSs in active or idle mode at least at the level of Location Areas (see Sect. 1). Moreover, although the position of active MSs is known at cell level, this information is not necessarily included in the VLR. This section aims to provide a rough estimation of the spatial granularity of localizations at LA level based on a large dataset of concurrent GPS position and Cell-ID (equivalently: CGI) recordings. Such a dataset is provided free of charge by OpenCellID [2].

The OpenCellID database

OpenCellID is a large collaborative project collecting GPS location data for cell identifiers (Cell-ID), with the main application of providing power-efficient and fast location information to mobile devices. As of August 2014, over a billion measurements were collected, which are publicly available under a free Creative Commons license [2]. The data is collected fully automatically by registered users via various smart phone apps. Although this database is primarily intended to provide a mapping from given cell ids to geo-locations, we use it in this study to estimate the localization error - i.e. the spatial granularity - of device localizations based on cells or Location Areas.

However, since OpenCellID data are collected by volunteers, compiled automatically and provided free of charge, there is no guarantee regarding the quality of the data. Typical errors in the OpenCellId database include

- *Erroneous cell-IDs*: occasionally the recorded cell ids are wrong. Typically in these cases, the cell id, local area code (LAC) and mobile network code (MNC) are mixed up. In practice, the erroneous cell-IDs do have only few GPS measurements attached and can be easily filtered out by introducing a threshold to the number of measurement records for each cell.
- *Unrealistic cell sizes:* The GPS measurements of some cells are distributed across a whole country which is clearly unrealistic. The reason for this type of errors is not known to the authors. One possible explanation for such phenomena are so-called “Cell-On-Wheels"

(COW) or “Cell In A Box" (CIAB). These mobile antennas are used by MNOs to provide temporary service with temporary equipment, e.g. to cover increased demand at specific events. Since mobile antennas can change their position and covered area, their ID will be attached to GPS measurements which vary greatly over time and can yield a distorted picture of actual cell sizes.

- *“Antenna dragging"*: This type of data artifact seems to be caused by devices not updating the cell id during a trip, reporting the original cell id throughout and wrongly attaching it to GPS measurements.
- *Outliers*: Often, antennas with a number of plausible and consistent measurement points have additional GPS positions attached that are far away from the other measurements and are obviously wrong (often they are often not even in the same country).

Moreover, dependencies between measurements collected by the same contributor can introduce distortions, but the OpenCellID database does not include any identifier of the device nor the person having collected the GPS and cell measurements. Robustly estimating LA sizes from the OpenCellID data in the presence of such errors and biases therefore involves an initial preprocessing step for data cleansing, and remaining noise is coped with by employing robust statistics to analyse the spatial extents of cells and LAs.

Analysis method

To alleviate biases towards “heavy contributors” and dependencies between successive measurements during a trip or repeated measurements at the same location, we apply the following filtering scheme:

- for each cell in the mobile network only one measurement per hour is retained, and all additional measurements are discarded;
- in a spatial 10m-by-10m grid only one measurement per grid-cell is retained, and all additional measurements are discarded.

Furthermore, cells and LAs with too few measurements are not included in the analysis:

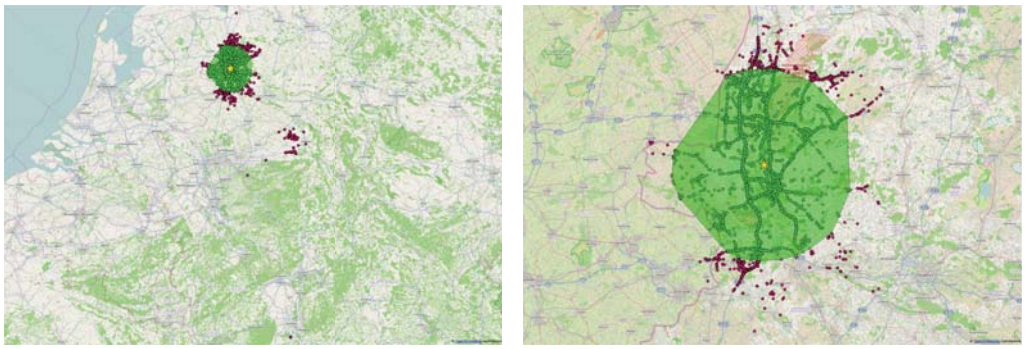
- Cells with less than 20 retained measurements are discarded;
- LAs with less than 10 different cells having a sufficient number of measurements are excluded from analysis.

For the remaining Location Areas we define a robust centroid using only the retained measurements. While the median is a common robust measure for one-dimensional location, it does not generalize easily to higher dimensions. Several such generalizations are known [7], and for our analyses we use the *centerpoint*, which is defined as a point for which each hyperplane through the centerpoint divides the point cloud into two subsets such that the smaller of these subsets has at least a $\frac{1}{d+1}$ fraction of the points. The algorithm provided in [6] provides a fast probablistic approach for computing centerpoints.

	Total		Urban Areas		Rural Areas	
Country	#LAs	#measurements	#LAs	#measurements	#LAs	#measurements
Germany	2028	122,253,897	334	14,252,655	1694	108,001,242
France	350	1,246,915	126	362,792	224	884,123
Italy	115	167,912	57	77,878	58	90,034
Austria	290	2,776,196	66	470,561	224	2,305,635

Table B.1: Number of LAs and measurements used for LA size estimation.

We characterize the size of a Location Area by the distances of each of the retained GPS measurement to the centerpoint, and to cope with outliers we use the 90^{th} percentile of these distances as robust statistic. The distribution of the obtained LA size estimations can help to get a picture of the spatial granularity of localizations based on Location Area IDs, e.g. for the purpose of estimating population densities. An example of a Location Area and its estimated size based on OpenCellID measurements is shown in Figure B.1. Figure B.2 shows the estimated spatial extents of all the LAs of one german mobile operator, which had enough data available to be included in this analysis.



(a) (b)

Figure B.1: Example of measurements of a single Location Area in the OpenCellID database: a) Zooming out reveals outlying GPS measurements (red dots) with large distances to the robust centerpoint (yellow star). b) The convex hull of the measurements within the 90^{th} percentile of distances to the centerpoint (green dots) approximates the spatial extent of the Location Area.

Results

We applied the analysis method described above to four different countries: Germany, France, Italy, and Austria. In our analysis we included the networks of all MNOs operating in these countries. Since the sizes of cells and Location Areas differ significantly between urban and rural areas, we computed the size distributions for urban and rural areas separately. Location Areas belonging to urban areas were identified by matching their center point to a map of densely populated areas, which is publicly available at [1] and depicted in Fig. B.3. The number of LAs and measurements used for this analysis in each of the four contries is shown in Table B.

The resulting LA size distributions are shown in Figure B.4. The median of the LA size in urban

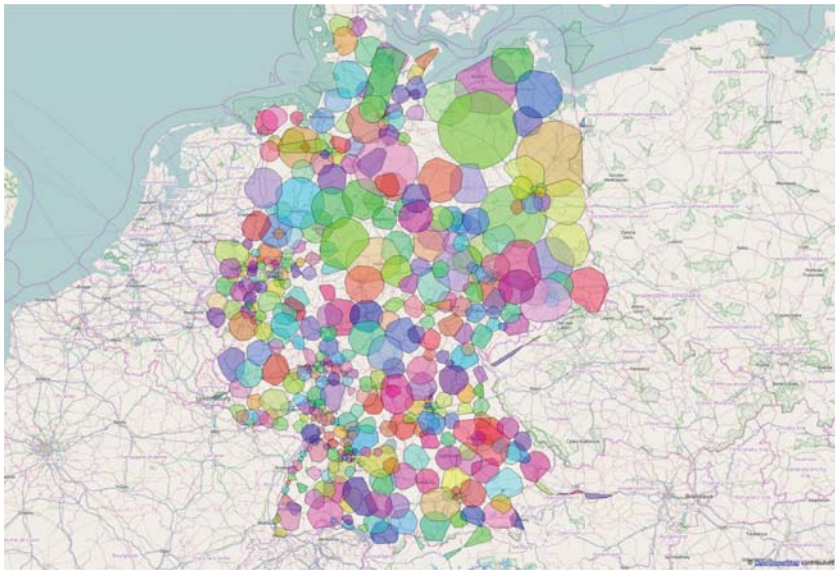


Figure B.2: Estimated spatial extents of all the LAs of one german mobile operator, which had enough data available to be included in this analysis.

areas is about 9km in Germany, about 10km in France and Italy, and about 6.5km in Austria. In rurals areas the median LA size estimation was about 18km in Germany, about 20km in Italy and Austria, and about 26km in France.

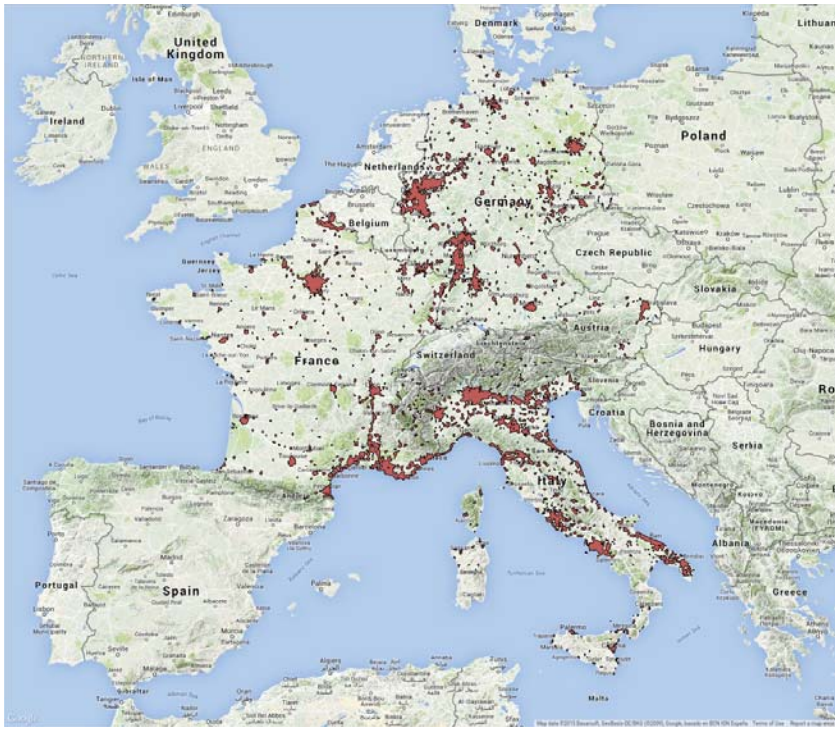


Figure B.3: Densely populated areas used for the analysis of Location Area sizes in urban and rural areas (data taken from [1].

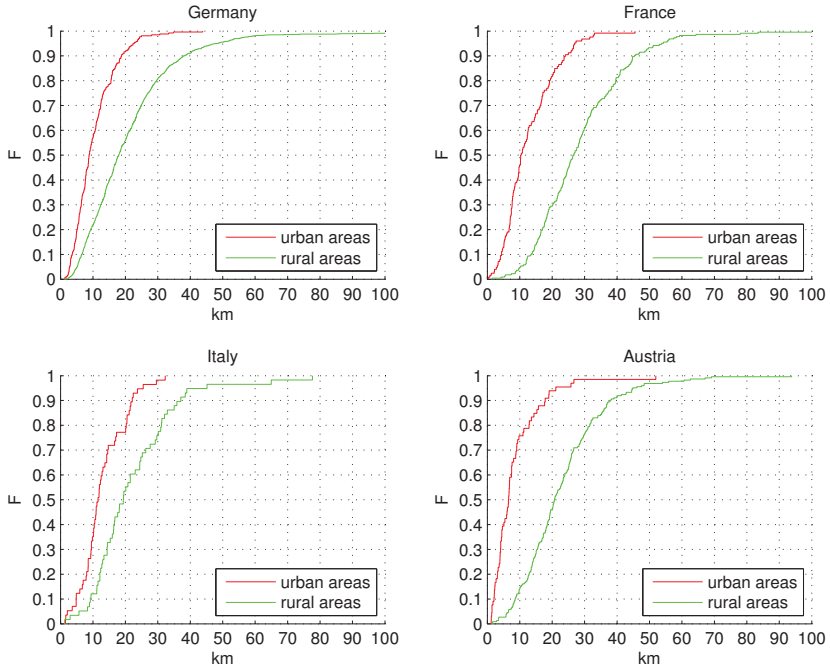


Figure B.4: Emprical CDF of Location Area sizes in Germany, France, Italy, and Austria.

List of Acronyms

APN	Access Point Name
BSC	Base Station Controller
BTS	Base Transceiver Station
CDR	Call Detail Record
CGI	Cell Global Identifier
CN	Core Network
CS	Circuit Switched
GGSN	Gateway GPRS Support Node
HLR	Home Location Register
IMEI	International Mobile Equipment Identity
IMSI	International Mobile Subscriber Identity
LA	Location Area
LAC	Location Area Code
LAI	Location Area Identity
LAU	Location Area Update
MCC	Mobile Country Code
MCD	Map/Counter Dataset (*)
MNC	Mobile Network Code
MNO	Mobile Network Operator
MS	Mobile Station
MSC	Mobile Switching Center
PLMN	Public Land Mobile Network
PS	Packet Switched
RAN	Radio Access Network
RA	Routing Area
RAC	Routing Area Code
RAI	Routing Area Identity
RNC	Radio Network Controller
SIM	Subscriber Identity Module
SGSN	Serving GPRS Support Node
SMS	Short Message Service
TA	Tracking Area
TAC	Type Allocation Code
T-IMSI	Temporary IMSI
VLR	Visiting Location Register

(*) This acronym was defined in this document and is not part of the standard 3GPP terminology.

Bibliography

[1] Natural earth data. www.naturalearthdata.com. Accessed: 2015-01-10.

[2] OpenCellID. http://opencellid.org. Accessed: 2014-10-20.

[3] ETSI TS 132 215. Charging data description for the packet switched (ps) domain. In *http://www.etsi.org/deliver/etsi_ts/132200_132299/132215/05.09.00_60/ts_132215v050900p.pdf*.

[4] F. Aurenhammer. Voronoi diagrams — a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3), 1991.

[5] Center for Spatial Information Science — Univ. of Tokyo. A study on urban mobility and dynamic population estimation by using aggregate mobile phone sources. http://www.csis.u-tokyo.ac.jp/dp/115.pdf.

[6] K. L. Clarkson, D. Eppstein, G. L. Miller, C. Sturtivant, and S.-H. Teng. Approximating center points with iterative radon points. *Int. J. Comput. Geom. Appl.*, 357(06), 1996.

[7] H. Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer Verlag, 1987.

[8] P. D. *et al.* Dynamic population mapping using mobile phone data. *PNAS*, 111(45), November 2014.

[9] F. Ricciato. Traffic monitoring and analysis for the optimization of a 3g network. *IEEE Wireless Communications — Special Issue on 3G/4G/WLAN/WMAN Planning*, 13(6), December 2006.

[10] F. Ricciato *et al.* Traffic monitoring and analysis in 3G networks: lessons learned from the METAWIN project. *Elektrotechnik & Informationstechnik*, 123/7/8, 2006.

[11] J. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1), 1997.

[12] INSPIRE Thematic Working Group Coordinate Reference Systems and Geographical Grid Systems. D2.8.I.2 Data Specification on Geographical Grid Systems — Technical Guidelines. http://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_GG_v3.1.pdf. Accessed: 2015-03-27.

[13] A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE Transaction on Intelligent Transportation Systems*, 2015.

[14] F. Pantisano and M. Craglia. Mobile network operator data to support urban planning and management. *JRC working document*, 2015.

[15] S. Tartarelli, N. d'Heureuse, and S. Niccolini. Lessons learned on the usage of call logs for security and management in ip telephony. *IEEE Communications Magazine*, 48(12), December 2010.

European Commission
Joint Research Centre – Institute for Environment and Sustainability

Title: Estimating population density distribution from network-based mobile phone data

Author(s): Fabio Ricciato, Peter Widhalm, Massimo Craglia and Francesco Pantisano

2015 – 65pp. – 21.0 x 29.7 cm

JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation