UTRECHT UNIVERSITY

Department of Social and Behavioural Sciences

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences (research) master thesis

Tackling uncertainty in the spatial density estimation from Mobile Network Operator data

First examiner:

Dr. Peter Lugtig

Candidate:

Marco Ramljak

Second examiner:

Dr. Fabio Ricciato

01.12.2021

Abstract

The processing pipeline from raw Mobile Network Operator (MNO) data to the final spatial density map requires modeling the (approximate) spatial footprint of cells – a task called "cell geo-location." Recent work has shown that, with appropriate estimation methods based on probabilistic models, the utilization of more detailed cell footprint information improves the final estimate's spatial accuracy considerably, compared with the simpler methods relying on Voronoi tessellations. However, such results were obtained (i) under the assumption of perfect cell footprint knowledge and (ii) limited to a single scenario characterized by a dense multilayer coverage pattern with a high degree of cell overlapping.

In this methodological sensitivity analysis, we investigate through simulations the robustness of probabilistic estimators to uncertainties and inaccuracies in the model input parameters, namely (i) the matrix of emission probabilities and (ii) prior information. To this aim, we develop parametric techniques that purposefully introduce inaccuracies into the estimation model with tunable magnitude. Also, we consider distinct prior information vectors with varying levels of informativeness.

To substantiate our findings, we research the estimators' sensitivity towards different network scenarios. Our results indicate that probabilistic estimators are robust towards inaccuracies in the emission probabilities. We find that probabilistic estimators deliver more accurate results than the Voronoi methods in all scenarios, even when confronted with extremely mismatched estimation models. For iterative estimators, we observe divergence, which occurs in some special cases under severe mismatching conditions, pointing to the need to improve further the numerical methods adopted by probabilistic estimators.

We expect our results to encourage further research on the probabilistic framework and novel estimation strategies.

1. Introduction

Mobile network operator (MNO) data represent a rich potential source for estimating the spatial distribution of mobile phones at a given time. From there, we can gain insight into the temporal variations of the spatial distribution of humans. In the past two decades, many research fields, e.g., in demography [1], [2], epidemiology [3], [4] urban planning [5], [6], tourism [7], etc., have developed applications with MNO data that promise high public value. According to UN sources, MNO data is one of the most promising and influential data sources within Official Statistics [8]. As such, utilizing MNO data (within Official Statistics) is no longer a question of relevance but of methodological and legal feasibility.

Moving from proof-of-concept towards the implementation of MNO data into official statistical production processes, [9] emphasize the importance of a reference methodological framework. Among discussing fundamental ideas and offering a first blueprint towards such a reference methodological framework, they introduce multiple MNO data sources [9]. We emphasize attention to the following two MNO data sources: (1) network event data, which logs every event (i.e., call, text message, data exchange) between any mobile phone and the radio network. And (2) network topology data, which contains information on the individual cells and their coverage within an operating area. Both data sources contain sensitive information and are not public. Even pseudonymized, network event data could disclose a mobile phone holder’s identity based on usage characteristics and mobility patterns; network topology data describes the network coverage characteristics, revealing the strengths and weaknesses of an individual MNO. For privacy protection reasons, data access and processing needs to happen at the MNO, by the MNO, and in agreement with Official Statistics’ actors [10].

Considering the task of spatial density estimation, two distinct approaches are currently in a pioneering phase, the static and the dynamic approach.

Their main difference lies in the processing of the first data source, network event data. The static approach, proposed by [11] and refined by [9], aggregates this data, resulting in a minimally privacy-invasive dataset because it only contains the number of phones connected to a specific cell at a reference time [10]. On the other hand, the dynamic approach, proposed by [12], assumes valuable information for the spatial density estimation within the temporal and spatial *trace* of mobile phones, which is processed into a time series of cell connections for every mobile device. Both approaches are still in their experimental phase and when they matured, future research should compare them in terms of methodological and legal feasibility, model robustness and parsimony, as well as, accuracy.

The other data source, network topology data, is the main data source for inferring the spatial coverage area (a.k.a. cell footprint or cell profile) of each cell. Both the dynamic and static approach require this task, a.k.a. geolocation of cells [9], [12], [13]. The accuracy of geolocation depends on the available data/parameters and the utilized method. Again, two approaches are identified in the literature: (1) *tessellations*, which divide the reference area into spatially disjoint coverage areas associated with different (groups of) cells. And (2) *overlapping cells*, which are methods that allow cell coverage areas to overlap. Most of the past applications utilize tessellations, therefore, assuming that every mobile device connects to its closest antenna, which is the inherent assumption of the popular Voronoi estimator [2], [14]. Essentially, any tessellation method reduces to a simple area-proportional computation and is, therefore, considered a deterministic estimation strategy. While its simplicity seems quite powerful in the context of scalability, the assumption of non-overlapping coverage areas does not match reality. MNO's purposefully architect their network in a way that their coverage areas overlap. Cell overlap reduces coverage holes, assuring well-enough reception quality across an operating area (in theory). Therefore, mobile phones have multiple cells they can connect to, and a recent experiment of a Swedish MNO has shown that in the case of 60% of all event records, mobile phones do not connect to its nearest cell [15].

Cell overlapping geolocation methods, proposed by [16], use radio propagation simulation techniques to actually model individual cell footprints. This allows the implementation of many cell-specific characteristics, if known, and can explicitly handle overlapping coverage areas by introducing a probabilistic framework. Recent simulation studies have shown that, with appropriate estimation methods based on probabilistic models, the availability of more detailed coverage area information allows to improve the spatial accuracy of the final estimate considerably, compared with the simpler traditional methods relying on Voronoi geolocations [9], [17]. However, such results were obtained (1) under the assumption of perfect knowledge on the coverage areas, i.e., omniscient network topology data, and (2) limited to a single simulated network scenario characterized by a dense, multi-layer coverage pattern with a high degree of cell overlapping. Under no realistic circumstances is it possible to model coverage areas perfectly accurately because they depend on many parameters, some of them being very volatile or immeasurable, such as weather conditions. It is questionable if we can expect such accurate spatial density estimations when using network topology data of a realistic certainty level, i.e., imperfect, and if they are robust across different network scenarios.

Furthermore, probabilistic estimators rely (implicitly or explicitly) on a prior assumption about the likelihood of each area to host mobile phones. Such initial assumption is then updated based on the information from the geolocation task. [13] suggest that this initial guess should be directly linked to the spatial density output indicator (e.g., working-day-time population) by utilizing multiple information sources to develop a so-called informative prior vector. The estimators' technical properties promise a more accurate estimation if the prior vector is a priori more spatially accurate.

This study draws directly on the work by [17] continuing the methodological research and focusing on the aspects of uncertainty within the spatial density estimation using the static approach. [9] point to three main dimensions of uncertainty in the overall estimation task: Spatial uncertainty, which relates to the fact that event locations are referred to extended cell coverage areas (not exact points) that are known only approximately. Temporal uncertainty, which relates to the sparseness of discrete event times.

And population coverage uncertainty, which relates to the differences between the observed population of mobile phones and the target population of humans. The term "uncertainty" is used here to refer to the causes of missing knowledge or incorrect information in input to the estimation task.

We dive deeper into the dimension of spatial uncertainty. In the probabilistic framework, spatial uncertainty translates into possible errors or missing knowledge in the setting of the emission probabilities, which is one of the model input parameters. Indeed also the other model input parameter, namely the prior vector, may be affected by a higher or lower degree of accuracy. We investigate the robustness of the final estimates towards inaccuracies in both model input parameters. To better understand how much each estimator gains from a (mildly) informative prior, we compute the rate of improvement compared to a non-informative, i.e., uniform prior vector. To deliver a holistic comparison between deterministic and probabilistic estimation strategies, we also present novel variants of the Voronoi methods that can implement prior information.

To substantiate the robustness of our findings, we assess the impact of different network scenario parameters (e.g., cell density, multiple cell layers) on the relative performances of the various estimators. For this, we design and conduct a simulation study, following the static approach, based on semi-synthetic data with all state-of-the-art estimation strategies containing deterministic and probabilistic cell geolocation methods. Owing to the spatial nature of the estimation problem, we use the Kantorovich-Wasserstein distance to measure the (dis)similarity between the estimated density and the true population distribution.

Our results show a robust relationship between the uncertainty of implemented coverage area information and the spatial accuracy of the resulting estimate. This means the probabilistic model's sophistication is dependent on the richness and accuracy of the network topology data. Furthermore, a more accurate prior vector leads to spatially more accurate results. Compared to the deterministic Voronoi estimators, the probabilistic estimators deliver a more reliable rate of improvement across all network scenarios.

Given that these observations are robust across different network scenarios, we advise investing further academic efforts in developing more performant estimation strategies based on probabilistic geolocation methods.

The remaining sections are: (2) a brief formalization of the estimation task with its relevant estimation strategies, (3) our research intent, (4) a detailed illustration of our empirical approach to this study, (5) the presentation of our results, (6) and finally, the discussion of concluding remarks.

2. The spatial density estimation task: a quick overview

This study is a direct follow-up work of [17] and we focus on extending their simulations, addressing research questions concerning probabilistic estimators towards model uncertainties. We follow their notation and agree with their formalization of the estimation task. For a more in-depth introduction and discussion between deterministic and probabilistic models, the derivation of the respective estimators as well as their concomitant assumptions, please refer to [17]. Nevertheless, for clarity reasons, we briefly recapitulate the estimation task and relevant estimation strategies, before outlining our dedicated research intent in the next section.

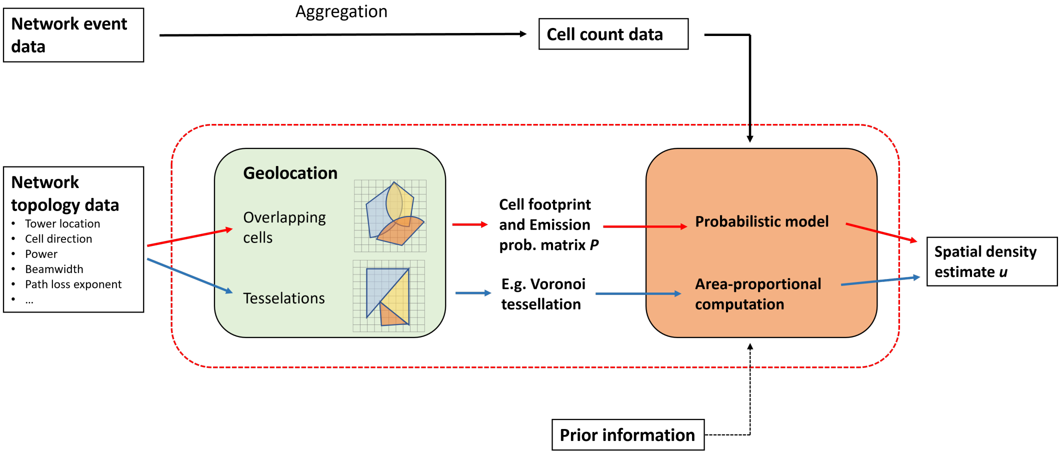


Figure 2.1: Data processing workflow for the spatial density estimation with the static approach.

Figure 2.1 shows a simplistic workflow of the spatial density estimation task within the static approach. We assume three distinct data sources:

- Operating area data: We assume the operating area to be discretized into a regular grid, e.g., 100m x 100m. Each square unit is termed as a tile, indexed in $j = 1, 2, \ldots, J$. u_j denotes the unknown non-negative number of mobile phones of the jth tile, stored in the column vector $u := [u_1 \dots u_J]^T$. This standardized unit helps us set a formalism framework that has been used in multiple studies beforehand [13], [17], [18].

- Aggregated network event data: Let the *i*th element c_i of the column vector $c := [c_1 \dots c_I]^T$ define the observed number of phones counted in cell $i = 1, 2, \dots, I$. Furthermore, we formalize the total number of phones across all cells by $C := \sum_{i=1}^{I} c_i = 1_I^T c$, the assumed total number of individual phones in the operating area. As in [17] c describes a cell count vector, aggregating all estimable phones around a reference time t^*. In this study we ignore the time dimension and focus only on the spatial density estimation of a certain point in time.
- Network topology data: This data source describes the configuration parameters of each relevant cell within the operating area. It can come in various levels of detail, ranging from only knowing the cell tower locations to having precise information on the azimuth direction, transmission power, beamwidth, etc. This data is used within MNO's to improve the network via radio propagation modeling.

The first task is to perform the task of cell geolocation, i.e., modelling/classifying the coverage region for each cell i in terms of the operating area. For this we combine the operating area data and the network topology data. This output is then used as an input parameter (next to others) in an estimator function. The geolocation method that one chooses has consequences on the subsequent model choice for spatial density estimation. Tessellations exclusively lead to deterministic models and their estimation equals in all cases a simple area-proportional computation of the respective cell count c_i across its geolocated coverage region S_i. Please refer to Figure 2.2 for an overview of popular Voronoi geolocation methods. The result of such a deterministic estimation strategy is a spatial density estimation vector, which contains a single point estimate $\hat{u}_j$ for each tile j.

While such simple point estimates are also the result for probabilistic models, their geolocation and their estimation are much more sophisticated and require more specification. Beginning again with the geolocation, previous

Deterministic estimation strategies (non-overlapping geolocation)

Compared to probabilistic estimators, deterministic estimators always reduce to a simple area-proportional computation, namely $\hat{u}_j = \frac{c_i}{S_i}$. S_i defines the size of the cell-specific area. Strictly put, deterministic estimators differ not within the estimation method but only in the geolocation method. A popular and often used deterministic geolocation method is the creation of Voronoi regions. So-called seed points (e.g., cell locations) are projected to the operating area, and each point in the area is classified with its closest seed. No overlapping cells nor holes can be modeled. This logic translates into the inherent assumption that each phone is connected to its closest cell. In our study, we consider three different Voronoi methods: (1) Voronoi-Tower uses the tower locations (blue triangle) as seed points, aggregating each cell count c_i and area S_i to its respective tower; (2) Voronoi-offset considers all cells moving each seed (red dots) by a small offset distance into the azimuth direction; and (3) Voronoi-barycenter defines seed points at the cell-specific location with the strongest signal (a.k.a. barycenter, red dots).

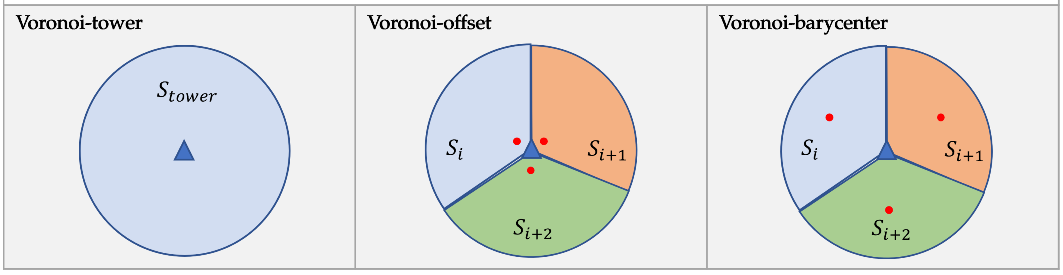


Figure 2.2: Voronoi geolocation methods.

research defines the coverage relation between any cell i and any tile j as s_{ij}, which is a non-negative quantity [9], [13], [17]. A suitable measure for this is the received signal strength (RSS), which is measured in dBm. The RSS within tile j from a certain cell i depends on many parameters, which we expect to derive from the network topology data. In reality, it is impossible to precisely predict the RSS of a specific cell i for any mobile phone within a specific tile j – it can only be approximated. As mentioned above, MNO's try to offer a good signal across an operating area by planning a network in a certain way: cell coverage areas are supposed to overlap. Especially in lively places, each phone might be confronted with up to a dozen cells that can offer acceptable reception, i.e., cells *compete* with each other to serve the mobile phone in that area. One reason for these multiple overlaps is that each cell has an individual capacity of phones it can connect to for a given time point. This is called load balancing, and it is not indicated by the RSS. To account for load balancing, previous research has introduced the signal dominance measure, which translates the RSS via a logistic function into the signal dominance [13]. This measure subsumes both the RSS and the concept of load balancing. It is also much more interpretable as it ranges between 0 (worse) and 1 (better).

Next, we need to model this cell competition, which essentially computes a connection likelihood for a generic phone within tile *j* considering all cells *i*. Cells that actually cover this tile *j* shall receive a non-zero value, cells that do not cover tile *j* shall receive the value zero, and logically, the sum of these likelihoods relating to a specific tile shall reach exactly 1. In statistical terms, we express the following conditional probability:

$$p_{ij} := \text{Prob}\{\text{detected in cell } i \mid \text{placed in tile } j\}. \tag{2.1}$$

For the rest of the paper we refer to p_{ij} as the *emission probabilities*. Using the signal dominance values we can model p_{ij}, formally:

$$p_{ij} = \frac{\mathrm{s_{ij}}}{\sum_{\mathrm{m=1}}^{\mathrm{k_j}} \mathrm{s_{mj}}}, \tag{2.2}$$

with k_j describing the number of cells covering tile *j*. For a more compact notation, we store p_{ij} into the so-called emission probability matrix $P_{[IxJ]}$, which is column stochastic, i.e., each column sums up to 1. Comparing the output objects from deterministic and probabilistic geolocation methods, Voronoi regions are considered a "[...] particular, degenerate case of emission probabilities, where *P* reduces to the identity matrix [...]" [17, pg.3].

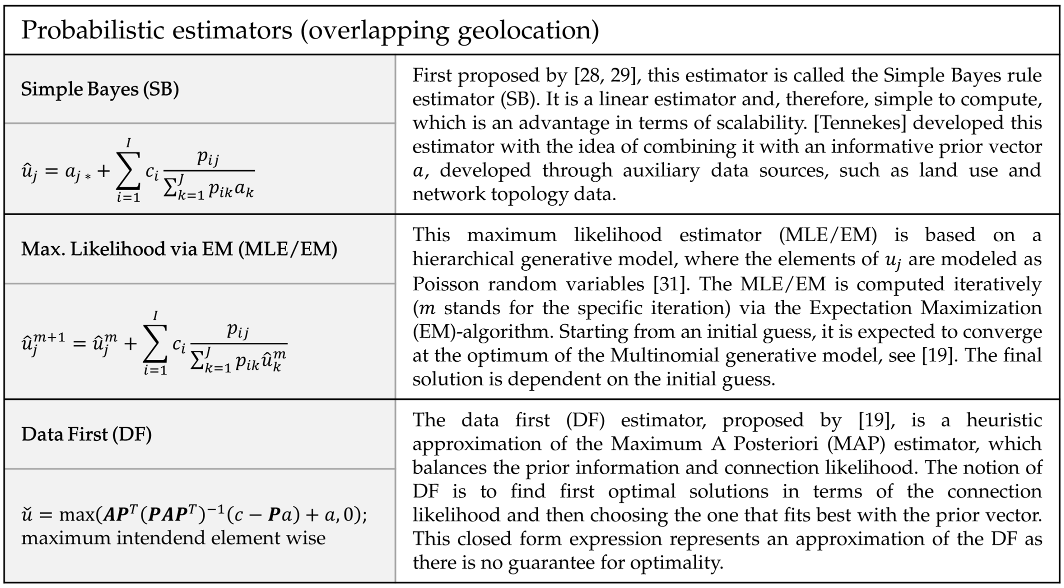
A (measured) cell count vector *c* can be interpreted as the single realization of a random vector $\tilde{c}$ whose expected value is given by:

$$\bar{c} := E[\tilde{c}] = Pu. \tag{2.3}$$

In the estimation problem, we must solve for the estimand *u* given the cell count vector *c*, representing the single available observation of $\tilde{c}$, and the estimation model matrix P. This is a type of inversion problem, therefore, the estimate $\hat{u}$ can be written in general as:

$$\hat{u} = g(P, c), \tag{2.4}$$

where $g(\cdot)$ defines the estimator of choice. This marks the final step within the probabilistic model: choosing and specifying an estimator $g(\cdot)$. Figure 2.3 lists the three different probabilistic estimators, which are discussed and chosen for their initial simulation in [17]. We shall end this section by pointing out some interesting user-related aspects that we deduce from previous experiments conducted by [17], [19].



Probabilistic estimators (overlapping geolocation)	
Simple Bayes (SB)	First proposed by [28, 29], this estimator is called the Simple Bayes rule estimator (SB). It is a linear estimator and, therefore, simple to compute, which is an advantage in terms of scalability. [Tennekes] developed this estimator with the idea of combining it with an informative prior vector a, developed through auxiliary data sources, such as land use and network topology data.
$\hat{u}_j = a_{j\,*} + \sum_{i=1}^{I} c_i \frac{p_{ij}}{\sum_{k=1}^{J} p_{ik} a_k}$	
Max. Likelihood via EM (MLE/EM)	This maximum likelihood estimator (MLE/EM) is based on a hierarchical generative model, where the elements of u_j are modeled as Poisson random variables [31]. The MLE/EM is computed iteratively (m stands for the specific iteration) via the Expectation Maximization (EM)-algorithm. Starting from an initial guess, it is expected to converge at the optimum of the Multinomial generative model, see [19]. The final solution is dependent on the initial guess.
$\hat{u}_j^{m+1} = \hat{u}_j^m + \sum_{i=1}^{I} c_i \frac{p_{ij}}{\sum_{k=1}^{J} p_{ik} \hat{u}_k^m}$	
Data First (DF)	The data first (DF) estimator, proposed by [19], is a heuristic approximation of the Maximum A Posteriori (MAP) estimator, which balances the prior information and connection likelihood. The notion of DF is to find first optimal solutions in terms of the connection likelihood and then choosing the one that fits best with the prior vector. This closed form expression represents an approximation of the DF as there is no guarantee for optimality.
$\breve{u} = \max(\boldsymbol{AP}^T(\boldsymbol{PAP}^T)^{-1}(c - \boldsymbol{P}a) + a, 0)$; maximum intendend element wise	

Figure 2.3: Comparing geolocation methods: tessellations and overlapping cells.

Two of the probabilistic estimators are expressed in vector form (SB, MLE/EM) and one of them in matrix form (DF). Each requires three input parameters, (1) the measurement data in the form of a cell vector (c_i; c), (2) an initial or prior guess in the form of a tile vector ($\hat{u}_j^m$; $\hat{a}_j$; a) or in the form of a diagonal matrix (A), and (3) the emission probability matrix (P; p_{ij}). We should point out that generally, the number of tiles is (much) larger than the number of cells, i.e., $J >> I$, while $J >> 1$. As a result, even if $\bar{c}$ were perfectly known, the direct inversion of Equation 2.3 would depict an underdetermined problem. Leading to issues of structural non-identifiability [20] , [17] suggest supporting the estimation process with auxiliary data such as prior information, spatial constraints, or structural properties of the desired solution. To overcome this problem, all listed estimators require the specification of a prior information vector.

The SB is the most computationally efficient probabilistic estimator. However, its authors,[13] also emphasize the importance of a (very) informative prior. This might not always be possible (e.g., if land-use data are not available) nor favorable (e.g., if the prior information vector is inaccurate). On the other hand, the DF estimator is computationally more demanding as one needs to perform a matrix inversion, which increases in complexity the more cells and tiles are involved. [17] suggest using the efficient Moore-Penrose pseudoinverse for this. The goal of DF is to prioritize the emission probabilities for its final solution, aiming at deriving as much information as possible from the geolocation. Finally, the computational efficiency of the MLE/EM depends on the number of iterations necessary for convergence. This depends on the informativeness of the prior. Preliminary experiments suggest around 200 iterations with an uninformative prior [9].

In terms of computational efficiency, all three estimators profit from consolidating the relevant input parameters. Consolidation refers to the creation of *supertiles*, first propsed by [9], where regular tiles j that are indifferently from each other in terms of prior value and emission probabilities are consolidated into one supertile. This requires the aggregation of prior values for the tiles j of a respective supertile and the dropping of equal instances within P.

Lastly, it is important to point out that the raw solution of this particular DF approximation requires the clipping of negative values and the renormalization via one MLE/EM iteration. Here, the raw DF estimation vector is specified as the initial guess vector, $\hat{u}_j^m$. This is necessary because – to receive a closed-form solution of the DF – the listed DF approximation allows for the relaxation of the following logical constraints: the non-negativity constraint on u and the sum of the estimation vector equaling the sum of the observed phones within the cell count vector constraint [9]. Initial tests have revealed that this DF approximation further gains in accuracy when conducting multiple iterations within the MLE/EM. One reason for this could be the approximate nature of this particular estimator [9].

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

3. Research Intent

Compared to deterministic estimators, probabilistic estimators consider much more information (if available) concerning the actual coverage of the cells, meaning they allow for flexible radio propagation modeling and cell overlap. [17] show that probabilistic estimators perform much better in terms of accuracy than deterministic estimators. However, they point out that their results have three important limitations concerning model inputs and context, which we address as our research questions.

First, in their simulation, they assume to know perfect network topology data, i.e., they use the generative model of their simulation as their estimation model [17]. Perfect network topology data is never available, as there are many parameters to be considered, some of them very difficult or even impossible to measure. In reality, the geolocation is performed with approximate network topology data, leading to inaccurate emission probabilities (P). It is important to know, how robust the performance of probabilistic estimators is towards inaccuracies. We formulate the following research question:

RQ1: How robust do probabilistic estimators perform towards errors in the estimation model matrix P**?** By systematically varying the error in P we can reveal the estimators' robustness to imperfect cell coverage information. We expect that probabilistic estimators gain in accuracy with increasing levels of accurate cell information knowledge, reaching peak accuracy with perfect knowledge.

Second, as mentioned above, all current probabilistic estimators solve the issue of non-identifiability by implementing a prior information vector. Generally, we can distinguish between two kinds of prior vectors, informative or non-informative. The essential impact of the prior vector within the estimators is an antecedent indication of how mobile phones are spatially distributed.[13] offer a modular Bayesian updating model for the development of an informative prior. They argue for a so-called *composite* prior, which combines land-use data and network topology data sources to develop an informative best guess before seeing the measurement data. For example, they assume a positive relationship with the number of cells and the number of people in an area. While a very informative prior can certainly be helpful if accurate, it can also bias the final estimation if being inaccurate. Essentially, the prior information vector should be linked to the final statistical output indicator, and accessing high quality data for the development of an informative prior might not always be feasible. We need to find the right balance between the influence of the prior vector and the influence of the observed measurement data on the estimators. Therefore, we suggest the investigation of a mildly informative prior. In the initial simulations by [17], only a non-informative, i.e., a uniformly distributed prior information vector, is considered. It is important to know how much the estimators profit from a mildly informative prior. Therefore, we formulate the following research question:

RQ2: How much do the estimators profit from increasing accuracy within the prior information vector? Varying the prior information vector will reveal the estimators' potential rate of improvement due to better prior information. We expect all estimators to improve when implementing an informative prior.

In this paragraph we explain the newly introduced variants of Voronoi estimators containing a prior information vector. Is this the correct location in the paper or should this be explained more formally in the section beforehand? In my opinion this is well placed here because we merely introduce these estimators as a helper for a more fair comparison on pior informativeness level. Also, I checkhed *For a better comparison between probabilistic and deterministic estimators, we implement novel variants of each Voronoi method, in which we can can consider a prior vector on the tile level. These variants were not* *included in the initial simulation by [17]. The prior vector describes the relative likelihood within each Voronoi polygon of hosting mobile phones. If the prior vector is uniform, then the cell-specific mobile phone count is disaggregated in a uniform fashion, which resembles the usual Voronoi tessellation in previous studies. If the prior vector is non-uniform, then the mobile phones connected to a specific cell are disaggregated according to the prior specific relative ratios.*

Third, the main argument against deterministic geolocation methods is the inability to model coverage areas that overlap. In reality, overlaps are constructed by design to prevent bad reception as the signal strength is not uniformly distributed within a coverage area. The propagation of the signal within its coverage area depends on the cell's parameters, starting with the type of the cell (omnidirectional or directional), its power level, beamwidth, etc. [13]. MNO's plan their network by establishing so-called cell layers. Cells within the same layer are very similar in propagation parameters. Cells from a macro layer, e.g., propagate their signal over large rural areas (i.e., large coverage areas). In contrast, micro or meso cells are often placed in urban areas and have small coverage areas while handling large volumes of mobile phones [13]. In the initial simulation of [17], only a single network scenario is considered characterized by a dense, multi-layer network topology with a high degree of cell overlap. From these initial results we cannot conclude the sensitivity of the estimators' relative performance towards these network characteristics (e.g., sparse vs. dense, single-layer vs. multi-layer). Therefore, we formulate the following research question:

RQ3: How sensitive are the estimators to network characteristics, such as cell density or differing number of cell layers? Varying the network characteristics through different network scenarios will further reveal the strengths and weaknesses of the different estimators in terms of accuracy. Furthermore, implementing multiple network scenarios will substantiate the findings from the other research questions, as we can investigate the influence of model inputs under different circumstances. We expect deterministic estimators to perform worse, (i) the higher the degree of cell coverage area overlap, and (ii) with the existence of multiple cell layers. Both of these network characteristics increase the probability of violating the prime Voronoi assumption: always connecting to the closest cell.

4. Methodology

To enforce the quality and trust within Official Statistics, not only the statistics and production processes need to be fully transparent and verifiable but also the (prior) methodological research substantiating these (future) production processes [21], [22]. Therefore, we have openly shared all simulations, models, and data in a git-repository[1] to assure full computational reproducibility, as well as publish the results in this open-access journal.

We are primarily interested in the robustness of the different estimators towards inaccuracies within model input parameters. Therefore, we need to ensure that we can correctly evaluate the spatial accuracy of certain model specification choices. This is only possible if one has access to a ground truth population (GTP) and a ground truth geolocation of the cell coverage areas, in the following the "generative model".

For this reason, a simulation study is most advantageous, as it allows us to have access to a reference GTP and complete control over the generative model.

To conduct our experiments and mimic MNO-like data, we use the MNO-simulator workflow, which is constructed within the open-source programming language R [23] and supports with development of the necessary non-trivial software.[2] This workflow contains three modules: (1) *generation*, which gives flexible options on simulating multiple scenarios concerning mobile phone density and network characteristics, as well as assigns mobile phones to cells based on an individually specified generative model P, (2) *estimation*, in which we specify our estimation models P^* and execute all

[1] The repository can be found here: https://github.com/R-ramljak/MNO_uncertainty.

[2] We created the MNO-simulator workflow within the development of this project, however, given its flexibility and potential usage for different projects, we have generalized it. An introductory presentation linking also to other resources can be found here: https://github.com/R-ramljak/MNO_simulator.

state of the art estimation strategies, and (3) *evaluation*, which offers multiple quantitative evaluation metrics, such as the Kantorovich-Wasserstein Distance [24]. For data handling and plotting the MNO-simulator workflow depends on the packages within the *tidyverse* [25], *data.table* [26] and *Matrix* [27], for spatial operations *sf* [28], *stars* [29] and *raster* [30], for radio propagation modelling *mobloc* [31] and for the computation of the KWD *SpatialKWD* [32].

The MNO-simulator enables the user to work with any kind of data availability – experiments can be conducted with complete synthetic scenarios (no real-world data available), semi-synthetic scenarios (some real-world data available, e.g., census data or (partial) cell coverage data), and real scenarios if actual data are available. The only "real" data source we use in this study is the operating area on a 100m*100m regular grid. We chose Munich's city and its near surroundings, encompassing 1,600 square kilometers (160,000 tiles of size 100m*100m), and used its census population to specify the GTP. We reduce the final GTP to about 1/3 of the true census population to mimic the customer basis of a single MNO with that market share (ca. 677,000 phones). Figure 4.1 visualizes the spatial density and the empirical cumulative (complementary) distribution (ECCDF) of the GTP. Population units within tiles cluster in space leading to the extreme right-tailed tile distribution.

Within the MNO-simulator, we can flexibly create various network scenarios, in the following called cellplans, based on many relevant cell parameters. The main functions for modelling cell coverage areas rely on the *mobloc* package [31]. We create four cellplans differing in the aspects of cell density and the number of cell layers. Cells are mounted on towers and each tower carries a triplet of 120° sector cells oriented in different azimuth directions (i.e., we only generate directional cells). Cellplan.1 is a rather sparse, one-layer scenario with a median coverage of four cells per tile, and cellplan.2 is an extremely dense, one-layer scenario with a median coverage of nine cells per tile. Cellplan.3 is a three-layer, dense scenario with a median coverage of eight cells per tile, and cellplan.4 is a two-layer, sparse scenario with a median coverage of five cells per tile. For comparison reasons, cellplan.3 is

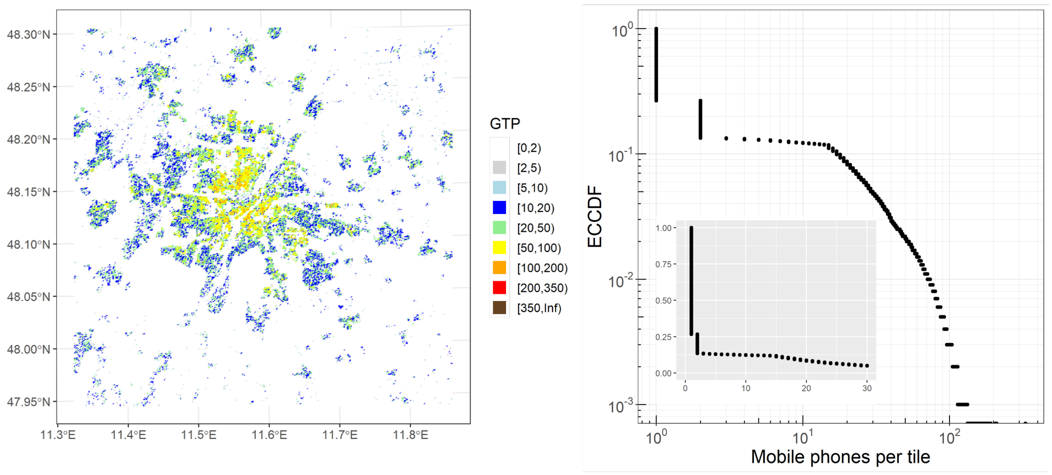


Figure 4.1: Spatial density map (left) and ECCDF (right) of the ground truth mobile phone population per tile.

identical to the network scenario used in the initial simulation of [17].

With *coverage* we define the relation between a tile j and cell i in terms of its signal dominance value. We store these relations in an edge list, typical for a two-mode network structure, and define a minimum signal dominance threshold of 0.05. Each relation must be above this value in order to be counted. Each cellplan assures full coverage of the complete operating area. This means every tile is covered by at least one cell in the generation phase (i.e., the generative model) with an signal dominance value of at least 0.05. We relax this requirement when creating mismatched estimation models, which we will explain in the next subsection. Please refer to the git-repository mentioned in footnote 1 for an in-depth (visual) description of the cellplan characteristics as well as their generation.

4.1 Model mismatch techniques

To implement erroneous cell coverage information, we define different emission probability matrices P^* that are of the same form as the generative model matrix P but differ in the level of accuracy. For this, we have come up with two so-called model mismatch techniques that purposefully introduce errors into the geolocation. We visualize the impact of these mismatch techniques by presenting an exemplary cell profile of a meso cell. One mismatch technique introduces random noise at the signal strength stage, while the other technique quantizes the variability level of the signal dominance. After we introduce the errors, we compute the emission probabilities p_{ij} and store them in an estimation model matrix P^*. In total, we create 11 estimation models – five of each technique introducing respectively varying magntiudes of mismatch – and one estimation matrix that is identical to the generative model matrix, i.e., $P^* = P$ as a benchmark. In the following, we explain each of these parametric mismatch techniques in more detail.

4.1.1 Model mismatch technique I: Spatially sensitive random noise

For this mismatch technique, we initialize the true cell profiles at the *signal strength* stage as a basis and introduce random noise, sampled from a uniform distribution with variable *min* and *max* values (e.g., min: -3 and max: +3, indicated as *no.03* in Figure 4.2 noise.dom.plot.ME.w.th). The most extreme case is min: -21, max: +21. Each sampled value is added to the true dBm value of the cell profile. This random noise is introduced in a spatially sensitive way, meaning cell coverage profiles are only being distorted at a realistic distance to their origin. We retain the true cell azimuth direction as this can be classified as certain knowledge. We transform the distorted signal strength values to signal dominance values, and because of the minimum signal dominance threshold, *coverage holes* (as seen in Figure 4.2) are possible. This mismatch technique resembles the signal strength volatility through extreme weather or the existence of high buildings – particularly the extremely noised versions can be interpreted as a "stress test" for the estimation strategies.

4.1.2 Model mismatch technique II: Quantization

For this model mismatch technique, we quantize – i.e., discretize in equal distances – the smooth signal dominance values into a defined number of

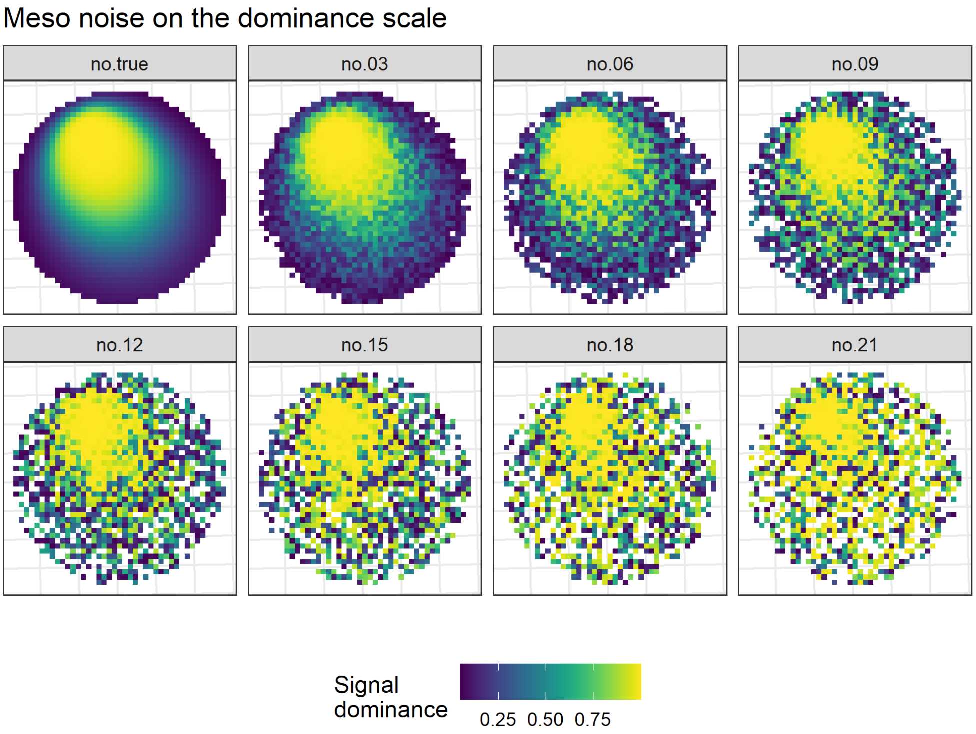


Figure 4.2: Example coverage areas visualizing the mismatch technique: spatially sensitive random noise.

value range categories. Each quantization version results in 2^n categories, where n is a variable parameter with values ranging between 0 and 10. Figure 4.3 shows the quantized cell profiles for each version (e.g., *quant.02* refers to 2^2 categories). For example, in the most extreme case, the complete, true coverage values on the signal dominance scale are discretized into one category – the complete coverage area has a uniform signal dominance value, a.k.a. as flat coverage. Compared to the first mismatch technique, we quantize at the signal dominance stage and not at the signal strength stage. Otherwise, the categories would get further distorted through the logistic model of the signal dominance transformation [13]. The quantization mismatch technique mimics the realistic situation that one only has access to the tower location, the cell azimuth, and an approximate power level of the cells, from which the coverage area can only be roughly modeled. Furthermore, through this mismatch technique, we can research the behavior of the estimators with almost equal emission probabilities p_{ij} within a tile.

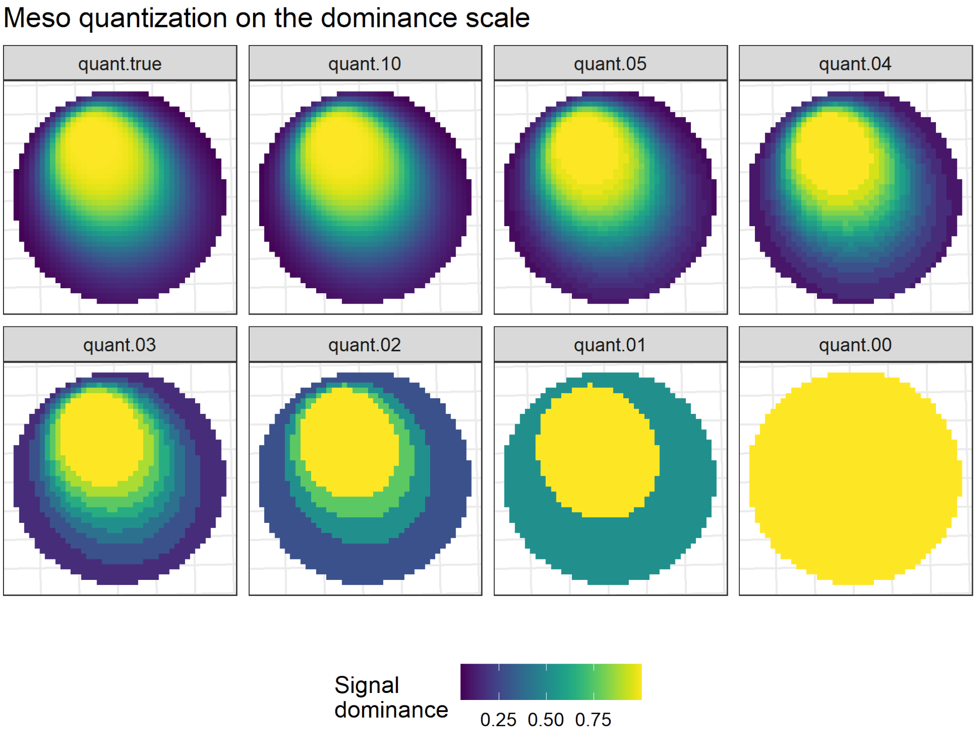


Figure 4.3: Example coverage areas visualizing the mismatch technique: quantization.

4.2 Prior

To mimic inaccuracies within the prior parameters, we specify two priors, a non-informative and a mildly informative prior. Both are defined on the tile level. The non-informative prior is characterized through a uniform vector of 1's. The mildly informative vector is a discretized version of the ground truth vector. We define three distinct levels according to the following rules:

- GTP <= 1 ~ 0.1,
- GTP > 1 & GTP <= 50 ~ 1,
- GTP > 50 ~ 100.

We want to emphasize again that the impact of the prior is not dependent on the actual values but the relative ratios. It is out of the scope of this study to evaluate the impact of purposefully wrong priors.

4.3 Estimation

Our simulation study compares the performance of the six different density estimation approaches described above, which [17] also use in their initial simulation. We consider the following Voronoi variants:

- **Vor-T** – Voronoi tessellation with one seed for each cell tower location.
- **Vor-O** – Voronoi tessellation with one seed for each cell placed at a small fixed distance (10 m) from the respective tower location in the azimuth direction.
- **Vor-B** – Voronoi tessellation with one seed for each cell placed at the barycenter (mean point) of the signal dominance profile for the given cell.

For a better comparison between probabilistic and deterministic estimators, we implement the novel computation of each Voronoi tessellation, in which we can specify a prior vector on the tile level. These variants were not included in the initial simulation by [17].

Concerning the probabilistic estimators, we consider the following:

- **SB** – the simple Bayesian estimator [13].
- **MLE/EM** – the MLE computed with the iterative EM procedure [33] after $n = \{10, 200\}$ iterations.
- **DF** – the approximated DF estimator [17], where the raw solution is renormalized and used as prior in a subsequent MLE/EM estimation. We report results for the iterations $n = \{1, 10, 200\}$.

All these estimators are executed with both prior vector specifications. For the probabilistic estimators, we also consider the eleven specified model matrices P^*, containing ten mismatched versions and the true version of emission probabilities. Each reported iteration is counted as one estimator version. In total we will compare for each probabilistic estimator *4 scenarios x 2 priors x 11 emission probability matrices = 88 versions*, and *4 scenarios x 2 priors = 8 versions* for each deterministic estimator. This leads to *6 probabilistic estimators x 88 versions + 3 deterministic estimators x 8 versions = 552 final* *estimations* on the tile level.

4.4 Evaluation criterion: Kantorovich-Wasserstein Distance

Concerning spatial density accuracy, the central question we aim to answer for each estimator is: how similar is our estimation to the ground truth? A first step to evaluate this question is to plot maps of the estimations' spatial density and compare it visually to the ground truth spatial density. However, this is merely a visual metric, not returning any objective measurement to assess the similarity properly.

For a quantitative measure of the similarity between two spatial density maps, we need to account for the spatial nature of the analysis problem (i.e., horizontal spatial errors and physical proximity between distribution masses). Therefore, we resort to the KWD with Euclidean ground distance, which was also used in previous work concerning the accuracy of spatial density estimates [9], [17]. For computing the KWD, we use the package SpatialKWD [32]. Its functionality revolves around a computationally efficient approximation method of the KWD [24]. The tunable coprime parameter L balances the computational resources with the KWD's estimation accuracy. We use $L = 3$ for this approximation parameter, which means the approximation error will be, in the worst-case 1.29%. We use the output value as an upper bound for the KWD estimate, and the worst-case adjustment as a lower bound.

As indicated in Section **??**, the properties of each estimator assure that the same mass (i.e., mobile phones/population units) as the ground truth mass is being distributed over the area. Using KWD, we primarily assess if the estimator distributes this mass in the same way over the area as the geographical distribution of the ground truth – and if not, how far any population unit needs to travel on average in terms of tiles in an optimized way to arrive at the ground truth, i.e., the average spatial error. Therefore, high KWD values are associated with a highly dissimilar spatial distribution, and low KWD values are associated with a highly similar spatial distribution to the ground truth.

Figure 4.4 recapitulates our complete analysis plan, following the modular structure of the MNO-simulator workflow.

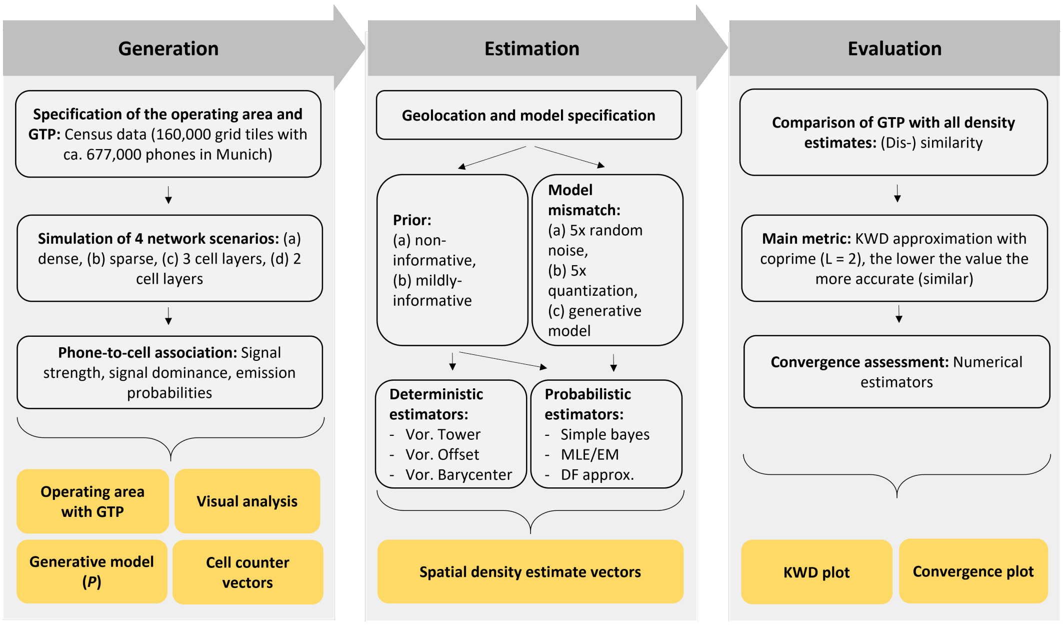

Generation
Specification of the operating area and GTP: Census data (160,000 grid tiles with ca. 677,000 phones in Munich)
Simulation of 4 network scenarios: (a) dense, (b) sparse, (c) 3 cell layers, (d) 2 cell layers
Phone-to-cell association: Signal strength, signal dominance, emission probabilities
Operating area with GTP
Visual analysis
Generative model (P)
Cell counter vectors
Estimation
Geolocation and model specification
Prior: (a) non-informative, (b) mildly-informative
Model mismatch: (a) 5x random noise, (b) 5x quantization, (c) generative model
Deterministic estimators: - Vor. Tower - Vor. Offset - Vor. Barycenter
Probabilistic estimators: - Simple bayes - MLE/EM - DF approx.
Spatial density estimate vectors
Evaluation
Comparison of GTP with all density estimates: (Dis-) similarity
Main metric: KWD approximation with coprime (L = 2), the lower the value the more accurate (similar)
Convergence assessment: Numerical estimators
KWD plot
Convergence plot

Figure 4.4: Analysis plan.

5. Results

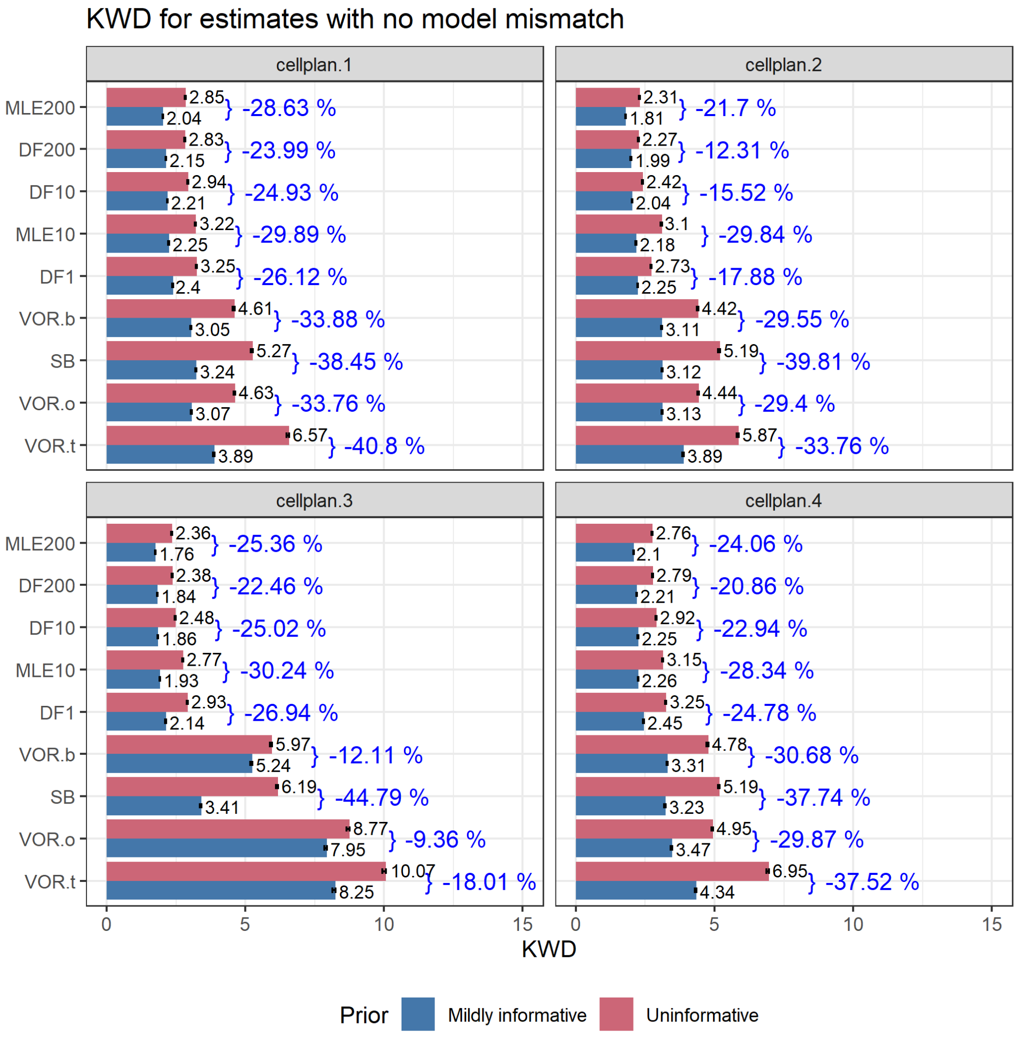


Figure 5.1: Accuracy evaluation (KWD) of estimates with no model mismatch per scenario: Each panel shows a network scenario, each estimate has two specifications, distinguishable by the specified prior (Uninformative, red; Mildly informative, blue). Percentages in blue indicate the potential reduction in terms of KWD between the differing prior specifications.

In the following we present our main results. Figure 5.1 presents the KWD results of each estimator for each cellplan, considering only emission probabilities P^* with *no* model mismatch. The red bars represent a non-informative prior specification, and the blue bars the mildly-informative prior specification. Our first impression reveals that the numerical probabilistic estimators, DF and MLE/EM, result in the most spatially accurate estimations. Their minimal difference in KWD is negligible, given the worst-case adjustment. This is the case for both prior specifications and all cellplans – most visibly in the cellplan.3, which is also the most realistic network because of its multi-layer configuration. Furthermore, even though the numerical estimators converge after around 200 iterations, their premature versions after ten iterations are still by more than one KWD unit more accurate than any Voronoi estimator. For example, the MLE/EM estimator specified with a non-informative prior leads in the third scenario to a KWD value of 2.36, which translates into an average spatial error of 236m. Vor-T with the same prior specification – the most prominent method in the past literature – yields a KWD value of 10.07, translating into an average spatial error of more than 1000m. Using the MLE/EM estimator compared to the simplest Voronoi estimator equals a reduction factor of more than x4. These results are in line with [9], [17].

In RQ3 we question the stability of the probabilistic estimators' performance across scenarios, differing in cell density and the presence of multiple cell layers. We can deduce that the performance superiority of the probabilistic estimators is robust across multiple network scenarios. As expected, all estimators profit from denser networks as the disaggregation task becomes less complex. Some probabilistic estimators can utilize this information more effectively, e.g., comparing DF1-cellplan.1 (3.25) with DF1-cellplan.2 (2.73). Specifically scenarios with more cells, e.g., cellplan.2, lead to higher spatial accuracy levels for all estimators compared to scenarios with fewer cells, e.g., cellplan.1.

Continuing with RQ2 (varying prior specifications) we hypothesized that utilizing prior knowledge – if informative – decreases the level of uncertainty and, therefore, increases the level of spatial accuracy. Our results indicate exactly this: All estimators improve their level of accuracy when implementing a mildly informative prior, compared to an uninformative prior. However, which estimator profits the most from informative and accurate prior knowledge? While some scenarios (cellplan.1 and cellplan.2) indicate a very large rate of improvement for the Voronoi estimators (29%-33%) compared to some of the probabilistic estimators (e.g. DF: 12%-29%), this effect is not constant across scenarios. The presence of multiple cell layers decreases this effect, which we subtly see in the results of cellplan.4, where we randomly placed a few micro-cells. In the most realistic scenario (cellplan.3) the large rate of improvement for Voronoi estimators completely disappears, while the rate of improvement for probabilistic estimators stays robust across all scenarios. A possible explanation for this is the differing signal dominance-to-distance relationship within multi-layer networks as the layer-specific cell footprints have different sizes. The increasing presence of multiple layers leads to a severe violation of the prominent Voronoi assumption: mobile phones *always* connect to the closest cell.

In RQ1, we question the probabilistic estimators' behavior and robustness towards mismatched emission probabilities. Current research on the technical properties of these estimators cannot reveal if their promising results solely rely on the perfect cell knowledge, i.e., $P^* = P$, or if they might "break" due to the mismatch, e.g., the EM procedure diverges. Figure 5.2 presents the KWD values for most estimators[1] in the form of a line plot, where the x-axis reports the increasing levels of implemented noise and the respective y-axis the KWD values. Each scenario is composed of two adjacent panels, differentiating the respective prior specification. Across almost all noised versions and throughout almost all scenarios, MLE/EM and DF yield lower KWD values than the best Voronoi method (VOR.b). This is especially visible in the most realistic scenario (cellplan.3), where the MLE and DF perform much better than the Voronoi methods, even with extremely noised models. Only specific constellations of the scenario and prior specification (cellplan.1, cellplan.2 and cellplan.4 with uninformative prior) and an extremely noised model (+/- 21dBm) lead to very similar KWD values between the best probabilistic estimators and the best Voronoi estimator (Vor-

[1] We experience some divergence in the quantized mismatched cell coverage areas, which are presented in the next paragraph. Therefore, we reduce the iterations for the numerical estimators to ten. To enable direct comparisons, we also reduce the number of iterations for the noised mismatched cell coverage. However, we want to emphasize that the noised versions fully converge.

B). As we hypothesized, the trend lines of the probabilistic estimators indicate a moderate increase of KWD, the higher the introduced noise. This is a promising finding for the probabilistic estimators, as it indicates a reliable spatial accuracy level, depending on the accuracy of the emission probabilities.

Moving on to the results with quantized emission probabilities, Figure 5.3 reports these in the same style as the noised versions. Introducing model mismatch through quantizing cell coverage areas leads to considerably different results concerning the robustness of the probabilistic estimators. Severe quantization leads to non-convergence of the MLE/EM, meaning the estimator does not reach an optimal result. This is also why we report for the MLE/EM, and the DF only results after ten iterations to avoid the distortion of the plot. We cannot deliver an exact explanation for this defect at this point. However, it can be argued that there is a link between the number of iterations and the level of model uncertainty, as the EM procedure tries to reach the optimal solution in terms of P^*. To avoid divergence, we recommend executing ten iterations at the most in practical cases, as the previous results suggest only a negligible potential spatial accuracy gain after further iterations. Future research has to scrutinize how the number of iterations of the MLE/EM is related to model uncertainty.

However, probabilistic estimators confronted with moderate quantization still lead to spatially more accurate results than Voronoi estimators. Moderate quantization resembles a high degree of mismatch as it significantly reduces the variability within the coverage area and, therefore, smoothes the elements of P^*. In the most realistic scenario (cellplan.3), cell coverage areas need at least 16 differing quantized levels (quant.04) to reach converged results. Further quantization leads to diverging results. Furthermore, decreasing the uncertainty through informative and accurate prior information can help mitigate the reduction of spatial accuracy caused by high model mismatch – this is the case for both forms of mismatch, noise and quantization. Therefore, it is a viable strategy to compensate for inaccuracies in emission probabilities by investing in the development of a more accurate prior.

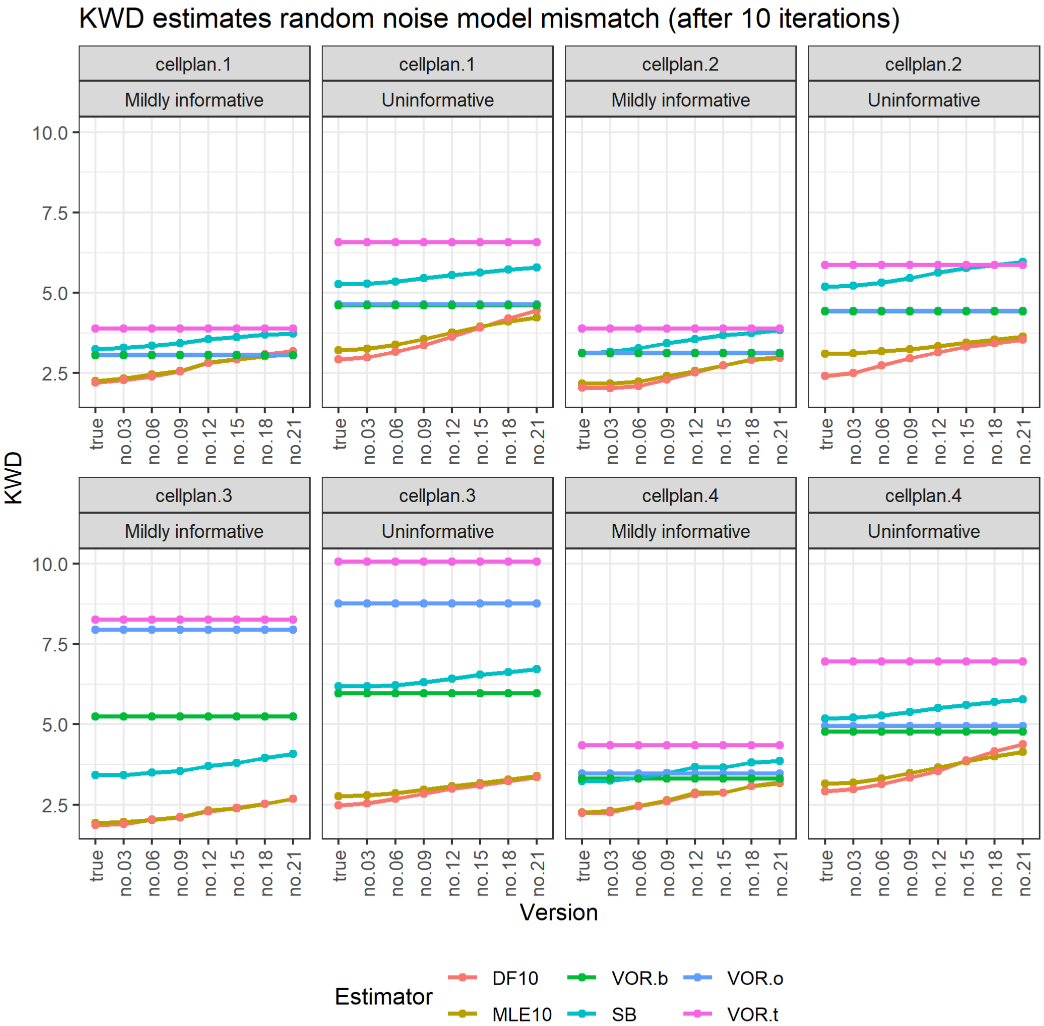


Figure 5.2: Accuracy evaluation (KWD) of estimates with noised model mismatch: Each panel shows the combination of a network scenario and a prior specification. The x-axis indicates the specific model mismatch version and the y-axis indicates the KWD value for the respective estimate, which is indicated by a line. Voronoi estimators stay constant throughout each mismatched version as they do not require this modeling information.

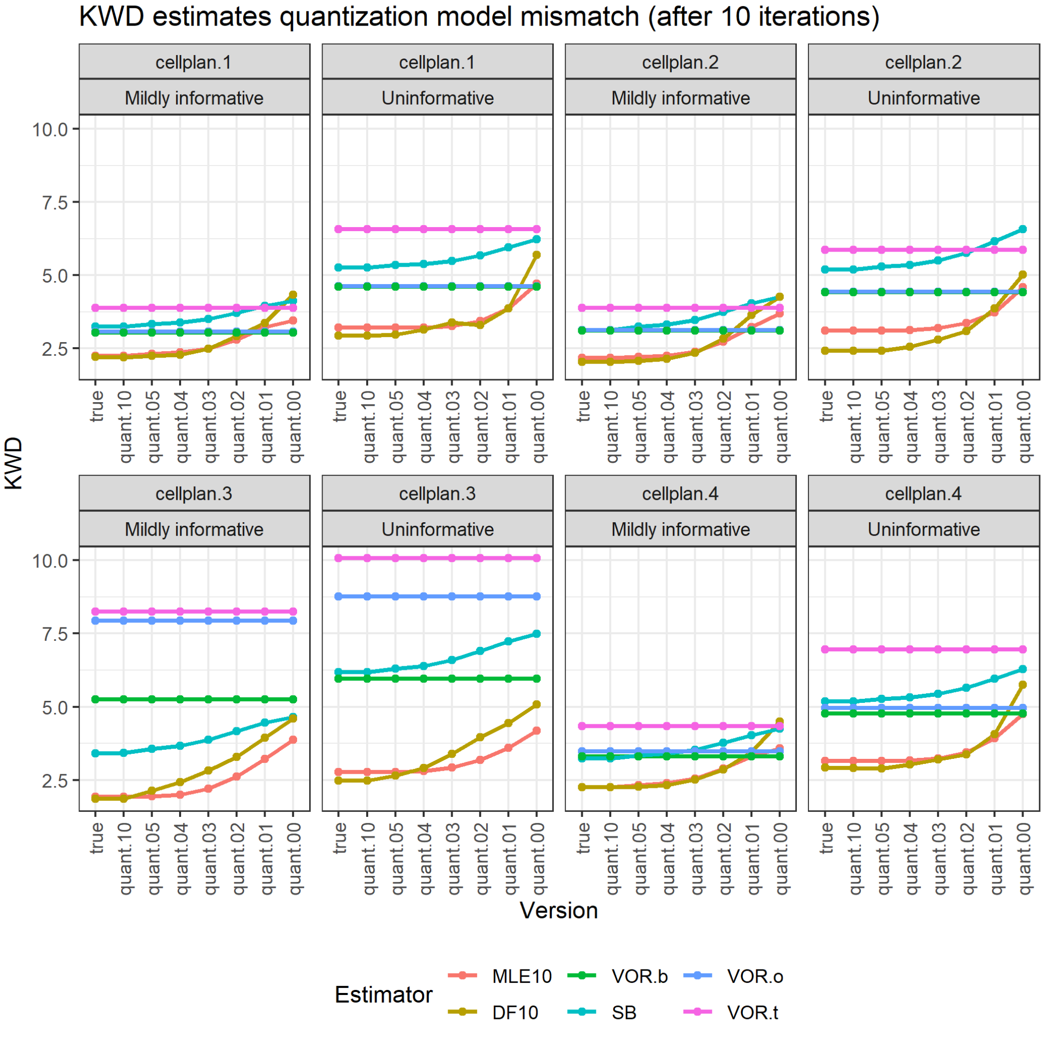


Figure 5.3: Accuracy evaluation (KWD) of estimates with quantized model mismatch: Each panel shows the combination of a network scenario and a prior specification. The x-axis indicates the specific model mismatch version, and the y-axis indicates the KWD value for the respective estimate, which is indicated by a line. Voronoi estimators stay constant throughout each mismatched version as they do not require this modeling information.

6. Discussion

In this paper we report on the results from our simulations concerning model uncertainties within the MNO spatial density estimation task. We test the robustness of probabilistic estimators towards model mismatch and different prior specifications and compare their performance to the traditional Voronoi estimators. To benchmark the estimators' accuracy, we resort to the Kantorovich-Wasserstein distance, taking the spatial nature of the estimation task into account. We substantiate our findings by executing our simulations across multiple network scenarios. Our extended simulations draw upon [17].

Even though one could argue that Voronoi estimators represent a parsimonious and therefore, scalable estimator, these estimation strategies (1) are based on a faulty assumption (i.e., mobile phones always connect to their closest cell), and (2) entail the unusual and unwanted behavior of an information implementation barrier. Therefore, Voronoi estimation methods offer limited flexibility in modeling cell coverage areas, even if further network topology information is available. With probabilistic estimators, on the other hand, all relevant available information concerning cell coverage modeling can be utilized. We show that these estimators entail the logical modeling property of improved performance through more and accurate information.

The numerical results indicate that even with severe model mismatch – purposefully implemented through spatially sensitive random noise – the MLE/EM and the DF still perform superior to all Voronoi methods. We can corroborate this finding across all tested network scenarios. In our most realistic network scenario, a three-layer network with high degree of overlap, we find for the MLE/EM and DF that our spatial error increases only by a factor of 0.8 when utilizing an extremely noised model compared to perfect cell knowledge. The spatial error doubles when comparing the extremely noised model MLE/EM or DF specification to the most performant Voronoi method (seed points are the barycenter). It needs to be acknowledged that the usage of the Voronoi barycenter is discouraged, as the availability of the barycenter location certainly gives the opportunity to perform a geolocation that considers overlapping cells, i.e., a probabilistic estimation strategy.

Merely heavily quantized cell profiles lead to misbehavior of the iterative approximations of these estimators. We cannot deliver a feasible explanation for this divergence. However, we offer a temporary workaround by using fewer iterations. The performance results from converged versions of the estimators suggest minor accuracy reductions between the 10th and 200th iterations. Future research should investigate the reason for this misbehavior as quantized cell footprints – to a certain degree – mimic quite realistic data availability. Within computing the iterative estimators we observed extreme speed increases through heavy quantization, as it lead to the creation of many supertiles. This reduces the computational burden and could be of special importance for applications with much greater operating areas.

Even though the performance reductions concerning model uncertainties are minor, our results confirm the performance boost through mildly informative priors. This is in line with the findings from [13], where the authors suggest investing substantial effort into the specification of an informative prior. These efforts can balance performance reductions through mismatch errors and vice versa: minimal informative priors can be balanced by more accurate cell coverage information. Further research should investigate the impact of wrong priors, which was out of the scope for this paper.

This research continues the overall (re)search for the best estimator(s) considering MNO data for spatial density estimation. Concluding from our robustness analysis, we suggest that a Voronoi estimator should only be chosen when merely the geographic point locations of the cells are available. However, a probabilistic estimator is advisable whenever the cell information on the azimuth propagation direction and power is available. Given the disadvantages of the Voronoi method mentioned above, we wonder: is it worth continuing academic efforts to further refine Voronoi estimation methods? We believe future research should focus more on developing novel estimation techniques that follow the probabilistic model.

Bibliography

[1] T. Koebe, "Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling," *PLoS ONE*, vol. 15, no. 11 November, 2020, ISSN: 19326203. DOI: `\mydoi{10.1371/journal.pone.0241981}`. arXiv: `2002.11618`. [Online]. Available: `http://dx.doi.org/10.1371/journal.pone.0241981`.

[2] B. Sakarovitch, M. P. de Bellefon, P. Givord, and M. Vanhoof, "Estimating the residential population from mobile phone data, an initial exploration," *Economie et Statistique*, vol. 2018, no. 505-506, pp. 109–132, 2018, ISSN: 17775574. DOI: `\mydoi{10.24187/ecostat.2018.505d.1968}`.

[3] K. H. Grantz, H. R. Meredith, D. A. Cummings, *et al.*, "The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology," *Nature Communications*, vol. 11, no. 1, pp. 1–8, 2020, ISSN: 20411723. DOI: `\mydoi{10.1038/s41467-020-18190-5}`. [Online]. Available: `http://dx.doi.org/10.1038/s41467-020-18190-5`.

[4] M. Szocska, P. Pollner, I. Schiszler, *et al.*, "Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021, ISSN: 20452322. DOI: `\mydoi{10.1038/s41598-021-81873-6}`.

[5] L. Galiana, B. Sakarovitch, and Z. Smoreda, "Understanding socio-spatial segregation in French cities with mobile phone data," pp. 1–12,

[6] J. Yang, Y. Shi, C. Yu, and S. J. Cao, "Challenges of using mobile phone signalling data to estimate urban population density: Towards smart cities and sustainable urban development," *Indoor and Built Environment*, vol. 29, no. 2, pp. 147–150, 2020, ISSN: 14230070. DOI: `\mydoi{10.1177/1420326X19893145}`.

[7] J. Raun, R. Ahas, and M. Tiru, "Measuring tourism destinations using mobile tracking data," *Tourism Management*, vol. 57, pp. 202–212, Dec. 2016, ISSN: 02615177. DOI: `\mydoi{10.1016/j.tourman.2016.06.006}`.

[8] UN, "Handbook on the Use of Mobile Phone Data for Official Statistics UN Global Working Group on Big Data for Official Statistics," Tech. Rep., 2019. [Online]. Available: `https://unstats.un.org/bigdata/blog/2019/mpd-task-team.cshtml`.

[9] F. Ricciato, G. Lanzieri, A. Wirthmann, and G. Seynaeve, "Towards a methodological framework for estimating present population den-

sity from mobile network operator data," *Pervasive and Mobile Computing*, vol. 68, pp. 1574–1192, 2020, ISSN: 15741192. DOI: \mydoi{10.1016/j.pmcj.2020.101263}. [Online]. Available: https://doi.org/10.1016/j.pmcj.2020.101263.

[10] F. Ricciato, A. Wirthmann, K. Giannakouris, F. Reis And, and M. Skaliotis, "Trusted smart statistics: Motivations and principles," *Statistical Journal of the IAOS*, vol. 35, no. 4, pp. 589–603, 2019, ISSN: 18747655. DOI: \mydoi{10.3233/SJI-190584}.

[11] M. Tennekes, Y. Gootzen, and S. H. Shah, "A Bayesian approach to location estimation of mobile devices from mobile network operator data," 2020.

[12] D. Salgado, L. Sanguiao, B. Oancea, S. Barragán, and M. Necula, "An end-to-end statistical process with mobile network data for official statistics," *EPJ Data Science*, vol. 10, no. 1, 2021, ISSN: 21931127. DOI: \mydoi{10.1140/epjds/s13688-021-00275-w}. [Online]. Available: http://dx.doi.org/10.1140/epjds/s13688-021-00275-w.

[13] M. Tennekes and Y. Gootzen, *A Bayesian approach to location estimation of mobile devices from mobile network operator data*, https://arxiv.org/abs/2110.00439, 2021. arXiv: arXiv:2110.00439v1.

[14] M. C. González, C. A. Hidalgo, and A. L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008, ISSN: 14764687. DOI: \mydoi{10.1038/nature06958}. arXiv: 0806.1256. [Online]. Available: https://www.nature.com/articles/nature06958.

[15] A. Ogulenko, I. Benenson, M. Toger, J. Östh, and A. Siretskiy, "The Fallacy of the Closest Antenna: Towards an Adequate View of Device Location in the Mobile Network," pp. 1–17, 2021.

[16] F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano, *Estimating Population Density Distribution from Network-based Mobile Phone Data*. 2015, pp. 1–90, ISBN: 9789279501937. DOI: \mydoi{10.2788/162414}. [Online]. Available: https://ec.europa.eu/eurostat/cros/system/files/Final-%20jrc-AIT-MNO-study-compressed%7B%5C_%7D1.pdf.

[17] F. Ricciato and A. Coluccia, *On the estimation of spatial density from mobile network operator data*, https://arxiv.org/abs/2009.05410, 2021. arXiv: arXiv:2009.05410v2.

[18] F. Ricciato, P. Widhalm, F. Pantisano, and M. Craglia, "Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation," *Pervasive and Mobile Computing*, vol. 35, pp. 65–82, 2017, ISSN: 15741192. DOI: \mydoi{10.1016/j.pmcj.2016.04.009}. [Online]. Available: http://dx.doi.org/10.1016/j.pmcj.2016.04.009.

[19] M. Ramljak, *Github Repository: MNO_mobdensity*, https://github.com/eurostat/MNO_mobdensity, 2021.

[20] A. Raue, C. Kreutz, F. J. Theis, and J. Timmer, "Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, 2013, ISSN: 1364503X. DOI: `\mydoi{10.1098/rsta.2011.0544}`. arXiv: `1202.4605`.

[21] J. Grazzini, P. Lamarche, J. Gaffuri, and J.-M. Museux, "Show me your code, and then I will trust your figures: Towards software-agnostic open algorithms in statistical production," in *New Techniques and Technologies for Statistics (NTTS) conference*, Jun. 2018. DOI: `\mydoi{10.5281/ZENODO.3240282}`. [Online]. Available: `https://zenodo.org/record/3240282`.

[22] M. Baker, *Why scientists must share their research code*, Sep. 2020. DOI: `\mydoi{10.1038/nature.2016.20504}`. [Online]. Available: `https://www.nature.com/articles/nature.2016.20504`.

[23] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: `https://www.R-project.org/`.

[24] F. Bassetti, S. Gualandi, and M. Veneroni, "On the computation of kantorovich-wasserstein distances between two-dimensional histograms by uncapacitated minimum cost flows," *SIAM Journal on Optimization*, vol. 30, no. 3, pp. 2441–2469, 2020, ISSN: 10526234. DOI: `\mydoi{10.1137/19M1261195}`. [Online]. Available: `https://doi.org/10.1137/19M1261195`.

[25] H. Wickham, M. Averick, J. Bryan, *et al.*, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019. DOI: `\mydoi{10.21105/joss.01686}`.

[26] M. Dowle and A. Srinivasan, *Data.table: Extension of 'data.frame'*, R package version 1.13.2, 2020. [Online]. Available: `https://CRAN.R-project.org/package=data.table`.

[27] D. Bates and M. Maechler, *Matrix: Sparse and dense matrix classes and methods*, R package version 1.3-2, 2021. [Online]. Available: `https://CRAN.R-project.org/package=Matrix`.

[28] E. Pebesma, "Simple Features for R: Standardized Support for Spatial Vector Data," *The R Journal*, vol. 10, no. 1, pp. 439–446, 2018. DOI: `\mydoi{10.32614/RJ-2018-009}`. [Online]. Available: `https://doi.org/10.32614/RJ-2018-009`.

[29] E. Pebesma, *Stars: Spatiotemporal arrays, raster and vector data cubes*, R package version 0.5-2, 2021. [Online]. Available: `https://CRAN.R-project.org/package=stars`.

[30] R. J. Hijmans, *Raster: Geographic data analysis and modeling*, R package version 3.3-7, 2020. [Online]. Available: `https://CRAN.R-project.org/package=raster`.

[31] M. Tennekes, *Mobloc: Mobile phone location algorithms and tools*, R package version 0.5-1, 2020. [Online]. Available: `https://github.com/MobilePhoneESSnetBigData/mobloc`.

[32] S. Gualandi, *Spatialkwd: Spatial kwd for large spatial maps*, R package version 0.4.0, 2021. [Online]. Available: `https://CRAN.R-project.org/package=SpatialKWD`.

[33] J. van der Laan and E. de Jongey, "Maximum likelihood reconstruction of population densities from mobile signalling data," in *Net-Mob'19*, 2019.