Large-Scale Cellular Network Modeling from Population Data: An Empirical Analysis

Andreas Achtzehn, Student Member, IEEE, Janne Riihijärvi, Member, IEEE, and Petri Mähönen, Senior Member, IEEE

Abstract—Accurate estimates of the spatial distribution of base stations (BSs) are crucial for the analysis of core performance and connectivity metrics in cellular networks. The appropriate tuning of these models, particularly through the fitting of point processes (PPs), enables practical simulations and theoretical analyses with realistic network configurations. However, the commonly made assumption of stationarity in model parameters does not hold for larger networks, with especially population inhomogeneities causing large network structure variations. We use BS location information from different operators in Germany, to explore the efficacy and limitations of using population data as a covariate to approximate cellular deployments. Our analysis shows that the overall network density is highly correlated already for small areas. Further, we find that for moderately populated areas purely population-driven PPs yield a good statistical match to our real data. The validity boundaries discussed in this paper provide a useful reference for determining at which length scales more complex, interaction-based, PPs are necessary.

I. INTRODUCTION

Research on cellular networks and their structures often requires for a precise understanding of the spatial configuration of real macro- and micro-cellular installations [1]. Such data is necessary to provide credible results on expected coverage, SIR distributions, and other key planning criteria for current and future mobile networks. Employing unsuited or wrongly calibrated models of these deployments bear the danger of highly inaccurate predictions, which makes finding good approximations of cellular network deployments a pressing challenge.

The most prominent way for studying network deployments is the fitting of point processes (PPs) to available data sets. The proponents of this approach thereby aim to derive universal rules on the configuration of real-world cellular networks that can be applied for practical scenario synthesis and analysis. While earlier works hereby mostly considered Poisson PPs due to their salient tractability features [2], more recent research also considers local interactions by applying more complex PP families [3]–[5]. Common to all these models is their stationarity assumption, i.e. that the locally derived parameters of the respective PP are valid at arbitrary locations and scales.

The density of a cellular network is generally not constant as the example of Fig. 1b for Germany readily shows. Instead, the planning goals of providing universal coverage and local capacity govern the base station (BS) configurations in such large-scale scenarios. The population density as shown in Fig. 1a is thereby often considered as a sensible covariate to model larger deployments [6]. This paper further validates this assumption using different network deployment and population data sets for Germany. Complementing earlier research such as [6], our aim is to identify the practical boundaries for which population density is a reasonable covariate for large-scale PP-based cellular network distribution models. The analysis presented in this paper confirms for Germany the general validity of population-based density predictions, i.e. estimating the number of BSs per unit area based on the population density is sensible, and a good prediction accuracy of purely population-driven PPs for moderately urbanized areas. We assume that our results are equally applicable also for other developed countries, and may provide a valuable reference for determining at which length scales more complex PPs need to be employed.

The rest of this paper is organized as follows: In Section II we introduce the reference data sets which we have used in this empirical analysis. Section III discusses the feasibility of using population counts to estimate the average density of macro- and microcellular installations. We further assess in Section IV how well suited interaction-less inhomogeneous PPs are in general for cellular network modeling. In Section V we discuss the practical boundaries for this class of PPs. Section VI concludes the paper.

II. REFERENCE DATA SETS

We use public records of 70093 base stations for our empirical analysis of base station distributions in Germany. The records list the GPS coordinates, sectorization, and antenna height information for all network operators. Furthermore, we use a 16367 sample subset which we were able to match...
with the deployment information of a single operator\(^1\). In our further analysis, we denote the respective collection of base station locations as \( \Gamma = \{ \gamma_i \}_{i=1}^{N} \), and the area of Germany as \( \mathcal{A} \).

We compare our cellular deployment information with two population data sets. The first originates from a recent census and provides residence information for 80.3 million registered inhabitants at the fine-grained resolution of a 100 m \( \times \) 100 m grid. It represents the most accurate publicly available information on where people live in the country. The coarser 30 arc-seconds resolution LandScan data set [7], which we use as a second reference, gives estimates on the average population count to offer an estimate on where people usually are. The latter better reflects population in non-residential environments, but is less accurate in sparse population environments. For our analysis, the population counts are represented by lattice data, \( \Theta = \{ (\zeta_i, c_i) \}_{i=1}^{M} \), where \( \zeta_i \in \mathcal{A} \) is the center of the count rectangle and \( c_i \) is the local population count.

### III. Base Station vs. Population Count

We start our analysis by studying whether the area population count is a sufficient metric to estimate the local number of base stations. While the country-wide match of population and deployment densities is rather apparent, see Figs. 1a and 1b, we are interested in how well this relationship is locally maintained, i.e. when the considered area is shrunk. When normalized by the area size, such count statistics yield the average density, \( \mathbb{E}(\lambda_c) \), for subsequent synthesis and verification of a PP.

We employ Monte-Carlo density sampling [8] for this purpose; we randomly select a closed-shaped study area of fixed size \( s \), \( A_s \subset \mathcal{A} \), and generate the respective base station counts \( \lambda_{i,c} \) and population counts \( \lambda_{i,p} \) as

\[
\lambda_{i,c} = \left| \{ \gamma_j | \gamma_j \in \Gamma, \gamma_j \in A_s \} \right|, \quad \lambda_{i,p} = \sum_{(\zeta_i, c_i) \in \Theta} c_i 1(\zeta_i \in A_s). \quad (1)
\]

\(^1\)For confidentiality reasons we cannot disclose the name of the operator.

This selection and counting is repeated \( N \) times to acquire bootstrap samples \( \Lambda_c(s) = \{ \lambda_{i,c} \}_{i=1}^{N} \) and \( \Lambda_p(s) = \{ \lambda_{i,p} \}_{i=1}^{N} \), respectively\(^2\). Note that the shape of \( \Lambda_c \) does not significantly affect the overall statistics of this estimation method [9]. We focus in the following on a potentially linear relationship between the values in \( \Lambda_c \) and \( \Lambda_p \). Thus we employ the Pearson correlation \( \rho(\Lambda_c, \Lambda_p) \) and coefficient of determination \( R^2 \) for a linear regression as metrics. In fact, for our application here \( R^2 \) corresponds directly to \( \rho^2 \), but in general differences arise if additional regressors (such as socio-economic indicators) were included in analysis.

The strong large-scale dependency between base station and population densities becomes immediately visible in Fig. 2, where we plot \( \rho \) when the study area size is varied. The calculation is based on 500,000 random selections of a sampling square with size \( s \). We find that for small study areas (\( s < 5 \text{ km}^2 \)) there is a higher correlation with the all-operators data set than with its single-operator equivalent. This is an expected outcome as the sampling error due to the overall larger number of base stations is reduced. Nevertheless, this discrepancy reduces when the study area size is increased; at approximately \( s = 50 \text{ km}^2 \) we observe for both data sets a high correlation of \( \rho > 0.9 \). Universally our data exhibits a better match to the census data, whereby the difference in correlation increases with larger area sizes. The census population data set is therefore better suited for the base station density estimation.

A first-order regression analysis with \( \Lambda_c(s) \) conditioned on \( \Lambda_p(s) \) allows to determine the domain of validity for the linearity assumption of the density estimate. With \( s = 25 \text{ km}^2 \) we find high values of \( R^2 = 0.80 \) and \( R^2 = 0.86 \) for the single and all operators data set, respectively, which indicate a good fit to the linear model. However, when further splitting the

\(^2\)We omit the step of normalizing the count statistics to the area size for sake of clarity, but remark their equivalence for the subsequent analysis.
When carrying out an independent regression on these high-density cases, the high $R^2$ can be retained, whereby linear regression provides good values of $R^2$ until $T \approx 2 \times 10^{-3}$, i.e. for even denser population areas linear approximations become unreasonable as BS density no longer follows the population density.

IV. INTERACTION-LESS BS APPROXIMATIONS

We will now assess to what extent non-stationary but interaction-less (i.e. inhomogeneous Poisson) PPs with population data as an input are sufficient for predicting the local structure of a cellular network. Our main quantitative metric for the fitness is the quasi log-likelihood estimate [10] of the actual base station data sets, which indicates how well the (potentially resampled) population distribution can explain the observed base station layout.

For this purpose, let us in a first step use the discrete population density $\Theta$ definition to derive an intensity function $f(\gamma)$ of an inhomogeneous Poisson PP [11] on $\mathcal{A}$ with

$$\tau = \eta \sum_{(\xi_i, c_i) \in \Theta} c_i,$$

where $\tau$ is Poisson distributed with $\mathbb{E}(\tau) = |\Gamma| = \int_{\mathcal{A}} f(\gamma) dy$, and $\eta$ is a normalizing constant. For $f(\gamma)$ with $\gamma \in \mathcal{A}$ we furthermore define it to be locally constant on each population count rectangle with value proportional to the count value $c_i$. Intuitively, this PP distributes on average $|\Gamma|$ base stations over $\mathcal{A}$, whereby the probability of finding a base station in the vicinity of population counting point $\xi_i$ is proportional to the count value $c_i$, but within the count rectangle each realization is homogeneous.

We generate $N$ realizations of the above PP, $\tilde{\Gamma}_j$, to acquire a quasi log-likelihood estimate for a smoothing bandwidth $\sigma$ as

$$L(\Gamma, \sigma) = \sum_{\gamma_i \in \Gamma} \log \left( \mathbb{E} \left[ \sum_{j=1}^{N} |\tilde{\Gamma}_j| \mathbb{E} \left[ \sum_{\gamma_k \in \tilde{\Gamma}_j} g(\gamma_i, \gamma_k, \sigma) \right] \right] \right).$$

The function $g(\cdot, \cdot)$ defines the density estimation kernel centered at the respective points in $\tilde{\Gamma}_j$. We use a Gaussian kernel with bandwidth $\sigma$, but note that our results are mainly sensitive to bandwidth used and not the precise kernel function selected [12]. The density estimate at the locations in $\Gamma$ is obtained by averaging over the realizations of the PP. The log-likelihood estimate of $\Gamma$ is finally obtained through summing over the (conditionally independent) logarithmic density estimates.

Interesting phenomena become visible when we vary the kernel size for our estimation, see Fig. 4. In order to maintain comparability, we show the likelihood estimates in this figure relative to the one-sigma kernel area, $A = \pi \sigma^2$. We find that for small bandwidths the average population data set provides significantly better estimates than its census equivalent. The reason for this apparent improvement is that the population is inherently less concentrated in the fewer areas of residence.\(^3\)

For both network data sets this difference diminishes when the kernel area approaches 1 km\(^2\) – census and population data become equally expressive at predicting the observed base station distributions. For all data set permutations we observe that already for very small kernel sizes the log-likelihood estimate of the smoothed inhomogeneous PP exceeds the log-likelihood of a homogeneous PP; using population density as intensity function is therefore generally superior. At $A \approx 1.5$ km\(^2\) the structure of the population data is lost as the convergence to the homogeneous PP indicates.

This result is further confirmed when we directly derive density estimates for the respective base station set $\Gamma$ and at the respective population count points $\xi_i$, again using the Gaussian kernel function. The correlation between density estimates and population counts $c_i$ reaches its maximum for kernel area sizes between 2 km\(^2\) ($\rho = 0.78$, average population, all operators) and 7 km\(^2\) ($\rho = 0.74$, average population, single operator), which underlines the good fit of the base station distribution to the population distribution. Contrary to the results for our overall density estimation, the average population data set is better suited to predict the local network structure, whereby the census data showed considerably worse fits ($\rho \leq 0.48$) at all kernel bandwidths.

V. DOMAIN OF VALIDITY FOR THE POISSON MODEL

In this section we shall briefly discuss in more detail where the inhomogeneous Poisson point process approximation is expected to break down, and where alternative models are needed should these regimes be necessary to model well for applications. Due to us using the inhomogeneous Poisson point process as a ‘baseline’ model, we use the distribution of nearest-neighbor distances $d_{nn}$ between base station locations as a descriptive summary statistic in our analysis.

\(^3\)For the example of Germany approximately 11 times more count points in the regular grid structure have non-zero population counts compared to the census data set.
For subsequent model validation work more advanced spatial statistics, such as $n$-point correlation functions will yield more information [11], but for our purposes the intuitively clear nearest neighbor distances will suffice.

We begin by contrasting the nearest neighbor distances in our single operator data set against those arising from the inhomogeneous Poisson point process defined by Eq. (2) that follows exactly the population density. Since the density of a Poisson point process is locally proportional to $1/d_{nn}^2$ we show in Fig. 5 the inverse squared empirical nearest neighbor distances as a function of the local population density. The results for the BS locations deviate from those of the inhomogeneous Poisson point process in two ways. First, the latter has many more nearby point pairs in dense regions. This is evidence of repulsion or local regularity in our data arising from network planning, as discussed in earlier work in the homogeneous point process context [3]–[5]. This local regularity is also partially responsible for the linear relationship between the population distribution and the modelled point process density breaking down in densely populated areas as discussed in Section III. Second, our data set features substantially more points in regions with very low population densities, and few points even in regions where the population density is zero. We believe these base station locations to arise from an almost population independent ‘overlay’, the purpose of which is to provide country-wide coverage for the customers of the operator. This belief is further supported by Fig. 1c which illustrates the points falling in these regions. We see that a very regular structure has emerged, encompassing approximately 19.2% of the original points, in total serving approximately 10.0% of the population.

In summary, we have seen that the local regularity in the homogeneous setting is clearly present also in our country-wide data set. Furthermore, inhomogeneous Poisson models lack the country-wide ‘overlay’ which again is probably better modeled by an almost homogeneous point process that is only weakly coupled to the population distribution. Nevertheless the inhomogeneous Poisson model has very wide degree of validity beyond these extreme scales. We also emphasize that the local regularity strongly indicates that the inhomogeneous Poisson model directly yields uppers bounds to distributions of quantities that are monotonic in local density of the point process, in particular interference statistics, further extending its usefulness. In this manner the inhomogeneous Poisson model is applicable to scenarios covering 90% of the German population.

VI. CONCLUSIONS

In this paper we have studied the feasibility of using population data to support the modeling of realistic cellular networks through point process models. Our analysis is particularly useful for simulation-driven studies, where there is only little practical earlier work on finding the validity region of this approach.

Our extensive analysis of a large country-wide data set of German deployments shows that for reasonably sized areas ($A \geq 20\text{km}^2$) the population density provides a sensible and accurate enough approximation of the average cellular network density. This relationship is well observable for small to medium populated areas, for which the large-scale structure of the cellular network is well approximated. We hence argue that with smoothed average population data sets a simple inhomogeneous Poisson PP already suffices to simulate real network distributions at a high significance level. Only in dense metropolitan areas we find a non-negligible level of local regularities in the cellular network. We presume that in those areas the non-Poisson modeling as discussed e.g. in [3]–[5] will be better suited for any practical analysis, but even there the simple population based model yields useful bounds. Our analysis of the country-wide data set has furthermore revealed an interesting overlay structure in the single operator case, which we plan to explore further in our future work.

REFERENCES