IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2023, accepted 11 September 2023, date of publication 13 September 2023,
date of current version 19 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3315129

METHODS

A Study on Simple Geometries for Modeling User Equipment Geospatial Attachment to Mobile Cells

DANNY QIU[1,2], ALASSANE SAMBA[1], HOSSAM AFIFI[2], AND YVON GOURHANT[1]
[1]Orange Innovation, 22300 Lannion, France
[2]UMR 5157, CNRS, Institut Polytechnique de Paris, Télécom SudParis Saclay, 91120 Palaiseau, France
Corresponding author: Danny Qiu (danny.qiu@orange.com)

ABSTRACT As data-driven solutions will become an important component of next-generation networks, we were faced with the difficulty of developing cross-domain datasets for training machine learning models. In order to understand how external sources of data should be associated with mobile network data at cell level, we have derived a method for splitting simple geometric coverage models for base stations to obtain coverage models for sectorized cells. Then, we developed a method to compare the coverage models with a ground truth of real measurements. We also proposed to use the notions of convex and concave hulls to quantify when a geometry is undersized or oversized. The two types of geometry that were evaluated were the Voronoi polygons and circular shapes, with sector split. All frequency bands considered, the results have shown that Voronoi sector split model combined with upscaling the coverage shapes was the closest to the ground truth, with an average recall of 0.74. Since upscaling Voronoi polygons should be a good practice to improve coverage modelling, we have also proposed an approach using more affordable data (i.e. cell level aggregates) instead of user locations to find the best scaling factor. All frequencies combined, we have observed an average increase of 0.13 points in the recall between the diagram with the default scale and the scale-tuned diagram.

INDEX TERMS Mobile network, cell coverage, geometric modeling, Voronoi, budget link.

I. INTRODUCTION

In the recent years, the ever growing usage of mobile networks in the professional sector and for individual usages have driven the development of new technology standards. The 5G and 6G technologies are designed to integrate data-driven models that make use of the big data collected by mobile network operators. Machine learning (ML) solutions are thoroughly investigated to propose intelligent components integrated to network optimization, anomaly detection, time series forecasting, planning tools or energy efficiency systems [1]. By definition, these algorithms are built to learn from a large dataset of real data. Their strength reside in their ability to extract latent variables embedded in the data, and to draw relationship out of complex phenomenons that may be difficult or time-consuming to notice for an expert.

A problem in particular retained our attention and motivated this study on cell coverage: the crossing of radio network configuration data with exogenous features. Exogenous data, also called "cross-domain" data [2] or "external information" [3] are data coming from sources that are unrelated to telecommunications, but are important to represent phenomenons in a way that mobile data cannot. The existing work combining exogenous and mobile network features is mainly dedicated to predict the network performance at user level or at base station level (see Section II for further details on the applications).

The quality of learnable information embedded in external features highly depend on the way they are processed. For example, what type of features should be kept? How do we cross external data with network infrastructure such that it is geographically relevant? The scope of this paper focuses on the second question, where a reasonable idea is to extract information specific to each base station (BS) or cell so that only objects spatially located in the vicinity of the

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

equipment are kept. To maximize the correlation between mobile traffic activity, human activity and urban fabric, the ideal extract area would correspond to the coverage of the said equipment. In theory, the exact coverage model for each cell is obtainable using electromagnetic propagation formula [4] combined with perfect knowledge of the environment (terrain elevation, 3D buildings). This is a time consuming process, both in terms of data acquisition and simulation. To keep it close to reality, a reasonable approximation can be derived from measurements of User Equipment (UE) locations and the cell of attachment. However it is difficult to obtain an exhaustive sampling with drive-test campaigns, and the treatment of the personal data that is collected can cause privacy issues. Therefore, coverage related studies often opt for simple geometries that require minimal parameters to model BS or cell coverage.

The most popular way to partition the geographic space into a map of BS coverage is the Voronoi diagram (or tessellation). The Voronoi diagram takes the location of all the BS in order to draw the regions of UE attachment. The underlying assumption is to consider that a UE is attached to the closest BS. However, studies about the validity of the Voronoi model seem almost non existent. Furthermore, while the recent radio infrastructure is tri-sectored and the urban fabric can be very heterogeneous, there is also very literature modelling sectorized cell coverage. This issue is particularly relevant for machine learning, because having a reliable cell coverage model would allow the training of models at a finer granularity than BS level, with a larger dataset available. Roughly, the training dataset could be at least three times larger with tri-sectored cells, and even more if we differentiate the coverage by band. Finally, to best of our knowledge, no study comparing Voronoi model with other types of geometries has been done before.

The final objective of this paper is to investigate if the Voronoi model, with or without variations, is in fact a sound simplification, or if another simple and naive model such as the circular shape could have been more adequate to use in the literature. The variations proposed do not constitute fundamentally new models, but we believe they may be good parameter tuning practices to adopt in order to be closer to the reality of cell coverage. With this being said, we propose the following two contributions.

The first contribution is to study the coverage accuracy of two simple geometric coverage models rescaled at different scaling factors. We compared how much the models overlapped a ground truth dataset describing the users locations and cell of attachment. The objective is to quantify how much a geometry modelling the coverage of a cell really overlaps with the locations of the UEs attached to the said cell. The geometric shapes are Voronoi polygons and circular shapes, split into sectors to model the cells coverage. The scaling factors are obtained through different strategies. For example, we tested whether the use of semi-empirical propagation models (Hata, UMa, RMa) to compute the radius of circular shapes for each cell was more accurate than re-scaling all the Voronoi polygons with the same scaling factor. The dataset of user location is geographically sampled in the French region of Île-de-France, and is comprised of 526, 000 user locations and 3, 255 cells of attachment. The data was collected between 2022-10-01 and 2023-03-01 and only concerned 4G users because it was the dominant technology at the time of the study. Therefore, it provided the empirical user distribution closest to reality. The location and serving cell measurements can be obtained via the Minimization of Drive Tests (MDT) mechanism [5]. Figure 1 shows the spatial distribution of the measurements colorized by geographical context. Details on the number of observations by context is given in Table 1. We propose a novel approach to study models accuracy, notably by introducing baselines to filter out under-scaled and over-scaled coverage shapes. With the ground truth measurements described previously, the results show that Voronoi-based cell shapes better approximate the spatial distribution of user-cell attachment than the circular shape, and uniform scaling of Voronoi polygons yielded better results than scaling each cell based on propagation models. The repository with the library for generating synthetic random cell topographies and coverage models is publicly available.[1]

The second contribution is to evaluate the efficiency of tuning Voronoi polygons on data aggregated at cell level that roughly approximate the distance distribution of the attached users. It is motivated by the awareness that the previous contribution requires difficult to acquire user measurements, which limits the possibility to apply the previous method to scale tune models. Cell radius estimations through user data aggregations are less accurate because we lose information about geographic positioning, however there are less privacy issues and it requires lesser data manipulation. The estimates of the cells maximum radius are computed from distance distributions obtained with timing advance data. Using the metrics of the previous contribution, the results show that polygons scale-tuned on partial information are always up-scaled to increase the number of overlaps with user locations.

TABLE 1. Size of the collected data used for ground truth.

Context	No. of user locations	No. of cells
Rural	82,901	418
Suburban	366,014	2,180
Urban	77,090	657
All	526,005	3,255

II. RELATED WORK

Having a simple coverage model is useful for merging different sources of spatial data with network equipment

[1] Available at: https://github.com/qiuda22/GeoLibCov.

TABLE 2. Methods applied and questions addressed.

Label	Equipment	Shape	Sector split	Scaling strategy	Hyperparameter	Questions addressed
voronoi-site	Base station	Voronoi	No	Voronoi uniform scaling	scaling factor in $[0.3, 3]$	2.a, 5
voronoi-cell-nosplit	Sector cell	Voronoi	No	Voronoi uniform scaling	scaling factor in $[0.3, 3]$	2.b, 5
voronoi-cell	Sector cell	Voronoi	Yes	Voronoi uniform scaling	scaling factor in $[0.3, 3]$	1.a, 2, 4.a, 5
uma/rma+circle	Sector cell	Circle	Yes	Scaling derived from UMa/RMa expressions	n_{PRB}	1.b, 4.b.ii, 5
uma/rma+voronoi	Sector cell	Voronoi	Yes	Scaling derived from UMa/RMa expressions	n_{PRB}	1a, 4.b.ii, 5
hata+circle	Sector cell	Circle	Yes	Scaling derived from Hata expressions	n_{PRB}	1b, 4.b.i, 5
hata+voronoi	Sector cell	Voronoi	Yes	Scaling derived from Hata expressions	n_{PRB}	1a, 4.b.i, 5
voronoi+circle	Sector cell	Circle	Yes	Circle scaling with Voronoi	scaling factor in $[0.3, 3]$	1.b, 4.c, 5
convexhull	Sector cell	Convex hull	Yes	-	-	3, 5
concavehull	Sector cell	Concave hull	Yes	-	-	3, 5
voronoi-cell	Sector cell	Voronoi	Yes	scaling factor tuning with cell-level data	scaling factor $s^* \in \mathbb{R}$	6

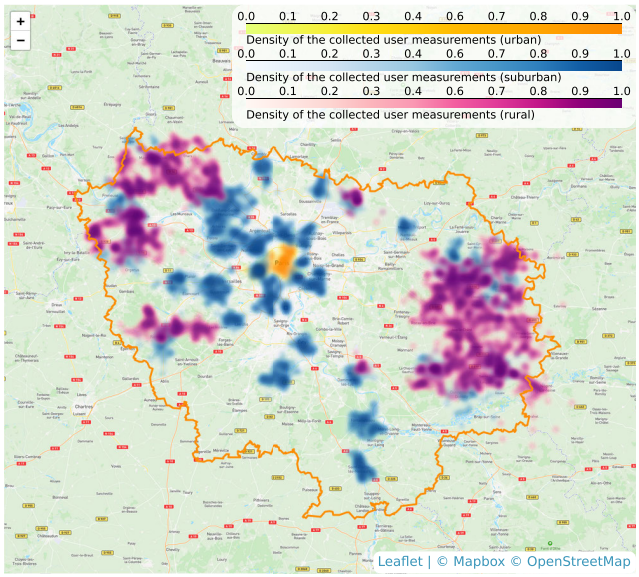

FIGURE 1. Heatmap of the user measurements collected to serve as ground truth in Île-de-France, France. Yellow: users attached to urban sites, Blue: to suburban sites, Purple: to rural sites. Orange line: Île-de-France boundary.

to create training datasets for ML models. The external sources are mainly related to topography (land uses, points of interests,...) or populations (distribution, demography). The existing literature on cross-domain ML training are focused on network performance prediction, either at user level [2], [6] or at base station level [3]. For each level of granularity, the method for merging exogenous features is different.

At user level granularity, the geographic space is discretized into a grid of small spatial units. Each unit is associated with the predicted network performance received by users inside it. In this case, there is no need for a coverage model: the exogenous data can be directly merged with the units using spatial intersection operations. The strength of this method lies in the fine-grained spatial accuracy, but it requires important resources to store, process and anonymize the data.

The second possibility is to study the network performance at base station level. While the spatial granularity is lower, it requires a smaller training dataset and preserves user privacy. The association of exogenous data with each equipment is less trivial than the previous case, and requires a coverage model to make the merge. As mentioned in the introduction, the Voronoi diagram is a widespread model of Base Station (BS) coverage. The diagram is the result of associating each point in space (location of a UE) to the closest BS. The notion of closeness is defined by the distance function chosen for the task. The default function is the Euclidean distance, but variants using the transmitting power of the BS or of a sectorized cell have been studied in [7]. In our previous work [8], we used the Voronoi diagram to associate relevant land use and PoI features to the base stations (BS). Outside of machine learning, Voronoi diagrams are widely used in stochastic geometry, where work references can be found in the survey [9]. Additionally, studies on human activities also use this tessellation to exploit mobile data. For example, the traffic of the mobile network is crossed with urban fabrics to estimate the real-time population distribution by functional regions [10], or to analyze the correlation with urbanization levels [11]. A few work subdivide the Voronoi polygons into sectorized cells regions: in [12], the authors split the Voronoi polygons using the azimuth, the tilt and the height of the antenna, in order to enhance the localization approximation of UEs. In [13], the authors also split the Voronoi polygon into finer coverage regions using the cells azimuth to assess the reliability of using mobile data as a proxy for population density. In these last two references, the studies consider that all cells on the same sector, independently of frequency band, share the same coverage.

Being very simple and not relying on radio configuration and propagation properties, it is expected for the Voronoi model to have accuracy limitations. However, it is not until very recently that this problem has been addressed. In [14], the authors show that on average, default scaled Voronoi

polygons does not spatially match the whole simulated probability distribution of user association to a base station. Rather, there is only a correspondence of 50% between the polygon area and the user distribution. In this paper, we reach the same conclusion that Voronoi polygons should be scaled to better reflect the reality of user distribution. However, we use a different ground truth constituted of real measurements, and a different approach to scaling as our study is not oriented towards probabilistic spatial diffusion. The accuracy of the ground truth in providing a realistic approximation of areas of user-cell attachment is assumed to be relatively high in urban areas. The quality of the approximation depends on the distribution of user locations, and will be higher if they are uniformly distributed. From our observations of the measurements, user locations overlapped with highways, residential roads and buildings which form a dense and continuous fabric. The level of urban fabric density is believed to be high enough to assume a uniform distribution of the measurements. Figure 2 shows the approximate shape (concave hulls) of the distribution of user measurements attached to three tri-sectored sites. The hull of user measurements is covering all the space, with a high level of overlap and leaving few gaps.

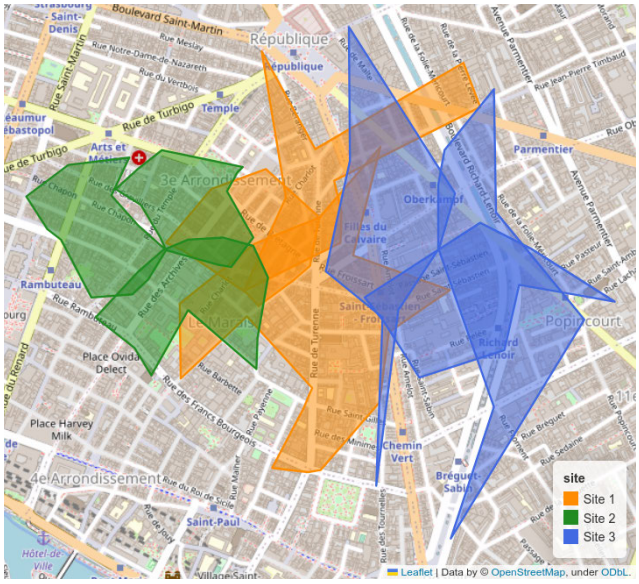

FIGURE 2. Concave hulls (α-shapes) of the ground truth user measurements attached to three tri-sectored LTE2600 sites (nine cells), using normalized length parameter $\lambda_p = 0.67$.

Furthermore, another objective of our study was to verify whether Voronoi was the best coverage shape, or if a circular geometry could perform better. In order to filter out shapes that are under-scaled or over-scaled, we propose to use the convex and concave hulls of user locations by cell attachment. The obtained hulls set a lower and upper bound for the precision metric (Section V-D). The convex hull of a set is unique [15] and the implementation that we used in PostGIS [16] is based on the Graham scan algorithm [17], [18]. On the contrary, a concave hull is not unique; different methods and constraints will generate different kinds of hulls. Examples of widespread hulls include α-shapes [19] and χ-shapes [20]. In the paper, χ-shapes were chosen because the algorithm was simpler to understand and parameterize.

The circle geometry was another classic model in the earlier days of cellular planning [21]. The idea is to make a simplification and consider that base station antennas form the same radiation pattern as an ideal isotropic antenna pattern, then to use path loss expressions to determine the limit coverage radius of the BS. There are several types of path loss models, from semi-empirical models used since GSM radio planning to analytical models developped in the recent years [22]. The latter are however very complex to integrate into research works because of the amount of terrain and network knowledge required. Thus, we chose to use simpler models that are classics in mobile planning: the Hata model [23] is known to work well in French urban areas [24] and the UMa and RMa models [25] which are suited for frequencies in the range $0.5 - 100$ GHz.

The coverage limit radius is defined as the distance at which the received power is equal to the sensitivity of the device. The method, detailed in Section VI-C is based on the works of [26] for establishing the budget link.

Finally, regarding the second contribution, the scaling factor of Voronoi polygons is tuned on cell radius approximated with timing advance (TA) data [27], [28]. Timing advance is a discretized measurement of the time taken by a signal to travel from a BS to a UE. It is used to synchronize the transmissions between both equipments.

III. METHOD

The objective of this work was to evaluate which shape was more representative of a sectorized cell coverage (Section V). Contrary to a map of best servers which gives the areas of cell attachment such that UEs have the best quality of service, in this study, the cell coverage is simply defined as the user attachment area to the said cell. Therefore, the map of cells coverage may present overlaps due to mechanisms such as handover, or gaps due to building blockage. We propose to scale the geometric shapes using different strategies detailed in Section VI to evaluate if up-scaled or down-scaled geometries increased the accuracy of the modelling on average.

In order to compare the geometric models, we use the precision and recall metrics also used in binary classification (Section IV). We defined a lower and upper bound for the precision using the precision of the convex and concave hulls of users locations partitioned by cell attachment (see Section V-D).

In Section VII we detail the method for finding the scaling factor that scales Voronoi polygons to the same dimension as the maximum user-cell distance estimated from TA data.

To summarize, our work can be structured around the following questions:

1) Which geometric shape best model the cell coverage: (a) Voronoi polygons, (b) circular shapes?

2) (a) How do the precision and recall vary between BS coverage modelling and cells coverage modelling? b) At cell level, what is the gain from splitting the BS coverage into sectors compared to not splitting it?
3) Which metrics and reference values should be used to compare the models?
4) Is it better to scale geometric shapes: (a) if Voronoi polygons, with a scaling factor identical for all the cells, (b) if circular shape, with a scaling factor such that it is resized to the dimensions of the corresponding Voronoi polygon, (c) with a scaling factor different for each cell, derived from a path loss formula, i.e. (i) Hata, (ii) or UMa/RMa?
5) Do the results vary significantly across different frequency bands?
6) What coverage efficiency can be obtained when tuning Voronoi polygons with cell radius data estimated from TA?

To answer these questions, we tested different combinations of geometries and scaling strategies described in Table 2. Note: The third and last row of the table shares the same label because we used the same model, but for different contributions. The interpretation and conclusions were given by observing the precision and recall averaged over all the cells.

IV. METRICS

A. NOTATION

For a combination of geometry and scaling strategy, the modelled coverage is evaluated by computing the mean precision and recall of the cells from the ground truth dataset. The ground truth dataset is made of users locations and their associated cells. We compute the precision/accuracy of models for each frequency band and partition the dataset accordingly.

Let $\mathcal{C} \subset \mathbb{N}$ be the set of identifiers of the cells in the ground truth dataset, $\mathcal{M} \subset \mathbb{R}^2$ be the set of user locations attached to any cell of $\mathcal{C}$, $\mathcal{F} \subset \mathbb{N}$ be the set of frequencies possibly deployed on a BS.

For band $f \in \mathcal{F}$, we define $\mathcal{C}_f \subset \mathcal{C}$ as the set of cells transmitting at frequency f, $\mathcal{M}_f \subset \mathcal{M}$ the subset of user locations attached to any cell of $\mathcal{C}_f$. For cell $c \in \mathcal{C}_f$, $M_c \subset \mathcal{M}_f$ is the set of user locations attached to that particular cell. Finally, we use $\mathrm{Cov}_{\mathrm{cell}}(c) \subset \mathbb{R}^2$ to denote the region covered by a cell as modelled by a geometric model.

B. PRECISION AND RECALL

We evaluate how a geometric model is close to the real coverage by computing the precision and recall of the shapes for each cell, then by taking the average. The precision and recall are metrics borrowed from binary classification.

Let $f \in \mathcal{F}$, $c \in \mathcal{C}_f$. The precision is defined as the ratio between the number of true positives and the sum of true positives and false positives. In the present case, the true positives are the locations of M_c contained inside $\mathrm{Cov}_{\mathrm{cell}}(c)$. The false positives are the locations contained in $\mathrm{Cov}_{\mathrm{cell}}(c)$ but not in M_c. The union of the set of true positives and false positives corresponds to $\mathrm{Cov}_{\mathrm{cell}}(c)$. In mathematical terms:

$$\mathrm{precision}(c) = \frac{|\{Q \in \mathcal{M}_f | Q \in M_c \cap \mathrm{Cov}_{\mathrm{cell}}(c)\}|}{|\{Q \in \mathcal{M}_f | Q \in \mathrm{Cov}_{\mathrm{cell}}(c)\}|} \quad (1)$$

The average precision over the cells of $\mathcal{C}_f$ is:

$$\overline{P} = \frac{1}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} \mathrm{precision}(c)$$

The recall is the ratio between the number of true positives and the sum of true positives and false negatives. The false negatives are the locations contained in M_c but not in $\mathrm{Cov}_{\mathrm{cell}}(c)$. The union of the set of true positives and false negatives corresponds to M_c.

$$\mathrm{recall}(c) = \frac{|\{Q \in \mathcal{M}_f | Q \in M_c \cap \mathrm{Cov}_{\mathrm{cell}}(c)\|}{|\{Q \in \mathcal{M}_f | Q \in M_c\}|} \quad (2)$$

The average recall over the cells of $\mathcal{C}_f$ is:

$$\overline{R} = \frac{1}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} \mathrm{recall}(c)$$

Since we want models to overlap with M_c as much as possible, the recall is the most important metric to maximize (Figure 3 (b)). However, only looking at recall may lead to choosing oversized coverage (Figure 3 (c)). A precision that is too low with respect to some threshold could detect this problem, but not only. It could also be used to detect undersized shapes (Figure 3 (a)) since the precision tends to be higher for smaller coverage shapes. Geometric models that have a precision too high with respect to some threshold could also be filtered out.

To fix a reasonable lower and upper precision bound, we propose to use the precision obtained with the convex and concave hulls of M_c (see Section V-D).

1) BASELINE VORONOI-SITE

To measure the loss in accuracy when going from BS to cell-level, we compute the precision and recall of the Voronoi tessellation. The precision and recall of a BS is expressed with different sets of true positives, false positives and false negatives. Let $\mathcal{B}$ be the set of BS identifiers. For a BS of id $b \in \mathcal{B}$, band $f \in \mathcal{F}$, $M'_{f,b}$ is the set of locations such that UEs are attached to any cell of b of frequency f, and $\mathrm{Cov}_{\mathrm{BS}}(b)$ is the modelled coverage of the BS (here a Voronoi polygon).

The precision and recall for a BS are given by:

$$\mathrm{precision}'(b) = \frac{|\{Q \in \mathcal{M}_f | Q \in M'_{f,b} \cap \mathrm{Cov}_{\mathrm{BS}}(b)\}|}{|\{Q \in \mathcal{M}_f | Q \in \mathrm{Cov}_{\mathrm{BS}}(b)\}|}$$

$$\mathrm{recall}'(b) = \frac{|\{Q \in \mathcal{M}_f | Q \in M'_{f,b} \cap \mathrm{Cov}_{\mathrm{BS}}(b)\}|}{|\{Q \in \mathcal{M}_f | Q \in M'_{f,b}\}|}$$

We then average the precision and recall over $\mathcal{B}$ restricted to the frequency f.

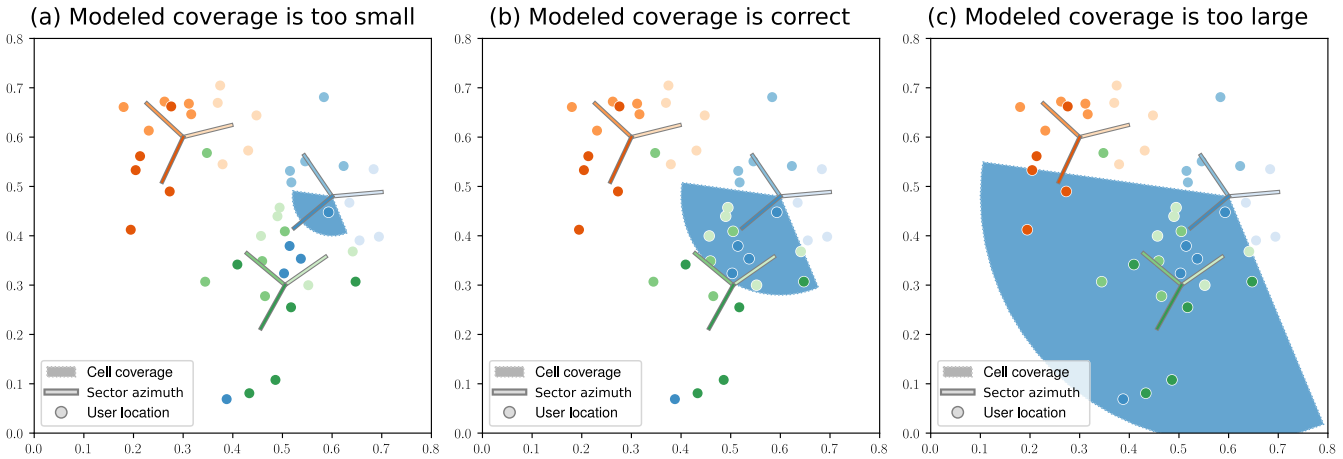

FIGURE 3. Coverage evaluation of one cell sector (in blue). (a) precision=1, recall=0.2; (b): precision=0.5, recall=0.8; (c): precision=0.23, recall=1. For illustration purposes, the radio topography and users locations in the figure were randomly generated and are not representative of the real cell and user distributions. The axes represent the axes of a map projection system for locating the cells, which were arbitrarily generated in the given range and unitless. In the real world, Cartesian coordinates used for map projections are expressed in meters.

2) BASELINE VORONOI-CELL-NOSPLIT

To measure the accuracy gain from splitting BS geometries into cell sectors, we compute the precision and recall of using Voronoi tessellation to model cell-level coverage. Given a band $f \in \mathcal{F}$ and a cell $c \in \mathcal{C}_f$, let $b \in \mathcal{B}$ be the BS where c is deployed. The expression of the precision and recall are the ones given by Equations 1 and 2, only that $\text{Cov}_{\text{cell}}(c) = \text{Cov}_{\text{BS}}(b)$, since we do not split the coverage of b.

V. SHAPES

To obtain cell coverage shapes, the first step is to create the coverage of their BS. Depending on the model chosen, BS coverage are either Voronoi regions or circles. Then, the regions are split into sub-regions, one for each sector, depending on antennas azimuth. Figure 4(a) illustrates the Voronoi-based coverage model and Figure 4(b) the circular-based model. Figure 4(c) and 4(d) respectively illustrate the convex and concave hulls of user location by cell attachment. Their use is developed in Section V-D.

A. VORONOI MODEL FOR BASE STATIONS

Let $n \in \mathbb{N}$ be the number of BS, $S = \{P_1, \ldots, P_n\} \subset \mathbb{R}^2$ be the set of their location coordinates.

The Voronoi region of a BS located at coordinates P is:

$$\text{Vor}(P) = \{Q \in \mathbb{R}^2 | \forall P' \in S \ \ ||\text{QP}|| \leq ||\text{QP}'||\}$$

where $||\text{QP}||$ is the two-dimension Euclidian distance between Q and P.

B. CIRCULAR MODEL FOR BASE STATIONS

Let P be the point location of a BS. The circular model, parameterized by radius R_l, defines the coverage region D:

$$D(P) = \{Q \in \mathbb{R}^2 | ||\text{QP}|| \leq R_l\}$$

The calculation of the limit radius R_l is explained in Section VI-C.

C. SPLITTING THE BS COVERAGE INTO SECTORS

Reusing the notations presented previously, we divide the BS modelled coverage into as many sub-regions as there are sectors deployed on the site. For an easier following on the notations, Figure 5 illustrates how a Voronoi polygon is split into cells coverage.

The cell coverage map can vary from one frequency to another, because not all frequencies are present on all the sites or on all the sectors.

Let $f \in \mathcal{F}$ be a frequency band deployed on a BS of id b located at coordinates P. $\text{Cov}_{\text{BS}}(b)$ is the geometric region modeling the coverage of b (Voronoi or circular).

The following explanations are given in an orthonormal coordinate system with a basis $(\vec{\mathbf{PX}}; \vec{\mathbf{PY}})$ (the location P of the BS is taken as the point of origin).

For any vectors $\vec{\mathbf{u}}, \vec{\mathbf{v}} \in \mathbb{R}^2$, we adopt the following notation:

- $\widehat{(\vec{\mathbf{u}}, \vec{\mathbf{v}})}$ denotes the geometric angle formed by two vectors with P as the angle origin,
- $\theta(\vec{\mathbf{u}}, \vec{\mathbf{v}})$ is the measure of the angle $\widehat{(\vec{\mathbf{u}}, \vec{\mathbf{v}})}$ in radians. Angles are measured counterclockwise in the interval $[0, 2\pi[$,
- $(P, \vec{\mathbf{u}})$ denotes a ray (half-line) of direction $\vec{u}$, with initial point P.

Let $\mathcal{A}_{P,b} = \{\vec{\mathbf{a}}_0, \ldots, \vec{\mathbf{a}}_{m-1}\} \subset \mathbb{R}^2$ be the set of distinct vectors representing the directions towards which the cells identified by ids $c_0, \ldots, c_{m-1} \in \mathbb{N}$ are deployed at location P, and transmitting at frequency f. The set $\mathcal{A}_{P,b}$ is ordered such that: $\theta(\vec{\mathbf{PX}}, \vec{\mathbf{a}}_0) < \ldots < \theta(\vec{\mathbf{PX}}, \vec{\mathbf{a}}_{m-1})$.

The region $\text{Cov}_{\text{BS}}(b)$ is partitioned into m subregions $\{\text{Cov}_{\text{cell}}(c_0), \ldots, \text{Cov}_{\text{cell}}(c_{m-1})\}$. Let $\vec{\mathbf{b}}_{k,k+1}$ be the direction vector such that $(P, \vec{\mathbf{b}}_{k,k+1})$ is the angle bisector of $(\vec{\mathbf{a}_k}, \vec{\mathbf{a}_{k+1}})$. Let $B_{k,k+1}$ (resp. $B_{k-1,k}$) be the point of intersection of

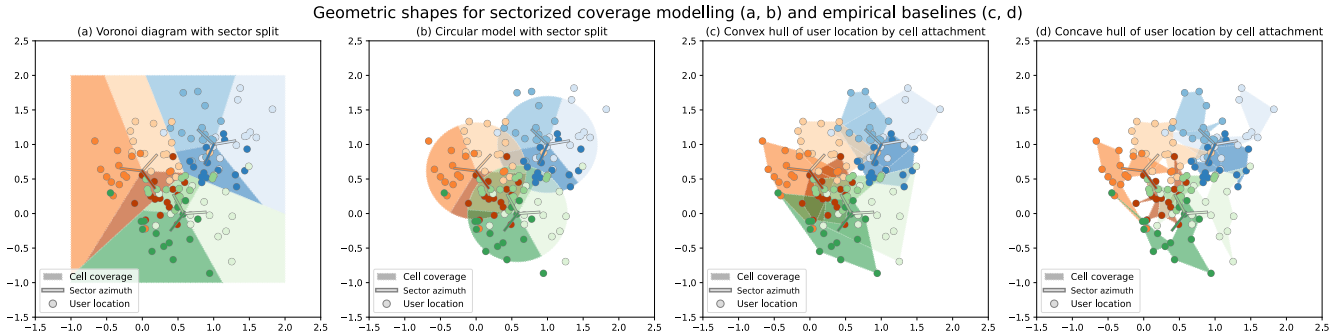

FIGURE 4. Figures (a) and (b) illustrate the geometric models that will be evaluated at different scales in the rest of the paper. Figures (c) and (d) illustrate the convex and concave hulls of user locations by cell, computed on the ground truth data. For illustration purposes, the radio topography and users locations in the figure were randomly generated and are not representative of the real cell and user distributions. The axis represent the axis of a map projection system for locating the cells and users, which are arbitrarily generated in the given range and unitless. In the real world, Cartesian coordinates used for map projections are expressed in meters.

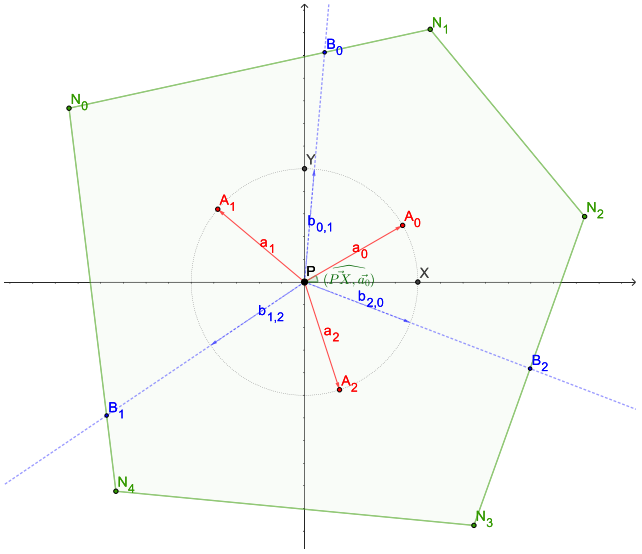

FIGURE 5. Example of a fictive Voronoi BS polygon divided into three sector areas in the Cartesian plane ($\vec{PX}$; $\vec{PY}$). $\text{Cov}_{\text{BS}}(b)$ is the polygon $N_0N_1N_2N_3N_4N_0$, and $\text{Cov}_{\text{cell}}(c_0)$ is the polygon $PB_0N_1N_2B_2P$. The axis ranges are arbitrary and unitless. In the real world, Cartesian coordinates used for map projections are expressed in meters.

$(P, \vec{\mathbf{b}}_{k,k+1})$ (resp. $(P, \vec{\mathbf{b}}_{k-1,k})$) with the boundary of $\text{Cov}_{\text{BS}}(b)$, denoted $\partial\text{Cov}_{\text{BS}}(b)$.

Let c_k be the id of the cell deployed on the k^{th} sector, and $\text{Cov}_{\text{cell}}(c_k) \subset \text{Cov}_{\text{BS}}(b)$ be the modelled region covered by c_k. The boundary is defined as:

$$\begin{aligned}\partial\text{Cov}_{\text{cell}}(c_k) = \text{PB}_{k-1,k} \cup \text{PB}_{k,k+1} \\ \cup \{Q \in \partial\text{Cov}_{\text{BS}}(b) \,|\theta(\vec{\mathbf{PX}}, \vec{\mathbf{b}}_{k-1,k}) \\ \leq \theta(\vec{\mathbf{PX}}, \vec{\mathbf{PQ}}) \leq \theta(\vec{\mathbf{PX}}, \vec{\mathbf{b}}_{k,k+1})\}\end{aligned}$$

For simplicity of notation, we omitted the modulo m term, but it applies when $k = 0$, i.e. $k - 1 \equiv m - 1 \pmod m$ and when $k = m - 1$, i.e. $k + 1 \equiv 0 \pmod m$.

D. EMPIRICAL HULLS AS REFERENCES

Let $c \in \mathcal{C}$ be a cell id, and $M_c = \{m_1, \ldots, m_N\} \subset \mathbb{R}^2$ the set of N user locations attached to c. Convex and concave hulls of M_c are polygons that contains all the points of the set, and which vertices are made of points from M_c. Following this definition, the recall of these hulls is always 1, so their precision can be used as a reference value. More precisely, we can set a lower and upper precision thresholds. These thresholds give indications of whether a modelled coverage is too small or too large.

1) CONVEX HULL

The convex hull of M_c is the minimal convex set that contains all the points of M_c (Figure 4(c)). In two dimension geometry, a region is said to be convex if, for any pair of points in the region, the line segment joining them is also contained in the region.

The convex hull serves as a lower bound precision indicator, because it has the largest area possible such that its vertices belong to M_c. So a coverage model that has a global precision $\overline{P}$ smaller than the precision of the convex hull is likely to be oversized.

2) CONCAVE HULL

As indicated by the name, concave hulls of M_c are not bound to be convex (Figure 4(d)). These hulls are used to capture the shape of a scatterplot more precisely than the convex hull. The χ-shapes form a family of concave hulls obtained with an algorithm based on eroding the edges of a Delaunay triangulation of M_c.

The function for creating χ-shapes is parameterized by the normalized length parameter $\lambda_P \in [0, 1]$, where a value of 0 produces maximal concaveness, and a value of 1 produces a convex hull (Figure 6). For further explanation about the algorithm, we refer the reader to [20], and to [29] for the Delaunay triangulation.

The precision of the χ-shape obtained with $\lambda_P = 0$ is one of the highest possible due to its maximal concaveness; the hull is made of almost (if not all) every point of M_c. Therefore, we use this definition of concave hull as an upper bound precision indicator. Any geometric coverage that has an average precision $\overline{P}$ higher than the precision of the χ-shape is likely to be undersized.

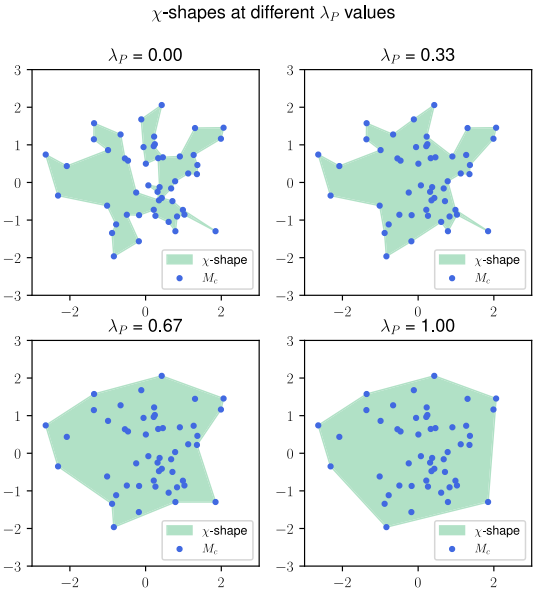

FIGURE 6. Illustration the impact of the normalized length parameter λ_P on the concaveness of χ-shapes. The users locations in the figure were randomly generated in an arbitrary range. The axis represent the axis of a map projection system. In the real world, Cartesian coordinates used for map projections are expressed in meters.

VI. SCALING STRATEGIES

The coverage models are scaled up and down to understand which range of scale parameters should be used to be closest to reality. The scaling is applied to the coverage model of the BS, before making the sector splits.

We evaluated three ways of scaling the geometries, each controlled by a hyperparameter:

1) For Voronoi polygons, we performed a uniform scaling of the coverage of all the BS using a single scaling factor, which is also the hyperparameter (Figure 7(b) and (c)).
2) For circular shapes, we scaled each BS coverage with a different scaling factor, such that the circle radius is equal to the radius of the bounding circle of the corresponding Voronoi polygon. When evaluating different scales, the Voronoi polygons are first scaled using the previous method. The hyperparameter is the scaling factor of the Voronoi polygons (Figure 7(f)).
3) For circular and Voronoi shapes, each BS coverage is scaled with a different scaling factor. The scaling factor is chosen such that the radius of its bounding circle is equal to the limit radius verifying the maximum allowable path loss equation. The hyperparameter controlling the maximum allowable path loss value is the number of physical resource blocs (PRB) received by the UE (Figure 7(e)).

A. VORONOI UNIFORM SCALING

Let $s \in \mathbb{R}$ be a scaling factor. Let $b \in \mathcal{B}$ be the id of a BS with point location $P \in \mathbb{R}^2$ and modelled coverage $\mathrm{Cov_{BS}}(b) \subset \mathbb{R}^2$. Any point $Q \in \partial\mathrm{Cov_{BS}}(b)$ of the boundary is transformed to Q' to form the scaled boundary.

The scale function $\mathrm{F_{scale}} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is in the form:

$$\mathrm{F_{scale}}(Q) = P + s(Q - P) = Q'$$

The transformation is done with the BS location P taken as origin. Any $s < 1$ downscales a shape (Figure 7(b)), and any $s > 1$ upscales it (Figure 7(c)). In our experiments, we observed empirically that an adequate search interval for s was [0.3, 3].

B. CIRCLE SCALING WITH VORONOI

Let r_c be the initial radius of the circular model. For a BS identified by id b, let r_v be the radius of the bounding circle of its corresponding Voronoi polygon at default scale. The bounding circle is the smallest circle containing the vertices of the polygon.

We first scale Voronoi polygons using the scale function $\mathrm{F_{scale}}$ defined previously with the scaling factor s. The radius of the scaled bounding circle equals sr_v. Let $s' \in \mathbb{R}$ be the scaling factor of the circular shape modelling the coverage of b. The expression of s' is:

$$s' = s\frac{r_v}{r_c}$$

The radius of the scaled circle is $r'_c = s'r_c = sr_v$, which is equal to the bounding circle radius of the scaled Voronoi polygon. Figure 7(f) is an example of scaling circles up to the bounding circle radius of the default Voronoi diagram.

C. SCALING DERIVED FROM PATH LOSS EXPRESSIONS

For a BS identified by id b with a modelled coverage $\mathrm{Cov_{BS}}(b)$, let r be the bounding circle of $\mathrm{Cov_{BS}}(b)$. We denote by R_l the limit radius such that the power received by a UE is equal to its sensitivity.

The coverage of b is scaled by a factor s where:

$$s = \frac{R_l}{r}$$

In the study, we vary the sensitivity threshold by changing the number of PRB n_{PRB} allocated to the user at the cell boundary. Figure 7(e) illustrates the scaling of circles with $n_{\mathrm{PRB}} = 1$. In the rest of the section, we present the expressions used to derive R_l from the maximum allowable path loss. We adopt most of the notation and budget link computation used in [26].

1) MAXIMUM ALLOWABLE PATH LOSS

The budget link is calculated on the constraint that the received power $\mathrm{P}_r(R)$ at a distance R must be higher or equal to the sensitivity S of the receiver. We assume that the transmission is limited in the uplink direction, so the transmitter is the UE and the receiver is the BS. The received power $\mathrm{P}_r(R)$ is equal to the transmitting power P_t augmented with the transmitter antenna gain G_{UE} and subtracted from

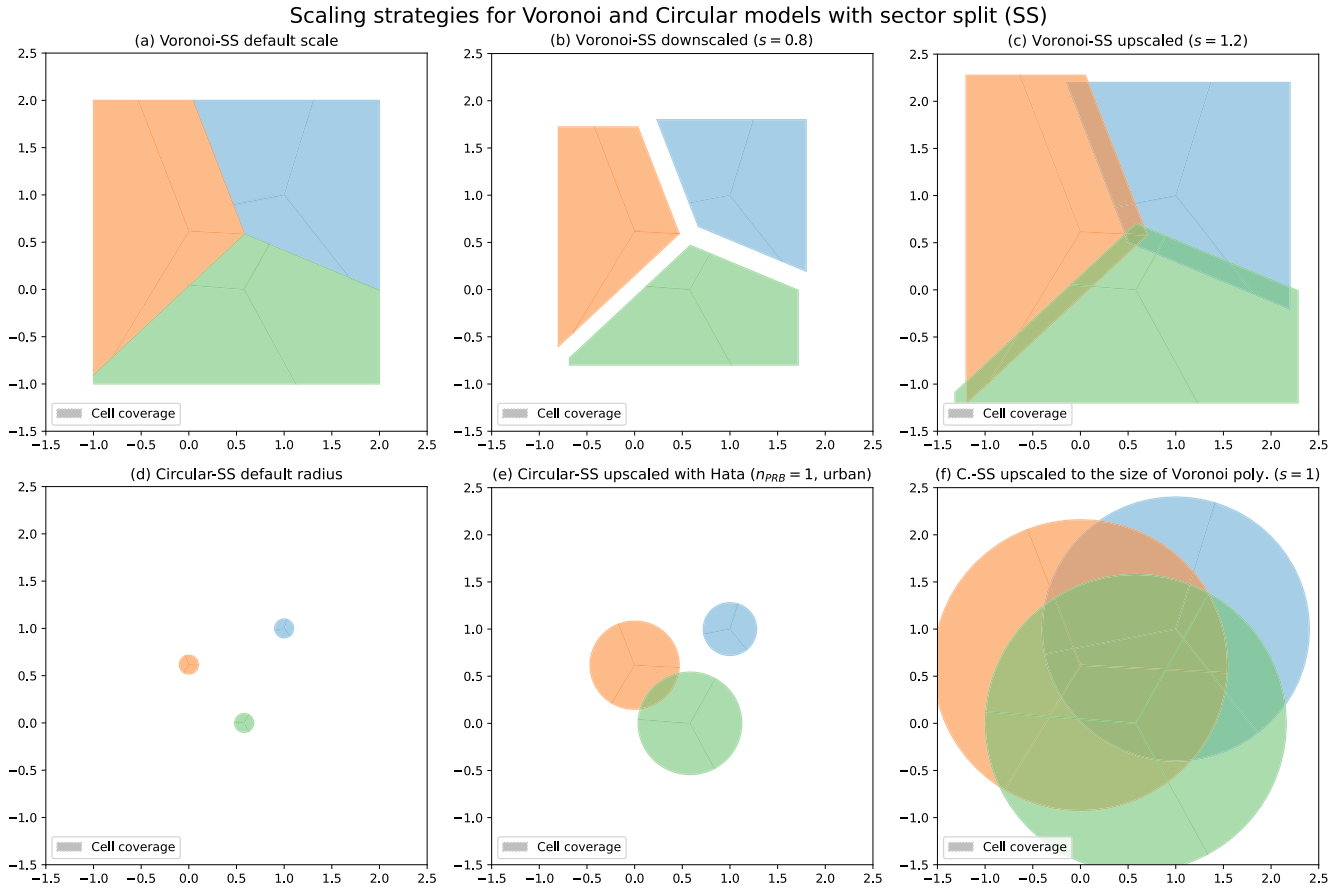

FIGURE 7. Geometric models (first row: Voronoi, second row: circular shape) scaled using different strategies. The axis represent the axis of a map projection system for locating the cells, which are arbitrarily generated in the given range and unitless. In the real world, Cartesian coordinates used for map projections are expressed in meters.

the path loss PL(R) and from the margins M:

$$\mathrm{P}_r(R) = P_t + G_{\mathrm{UE}} - \mathrm{PL}(R) - M \geq S$$

In reality, G_{UE} varies depending on the direction of arrival of the signal. In the study, the UE antenna is approximated as isotropic antenna which emits an effective isotropic radiated power $\mathrm{EIRP} = P_t + G_{\mathrm{UE}}$ in all directions. Consequently, the previous equation is simplified as:

$$\mathrm{P}_r(R) = \mathrm{EIRP} - \mathrm{PL}(R) - M \geq S$$

The maximum allowable path loss (MAPL) is the value such that the received power is equal to the sensitivity of the device. The limit radius achieving the MAPL is denoted R_l. In other words:

$$\begin{cases} \mathrm{PL}(R_l) = \mathrm{MAPL} \\ \mathrm{P}_r(R_l) = \mathrm{EIRP} - \mathrm{MAPL} - M = S \end{cases} \tag{3}$$

Interested in finding R_l, we rearrange the previous expressions and take the inverse function of PL, denoted PL^{-1}:

$$\begin{cases} R_l = \mathrm{PL}^{-1}(\mathrm{MAPL}) \\ \mathrm{MAPL} = \mathrm{EIRP} - S - M \end{cases} \tag{4}$$

The Table 3 indicates the values of EIRP, S, M as well as the path loss models. Unless stated otherwise in the paper, the default unit measurements in use are those indicated in the table.

2) CELL CONTEXT ATTRIBUTION

The context of signal propagation (urban, suburban, rural) is determined with the knowledge of the geography of the region Île-de-France. We use the administrative departments to separate the urban areas from the suburban areas, and use the notion of urban unit slice [30] to identify rural areas.

The urban unit slice (fr: *tranche d'unité urbaine*) is a statistical data provided by INSEE. It is based on the concept of urban unit. The urban unit is defined by INSEE as *"a commune or group of communes with a continuous built-up area (no more than 200 metres between two buildings) with at least 2,000 inhabitants."* [31] The "commune" is the smallest level of administrative division in France, governed by a municipality. Based on their population, an urban unit slice is assigned to each urban unit. An urban unit slice is a numeric category between 0 and 8, where the higher the number is, the more populated the urban unit is. The slice 0 is attributed

to communes outside of an urban unit, which we interpret as rural areas.

The following context attribution rules were applied:

- if the cell is located in department 75 (Paris), then the propagation context is defined as urban.
- if the cell is deployed in department 77, 78, 91 or 95, and if the urban unit slice of the commune is 0, then the context defined as rural.
- otherwise, the propagation context is defined as suburban.

3) HATA MODEL

The Hata path loss model has three variants: urban, suburban and rural. Using the variable notations in Table 3, we reverse the Hata path loss functions to get the expressions of R_l.

Urban: Since Île-de-France is a heavily urbanized region, and the studied frequencies are superior to 400 MHz, the antenna height correction factor $\mathrm{a}(h_{\mathrm{UE}})$ is equal to:

$$\mathrm{a}(h_{\mathrm{UE}}) = 3.2[\log_{10}(11.75\, h_{\mathrm{UE}})]^2 - 4.97$$

With R_l expressed in km, using the Hata path loss model $\mathrm{H_U}$, the Equation 3 becomes:

$$\begin{aligned}\mathrm{MAPL} &= \mathrm{H_U}(R_l)\\ &= 69.55 + 26.16\log_{10}(f) - 13.82\log_{10}(h_{\mathrm{BS}})\\ &\quad - \mathrm{a}(h_{\mathrm{UE}}) + [44.9 - 6.55\log_{10}(h_{\mathrm{BS}})]\log_{10}(R_l)\end{aligned}$$

Reversing $\mathrm{H_U}$, we get:

$$\begin{aligned}\mathrm{H_U^{-1}(MAPL)} &= R_l\\ &= 10^{\frac{\mathrm{MAPL} - 69.55 + 13.82\log_{10}(h_{\mathrm{BS}}) + \mathrm{a}(h_{\mathrm{UE}}) - 26.16\log_{10}(f)}{44.9 - 6.55\log_{10}(h_{\mathrm{BS}})}}\end{aligned} \tag{5}$$

Suburban: The Hata suburban model $\mathrm{H_{SU}}$ is formulated as:

$$\mathrm{H_{SU}}(R_l) = \mathrm{H_U}(R_l) - 2\left[\log_{10}\left(\frac{f}{28}\right)\right]^2 - 5.4$$

Reversing $\mathrm{H_{SU}}$:

$$\begin{aligned}\mathrm{H_{SU}^{-1}(MAPL)} &= R_l\\ &= \mathrm{H_U^{-1}(MAPL)} \times 10^{\frac{5.4 + 2\left[\log_{10}\left(\frac{f}{28}\right)\right]^2}{44.9 - 6.55\log_{10}(h_{\mathrm{BS}})}}\end{aligned} \tag{6}$$

Rural: The Hata rural quasi-open model $\mathrm{H_{RU}}$ is formulated as:

$$\begin{aligned}\mathrm{H_{RU}}(R_l) &= \mathrm{H_U}(R_l) - 4.78[\log_{10}(f)]^2\\ &\quad + 18.33\log_{10}(f) - 35.94\end{aligned}$$

Reversing $\mathrm{H_{RU}}$:

$$\begin{aligned}\mathrm{H_{RU}^{-1}(MAPL)} &= R_l\\ &= \mathrm{H_U^{-1}(MAPL)} \times 10^{\frac{35.94 - 18.33\log_{10}(f) + 4.78[\log_{10}(f)]^2}{44.9 - 6.55\log_{10}(h_{\mathrm{BS}})}}\end{aligned} \tag{7}$$

4) 0.5-100 GHz MODELS

The 3GPP standard has defined two path loss formula for macro cells: Urban Macro (UMa) and Rural Macro (RMa). We used UMa model for urban and suburban cells and RMa model for rural cells. Instead of expressing the path loss as a function of the distance R_l between the BS and the user, UMa and RMa are expressed as a function of the distance between the user and the top of the antenna, noted $d_{3\mathrm{D}}$.

Using the Pythagorean theorem, the relationship between R_l, $d_{3\mathrm{D}}$ and h_{BS} is:

$$R_l = \sqrt{d_{3\mathrm{D}}^2 - h_{\mathrm{BS}}^2} \tag{8}$$

Urban Macro (UMa): We consider a pessimistic scenario where it is not possible to have LoS communication at the limit radius.

The path loss model UMa is formulated as:

$$\begin{aligned}\mathrm{MAPL} &= \mathrm{UMa}(d_{3\mathrm{D}})\\ &= 13.54 + 39.08\log_{10}(d_{3\mathrm{D}})\\ &\quad + 20\log_{10}(f)\ - 0.6(h_{\mathrm{UE}} - 1.5)\end{aligned}$$

Rearranging the previous expression, $d_{3\mathrm{D}}$ is expressed as follows:

$$d_{3\mathrm{D}} = 10^{\frac{\mathrm{MAPL} - 13.54 + 0.6(h_{\mathrm{UE}} - 1.5) - 20\log_{10}(f)}{39.08}}$$

Rural Macro (RMa) We suppose that the limit radius is always beyond the breakpoint distance d_{bp} defined in the standard, so we use the following formulation:

$\mathrm{RMa}(d_{3\mathrm{D}}) = a\log_{10}(d_{3\mathrm{D}}) + bd_{3\mathrm{D}} + c$

where:

$$\begin{cases} a = 60 + 0.03h^{1.72}\\ b = 0.002\log_{10}(h)\\ c = 20\log_{10}\left(40\pi\dfrac{f}{3}\right) - 0.044h^{1.72} - 40\log_{10}(d_{\mathrm{bp}})\end{cases}$$

The reverse function RMa^{-1} is a function of the principal branch (0) of the Lambert W function W_0.

$$\begin{aligned}d_{3\mathrm{D}} &= \mathrm{RMa}^{-1}(\mathrm{MAPL})\\ &= \frac{a}{b\log(10)}W_0\left(\frac{10^{(\mathrm{MAPL} - c)/a}b\log(10)}{a}\right)\end{aligned} \tag{9}$$

VII. SCALING FACTOR TUNING WITH CELL-LEVEL DATA

This section details the second contribution consisting in using user aggregated data at cell level to tune the scaling factor of Voronoi polygons. The efficiency of this method is examined using the same metrics and ground truth as before.

The dataset for tuning the scaling factor is the empirical cumulative distribution function of user distances by cell. For a cell of id $c \in \mathcal{C}$, let $R_{\mathrm{TA}} : \Omega \to \mathbb{R}$ be the random variable corresponding to a user's distance to c. The space of events $\Omega \subset \mathbb{N}$ is the set of possible timing advance (TA) values measured between the user and c.

Let $(r_{c,1}, \ldots, r_{c,n}) \in \mathbb{R}$ be n user distances sampled from R_{TA}. For $t \in \mathbb{R}$, $\mathrm{F}_n(t)$ is the empirical cumulative distribution

TABLE 3. Budget link in the uplink direction. In the first column, the terms in bold indicate a section, and the terms in italic are the result of a formula involving the variables presented in the section.

	Unit	Variable	Formula	Value
Parameters				
Frequency	MHz	f	Hyperparameter	$\{700, 800, 1800, 2100, 2600\}$
Number of PRBs	-	n_{PRB}	Hyperparameter	[1;50]
Bandwidth	MHz	W	$0.180 \times n_{\text{PRB}}$	[0.180;9]
BS Height	m	h_{BS}	Value from dataset	[1.3, 98.3]
UE Height	m	h_{UE}	Typical value	1.7
Temperature	K	T	Typical value	290
Boltzman constant	J·K^{-1}	k_B	Constant	1.38×10^{-23}
Noise density	dBm/Hz	N_d	$10\log_{10}(k_B T \times 10^3)$	-174
Area coverage probability	-	P_{cov}	Typical value	0.95
Quantile function of the normal distribution	-	Q	$\sqrt{2}\text{erf}^{-1}(2p-1)$	-
Context	-	-	Value defined from dataset	{Urban, Suburban, Rural}
Slow Fading SD	dB	σ_{SF}	Typical values	8 if Context = Urban 7 if Context = Suburban 6 if Context = Rural
Transmission				
Power	dBm	P_t	Typical value	23
Transmitter antenna gain	dBi	G_{UE}	Typical value	0
EIRP	dBm	EIRP	$P_t + G_{\text{UE}}$	23
Reception				
Target SINR	dB	SINR	Typical value	-6
Noise power	dBm	N_{pow}	$N_d + 10\log_{10}(W \times 10^6)$	[-121.45; -104.46]
Noise figure	dB	N_{fig}	Typical value	5
Receiver antenna gain	dBi	G_{BS}	Typical values	16 if $f \in \{700, 800\}$ 18 if $f \in \{1800, 2100\}$ 19 if $f = 2600$
Cable loss	dB	L_{cable}	Typical value	2
Sensitivity	dBm	S	$\text{SINR} + N_{\text{pow}} + N_{\text{fig}} - G_{\text{BS}} + L_{\text{cable}}$	[-139.45, -119.46]
Margins				
Interference	dB	M_I	Typical value	3
Slow Fading	dB	M_{SF}	Using Jakes formula: $\sigma_{\text{SF}} Q(P_{\text{cov}})$	[9.84; 13.12]
Body loss	dB	M_{body}	Typical value	1
Building penetration	dB	M_{building}	Typical values	18 if Context = Urban 15 if Context = Suburban 12 if Context = Rural
Total margins	dB	M	$M_I + M_{\text{SF}} + M_{\text{body}} + M_{\text{building}}$	[25.84; 35.12]
Cell radius				
Maximum allowable path loss	dB	MAPL	$\text{EIRP} - S - M$	[107.34; 136.61]
Propagation model	-	PL $\in$ {H, UMa, RMa}	{Hata, UMa, RMa}	-
Radius	m	R_l	$\text{PL}^{-1}(\text{MAPL})$	[250; 13000] (See Section VIII.)

function of R_{TA}. With $\mathbb{1}_A$ as the indicator function of event A, the expression of $\text{F}_n(t)$ is:

$$\text{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{r_{c,i} \le t}$$

The empirical limit radius $\hat{R}_c \in \mathbb{R}$ is the smallest value verifying $\text{F}_n(\hat{R}_c) \ge 0.95$. We leave out 5% of the measures to filter out outlier measurements.

Let $(c_1, \ldots, c_m)$ be the m cells constituting the dataset, $\hat{\mathbf{R}} = (\hat{R}_{c_1}, \ldots, \hat{R}_{c_m})$ is the vector such that the i^{th} element is the empirical limit radius of c_i, and $\mathbf{r} = (r_{c_1}, \ldots, r_{c_m})$ is the vector such that the i^{th} element is the bounding circle radius of the Voronoi polygon of c_i. When the polygons are scaled by a scaling factor s, the corresponding bounding circle radii are described by the vector $\mathbf{r}' = s\mathbf{r}$. The optimal scaling factor is the value that minimizes the mean average error (MAE) between $\mathbf{r}'$ and $\hat{\mathbf{R}}$:

$$\begin{cases} \text{MAE}(\hat{\mathbf{R}}, \mathbf{r}') = \dfrac{1}{n}\displaystyle\sum_{i=0}^{n} |\hat{\mathbf{R}}_{c_i} - s\mathbf{r}_{c_i}| \\ s^* = \arg\min\limits_{s \in \mathbb{R}} \text{MAE}(\hat{\mathbf{R}}, s\mathbf{r}) \end{cases} \tag{10}$$

VIII. RESULTS

We first present the results of the comparison between the Voronoi polygons and the circular shapes for modelling cell coverage combined with different scaling strategies. Then we present the precision and recall achieved by scale-tuning Voronoi polygons compared with the default scale.

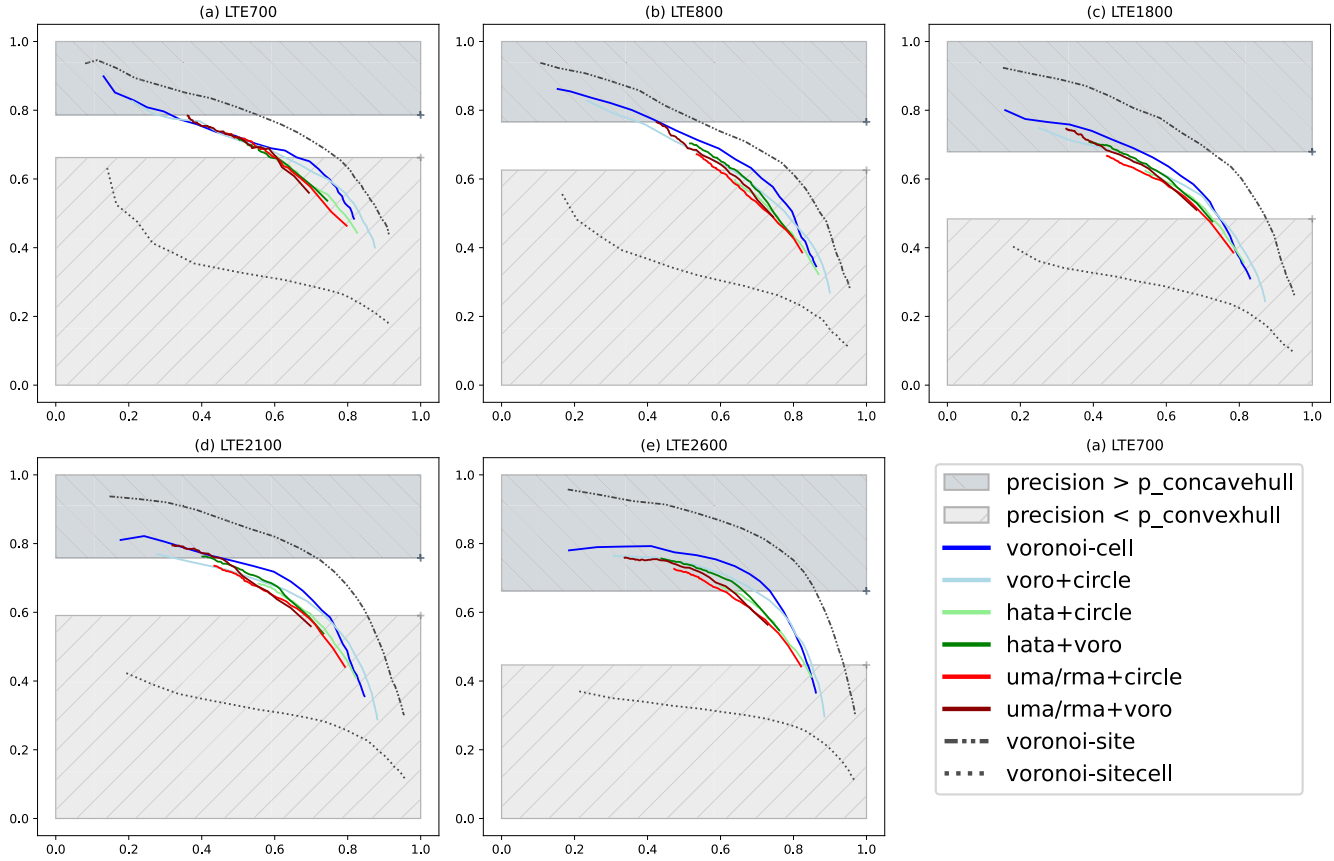

FIGURE 8. Geometric models precision (vertical axis) and recall (horizontal axis) obtained with the ground truth data. One subfigure is the evaluation of several models for the same frequency band, and one line is one model evaluated at different scales.

A. GEOMETRIC MODEL AND SCALING STRATEGY COMPARISON

Figure 8 shows the average precision $\overline{P}$ and recall $\overline{R}$ of the geometric models by frequency. One curve corresponds to a combination of a geometric model with a scaling strategy. One point on a curve corresponds to the precision and recall of a geometric model scaled with the value of a scaling factor. We refer the reader to Table 2 for a description of the combination that were tested and the corresponding labels in the legend.

While it looks like the precision-recall curves of binary classifiers, the graphs presents major differences in how their design and interpretation. In our case, it is unlikely for the precision or recall to be equal 1. This is because we look at the average precision and recall averaged over all the cells at once. Also, the hyperparameters controlling the scaling factors (number of PRB or scaling factor) vary in ranges that does not guarantee that the scaled shapes will cover all the positives (recall $= 1$) or just one positive (precision $= 1$).

The hyperparameter of the scaling strategies influences the precision and recall. For a scaling strategy parameterized by a scaling factor (voronoi-cell, voro+circle, voronoi-site, voronoi-sitecell), the larger the factor is, the larger the geometry is. Consequently, the shape associated to a cell will overlap more users attached to that cell, and the recall increases. At the same time, the shape will likely overlap more users that are not attached to the cell, so the precision decreases. This means that the scaling factor increases when we read the curves from left to right. For scaling strategies parameterized by the number of PRB n_{PRB} received by a UE (hata+circle, hata+voro, uma/rma+circle, uma/rma+voro), the smaller the number of PRB is required, the larger the limit radius R_l is. This means that n_{PRB} decreases when we read the curves from left to right.

The efficiency of two combinations (model, scaling strategy) A and B can be compared by observing the positioning of their curves. If the curve of A is above B, then the combination A is closer to the real coverage of the cell.

With these elements of interpretation, we now answer the questions 1-5 of Section III based on the graph results.

1) WHICH GEOMETRIC SHAPE BEST MODEL THE CELL COVERAGE: VORONOI POLYGONS OR CIRCULAR SHAPES?

The geometric model that best follows the distribution of users by cell attachment is the Voronoi sector split model labeled voronoi-cell. This can be observed for all the bands,

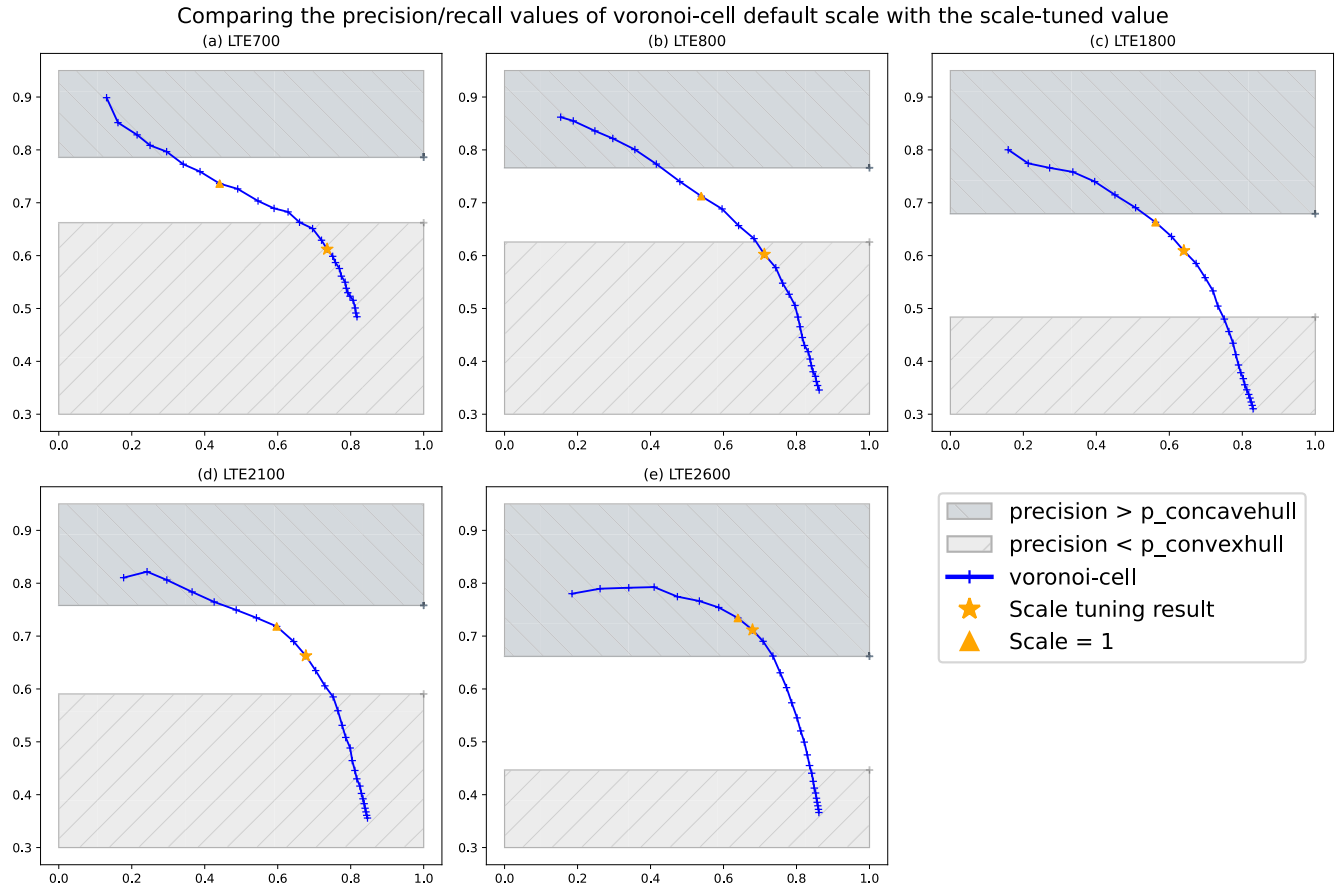

FIGURE 9. Voronoi precision/recall at different scales, overlayed with the default scaled factor (triangle) and the tuned scaling factor (star). The scatter plot of each tested scale is also displayed in blue.

especially the higher ones. For the lowest band (LTE700), all the models were almost equivalent. When the scaling factor derives from path loss equations, the Voronoi-based geometries (hata+voro, uma/rma+voro) are slightly better than circular-based shapes (hata+circle, uma/rma+circle).

2) A) HOW DO THE PRECISION AND RECALL VARY BETWEEN BS COVERAGE MODELLING AND CELLS COVERAGE MODELLING? (B) AT CELL LEVEL, WHAT IS THE GAIN FROM SPLITTING THE BS COVERAGE INTO SECTORS COMPARED TO NOT SPLITTING IT?

(a) There is a decrease in both precision and recall when sector splitting the BS coverage, as we can see that the coverage model voronoi-site for BS is above voronoi-cell. This can be explained by intra-site sector coverage overlapping, meaning that users are not always attached to the closest sector. (b) If we do not apply sector split, and consider the coverage of each cell to be equal to the coverage of the BS, then the precision and recall values drop significantly. This shows that while there may be intra-sector overlaps, most users are connected to the closest sectors. It also confirms the usefulness of doing sector-split when modelling the cells coverage.

3) WHICH REFERENCE VALUES SHOULD BE USED TO COMPARE THE MODELS?

Supposing that the real coverage of a cell is somewhere between the convex and the concave hull of its users locations, we use these hulls average precisions as reference values. The higher bound p_concavehull varies between 0.65 and 0.8, and the lower bound p_convexhull between 0.45 and 0.65 depending on the frequency. These values show a very high overlap between inter-sites sectors, and the importance of upscaling the geometric shapes to render these overlaps.

4) IS IT BETTER TO SCALE GEOMETRIC SHAPES: FOR VORONOI POLYGONS WITH A SCALING FACTOR IDENTICAL FOR ALL THE CELLS, FOR CIRCULAR SHAPES WITH A SCALING FACTOR SUCH THAT IT IS RESIZED TO THE DIMENSIONS OF THE CORRESPONDING VORONOI POLYGON, OR WITH A SCALING FACTOR DIFFERENT FOR EACH CELL? FOR THE LAST CASE, SHOULD WE USE HATA OR UMA/RMA MODEL?

The best scaling strategy was the uniform scaling of Voronoi cells (voronoi-cell), followed by scaling circles up to scaled Voronoi polygons dimensions (voro+circle), then followed

by scaling circles with the scaling factor derived from Hata model (hata+voro, hata+circle), then derived from UMa/RMa (uma/rma+voro, uma/rma+circle), although the last four combinations are often close.

5) DO THE RESULTS VARY SIGNIFICANTLY ACROSS DIFFERENT FREQUENCY BANDS?

While references, precision and recall values may vary across the frequencies, the curves pattern and their positioning are consistently the same.

6) WHAT COVERAGE EFFICIENCY CAN BE OBTAINED WHEN TUNING VORONOI POLYGONS WITH CELL RADIUS DATA ESTIMATED FROM TA?

The data we used for finding the scaling factor of Voronoi polygons is a dataset of empirical cumulative distribution of user distances by cells. Prior to data acquisition, the distances of the cumulative distribution function had been discretized into 12 values. The possible values of $\hat{R}$ in km were $I = \{0.35, 0.7, 1.1, 2.2, 3.6, 5.8, 8.0, 10.0, 15.0, 20.0, 25.0, 30.0\}$.

TABLE 4. Optimal scale and corresponding MAE in meters.

Band (MHz)	s^*	MAE($\hat{\mathbf{R}}, s^*\mathbf{r}$)
700	2.1	1159
800	1.7	1001
1800	1.6	939
2100	1.6	888
2600	1.4	726

Table 4 shows the tuned scales s^* and the minimized MAE($\hat{\mathbf{R}}, s^*\mathbf{r}$). The MAE, expressed in meters, is relatively high but the results are affected by the coarse spatial granularity of the binned distances from the dataset. Indeed, the average distance between two consecutive values of I is 2.7 km, 1.4 km if we only consider distances < 10 km. The MAE can be interpreted in the following way: on average, the scaled polygons bounding circle radius is off by one consecutive discrete value $\hat{R}$.

The table also shows that s^* is always superior to 1 meaning that according to the method, the Voronoi polygons should be upscaled. Figure 9 plots the curve voronoi-cell with the precision-recall of all the tested scaling factors between [0.3, 3] with a step of 0.1. The default scale is marked with a triangle and s^* is marked with a star. The graph shows that the recall of polygons scaled by s^* is always higher. In two cases (LTE700, LTE800), the precision is slightly below the lower bound and in one case it is still above the upper bound.

IX. DISCUSSION

From the previous answers, it appears that the Voronoi diagram, despite its simplicity, has the best results in modelling the cells coverage. The polygons can be uniformly scaled to be closer to the reality by rendering the cells overlaps. While the study cannot provide explanations to why this strategy is the best, we can still propose an interpretation.

TABLE 5. For LTE800 cells, the hyperparameters of each model such that the target precision is chosen as the precision of the convex hull.

Model	Hyperparameter	Precision	Recall
convexhull	-	*0.63*	*1.0*
voronoi-cell	$s = 1.3$	0.63	**0.68**
hata+voro	$n_{\text{PRB}} = 15$	0.63	0.64
hata+circle	$n_{\text{PRB}} = 50$	0.60	0.64
voro+circle	$s = 0.8$	0.64	0.62
uma/rma+voro	$n_{\text{PRB}} = 8$	0.63	0.61
uma/rma+circle	$n_{\text{PRB}} = 29$	0.63	0.59

The mobile network is designed so that a user equipment is attached to the cell from which the receiving signal is the strongest. In the studied area, it is likely that the density of base stations is so high that the limit radius computed using the MAPL is oversized. In other terms, the range of theoretic signal reception may be larger than the range of user attachment seen in the ground truth data. Furthermore, the high density of base stations could also lead the real boundaries of user attachment to be ''sharper'' than a circle. We think the shape of the Voronoi polygons might be more representative of the boundary in this case. The uniform scaling, when combined with Voronoi based shapes, preserve the information of cell densities and how close they are to each other. For urban areas (i.e. big cities), we thus recommend for people using Voronoi based cell coverage models to resize the shapes with a scaling factor in the range [1.0, 2.0]. The coverage is larger than the default diagram due to various phenomenons, such as the handover mechanism. In the present case, if we were to choose a scaling parameter s or n_{PRB}, one way could be to choose the hyperparameter such that the precision of the model equals the precision of the convex hull. That way, we try to have a recall as high as possible without taking the risk of oversizing the model. Table 5 shows the retained hyperparameters for LTE800. Tables for the other frequencies are given in the Appendix. In case where user measurements are difficult to acquire, we have shown that it was possible to get a fair to good approximation of the optimal scaling factor by tuning it with user distances aggregated at cell level, even if the data is coarse. To avoid overscaling, a compromise is to take the mean or a weighted mean between 1 (default scale) and s^*. That way, one will end between the triangle marker and the star marker on the graphs of Figure 9.

For less populated areas, the results may be biased by the accuracy of the ground truth. As stated in the introduction, the urban fabric and the distribution of the population was dense enough to assumed that distribution of measurements was uniform. However, less populated areas do not have the same continuous distribution of population and infrastructure. In many rural areas, people are concentrated in small towns and villages surrounded by fields. This can lead to coverage models underestimating how far a user could be attached, because no user has been far enough to provide a measurement in areas distant from dwellings. This aspect

can be limiting for studies aiming at analyzing the network performance at every point where a user can potentially be, independently of how likely they would be there. But the resulting tuned coverage remain relevant if we want to model the coverage area such as it is closest to human activity and network usage.

X. CONCLUSION

Throughout this study, we have developed a method to compare geometric coverage models with a ground truth of real measurements, using precision and recall metrics. We proposed to use the notions of convex and concave hulls to define a reference interval of precision to avoid undersizing or oversizing coverage models.

Two geometries were evaluated for modelling sectorized cell coverage: Voronoi polygons and circles, split into sectors. Three scaling strategies were also tested: scaling the shapes with an arbitrary scaling factor uniform for all the Voronoi cells, using a scaling factor different for each cell and dependent on the conditions of signal propagation, or using a scaling factor so that circles are the same dimensions as Voronoi cells. The results have shown that Voronoi sector split model combined with uniform scaling was the closest to the ground truth. This demonstrates the simplicity, yet strength of Voronoi-based solutions, and confirms the relevance of using this model when complex simulators and network knowledge are out of reach.

Since we have shown that upscaling Voronoi polygons should be a good practice to improve coverage modelling, we have also proposed an approach using more affordable data (i.e. cell level aggregates) to tune the scaling factor.

TABLE 6. For LTE700 cells, the hyperparameters of each model such that the target precision is chosen as the precision of the convex hull.

Model	Hyperparameter	Precision	Recall
convexhull	-	*0.66*	*1.0*
voronoi-cell	$s = 1.5$	0.66	**0.66**
voro+circle	$s = 1.0$	0.66	0.63
hata+circle	$n_{\text{PRB}} = 38$	0.66	0.61
uma/rma+voro	$n_{\text{PRB}} = 4$	0.66	0.61
uma/rma+circle	$n_{\text{PRB}} = 14$	0.66	0.60
hata+voro	$n_{\text{PRB}} = 10$	0.66	0.59

TABLE 7. For LTE800 cells, the hyperparameters of each model such that the target precision is chosen as the precision of the convex hull.

Model	Hyperparameter	Precision	Recall
convexhull	-	*0.63*	*1.0*
voronoi-cell	$s = 1.3$	0.63	**0.68**
hata+voro	$n_{\text{PRB}} = 15$	0.63	0.64
hata+circle	$n_{\text{PRB}} = 50$	0.60	0.64
voro+circle	$s = 0.8$	0.64	0.62
uma/rma+voro	$n_{\text{PRB}} = 8$	0.63	0.61
uma/rma+circle	$n_{\text{PRB}} = 29$	0.63	0.59

This work was initially motivated by the need to better understand how external sources of data should be associated with mobile network data at cell level to create informative machine learning training sets. We hope this study helps the development of cross-domain ML applications, or any kind of data-driven solution that could be integrated into next-generation networks.

Further work could be aimed at evaluating more complex coverage models to improve our knowledge about their accuracy in comparison to ground truth data. For example, different scaling parameters and/or models which take antenna radiation patterns into account can be considered. The cellular coverage models described in this study could also be improved by modelling the existing intra-sites sector overlapping.

TABLE 8. For LTE1800 cells, the hyperparameters of each model such that the target precision is chosen as the precision of the convex hull.

Model	Hyperparameter	Precision	Recall
convexhull	-	*0.48*	*1.0*
voronoi-cell	$s = 1.7$	0.48	**0.75**
voro+circle	$s = 1.1$	0.48	0.75
hata+circle	$n_{\text{PRB}} = 6$	0.48	0.73
hata+voro	$n_{\text{PRB}} = 1$	0.48	0.73
uma/rma+circle	$n_{\text{PRB}} = 3$	0.48	0.72
uma/rma+voro	$n_{\text{PRB}} = 1$	0.51	0.68

TABLE 9. For LTE2100 cells, the hyperparameters of each model such that the target precision is chosen as the precision of the convex hull.

Model	Hyperparameter	Precision	Recall
convexhull	-	*0.59*	*1.0*
voronoi-cell	$s = 1.5$	0.58	**0.75**
voro+circle	$s = 0.9$	0.60	0.73
hata+circle	$n_{\text{PRB}} = 11$	0.59	0.70
uma/rma+circle	$n_{\text{PRB}} = 5$	0.58	0.69
hata+voro	$n_{\text{PRB}} = 2$	0.59	0.69
uma/rma+voro	$n_{\text{PRB}} = 2$	0.61	0.64

TABLE 10. For LTE2600 cells, the hyperparameters of each model such that the target precision is chosen as the precision of the convex hull.

Model	Hyperparameter	Precision	Recall
convexhull	-	*0.45*	*1.0*
voro+circle	$s = 1.5$	0.44	**0.85**
voronoi-cell	$s = 2.2$	0.44	0.84
hata+circle	$n_{\text{PRB}} = 2$	0.46	0.82
uma/rma+circle	$n_{\text{PRB}} = 1$	0.44	0.82
hata+voro	$n_{\text{PRB}} = 1$	0.55	0.76
uma/rma+voro	$n_{\text{PRB}} = 1$	0.56	0.73

APPENDIX I. TABLES OF PRECISION AND RECALL FOR CONVEX HULL PRECISION TARGET

See Tables 6–10.

APPENDIX II. MAE CURVES FROM TUNING THE SCALING FACTOR

See Figure 10.

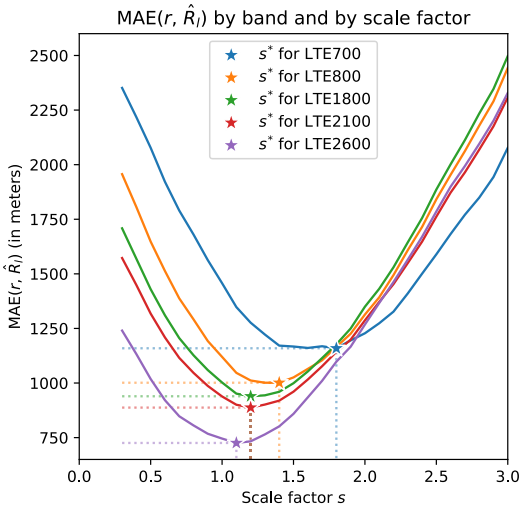

FIGURE 10. MAE curves from tuning the scaling factor.

APPENDIX III.
PRECISION-RECALL OF DEFAULT VERSUS SCALE-TUNED VORONOI POLYGONS

TABLE 11. Precision-recall of the default scale (s=1) vs the tuned scale (s*).

Band (MHz)	Scale	Recall	Precision
700	$s = 1$	0.44	0.74
	$\mathbf{s^* = 2.1}$	**0.74**	0.61
800	$s = 1$	0.54	0.71
	$\mathbf{s^* = 1.7}$	**0.71**	0.60
1800	$s = 1$	0.56	0.66
	$\mathbf{s^* = 1.6}$	**0.64**	0.61
2100	$s = 1$	0.60	0.72
	$\mathbf{s^* = 1.6}$	**0.68**	0.66
2600	$s = 1$	0.64	0.73
	$\mathbf{s^* = 1.4}$	**0.68**	0.71

REFERENCES

[1] J. Kaur, M. A. Khan, M. Iftikhar, M. Imran, and Q. E. U. Haq, "Machine learning techniques for 5G and beyond," *IEEE Access*, vol. 9, pp. 23472–23488, 2021.

[2] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1389–1401, Jun. 2019.

[3] J. Feng, X. Chen, R. Gao, M. Zeng, and Y. Li, "DeepTP: An end-to-end neural network for mobile cellular traffic prediction," *IEEE Netw.*, vol. 32, no. 6, pp. 108–115, Nov. 2018.

[4] Z. Yun and M. F. Iskander, "Ray tracing for radio propagation modeling: Principles and applications," *IEEE Access*, vol. 3, pp. 1089–1100, 2015.

[5] *Technical Specification Group Radio Access Network; Radio Measurement Collection for Minimization of Drive Tests (MDT)*, document TS 37.320, V17.3.0, 3GPP, 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/37_series/37.320/37320-h30.zip

[6] L. Dai and H. Zhang, "Propagation-model-free base station deployment for mobile networks: Integrating machine learning and heuristic methods," *IEEE Access*, vol. 8, pp. 83375–83386, 2020.

[7] J. N. Portela and M. S. Alencar, "Cellular coverage map as a Voronoi diagram," *J. Commun. Inf. Syst.*, vol. 23, no. 1, pp. 22–31, Apr. 2008.

[8] D. Qiu, A. Samba, H. Afifi, and Y. Gourhant, "Classifying urban fabrics into mobile call activity with supervised machine learning," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2021, pp. 1948–1953.

[9] Y. Hmamouche, M. Benjillali, S. Saoudi, H. Yanikomeroglu, and M. D. Renzo, "New trends in stochastic geometry for wireless networks: A tutorial and survey," *Proc. IEEE*, vol. 109, no. 7, pp. 1200–1252, Jul. 2021.

[10] F. Xu, P. Zhang, and Y. Li, "Context-aware real-time population estimation for metropolis," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2016, pp. 1064–1075.

[11] S. Mishra, Z. Smoreda, and M. Fiore, "Second-level digital divide: A longitudinal study of mobile traffic consumption imbalance in France," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2532–2540.

[12] E. Graells-Garrido, O. Peredo, and J. García, "Sensing urban patterns with antenna mappings: The case of Santiago, Chile," *Sensors*, vol. 16, no. 7, p. 1098, Jul. 2016.

[13] 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 *Proc. Eur. Conf. Qual. Off. Statist.*, 2016, pp. 1–16.

[14] O. E. Martínez-Durive, T. Couturieux, C. Ziemlicki, and M. Fiore, "VoronoiBoost: Data-driven probabilistic spatial mapping of mobile network metadata," in *Proc. 19th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Sep. 2022, pp. 100–108.

[15] R. T. Rockafellar, *Convex Analysis*, vol. 11. Princeton, NJ, USA: Princeton Univ. Press, 1997.

[16] PPS Committee. (2023). *PostGIS, Spatial and Geographic Objects for PostgreSQL*. [Online]. Available: https://postgis.net

[17] Open Source Geospatial Foundation, GEOS Contributors. (2021). *GEOS Coordinate Transformation Software Library*. [Online]. Available: https://libgeos.org/

[18] R. L. Graham, "An efficient algorithm for determining the convex hull of a finite planar set," *Inf. Process. Lett.*, vol. 1, pp. 132–133, Jun. 1972.

[19] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 4, pp. 551–559, Jul. 1983.

[20] M. Duckham, L. Kulik, M. Worboys, and A. Galton, "Efficient generation of simple polygons for characterizing the shape of a set of points in the plane," *Pattern Recognit.*, vol. 41, no. 10, pp. 3224–3236, Oct. 2008.

[21] V. H. M. Donald, "Advanced mobile phone service: The cellular concept," *Bell Syst. Tech. J.*, vol. 58, no. 1, pp. 15–41, Jan. 1979.

[22] C. Phillips, D. Sicker, and D. Grunwald, "A survey of wireless path loss prediction and coverage mapping methods," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 255–270, 1st Quart., 2013.

[23] *Technical Specification Group Radio Access Network; Radio Network Planning Aspects*, document TR 43.030, V17.0.0, 3GPP, 2022.

[24] M. Coupechoux, "Performances 5G: Étude comparée en zones rurales et urbaines," Telecom Paris, Institut Polytechnique de Paris, Paris-Saclay, France, Tech. Rep., Oct. 2021. [Online]. Available: https://telecom-paris.hal.science/hal-03367918/file/rapport.pdf

[25] *Technical Specification Group Radio Access Network; Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 38.901, 3GPP, V17.0.0, 2022.

[26] M. Coupechoux, "Bilans de liaison: De la 2G à la 5G," Telecom Paris, Institut Polytechnique de Paris, Paris-Saclay, France, Tech. Rep., 2021. [Online]. Available: https://marceaucoupechoux.wp.imt.fr/files/2021/05/RIO207-BdL-2G-4G-5G.pdf

[27] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, document TS 36.211, 3GPP, V17.2.0, 2022.

[28] J. D. Roth, M. Tummala, and J. W. Scrofani, "Cellular synchronization assisted refinement (CeSAR): A method for accurate geolocation in LTE-A networks," in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 5842–5850.

[29] B. Delaunay, "Sur la sphere vide," *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, vol. 7, nos. 793–800, pp. 1–2, 1934.

[30] INSEE. (2023). *Découpage Communal. Table d'Appartenance Géographique des Communes et Tables de Passage*. [Online]. Available: https://www.insee.fr/fr/information/2028028

[31] INSEE. (2023). *Urban Unit*. [Online]. Available: https://www.insee.fr/en/metadonnees/definition/c1501

DANNY QIU received the master's degree in computer science and applied mathematics from ENSIMAG, France, in 2020. She is currently pursuing the Ph.D. degree with the Institut Polytechnique de Paris, France. She is currently a Research Engineer with Orange Innovation, Lannion, France. Her research interests include data mining, machine learning, and geographic information systems.

ALASSANE SAMBA received the master's degree in data science from the National School of Statistics and Information Analysis (ENSAI), Rennes, France, in 2014, and the Ph.D. degree in computer science from IMT Atlantique, Institut Mines-Telecom, France, in 2018. In 2015, he joined Orange Innovation as a Ph.D. Student. He is currently a Researcher at Orange Innovation, where he currently leads a research project for network automation and intelligent control loops. His current research interests include data science and network management.

HOSSAM AFIFI received the Ph.D. degree from INRIA Sophia Antipolis. After his postdoctoral training at Washington University, St. Louis, MO, USA, he joined Mines Atlantique as an Assistant Professor. His tenure was obtained after a sabbatical stay at Nokia Research Mountain View, Mountain View, CA, USA. He joined the Samovar Research Laboratory, Télécom SudParis Saclay. He is currently a Full Professor at the Institut Polytechnique de Paris, Télécom SudParis Saclay. His current research interests include applied machine learning on computer science hard problems and energy management.

YVON GOURHANT received the Ph.D. degree in computer science from INRIA, Rocquencourt, France, in 1991. He joined Orange Labs, Lannion, France, in 1993, as a Researcher in distributed systems. From 1998 to 2009, he was leading a Research and Development Team on Programmable and Adaptable Networks. Since 2009, he has been leading and acting in a research program on low-cost network infrastructures for emerging countries (e.g., cellular radio network coverage over large territories and energy efficiency).