This is the author's accepted version of the article. The final version published by IEEE is O. E. Martínez-Durive, T. Couturieux, C. Ziemlicki and M. Fiore, "VoronoiBoost: Data-driven Probabilistic Spatial Mapping of Mobile Network Metadata," 2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), 2022, pp. 100-108, doi: 10.1109/SECON55815.2022.9918610..

VoronoiBoost: Data-driven Probabilistic Spatial Mapping of Mobile Network Metadata

Orlando E. Martínez-Durive[⋆†], Theo Couturieux[*], Cezary Ziemlicki[‡], and Marco Fiore[⋆]
[⋆]IMDEA Networks Institute, Spain, [†]Universidad Carlos III de Madrid, Spain,
[*]INP-ENSEEIHT, France, [‡]Orange Innovation, France
{orlando.martinez, marco.fiore}@imdea.org, theo.couturieux@etu.enseeiht.fr, cezary.ziemlicki@orange.com

***Abstract*—Mapping information collected at the level of individual base stations onto the geographical space is a required operation for many works relying on mobile network metadata. The common practice is to represent base station coverage as Voronoi cells, and assume that users are uniformly distributed therein. In this paper, we leverage a large-scale dataset of realistic spatial association probabilities to over 5,000 operational base stations, and quantify the substantial problems of such a simplistic mapping approach. To address the limitations of legacy Voronoi representations, we develop VoronoiBoost, a data-driven model that scales Voronoi cells to match the probabilistic distribution of users associated to each base station. VoronoiBoost relies on the same input as traditional Voronoi decompositions, but provides a richer and more accurate rendering of where users are located: hence, it can be readily used by researchers to substantially improve the spatial representation of mobile network metadata. Our experiments demonstrate that VoronoiBoost improves the quality of mapping by 44% on average over standard Voronoi cells. We also showcase the utility of our model in a practical Edge network planning use case, where the information produced by VoronoiBoost drives a deployment up to 28% more accurate than that obtained with Voronoi cells.**

***Index Terms*—Remote sensing, Data analysis, Mobile communication, Cellular technology, Machine learning algorithms**

I. Introduction

Metadata collected in mobile networks is being employed to support a wide range of research efforts across many knowledge domains [1], [2]. Notable examples of usages include mobile service traffic measurements in networking research [3], [4], signalling data in transportation research [5], location metadata in sociology research [6], [7], activity information in urban planning research [8], or presence statistics in geography research [9]. The pervasiveness of the spatial and population coverage, the high temporal granularity at which the metadata can be recorded, and the relative ease of the data collection are all reasons of the success of mobile networks as a source of rich information that enables studies at scale.

In terms of spatial representation, mobile network metadata is typically geo-referenced at the level of individual Base Stations (BSs). This is because BSs are the part of the network infrastructure that is the closest to the User Equipment (UE), hence the most granular; and, information about which BS is managing specific users or traffic flows is part of the control data flowing through the system, which makes locating the metadata at the BS level straightforward for the operator.

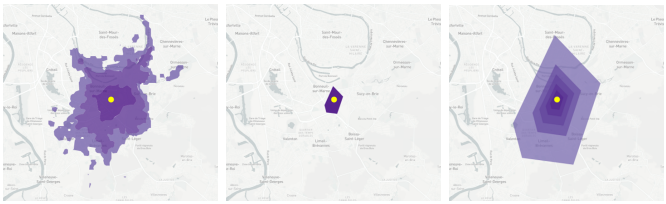
However, many studies require to map the metadata (*e.g.*, users, traffic, or activity levels) over the geographical space, rather than to the exact latitude and longitude coordinates of the BSs. An ideal spatial mapping would be based on the *probabilistic spatial diffusion* of each BS, which we define as *the likelihood that mobile subscribers associated to a given BS are located at a precise high-resolution position*. Based on such information, one can distribute the metadata over the territory in a dependable manner. An example of probabilistic spatial diffusion for a real-world BS is in Figure 1a.

Unfortunately, in practice, probabilistic spatial diffusion information is very rarely accessible by researchers. In the vast majority of the literature, the positions of BSs are the only input available for geo-referencing the network metadata. In such a situation, a very popular strategy for spatial mapping are *Voronoi diagrams* [10]. These are tessellations of the target region that can be computed from the BS deployment and associate each spatial point to the closest BS. The mobile network metadata associated to one BS can then be uniformly spread over the surface of the corresponding Voronoi cell. Therefore, using a Voronoi representation for spatial mapping corresponds to making two assumptions: (i) each UE attaches to the geographically closest BS; and (ii) UEs are perfectly and evenly distributed over the target geographical area. An example of Voronoi cell is shown in Figure 1b.

As apparent by juxtaposing Figure 1a and Figure 1b, a Voronoi representation has clear limitations for spatial mapping. In this paper, we first quantify such limitations, and then present an improved model of the probabilistic spatial diffusion of individual BSs. Importantly, a hard requirement for our model is that it shall only rely on *input derived from the BS deployment*, exactly like legacy Voronoi diagrams: this makes the tool readily usable by the interdisciplinary community analysing mobile network metadata[1], improving the quality of research and results. We acknowledge that better spatial mappings may be enabled by models fed with data on, *e.g.*, antenna azimuth, BS hardware, geographical topography of the area, or detailed radio signal propagation environment; yet, such information is again typically hard to acquire, making the model substantially less interesting for researchers. Overall, achieving our goal entails four contributions.

- We shed light on the severe limitations of Voronoi cells for spatial mapping of mobile network metadata. To this end, we take advantage of a large-scale realistic dataset

[1]https://github.com/nds-group/voronoiBoost



(a) Ground-truth (b) Voronoi (c) VoronoiBoost

Fig. 1: (a) Probabilistic spatial diffusion from our reference dataset for one representative BS. (b) Legacy Voronoi cell for the same BS. (c) Probabilistic spatial diffusion output by our proposed model, VoronoiBoost, for the same BS. Darker colors denote a higher probability UEs localization.

of the probabilistic spatial diffusion of thousands of 4G outdoor BSs in multiple urban areas, and perform a quantitative analysis of the quality of Voronoi cells as approximations of such data, highlighting their limitations.

- We propose VoronoiBoost, a data-driven model based on Gradient Boosting that produces multiple scalings of the original Voronoi cells to match the actual BS diffusion. Our model meets the requirement of using only BS location input, and produces a probabilistic representation that is much more informative than that of Voronoi cells.
- We demonstrate the advantages of VoronoiBoost via extensive tests with real-world BS deployments and ground-truth spatial diffusion data, showing that our model can improve the mean quality of the mapping by 44% over standard Voronoi cells. An example of the probabilistic output of VoronoiBoost is in Figure 1c.
- We showcase the utility of our model for practical tasks with an Edge network planning problem, and use VoronoiBoost-generated spatial mapping to inform an Edge facility positioning algorithm. Here, VoronoiBoost drives a deployment that is up to 28% more accurate than that obtained with a legacy Voronoi representation.

II. Related work

Measuring the real-world spatial diffusion of BSs is a task that requires UE-side information; yet, experimental mappings based on UE data collection do not scale to the number of BSs composing the radio access network of each operator, or to the surface they service [11]. Techniques to build complete maps of network performance indicators from limited UE measurements may alleviate the effort [12]–[14]. However, these approaches are not designed for spatial diffusion, rather for signal quality metrics; as detailed in Section III, significant post-processing is required to translate the latter into the former. More importantly, signal map completion techniques still require performing a large number of client-side field tests encompassing the same area where the mobile traffic metadata was collected, which would not be feasible in large-scale multi-city scenarios like those we target in our experiments.

A less expensive approach to estimate the diffusion of a cellular BS is using a propagation model to compute the strength of the received signal at each point in space, and then employing such information to estimate the probability that UEs are located at each position in space: this approach represents in fact a subset of the operations carried out to generate the ground-truth data we use in our study, as later elaborated in Section III. The literature on modelling RF signal propagation is in fact very vast. Limiting ourselves to models that have been used in the context of spatial mapping of mobile network metadata, representative examples are the classical logarithmic path loss model [15], or the Okumura-Hata model [16]. These models require knowledge of, *e.g.*, the height, azimuth, radiated power, operating frequency, or tilt of each BS, possibly complemented with topographical information; all these details are generally not available publicly. Similar considerations apply to complex tools for RF signal propagation estimation, *e.g.*, based on ray tracing [17], which need a thorough configuration of the network parameters and detailed information of the propagation environment. Unlike the approaches above, our aim is to directly generate a fair approximation of BS diffusion for spatial mapping, using only minimum information on the network infrastructure.

In the usual case where the available data about the radio access network is limited to BS locations, Voronoi diagrams are the de-facto default option for representing spatial diffusion. As indicated in Section I, many studies in and beyond networking employ this strategy for its simplicity, and very recent examples also abound [18]–[21]. Yet, Voronoi cells suffer from obvious limitations, which we will later expound in Section IV. Minor variations of the approach, which shift the Voronoi anchors along the BS azimuth or use side information to improve the anchor positioning [22]–[24], suffer from the same problems of the original tessellation. A more significant upgrade are complex, multiplicatively weighted Voronoi tessellations [25], which however rely on sectoring, transmission power, or path loss information to weight distance measures; again, such data is difficult to access in practice, and the resulting mapping is still uniform over space.

Bayesian inference has been proposed as a strategy to partially overcome the limits of a Voronoi tessellation [26]. By generating a statistical representation of the location of a UE given its distance from the BS of attachment, the approach allows computing an approximate probabilistic spatial diffusion of the BS. Yet, this needs fine-tuning with measurement data collected in the target network, about the signal-to-noise and interference ratio (SINR) thresholds, interference from non-neighboring base stations, or round trip time (RTT), which are typically not known. A simplified version of this method, which only requires BS locations, has been also proposed [27]. However, this more essential model results in a Voronoi cell smoothing controlled by a power attenuation parameter, whose proper setting again requires field measurements.

Overall, our work fills a gap in the literature, as it sets forth an approximate model of probabilistic diffusion that only requires BS location data, and can help researchers to sensibly enhance the spatial rendering of mobile network metadata. In addition, this is the first study to validate a diffusion model against dependable large-scale multi-city ground-truth data.

III. Ground-truth diffusion data

Our study builds upon a large-scale dataset of *probability of association* to $5,773$ 4G outdoor BSs deployed by a major network operator in 14 urban and suburban regions of France, each comprising a geographical surface between 576 and $2,500$ km^2. For each BS i, the dataset describes the probability $P(i|\ell)$ that a UE at location ℓ is associated[2] to i; therefore, $\sum_i P(i|\ell) = 1$. Specifically, $P(i|\ell)$ is reported over an area of $3,600$ km^2 centered at the location of BS i, and tessellated into regular grid cells ℓ measuring 100×100 m^2 each.

The probabilities of association have been computed internally by the network operator to support its valued-added commercial services. More precisely, a proprietary software that combines models of electromagnetic fields propagation, technical network parameters (including the technology, frequency, power, height of tower, azimuth and tilt of each antenna deployed at the BS site) and information about the surrounding terrain shape was first used to accurately estimate the received signal strength from a BS i to each location ℓ. Then, custom priority tables based on experimental assessments were used to determine $P(i|\ell)$ from the technology, frequency and signal strength of all BS i providing coverage to a same location ℓ. Finally, extensive field tests were realized to validate and fine-tune the resulting probabilities.

An important remark is that the $P(i|\ell)$ values above do not directly match our definition of probabilistic spatial diffusion. Indeed, the probability that a UE associated to BS i is located at ℓ is $P(\ell|i)$, and $\sum_\ell P(\ell|i) = 1$. Translating the probabilities in the dataset provided by the operator into those we need is straightforward using Bayes' theorem:

$$P(\ell|i) = \frac{P(i|\ell)P(\ell)}{P(i)} = \frac{P(i|\ell)P(\ell)}{\sum_{\ell'} P(i|\ell')P(\ell')}. \tag{1}$$

Without any prior knowledge about the likelihood that a UE is at a specific location ℓ, the safest and common assumption is that $P(\ell)$ follows a uniform distribution [15]. Hence, $P(\ell)$ is a constant across all locations, and the computation of $P(\ell|i)$ is reduced to a normalization of the $P(i|\ell)$. Formally,

$$P(\ell|i) = \frac{P(i|\ell)}{\sum_{\ell'} P(i|\ell')}. \tag{2}$$

Figure 1a illustrates an example of probabilistic spatial diffusion in (2) for one specific BS in our reference dataset.

IV. Problem and concept

We leverage the data presented in Section III to perform a first-of-its-kind quantitative analysis of the quality of Voronoi tessellation as an approximation for cellular BS diffusion. The results, in Section IV-A, highlight evident limitations of such an approach. We then lay the foundations to our model in Section IV-B, showing how a simple scaling of the Voronoi cells can substantially enhance the representation of diffusion.

[2]Note that i is a BS, hence $P(i|\ell)$ refers to the total probability that UEs at ℓ associate to any of the antennas co-located at the BS site.

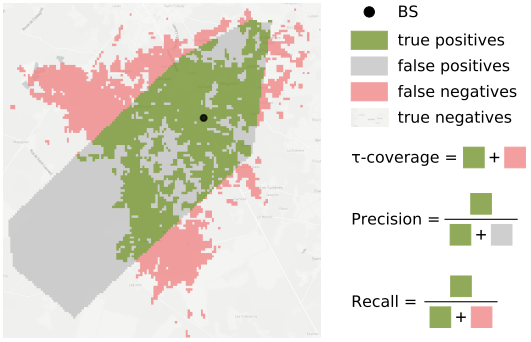


Fig. 2: Example of τ-diffusion of one BS, for $\tau = 0.75$, superposed to the corresponding Voronoi cell. Colors highlight true/false positives/negatives in the locations ℓ, used to compute precision, recall, and F-score.

A. A quantitative analysis of Voronoi accuracy

We start by defining a suitable metric that quantifies the representativeness of Voronoi cells as a proxy for the actual diffusion of individual BSs. Let us consider a given BS i, with a probabilistic spatial diffusion $P(\ell|i)$ over locations ℓ. We create a set of diffusion maps for specific probability thresholds τ, by selecting the smallest set $\mathcal{T}(i)$ of locations such that $\sum_{\ell\in\mathcal{T}(i)} P(\ell|i) \geq \tau$. We name $\mathcal{T}(i)$ the *τ-diffusion* of BS i: this is the most compact spatial area where a fraction τ of the total UE associations with BS i is generated. A visual example of τ-diffusion for one BS and $\tau = 0.75$ is in Figure 2.

We then compute a Voronoi tessellation of space based on the latitude and longitude coordinates of all BSs in the target geographical area. Let $\mathcal{V}(i)$ be the set of locations ℓ that are (at least partially) contained in the Voronoi cell of BS i. Also, $v_\ell(i) \in (0, 1]$ is the fraction of the surface of location $\ell \in \mathcal{V}(i)$ intersecting with the Voronoi cell of BS i.

For the generic BS i, we define a location ℓ as: (i) true positive, if $\ell \in \mathcal{T}(i) \cap \mathcal{V}(i)$; (ii) true negative, if $\ell \notin \mathcal{T}(i)$ and $\ell \notin \mathcal{V}(i)$; (iii) false positive, if $\ell \notin \mathcal{T}(i)$ but $\ell \in \mathcal{V}(i)$; and, (iv) false negative, if $\ell \in \mathcal{T}(i)$ but $\ell \notin \mathcal{V}(i)$. Based on the above, we can quantify the quality of the Voronoi cell in rendering the τ-diffusion in terms of precision and recall.

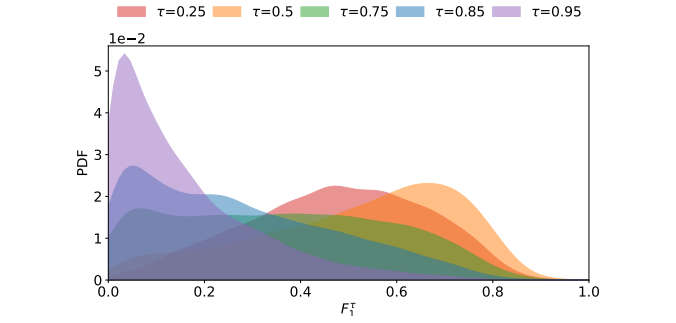
- *Precision* is the weighted fraction of locations in the Voronoi cell that actually belong to the BS τ-diffusion:

$$\mathsf{P}^\tau(i) = \sum_{\ell\in\mathcal{T}(i)\cap\mathcal{V}(i)} v_\ell(i) \Big/ \sum_{\ell\in\mathcal{V}(i)} v_\ell(i)\ . \tag{3}$$

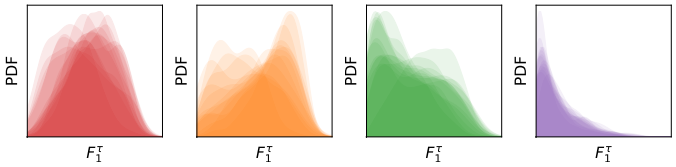
- *Recall* corresponds to the weighted fraction of locations in the τ-diffusion of the BS that are within the Voronoi cell:

$$\mathsf{R}^\tau(i) = \sum_{\ell\in\mathcal{T}(i)\cap\mathcal{V}(i)} v_\ell(i) \Big/ |\mathcal{T}(i)|, \tag{4}$$

where $|\cdot|$ denotes the cardinality of the argument set. Note that (3) and (4) imply that only a portion $v_\ell(i)$ of the location contributes to the definitions true and false positives. An illustrative example of these measures is in Figure 2 for one BS and $\tau = 0.75$. In order to summarize the performance in



(a) Overall



(b) $\tau = 0.25$ (c) $\tau = 0.50$ (d) $\tau = 0.75$ (e) $\tau = 0.95$

Fig. 3: PDFs of the F-score $\mathrm{F}_1^\tau(i)$ obtained with Voronoi cells. (a) PDFs for different diffusion probabilities τ, computed over all BSs i. (b)–(e) Each plot breaks down results for a given τ value across the 14 geographical areas in our reference dataset.

both metrics, the *balanced F-score* $\mathrm{F}_1^\tau(i)$ can be computed as the harmonic mean of precision and recall, *i.e.*,

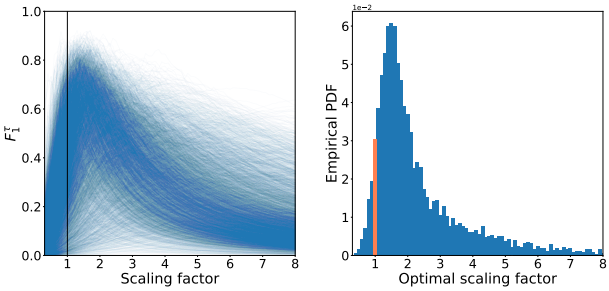
$$\mathrm{F}_1^\tau(i) = 2 \cdot \frac{\mathsf{P}^\tau(i) \cdot \mathsf{R}^\tau(i)}{\mathsf{P}^\tau(i) + \mathsf{R}^\tau(i)}. \tag{5}$$

The F-score in (5) ranges from 0, if the intersection of the Voronoi cell and τ-diffusion is empty and the precision or recall is zero, to 1, if the Voronoi cell and τ-diffusion perfectly match and both measures take value one.

Computing the F-score $\mathrm{F}_1^\tau(i)$ for different values of τ returns the Probability Density Functions (PDFs) over all BSs depicted in Figure 3a. The PDFs show a highly heterogeneous quality of the diffusion rendering. This diversity is observed both within and across τ-diffusions. On the one hand, F-scores for a same value of the diffusion probability τ vary substantially depending on the considered BS, as all PDFs include $\mathrm{F}_1^\tau(i)$ samples that range from 0 to 0.95. On the other hand, the statistics of the PDFs for different values of τ are unalike, with, *e.g.*, sensibly diverse mean and dispersion.

In all cases, the typical $\mathrm{F}_1^\tau(i)$ values reveal that Voronoi cells have substantial limitations in mimicking the actual diffusion of BSs. A regular Voronoi diagram tends to approximate more closely the spatial areas where 50% of UE associations are generated (*i.e.*, $\tau = 0.50$): yet, even in this best-case scenario, thew median F-score is just 0.56. And, when changing the value of τ, such values drop substantially: the very low median values of $\mathrm{F}_1^\tau(i)$ at 0.35, 0.23 and 0.11 for τ equal to 0.75, 0.85 and 0.95, respectively, highlight how a Voronoi tessellation returns a largely unreliable spatial mapping of BS metadata. Figures 3b–3e confirm that the conclusion above holds across the 14 geographical areas present in our ground-truth dataset, with minor variance due to the specificity of each topography.

Our quantitative analysis exposes problems that sum to inherent qualitative issues of the Voronoi diagrams: (i) the uniform spatial mapping over the Voronoi cell polygon cannot capture the inherent probabilistic nature of BS diffusion; and, (ii) the Voronoi tessellation leads by construction to disjoint cells, which cannot render the natural overlap among areas covered by nearby BSs. Together, these issues can easily bias the outcome of analyses that rely on a Voronoi tessellation to map network traffic metadata onto the geographical space.



(a) (b)

Fig. 4: (a) F-score $\mathrm{F}_1^\tau(i)$ of the scaled Voronoi cells, versus the employed scaling factor. Each curve portrays the result for one BS. (b) Distribution of the optimal scaling factor that maximizes the $\mathrm{F}_1^\tau(i)$ for each BS i. Results are for $\tau = 0.75$.

B. Scaling as an approach to improve the Voronoi model

The limitations outlined above call for improved representations of the BS diffusion, which can then enhance the dependability of spatial analyses based on mobile network metadata. As anticipated in Section I, we are interested in a model that is as practical as the original Voronoi tessellation; namely, it shall solely build on the same information that underpins the Voronoi representation, *i.e.*, the latitude and longitude coordinates of BSs. Abiding by this requirement, we propose *scaling Voronoi cells* as a simple strategy that allows overcoming at once all limitations listed in Section IV-A. The basic concept is that by up- or down-scaling the Voronoi polygon around its barycenter one can control the precision and recall of the Voronoi cell for a specific τ-diffusion of the BS. At the same time, merging scaled cells for different τ results in a probabilistic mapping; also, up-scaling automatically leads to overlapping BS diffusion regions. As an illustrative example, the multiple polygons composing Figure 1c are the result of different scalings of the original Voronoi cell of the BS.

In order to validate the concept, we perform an exhaustive search over a full range of scaling factors, and observe their impact on the F-score in (5), for all combinations of probability threshold τ and BS i. The result is depicted in Figure 4a, for $\tau = 0.75$ and all BSs in our dataset; all values of τ yield equivalent behaviors, and are not shown for the sake of brevity. We observe that $\mathrm{F}_1^\tau(i)$ varies dramatically under different scaling factors for all BSs, according to a fairly consistent bell-shaped convex function. However, the curve dynamics are heterogeneous across BSs. We are especially interested in the *optimal* scaling factor that achieves the highest F-score, *i.e.*, the peak of the bell-shaped curve, for each BS. As shown by the PDF in Figure 4b, such values are varied,

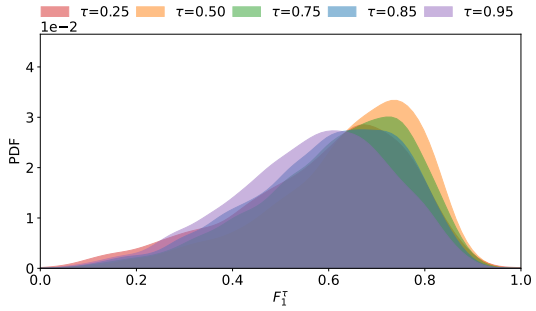


Fig. 5: PDFs of the F-score $\mathrm{F}_1^\tau(i)$ upon optimal scaling of the Voronoi cell for each BS. Each PDF refers to one value of the total diffusion probability τ, and is computed over all BSs.

and range from a down-scaling of 0.4 to an up-scaling of 15.9, depending on the considered BS. In fact, less than 5% of BSs have optimal $\mathrm{F}_1^\tau(i)$ at a scaling factor of 1 for $\tau = 0.75$, as shown in Figure 4b; that is, the original Voronoi cell is not the best scaling choice for this τ-diffusion in over 95% of cases.

Figure 5 shows the PDF of the F-scores $\mathrm{F}_1^\tau(i)$ attained by the optimally scaled Voronoi cells of all BSs, for different τ's. By comparing the distributions with those for the original Voronoi cells, in Figure 3a, the potential gain of a simple scaling strategy is evident. The PDFs are now substantially shifted towards higher F-scores, with a much smaller fraction of BS suffering from low-quality representations. The result is now very consistent for all probability thresholds τ, with median $\mathrm{F}_1^\tau(i)$ values steadily in the 0.55–0.67 range.

V. Practical Scaling of Voronoi Cells

We develop a practical data-driven model that leverages on the potential of Voronoi cell scaling for diffusion rendering, as illustrated by our previous results. As set forth in Sections I and II, we require that our model solely builds on BS locations; the reason is that such information is typically the only available to researchers working with mobile network metadata, hence the choice makes the model readily usable in practice.

The modeling task is not trivial, since, as shown in Figure 4, the optimal scaling factors can be very diverse, which calls for BS-specific decisions. We tackle the problem as follows. First, we extract a comprehensive set of geometrical features from the simple locations of BSs. Second, we test a number of supervised machine learning methods, and compare their capability to infer the per-BS scaling factor from the geometrical features. Third, we fine-tune the selected method to derive our final model. These steps are detailed in the following.

A. Feature engineering

While the location information typically available are BS latitude and longitude coordinates (or transformations thereof), much richer information can be extrapolated from such data by considering the whole geometry of the BS deployment. For instance, new features can be obtained by relating the position of each BS to those of the surrounding BSs (*e.g.*, distance to the closest BS); others can leverage more elaborate representations, such as the Delaunay triangulation graph built on BS locations (*e.g.*, number of one-hop neighbors), or the Voronoi diagram itself (*e.g.*, surface of the Voronoi cell).

Taking this approach, we extract a comprehensive set of 134 features, whose complete list is provided in Table I. An important remark is that we also consider the target diffusion probability as a feature, so that τ is part of the learning process, and the resulting model works for any τ-diffusion. Table I also reports the Pearson's correlation coefficient ρ of all features to the optimal scaling factor. The value is consistently low, indicating that the features are not good predictors if taken individually. This motivates the use of machine learning methods that can combine such features to improve the result.

B. Learning the optimal scaling factor

The availability of ground-truth data about BS diffusion, hence of labels about the optimal scaling factor for each BS, makes *supervised learning* an obvious choice for the task. As the scaling factor is a continuous variable, ours is basically a regression problem. We thus experiment with a range of classical[3] linear and non-linear regression methods, as follows.

Linear Regression (LR). We use an ordinary least squares linear regression, which fits a linear model to minimize the residual sum of squares between the optimal and predicted scaling factors. We test both with and without intercept term.

LASSO Regression. This is a LR enhanced by a variable selection and an L_1 regularization term. The model is tested with different regularization terms from 0.1 up to 0.7, and both with and without intercept term.

Random Forests (RF). This model creates a forest of decision trees, each using a random subset of the features. We tested with $\{500, 1000, 1500, 2000, 2500\}$ estimators, maximum tree depth of 5 and 10, *MSE* as the loss function, and each tree built with a random subset of $\sqrt{135}$ features.

Support Vector Regression. This model learns hyperplanes in higher dimensions that separate efficiently the target space. We tested different regularization parameters in $\{0.1, 1, 10, 100\}$ and kernels, where radial basic is the one with higher performance.

K-nearest Neighbors (KNN). This model is based on the average value of the k closest neighbords weighted by the inverse of the distance respect to the input vector. We tested different number of neighbors in $\{5, 25, 50, 100\}$.

Gradient Boosting Regressor (GBR) This method starts from an initial set of decision trees, and adds new trees with the objective of minimizing the loss function in the training set. We test this model with the same parameters as the RF.

In order to train the methods above, we first isolate a training dataset, consisting of 80% of the BSs in our diffusion data, which are randomly selected and associated with their ground-truth optimal scaling factors. The remaining 20% of the data will be exclusively used to assess the final model performance with previously unseen BSs, later in Section VI-A.

[3]As we use well-known learning tools, we do not discuss extensively their operation, which can be found in any textbook. Instead, we take care of detailing their parameterization, so that our methodology is reproducible.

Notation	Description	ρ	Selected
S_v	Voronoi cell surface area of BS i	-0.15	
P_v	Voronoi cell perimeter of BS i	-0.19	
Δ_v	Voronoi cell diameter, *i.e.*, the largest distance between two cell vertices, of BS i	-0.19	
W_v	Voronoi cell width, *i.e.*, the height of the smallest rectangle enclosing the Voronoi cell, of BS i	-0.19	
$\bar{\mathcal{E}}_v$	Voronoi cell mean edge length of BS i	-0.17^*	
$\mathcal{V}_v$	Distance of BS i to the closest vertices of its Voronoi cell, $\forall v \in \{1,2,3\}$	-0.18^*	
$\bar{\mathcal{V}}_v$	Mean distance of BS i to all vertices of its Voronoi cell	-0.19^*	
$\mathcal{V}_v^+$	Maximum distance of BS i to all vertices of its Voronoi cell	-0.19	✓
Θ_v	Distance of BS i to the barycenter of its Voronoi cell	-0.15	
$\mathcal{D}_{v,k}$	Distance between BS i and its k^{th} closest BS, $\forall k \in \{1,\dots,7\}$	$+0.18^*$	
$\bar{\mathcal{D}}_{i,k}$	Average distance between BS i and all of its first k closest neighbors, $\forall k \in \{1,\dots,12\}$	-0.19^*	
$\mathcal{F}_v$	Flag taking value 1 if the Voronoi cell of BS i touches the border of the geographical area	-0.02	
$\bar{\Theta}_k$	Average distance of all BSs in k-hop to their Voronoi cell barycenter $\forall k \in \{1,\dots 10\}$	-0.04^*	
$\bar{S}_k$	Average surface area of the Voronoi cells in k-hop neighbors of BS i $\forall k \in \{1,\dots,10\}$	-0.05^*	
$\bar{\mathcal{D}}_k$	Average distance between BS i and k-hop neighbors $\forall k \in \{1,\dots,\dots,10\}$	-0.18^*	
$\underline{\mathcal{D}_k}$	Minimum distance between BS i and k-hop neighbors $\forall k \in \{1,\dots,10\}$	-0.18^*	8
$\bar{\mathcal{I}}_k$	Average distance among adjacent BSs in k-hop neighbors of BS i $\forall k \in \{1,\dots,10\}$	-0.18^*	
λ_i	Clustering coefficient of BS i in the Delaunay triangulation	-0.17	
$\mathcal{C}_i$	Constraint coefficient of BS i in the Delaunay triangulation	$+0.06$	
Ξ_i	Effective size coefficient of BS i in the Delaunay triangulation	$+0.10$	
$Dg_{i,k}$	Number of nodes in the Delaunay triangulation in the k-hop of BS i $\forall k \in \{1,\dots,10\}$	-0.26^*	
$\omega_{i,k}$	Width of the convex hull, that contain the BS in k-hop $\forall k \in \{1,\dots,10\}$	-0.19^*	

(a) Basic features

Notation	ρ	Selected
$S_v \cdot P_v$	-0.11	
$S_v \cdot \Delta_v$	-0.11	
$P_v \cdot W_v$	-0.15	
$\Delta_v \cdot W_v$	-0.15	
S_v / P_v	-0.19	
S_v / Δ_v	-0.19	
P_v / W_v	-0.01	
Δ_v / W_v	-0.01	✓
$\bar{\Theta}_k / \bar{\Theta}_1$	$+0.32$	10
$\bar{\mathcal{D}}_k / \bar{\mathcal{D}}_1$	$+0.35$	$6, 8, 10$
$\bar{\mathcal{I}}_k / \bar{\mathcal{I}}_1$	$+0.35$	$6, 8, 10$
$\bar{S}_k / \bar{S}_1$	$+0.31$	$4, 6, 8$
ω_k / ω_1	$+0.31$	8

(b) Composite features

Notation	ρ	Selected
τ	$+0.48$	✓

(c) Diffusion feature

TABLE I: Complete list of geometrical features used for Voronoi cell scaling factor inference, and their Pearson's correlation ρ with the optimal scaling factor. The asterisk * indicates that the row combines multiple features, hence ρ is the maximum correlation; the measured standard deviation in these cases is always lower than 0.01. The last column indicates if the feature is used by the proposed VoronoiBoost model, and specifies the exact indices k when applicable. (a) Basic features obtained from the BS deployment. (b) Composite features obtained via operations on the basic features. (c) Target τ-diffusion feature.

Model	Features	R^2	$\text{F}_1^\tau(i)$
Voronoi			0.34 ± 0.235
Optimal			0.61 ± 0.172
LASSO Regression	135	0.31 ± 0.002	0.42 ± 0.176
Linear Regression	135	0.42 ± 0.011	0.41 ± 0.197
Random Forest	135	0.55 ± 0.010	0.43 ± 0.192
Support Vector Regression	135	0.56 ± 0.011	0.52 ± 0.188
KNN	135	0.60 ± 0.007	0.50 ± 0.183
GBR	135	0.85 ± 0.005	0.55 ± 0.183
VoronoiBoost	15	0.84 ± 0.004	0.55 ± 0.182

TABLE II: Performance summary of regression methods.

The models are then trained via *k-fold cross-validation*, which is known to reduce the impact of possibles biases in the data [28]. We therefore split the training data into k random subsets, train each of the regression methods with $k-1$ subsets, and measure the model performance with the remaining subset. The procedure is repeated k times, every time using a different subset for validation. The whole cross-validation can be repeated for multiple iterations by selecting diverse random partitionings of the training data into k subsets.

Precisely, we set an hyper-parameters space for each model and for each combination of hyper-parameter we run 5-fold cross-validation, and compute the R^2 between the optimal and predicted scaling factor across all BSs. This allows to perform hyper-parameter tuning and model selection at the same time. Table II summarizes the result, showing that GBR is the best option, attaining a good $R^2 = 0.85$ in the prediction of optimal scaling. In addition, the table also reports the median and standard deviation of the F-score $\text{F}_1^\tau(i)$ in (5), computed using the model-scaled Voronoi cells and the ground-truth τ-diffusion data. The higher R^2 reflects into a superior F-score, confirming that GBR gives the best approximation of the actual diffusion, with an overall median $\text{F}_1^\tau(i)$ around 0.55.

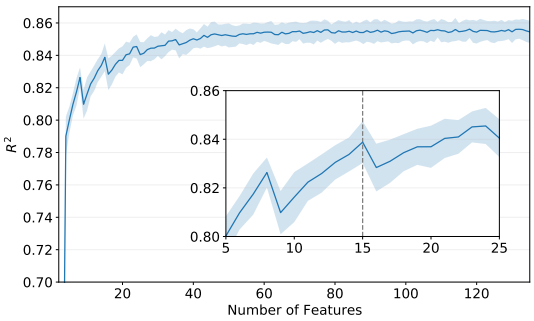


Fig. 6: Mean R^2 and deviation between the GBR-predicted and optimal scaling factors, with an increasing set of features.

C. Voronoi scaling with Gradient Boosting Regressor

Based on the results above, we select GBR as the engine for our model. We then fine-tune the GBR method by exploring the effect of a reduced input feature set on its performance. Indeed, limiting the number of features lowers the complexity of the model and, more importantly, eases its interpretability.

To this end, we train a GBR instance with the full 135-features set, and extract the feature importance list. Then, 135 different models are trained with a decreasing number of features, and record their associated R^2 values. This procedure is repeated 20 times, allowing to compute the mean R^2 and the standard deviation given a number of features and the average feature importance. Figure 6 depicts the evolution of the mean R^2 between the optimal and predicted scaling factors, versus the size of the feature set. The quality of the model inherently decreases as the number of features is reduced. However, there is just a 0.01 difference in R^2 between training the method with a 15-feature set or the full list. Reducing the number of features below 15 causes instead the accuracy to drop rapidly.

The retained 15 features are marked in the last column of Table I. A closer inspection lets us speculate on GBR choices.

- The target probability threshold is an obvious inclusion, as it is critical to disambiguate across different τ-diffusion targets, and forces larger factors for higher τ's.
- Δ_v/W_v and $\mathcal{V}_v^+$ provide information about the geometry of the target Voronoi cell, indicating that the scaling factor must be adjusted to the *shape* of the cell.
- The other selected features $\bar{\Theta}_{10}/\bar{\Theta}_1$, ω_8/ω_1, $\bar{S}_k/\bar{S}_1$ for $k \in \{4, 6, 8\}$, and both $\bar{\mathcal{D}}_k/\bar{\mathcal{D}}_1$ and $\bar{\mathcal{I}}_k/\bar{\mathcal{I}}_1$ for $k \in \{6, 8, 10\}$ are proxies for the density and shape of the neighbours of the target BS at different hops; in line with real-world radio planning practices, features denoting denser deployments curb the scaling, *i.e.*, packed BSs cover smaller areas.

Interestingly, the selected features underscore that local properties of the Voronoi cell alone are not sufficient to describe the BS diffusion, and BS deployment data beyond the first order neighborhood is instrumental to a correct modelling.

Ultimately, our proposed model is a Gradient Boosting Regressor method fed with the reduced set of 15 features mentioned above. We name it VoronoiBoost, for Voronoi scaling with Gradient Boosting Regressor, and report its overall performance in the last row of Table II. Comparing F-scores, VoronoiBoost yields a gain of 44% over the original Voronoi cells, and a less than 10% gap to optimal scaling.

D. Probabilistic spatial diffusion with VoronoiBoost

VoronoiBoost can be finally used to produce complete approximations of probabilistic spatial diffusion. This is done by: (i) superposing all scaled Voronoi polygons for different values of τ; (ii) computing the total probability in each scaled Voronoi contour (*e.g.*, a probability 0.10 would be associated to the contour given by the difference between the scaled Voronoi polygons for $\tau = 0.85$ and $\tau = 0.75$); and, (iii) computing $P(\ell|i)$ by uniformly distributing the total probability above across locations in the contour. An example of the outcome is in the right-hand-side plot of Figure 1.

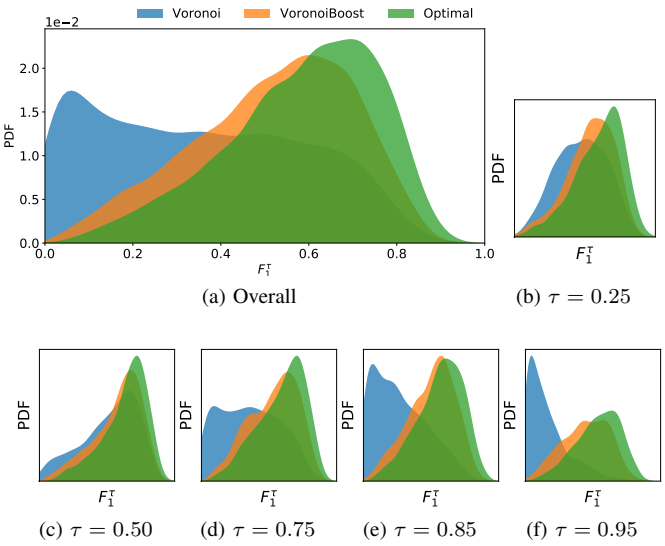
VI. VoronoiBoost Performance Evaluation

We explore VoronoiBoost performance in details, breaking it down along a variety of dimensions. Next, we assess the tool in terms of (i) operation with previously unseen data, (ii) support for mobile traffic maps, and (iii) application in a practical use case of Edge network planning.

A. Overall test performance

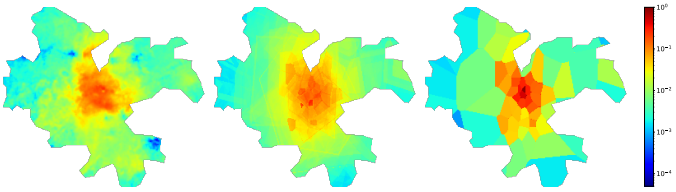
We run the trained VoronoiBoost on the 20% test portion of the dataset that was not used in prior experiments, as described in Section V-B. This ensures that VoronoiBoost produces scaling factors for BSs it did not see before, and that the performance is representative of practical cases where the model is used on a completely new BS deployment environment. Results are summarized in Figure 7.

We first look at the distributions of the F-score obtained by VoronoiBoost on all BSs i in the test data, and aggregated over all τ-diffusion cases, in Figure 7a. To ease the comparison, the distribution is superposed to those yielded by the original Voronoi cells, and by the optimal scaling identified via exhaustive search. The improvement of VoronoiBoost over a legacy Voronoi tessellation is evident; also, the result is relatively close to that granted by an optimal scaling. The actual numbers are indistinguishable from those observed for the training data, as $\mathrm{F}_1^\tau(i) = 0.55 \pm 0.182$ also with the test data. Interestingly, the gain of VoronoiBoost is not homogeneous across values of τ. This is evident in Figures 7b to 7f, with PDFs for different τ-diffusion objectives. More precisely, when the original Voronoi diagram performs well, *e.g.*, for $\tau = 0.50$, our model only yields a mild improvement; however, it dramatically enhances the quality of the rendering when regular Voronoi cells perform poorly under other τ's.



(a) Overall (b) $\tau = 0.25$ (c) $\tau = 0.50$ (d) $\tau = 0.75$ (e) $\tau = 0.85$ (f) $\tau = 0.95$

Fig. 7: PDFs of F-score $\mathrm{F}_1^\tau(i)$ with (i) original Voronoi cells, (ii) VoronoiBoost, and (iii) optimal scaling. (a) Total PDF over all τ's. (b)-(f) Breakdown over individual values of τ.



(a) Ground truth (b) VoronoiBoost (c) Voronoi

Fig. 8: Maps of mobile data traffic computed via a spatial distribution of the traffic volume recorded at the 55 test BSs in the metropolitan area of Le Mans, France. Spatial mappings are performed based on diffusion from (a) data, (b) VoronoiBoost, and (c) Voronoi cells.

The results prove that VoronoiBoost can homogenize the quality of the diffusion across values of τ, aligning it to the accuracy of best-case scenarios with legacy Voronoi.

Model	MAE	ρ	KL	SSIM
Voronoi	6.83×10^{-3}	0.809	0.073	0.883
VoronoiBoost	4.01×10^{-3}	0.884	0.040	0.918

TABLE III: Similarity of traffic maps obtained with Voronoi cells and VoronoiBoost against the ground-truth map, measured with MAE, Pearson's correlation ρ, Kullback–Leibler (KL) divergence, and SSIM ($\in [0, 1]$, higher is better).

B. Mobile traffic maps

As a second experiment, we assess the capability of VoronoiBoost to generate realistic spatial maps of mobile traffic consumption. For this test, we adopt a different training strategy, based on a *leave-one-out* approach: we use ground-truth data from all BSs in 13 geographical regions of our dataset to train VoronoiBoost, and then employ the trained model to predict the probabilistic spatial diffusion of all BSs in the remaining region. Interestingly, this modus operandi lets us prove the effectiveness of VoronoiBoost also in rendering diffusion in whole new areas not seen during training.

Due to space limitations, we focus on BSs servicing the metropolitan area of Le Mans, France. We compute diffusion probabilities $P(\ell|i)$ for the 55 BSs in the target area using three approaches: (i) we take the $P(\ell|i)$ extrapolated from the diffusion data provided by the operator, as described in Section III, which we treat as ground truth; (ii) we use VoronoiBoost to derive $P(\ell|i)$ as described in Section V-D; (iii) we perform a legacy Voronoi tessellation, and then uniformly distribute[4] the diffusion probability of each BS i over its Voronoi cell. Note that the latter approach is the current common practice adopted in the literature. We then use the three definitions of $P(\ell|i)$ to realize a spatial mapping of the mobile data traffic load recorded by the network operator at each BS in the target area[5]. To this end, we allocate to each location ℓ a fraction $P(\ell|i)$ of the total traffic volume $t(i)$ of each BS i; therefore, the aggregate traffic located at ℓ is the sum of the contributes of all BSs, or $t_\ell = \sum_i P(\ell|i) \cdot t(i)$.

The resulting three spatial traffic maps[6] are portrayed in Figure 8. A visual comparison already highlights the higher quality of the traffic mapping entailed by VoronoiBoost, which is closer to that granted by our ground-truth diffusion data. A quantitative assessment with classical vector and computer vision metrics, in Table III, supports the claim. Our model reduces the MAE by over 40%[7], and yields higher correlation and lower KL divergence. The Structural Similarity Index (SSIM) is closer to the maximum value of 1 with VoronoiBoost, meaning that the structures in the traffic map produced by our model better resemble those in the ground truth.

C. Application use case

We demonstrate the practical advantages of the diffusion representation produced by VoronoiBoost for data-driven analyses via a representative application use case, *i.e.*, Far Edge Site planning in virtualized RANs (vRANs). Specifically, we consider a modern vRAN environment where many cheap Radio Units (RUs) performing minimum radio operations (*e.g.*, FFT, Cyclic Prefix, or P/S) are densely deployed across the territory; more complex functionalities (*e.g.*, coding, modulation, or FEC) are instead delegated to Distributed Units (DUs) located in Far Edge Sites (ESs) with substantial compute resources [29]. In this setting, the ESs deployment must be designed by accounting for the traffic demands at RUs, so as to optimize the utilization of the ES capacity, and reducing their energy footprint [30]. Our aim is studying how the accuracy of diffusion data can affect the quality of the ES planning.

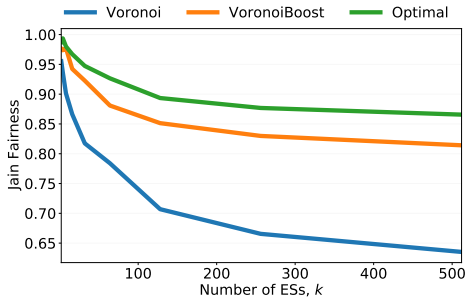


Fig. 9: Jain's fairness index of the traffic served by different ESs, versus their deployment density, for diverse models.

We use the traffic maps derived in Section VI-B to plan for the deployment of k ESs and their association with RUs in the area. We assume a dense and regular deployment of RUs $r \in \mathcal{R}$, each serving one location ℓ: the load of each RU is thus $t_r = t_\ell$. Then, we formulate ES deployment as an optimization problem that aims at balancing the volume of traffic handled by each ES, while minimizing the geographical distance between RUs and their serving ES, so as to reduce fronthaul latency [31]. Formally, we represent the vRAN as a graph $G(\mathcal{R}, \mathcal{E})$, where an edge $e_{r,r'} \in \mathcal{E}$ only exists if RUs r and r' serve adjacent locations. Now, we have to solve

$$\min \sum\nolimits_{e_{r,r'} \in \mathcal{E}} x(e_{r,r'}), \quad \text{s.t.} \tag{6}$$

$$1 - \epsilon \leq \sum_{r \in \mathcal{R}} y(r,s) \cdot t_r \Big/ \sum_{r \in \mathcal{R}} \frac{t_r}{|\mathcal{S}|} \leq 1 + \epsilon, \quad \forall s \in \mathcal{S}, \tag{7}$$

$$\sum\nolimits_{s \in \mathcal{S}} y(r,s) = 1, \quad \forall r \in \mathcal{R}, \tag{8}$$

$$x(e_{r,r'}) \geq y(r',s) - y(r,s), \quad \forall e_{r,r'} \in \mathcal{E}, \ \forall s \in \mathcal{S}, \tag{9}$$

$$x(e_{r,r'}) \geq y(r,s) - y(r',s), \quad \forall e_{r,r'} \in \mathcal{E}, \ \forall s \in \mathcal{S}, \tag{10}$$

where $|S| = k$ is the number of sites, and $x(e_{r,r'})$ and $y(r,s)$ are decision variables equal to one if edge $e_{r,r'}$ is cut by a partition, and if RU r is associated with ES s, respectively.

The problem in (6) can be efficiently solved via the KaFFPa heuristic [32]. The result partitions the RUs r across ESs s, so that the ES traffic load is equivalent by a margin ϵ, as required by (7). Also, RUs served by a same ES are geographically adjacent, which reduces fronthaul latency.

We solve the problem in (6) using values of $t_r = t_\ell$ from the traffic maps in Figures 8b and 8c. We assess the quality of the resulting deployments by measuring their fairness, *i.e.*, whether the load is actually balanced across ESs as expected

[4]Formally, using the notation in Section IV-A, $P(\ell|i) = v_\ell(i) / \sum_\ell v_\ell(i)$.
[5]We employ per-BS traffic loads measured during one random working day.
[6]Traffic is normalized to the maximum BS daily load for confidentiality.
[7]Note that MAE values refer to normalized traffic, and correspond to errors in the order of tens of MB per BS and day, which are worth being reduced.

from the optimization problem. To this end, we compute the Jain's fairness index of the traffic served by each ES s, which is $\sum_{r\in\mathcal{R}} y(r,s)\cdot t_r$. An important remark is that at this stage *we use the actual traffic t_r obtained from ground-truth diffusion data*, shown in Figure 8a, to evaluate all deployments. The rationale is that we want to assess how an ES deployment would perform in the field (which, in our case, coincides with the ground-truth traffic map), even if the deployment was driven by diffusion estimated via VoronoiBoost or Voronoi cells. We repeat the experiment for multiple planning configurations characterized by a different number of ESs k.

Figure 9 shows that relying on a Voronoi diagram to map traffic over space only works for naive settings where the whole area must be divided among a small number of ES, which allows for a large error margin; with larger and more challenging deployments, the Voronoi representation causes the quality of the planning to drop quickly to fairness lower than 0.7. On the contrary, a geographical mapping of traffic based on VoronoiBoost diffusion drives more dependable deployments, whose fairness remains consistently above 0.8, with gains up to 28% over legacy Voronoi cells.

VII. Conclusions

We presented VoronoiBoost, a data-driven model that, despite its simple concept and implementation, can largely improve the quality of traditional BS diffusion representations. We demonstrated the significant advantages of VoronoiBoost via a thorough performance analysis and with a practical use case, where the spatial mapping enabled by the model grants fairer load balancing to an Edge network deployment. The model is primarily intended to support researchers under typical data access conditions, hence it only needs minimal information about BS locations in order to be fully operational.

Acknowledgment

This work has been supported by the research project CANCAN (Content and Context based Adaptation in Mobile Networks), grant no. ANR-18-CE25-0011, funded by the French National Research Agency (ANR). The work of O.M. and M.F. was supported by NetSense, grant no. 2019-T1/TIC-16037 funded by Comunidad de Madrid.

References

[1] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Science*, vol. 4, no. 1, 2015.

[2] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, 2015.

[3] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *IEEE INFOCOM*, 2011.

[4] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in *ACM CoNEXT*, 2017.

[5] A. Janecek, K. A. Hummel, D. Valerio, F. Ricciato, and H. Hlavacs, "Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation," in *ACM UbiComp*, 2012.

[6] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, 2008.

[7] C. Sarraute, J. Brea, J. Burroni, and P. Blanc, "Inference of demographic attributes based on mobile phone usage patterns and social network topology," *Social Network Analysis and Mining*, vol. 5, no. 1, 2015.

[8] M. De Nadai, J. Staiano, R. Larcher, N. Sebe, D. Quercia, and B. Lepri, "The death and life of great italian cities: a mobile phone data perspective," in *ACM WWW*, 2016.

[9] G. Khodabandelou, V. Gauthier, M. Fiore, and M. A. El-Yacoubi, "Estimation of static and dynamic urban populations with mobile network metadata," *IEEE Trans. on Mobile Computing*, vol. 18, no. 9, 2018.

[10] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Comput. Surv.*, vol. 23, sep 1991.

[11] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: A survey of research," *ACM Comput. Surv.*, vol. 47, no. 2, 2014.

[12] J. Robinson, R. Swaminathan, and E. W. Knightly, "Assessment of urban-scale wireless networks with a small number of measurements," in *ACM MobiCom*, 2008.

[13] E. Alimpertis, A. Markopoulou, C. Butts, and K. Psounis, "City-wide signal strength maps: Prediction with random forests," in *ACM WWW*, 2019.

[14] E. Alimpertis, A. Markopoulou, C. T. Butts, E. Bakopoulou, and K. Psounis, "A unified prediction framework for signal maps," *arXiv preprint arXiv:2202.03679*, 2022.

[15] M. Tennekes and Y. A. P. M. Gootzen, "A bayesian approach to location estimation of mobile devices from mobile network operator data," *Journal of Spatial Information Science*, 2021.

[16] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, "Quantifying urban attractiveness from the distribution and density of digital footprints," 2009.

[17] Z. Yun and M. F. Iskander, "Ray tracing for radio propagation modeling: Principles and applications," *IEEE Access*, vol. 3, 2015.

[18] S. Mishra, Z. Smoreda, and M. Fiore, "Second-level digital divide: A longitudinal study of mobile traffic consumption imbalance in france," in *ACM WWW*, 2022.

[19] S. Scepanovic, S. Joglekar, S. Law, and D. Quercia, "Jane jacobs in the sky: Predicting urban vitality with open satellite data," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, 2021.

[20] F. Batista e Silva, S. Freire, M. Schiavina, K. Rosina, M. A. Marín-Herrera, L. Ziemba, M. Craglia, E. Koomen, and C. Lavalle, "Uncovering temporal changes in europe's population density patterns using a data fusion approach," *Nature communications*, vol. 11, no. 1, 2020.

[21] H. Salat, Z. Smoreda, and M. Schläpfer, "A method to estimate population densities and electricity consumption from mobile phone data in developing countries," *PloS one*, vol. 15, no. 6, 2020.

[22] H. Teerayut, *A study on urban mobility and dynamic population estimation by using aggregate mobile phone sources*. PhD thesis, Tokyo University, 2010.

[23] F. de Meersman, G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, and H. I. Reuter, "Assessing the quality of mobile phone data as a source of statistics," in *European Conference on Quality in Official Statistics, Eurostat*, 2016.

[24] E. Graells-Garrido, O. Peredo, and J. García, "Sensing urban patterns with antenna mappings: The case of santiago, chile," *Sensors*, vol. 16, no. 7, 2016.

[25] J. Portela and M. Alencar, "Cellular network as a multiplicatively weighted voronoi diagram," in *IEEE CCNC*, 2006.

[26] H. Zang, F. Baccelli, and J. Bolot, "Bayesian inference for localization in cellular networks," in *IEEE INFOCOM*, 2010.

[27] V. A. Traag, A. Browet, F. Calabrese, and F. Morlot, "Social event detection in massive mobile phone data using probabilistic location inference," in *IEEE SocialCom*.

[28] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[29] 3GPP, "Release description; release 15," Technical report (TR) 21.915, 3rd Generation Partnership Project (3GPP), 06 2019. Version 1.1.0.

[30] R. Singh, C. Hasan, X. Foukas, M. Fiore, M. K. Marina, and Y. Wang, "Energy-efficient orchestration of metro-scale 5g radio access networks," in *IEEE INFOCOM*, 2021.

[31] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How should i slice my network? a multi-service empirical evaluation of resource sharing efficiency," in *ACM MobiCom*, 2018.

[32] C. S. Pan and M. L. Zymbler, "Very large graph partitioning by means of parallel dbms," in *East European Conference on Advances in Databases and Information Systems*, Springer, 2013.