Research Report

Marco Ramljak

2021-10-13

I Introduction

The following paper represents a so-called interim research report. It is written in the form of a mini-thesis and therefore includes most of the relevant parts of the final thesis, however, in a preliminary fashion. This applies particularly to the Theory and Results/Discussion section. As an aid to understanding between this preliminary work and the final thesis some author's notes are written to the reader, which is marked in cursive.

Data generated by the cellular network of a mobile network operator (MNO) represent a rich potential source for estimating the spatial distribution of mobile phones at some given time and, from there, gain insight into the temporal variations of the spatial distribution of humans – relevant for applications in, e.g., demography (e.g., Koebe 2020; Sakarovitch et al. 2018), epidemiology (e.g., Grantz et al. 2020; Szocska et al. 2021) and urban planning (e.g., Galiana, Sakarovitch, and Smoreda, n.d.; Yang et al. 2020). One popular data source resulting from MNO data are Call Detail Records (CDR). This data source structures all communications between a mobile device and the MNO network in a database. It logs the user (mobile device holder, anonymized), the radio cell, the communication form (e.g., call, text message), and the time - i.e., Big data. MNO's can aggregate this data for a specific time range to develop a cell counter dataset that counts the mobile phones "connected" to a certain cell at a specific point in time. The resulting dataset can be used - in a privacy preserving way because of the aggregation - to estimate the spatial density of mobile phones within the operating area by "distributing" the respective number of mobile phones over the coverage area of the cell in a reasonable way (Ricciato et al. 2015, 2020).

Modeling these cell-specific coverage areas is a task called "geo-location," and it is a complex task as the coverage area of a particular cell is - in reality - never known precisely - but only approximated. The coverage area of a cell is dependent on many parameters, some of them being very volatile such as the weather conditions. Currently, there are two distinct strategies in the literature performing this task. The traditional approach is coined as "Voronoi geo-location" (deterministic). It classifies each point in the operating area to its nearest cell - therefore allowing for no overlap between the cell coverage areas. While this approach is very simple as it requires minimal information on the cells to perform, it inherently assumes that each mobile phone always connects to its closest cell, which - shown in a recent empirical analysis of a Swedish MNO - is not the case for 60% percent of all connections (Ogulenko et al. 2021). In reality, many types of cells have different characteristics, and coverage areas are planned in a way that they overlap to improve the overall reception quality. Voronoi geo-locations, therefore, do not offer the flexibility of accurately modeling cell overage areas.

The current experimental approach is based on radio propagation modeling. It allows for the implementation of many cell-specific characteristics, if available, and can explicitly handle overlapping coverage areas by introducing a probabilistic logic - however, its flexibility and high level of sophistication are dependent on much more cell information. Currently, this poses a high barrier as MNO's do not want to share data on their network, fearing it might reveal competitive advantages. Recent simulation studies have shown that, with appropriate estimation methods based on stochastic 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 geo-locations (Ricciato et al. 2020; Ricciato and Coluccia

2021). However, such results were obtained (i) under the assumption of perfect coverage area knowledge and (ii) limited to a single scenario characterized by a dense, multi-layer radio coverage pattern with a high degree of cell overlapping. It is questionable if such spatial accuracy gains can be expected when using information on the cells of a realistic quality level, i.e., not perfect, and if they are robust across different network topologies. Reasonably it is expected that probabilistic estimation strategies gain in spatial accuracy with increasing quality levels of cell information knowledge, reaching peak accuracy levels with perfect knowledge. However, this expectation needs to be tested empirically, as a consistent relationship between the quality level of cell information and the spatial accuracy of the resulting estimate has not been theoretically proven.

In the present work, we want to further scrutinize this potential trade-off between spatial accuracy gain and information revealment. We conduct an extensive sensitivity analysis based on semi-synthetic data of all state-of-the-art estimation strategies containing both cell geo-location methods. We investigate the sensitivity of estimation accuracy (i) to uncertainties in coverage area modeling due to imperfect knowledge of the radio propagation environment and (ii) to various scenario parameters concerning the radio network topologies (e.g., cell size, density, degree of overlapping). 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. Preliminary results show that a somewhat stable relationship between the quality level of implemented cell information and the spatial accuracy of the resulting estimate is empirically observed. These observations are also robust across different network topologies making it advisable to further invest academic efforts in developing more performant estimation strategies based on a probabilistic geo-location. Furthermore, academic and political efforts are needed in creating workflows for accessing high-quality cell information knowledge in secure ways.

The remaining sections contain: (II) the formalization of the estimation task and the discussion of all relevant estimation strategies, (III) the introduction of the empirical approach to this study, (IV) the illustration of preliminary results, (V) and finally the discussion of concluding remarks.

II Theoretical Framework

We assume the operating area to be discretized into a regular grid, e.g., 100m*100m. Each individual square unit is termed as a tile (formalized as j), which can contain a certain non-negative number of mobile phones to be estimated (formalized as u_j). This standardized unit helps us set a formalism framework that has been used in multiple studies beforehand (e.g., Tennekes and Gootzen 2021; Ricciato and Coluccia 2021; Ricciato et al. 2017). Let the ith element c_i of the column vector $c := [c_1 \ldots c_I]^T$ define the observed number of mobile phones counted in cell $i = 1, \ldots, I$ (cell count vector). We define 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 phones in the operating area.

Connecting to a defined cell i primarily depends on the signal strength, which is measured on the dBm scale. For this study, its feasible values range between -70 dBm (perfect signal) and -120 dBm (very bad signal). The received signal strength from a certain cell depends on a multitude of parameters, some of which are not even measurable. Therefore, in reality, it is impossible to precisely predict the signal strength of a specifc 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 12 cells that can offer reception within the defined range (also called "cell competition"). 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 through the signal strength. To account for load balancing, previous research has introduced the signal dominance measure, which translates signal strength via a logistic function into signal dominance (Tennekes and Gootzen 2021). This measure contains both the received signal strength and a model for load balancing and is much more interpretable as it ranges between 0 (worse) and 1 (better).

Both signal strength and signal dominance are measures that are computed in a network-independent way. A network-dependent measure is the emission probability p_{ij}, which rates all available cells for the phones

within a specific tile based on their connection likelihood. It is defined as the following conditional probability:

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

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.

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

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

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

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

where $g(\cdot)$ defines the estimator of choice. These will be described in the following subsection.

In simple terms, the estimation problem can be defined as distributing the connected phones of each cell across the operating area in a most likely way in terms of the connection likelihood.

(Author's note: These are the most important definitions and formulas (adapted from (Ricciato and Coluccia 2021)) *one needs to know about currently. In the final thesis, this whole "formula chain" will be described in graphical form to save "official" words.)*

Estimation Methods

(Author's note: The following subsection describes all field-relevant estimation methods based on the accompanying geo-location logic introduced in the Introduction. The same notation from above applies unknown symbols are briefly explained. Only the most important aspects have been listed as an aid for further understanding.)

1. **Voronoi (Deterministic)**

    1. Formula: $\hat{u}_j = \frac{c_i}{S_i}$
       where $\hat{u}_j$ represents the tile specific estimate, c_i the cell-specific number of connected cell phones, and S_i the size of the cell-specific area. If a tile intersects with multiple Voronoi regions (not because of overlap but because the Voronoi coordinate system does not have to be aligned with the regular grid), a weighted average based on the sizes of the intersections is computed.

    2. Logic: Seed points (e.g., geographic cell location) are projected to the focus area, and each point in the area is classified with its closest seed. By definition, no overlapping cells nor holes can be modeled. The inherent assumption is that each phone is connected to its closest cell.

    3. Further notes:

        1. Seed options
            - Seed: Tower
            - Seed: Cell offset
            - Seed: Cell barycenter

2. **Simple Bayes (Probabilistic)** (Tennekes and Gootzen 2021)

    1. Formula: $\hat{u}_j = a_j * \sum_{i=1}^{I} c_i \frac{p_{ij}}{\sum_{k=1}^{J} p_{ik} a_k}$

2. Logic: Direct computation based on Bayes rule

3. Further notes: a_j refers to a prior value for the specific tile j.

3. **MLE/EM (Probabilistic)** (Ricciato et al. 2020)

1. Formula: $\hat{u}_j^{m+1} = \hat{u}_j^m * \sum_{i=1}^{I} c_i \frac{p_{ij}}{\sum_{k=1}^{J} p_{ik} \hat{u}_k^m}$

2. Logic: Hierarchical generative model where the elements of u are modeled as Poisson random variables. MLE is computed iteratively (m stands for the specific iteration) via the Expectation Maximization (EM)-algorithm.

4. **DF approximation (Probabilistic)** (Ricciato and Coluccia 2021)

1. Formula: $\hat{u} := max(AP^T(PAP)^{-1}(c - Pa) + a; 0)$ maximum intended element-wise

2. Logic: Estimating via maximum likelihood candidate estimates that fit best based on the emission probabilities and choosing the one that fits best with a set prior vector.

3. Further notes: a_j refers to a prior value for the specific tile j and $A = diag(a)$.

Sensitivity towards cell coverage model mismatch

Compared to deterministic geo-locations, probabilistic geo-location strategies consider much more information concerning the actual coverage of the cells, meaning they allow for flexible radio propagation modeling. Furthermore, previous research has shown that probabilistic estimators perform better in spatial accuracy than deterministic estimators, assuming that perfect cell coverage information is available (Ricciato et al. 2020; Ricciato and Coluccia 2021). Perfect cell coverage information is never available, as there are many parameters to be considered, some of them very difficult or even impossible to measure. The question arises, how stable these estimators perform, when only an approximation of the accurate coverage information is available?

Robustness towards network topology (scenarios)

The main argument against a deterministic geo-location is the inability to model coverage areas that overlap. In reality, overlaps are necessary to prevent bad reception as the radio signal is not uniformly distributed within a coverage area. How the signal is propagated within a coverage area is dependent on the cell's parameters, starting with the type of the cell (omni-directional or directional), its power level, its beam width, etc.[1] MNO's plan their network by establishing so-called cell layers. Cells within the same layer are pretty similar in propagation characteristics to cells from other layers. Cells from a Macro layer, e.g., try to propagate some signal over large rural areas (i.e., large coverage areas). In contrast, Micro cells are often placed in urban areas and have small coverage areas to handle large volumes of mobile phones. In modeling terms, we can translate these differences into differing signal dominance-to-distance relations between layers. Differing signal dominance-to-distance relations are one of the reasons why mobile phones do not always connect to their closest cell.

Research Questions

RQ1: How robust are estimation strategies to network model mismatching errors (information quality)?

---

[1]It is out of the scope of this report to introduce the basics of radio propagation modeling. An extensive yet compact introduction on the topic can be found here (Tennekes and Gootzen 2021).

Varying the coverage information quality within probabilistic estimators will reveal the estimators' sensitivity to imperfect cell coverage information. We assume that the spatial accuracy of the estimate derived from a probabilistic estimator is positively related to the level of quality of the coverage information implemented.

RQ2: How sensitive are estimation strategies to network characteristics, such as cell density or differing cell layers?

Varying the network characteristics through different network topologies will increase the final performance results' robustness as well as indicate the necessary quality degree of cell coverage information when comparing probabilistic estimators to deterministic estimators. We expect estimators based on a Voronoi geo-location approach to perform worse, (i) the higher the degree of coverage area overlap, and (ii) with the existence of multiple cell layers. Furthermore, it is important to find out if potential spatial accuracy gains through probabilistic estimation strategies are consistent across differing network topologies.

III Empirical Approach

The estimation methods under scrutiny expect two primary data sources: (Aggregated) Call Detail Records (CDR) data, which are classified as sensitive for personal information reasons, and data on the spatial coverage of an MNO, which are classified as sensitive for competition reasons. Accessing "real" data for this study is therefore not feasible. However, we are primarily interested in the robustness of the different estimators. Therefore, we need to verify that we can correctly evaluate the spatial accuracy of certain model specification choices. This can only be done if one has access to a ground truth population (GTP) and a ground truth geo-location (in the following "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 execute our simulations, we use the MNO-simulator workflow[2], which was constructed within the programming language R and helps with developing the non-trivial software that is necessary to "mimic" MNO-like data. This workflow contains three modules: (1) *Toyworld Generation*, which gives flexible options of generating multiple scenarios concerning mobile phone density and network density, as well as assigns mobile phones to cells based on an individually specified generative model P; (2) *Estimation*, which helps with specifying geo-location modules (creating estimation models P^*) and executing all state of the art estimation strategies; and (3) *Evaluation*, which offers multiple quantitative evaluation metrics, such as the Kantorovich-Wasserstein Distance (KWD) from the `SpatialKWD` (Bassetti, Gualandi, and Veneroni 2020) package taking the spatial nature of the estimation problem into account.

Data sources and simulations

To create a toyworld within the MNO-simulator we need:

- a reference area on a grid with a mobile phone population assigned to each tile j within the grid,

- a network topology consisting of multiple cells i specified with the necessary parameters,

- a process that creates cell counters c based on a generative model P.

The only "real" data source we use in this study is the reference area on a 100m*100m regular grid. We chose the city of Munich and its near surroundings, which encompasses 1,600 square kilometers (160,000 tiles of size 100m*100m) and used its census population for the generation of the GTP. The final GTP in our study was reduced to about 1/3 of the true census population to mimic the customer basis of a single MNO with that market share. Fig. 1 visualizes the spatial density of the GTP.

---

[2]The MNO-simulator was developed by the same authors of this paper and has been used already in this preprint. It will be formally presented for the first time at uRos 2021 and within a separate paper early next year. For general data handling and visualization, it relies on the `tidyverse` and `data.table`, for spatial data operations on `sf`, for radio propagation modeling on `mobloc`, and for approximating the KWD it relies on `SpatialKWD`.

GTP
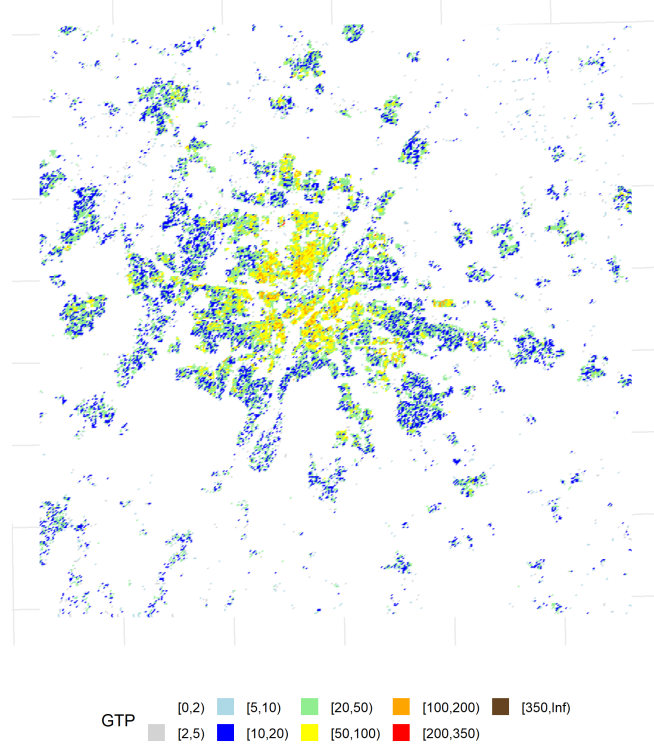| | [0,2) | | [5,10) | | [20,50) | | [100,200) | | [350,Inf) |
| | [2,5) | | [10,20) | | [50,100) | | [200,350) | |

Figure 1: Spatial density map of the ground truth mobile phone population

Within the MNO-simulator we can flexibly create various network topologies (in the following "toyworld" or "network") based on a multitude of relevant cell parameters. We are especially interested in the effects on the estimators based on particular dense and sparse networks, and multi-layer networks. Therefore, we create four toyworlds, differing in these aspects.

Author's note: The (visual) description of the four toyworlds will be implemented at a later stage. Toyworld 1/Toyworld 2 is a one-layer network with a particularly sparse/dense pattern, meaning it has a low/high number of cells and, therefore, low/high coverage area overlap. Toyworld 3 is a two-layer network, and Toyworld 4 is a three-layer network, which in both cases leads to multiple differing signal dominance-to-distance relations.

Model mismatch techniques

As we are interested in implementing cell information knowledge of varying quality, we define different estimation model matrices P^* that are of the same form (i.e., dimension) as the generative model matrix P, but differ in the level of information or quality. For this, we have come up with two so-called model mismatch techniques that purposefully introduce errors into the cell information. One mismatch technique introduces random noise at the signal strength stage, while the other technique quantizes the variability level of the signal dominance. After the errors are introduced, the emission probabilities p_{ij} are calculated and stored in an estimation model matrix P^*. In total, we create 11 estimation matrices - five of each mismatch technique - and one estimation matrix that is identical to the generative model matrix ($P^* = P$) as a benchmark. In the following, each mismatch technique is further explained.

6

Model mismatch technique I: Spatially sensitive random noise

We take the true signal strength (measured in dBm) as a basis for this mismatch technique and introduce random noise sampled from a uniform distribution with variable min-max values (e.g., min: -3; max: 3), see Fig. 2. The most extreme case is min: -21, max: 21. Each sampled value is added (or subtracted) 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. Furthermore, the cell coverage direction is retained as this can be classified as secure knowledge. Distorted signal strength values are transformed to signal dominance values in the usual way, and through the minimum signal dominance threshold, "coverage holes" (as seen in Fig. 2 below) are possible. For example, this mismatch technique resembles the signal strength volatility through weather conditions - particularly the extremely noised versions can be interpreted as a "stress test" for the estimation strategies.
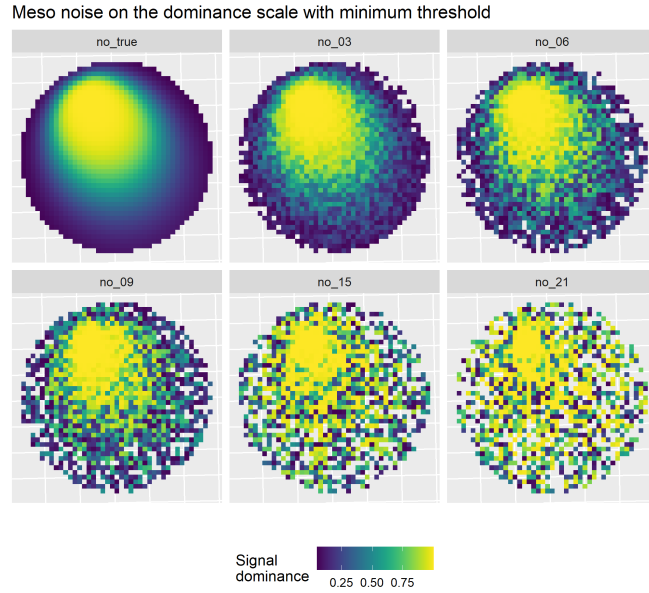

Figure 2: Example coverage areas visualizing the mismatch technique: spatially sensitive random noise

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 value range categories. Fig. 3 shows the quantized coverage areas, and Fig. 4 shows the translation of the true signal dominance value range to the final quantized signal dominance value, respectively, for each version. The number of categories is the variable parameter (2^n). 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. We quantize at the signal dominance stage and not at the signal strength stage, as otherwise, the categories would get distorted through the signal dominance computation. This mismatch technique resembles different levels of quality of cell coverage data - a realistic situation is if one only has access to the location, the direction, and an approximate power level of the cells, from which the coverage area can only be very roughly approximated. Furthermore, through this mismatch technique, we can research the behavior of the estimation strategies with almost equal emission probabilities p_{ij}.
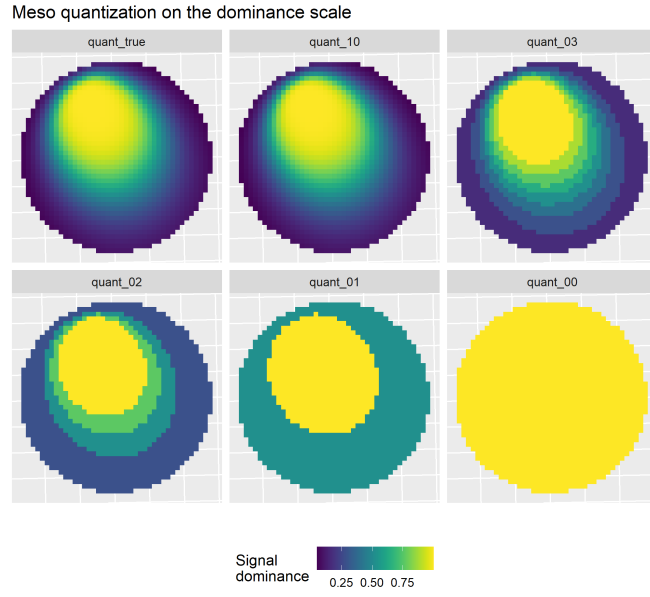
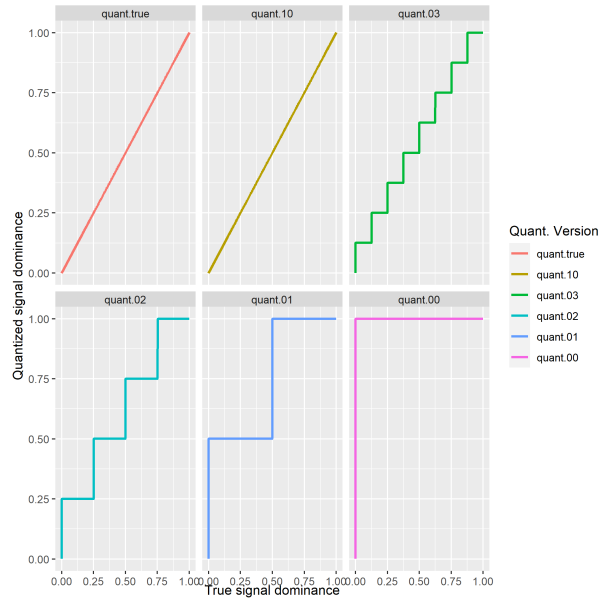Figure 3: Example coverage areas visualizing the mismatch technique: quantization

Figure 4: Translation of the true signal dominance value range to the final quantized signal dominance value

Evaluation criterion: Kantorovitch-Wasserstein Distance

Concerning spatial density accuracy, the central question we ask ourselves in our experiments 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 quantitatively measuring the similarity between two spatial density maps, we need to account for the spatial nature of the analysis problem. Therefore, we resort to the Kantorovitch-Wasserstein Distance (KWD), which was also used in previous work concerning the accuracy of spatial density estimates (Ricciato and Coluccia 2021; Ricciato et al. 2020). For computing the KWD we use the R-package `SpatialKWD` (Bassetti, Gualandi, and Veneroni 2020). This package results from the referenced article, which develops a computationally efficient approximation method of the KWD. A tunable integer parameter L balances the computational resources with the KWD's estimation accuracy. We use 'L = 2' for this approximation parameter, which means that the approximation error will be in the worst case 2.68%. The output value is used as an upper bound for the KWD estimate, and the worst-case adjustment acts as a lower bound.

As indicated in the section describing the estimation methods, 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 this mass is distributed in the same way over the area as the geographical distribution of the ground truth - and if not, how far any population unit (mobile phone) needs to travel on average in terms of tiles. Therefore, high KWD values are associated with a highly dissimilar distribution, and low KWD values are associated with a highly similar distribution compared to the ground truth.

Analysis plan

Authors note: The following will be contained in a graphic or flowchart

- For the same area with the same population values per tile, we created four different networks differing in cell density and/or the number of layers.
- We use the generative model information concerning the cell coverage to create ten mismatched models and one true model for the estimation phase.
- We run all probabilistic estimation strategies with all estimation models P^* to receive the respective final estimation distribution. We also run all deterministic estimation strategies as a benchmark.
- Each final estimation distribution is quantitatively evaluated using KWD. We are interested in the estimated distributions with the lowest KWD values.
- Further relevant statistics are calculated, which assess, e.g, the convergence of numerical estimation strategies.

IV Results

The current results only entail the most crucial result of the two one-layer toyworlds (sparse and dense) for the noise mismatch technique (Fig. 5). More descriptive results, and the other toyworlds and the quantization mismatch, will be implemented in the final thesis.

1. Mismatch (RQ1)

    - Even with a severe model mismatch, probabilistic estimators always lead to spatially more accurate results than any Voronoi estimator. The KWD values are consistently lower the less noise is introduced. Furthermore, all "noised" versions of the probabilistic estimators (MLE, DF) are lower than any Voronoi estimator (Vor_t, Vor_o, Vor_b).
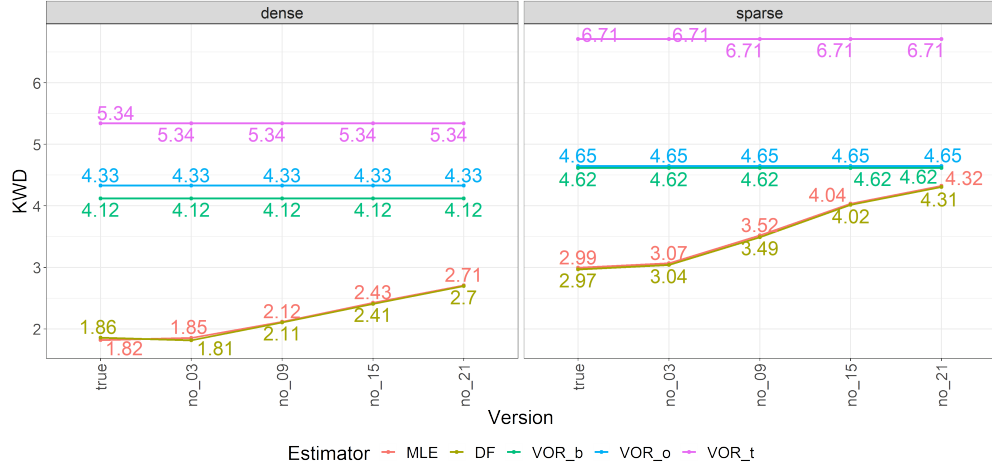
Figure 5: This plot shows each KWD value for every estimator, with every noised estimation model for the sparse and dense network topology. The colors (estimators) are classified in the bullet point below.

2. Differing networks (RQ2)

   - Higher network density offers more information and can therefore lead to spatially more accurate results than sparse networks. Probabilistic estimators (MLE, DF) utilize this information better than Voronoi estimators (Vor_t, Vor_o, Vor_b) as they have lower KWD values.

V Discussion / Conclusion

This research project scrutinizes the trade-off between spatial estimation accuracy and model parsimony concerning the task of deriving mobile phone density from MNO data. One could argue that Voronoi estimators, i.e. deterministic estimators, represent a parsimonious estimator. However, these estimation strategies (1) are based on a faulty assumption (mobile phones always connect to its closest cell), and (2) entail the unusual and unwanted barrier of an information implementation barrier. Therefore, they offer limited flexibility in modeling cell coverage areas, even if further helpful information would be available. With probabilistic estimators, on the other hand, all relevant available information concerning cell coverage modeling can be implemented/used, and these estimators entail the logical modeling property of improved performance through better/more information.

Concluding from our sensitivity analysis results, we argue that a Voronoi estimator should only be chosen when merely the geographic cell locations are available. However, a probabilistic estimator is advisable whenever the cell information on their propagation direction and/or power is available.

(Author's note: Limitations and further aspects for the Discussion / Conclusion will be added when all results have been observed and interpreted.)

VI References

Bassetti, Federico, Stefano Gualandi, and Marco Veneroni. 2020. "On the computation of kantorovich-wasserstein distances between two-dimensional histograms by uncapacitated minimum cost flows." *SIAM Journal on Optimization* 30 (3): 2441–69. https://doi.org/10.1137/19M1261195.

Galiana, Lino, Benjamin Sakarovitch, and Zbigniew Smoreda. n.d. "Understanding socio-spatial segregation in French cities with mobile phone data Year = 2018," 1–12.

Grantz, Kyra H., Hannah R. Meredith, Derek A. T. Cummings, C. Jessica E. Metcalf, Bryan T. Grenfell, John R. Giles, Shruti Mehta, et al. 2020. "The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology." *Nature Communications* 11 (1): 1–8. https://doi.org/10.1038/s41467-020-18190-5.

Koebe, Till. 2020. "Better coverage, better outcomes? Mapping mobile network data to official statistics using satellite imagery and radio propagation modelling." *PLoS ONE* 15 (11). https://doi.org/10.1371/journal.pone.0241981.

Ogulenko, Aleksey, Itzhak Benenson, Marina Toger, John Östh, and Alexey Siretskiy. 2021. "The Fallacy of the Closest Antenna: Towards an Adequate View of Device Location in the Mobile Network," 1–17.

Ricciato, Fabio, and Angelo Coluccia. 2021. "On the estimation of spatial density from mobile network operator data," 1–19. http://arxiv.org/abs/arXiv:2009.05410v2.

Ricciato, Fabio, Giampaolo Lanzieri, Albrecht Wirthmann, and Gerdy Seynaeve. 2020. "Towards a methodological framework for estimating present population density from mobile network operator data." *Pervasive and Mobile Computing* 68: 1574–1192. https://doi.org/10.1016/j.pmcj.2020.101263.

Ricciato, Fabio, Peter Widhalm, Massimo Craglia, and Francesco Pantisano. 2015. *Estimating Population Density Distribution from Network-based Mobile Phone Data.* https://doi.org/10.2788/162414.

Ricciato, Fabio, Peter Widhalm, Francesco Pantisano, and Massimo Craglia. 2017. "Beyond the 'single-operator, CDR-only' paradigm: An interoperable framework for mobile phone network data analyses and population density estimation." *Pervasive and Mobile Computing* 35: 65–82. https://doi.org/10.1016/j.pmcj.2016.04.009.

Sakarovitch, Benjamin, Marie Pierre de Bellefon, Pauline Givord, and Maarten Vanhoof. 2018. "Estimating the residential population from mobile phone data, an initial exploration." *Economie Et Statistique* 2018 (505-506): 109–32. https://doi.org/10.24187/ecostat.2018.505d.1968.

Szocska, Miklos, Peter Pollner, Istvan Schiszler, Tamas Joo, Tamas Palicz, Martin McKee, Aron Asztalos, et al. 2021. "Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis." *Scientific Reports* 11 (1): 1–9. https://doi.org/10.1038/s41598-021-81873-6.

Tennekes, Martijn, and Yvonne Gootzen. 2021. "A Bayesian approach to location estimation of mobile devices from mobile network operator data." http://arxiv.org/abs/arXiv:2110.00439v1.

Yang, Junyan, Yi Shi, Chuck Yu, and Shi Jie Cao. 2020. "Challenges of using mobile phone signalling data to estimate urban population density: Towards smart cities and sustainable urban development." *Indoor and Built Environment* 29 (2): 147–50. https://doi.org/10.1177/1420326X19893145.