

City-Wide Signal Strength Maps: Prediction with Random Forests

Emmanouil Alimpertis, Athina Markopoulou, Carter Butts and Kostantinos Psounis

ABSTRACT

Signal strength maps are of great importance to cellular providers for network planning and operation, however they are expensive to obtain, inherently limited in scale, and possibly inaccurate in some locations. In this paper, we develop a rich prediction framework based on random forests to improve signal strength maps from limited measurements. First, we propose a random forests (RFs)-based predictor, with a rich set of features including location as well as time, cell ID, device hardware and others, which are considered jointly for the first time. We show that our RFs-based predictor can significantly improve the tradeoff between prediction error and number of measurements needed compared to state-of-the-art data-driven predictors, requiring 80% less measurements for the same prediction accuracy. Second, we show that our framework naturally extends to combine RFs with wireless propagation models to further improve prediction. Third, we leverage two types of real-world LTE RSRP datasets to evaluate and gain insights into the performance of different prediction methods: (i) a small but dense **Campus dataset**, collected on a university campus and (ii) several large but sparser **NYC and LA datasets**, provided by a mobile data analytics company.

1 INTRODUCTION

Cellular providers rely on key performance indicators (*a.k.a.* KPIs) to understand the performance and coverage of their network, as well as that of their competitors, in their effort to provide the best user experience. KPIs usually include wireless channel measurements (the most important of which for LTE is arguably the reference signal received power, *a.k.a.* RSRP) as well as other performance metrics (*e.g.*, throughput, delay, jitter), together with the frequency band, location of receiver, time and other information associated with the measurement. Signal maps consist of a large number of measurements of KPIs in several locations and are of crucial importance to cellular operators, for network management, maintenance, upgrades, and operations, *e.g.*, in order to determine if and where to deploy more cells, to identify problems and troubleshoot.

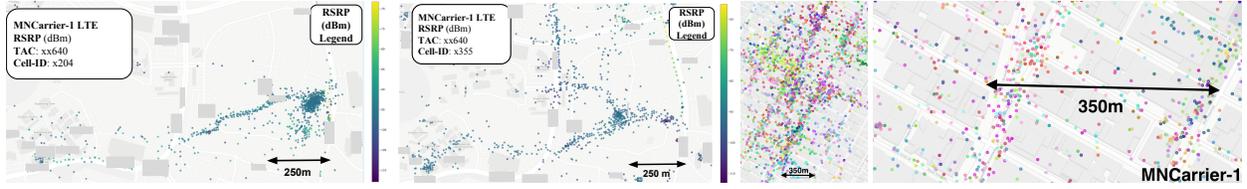
Although cellular providers can collect measurements on the network edge themselves, they increasingly choose to outsource the collection of signal maps to third parties for a variety of reasons, including: cost, liability related to privacy concerns of collecting data on end-user devices, and lack of access to competitor networks. One way that operators obtain detailed and accurate measurements is by hiring dedicated vans (*a.k.a.* wardriving [37]) with special equipment, to drive through, measure and map the received signal strength (RSS) in a particular area of interest. However, this method cannot provide wide scale (city- or country-wide) measurements. Another practice of operators is to buy signal map data from specialized mobile analytics companies, such as OpenSignal [27], RootMetrics [33], Tutela [36], and others. These companies crowdsource measurements directly from end-user devices, via standalone mobile apps [27], or measurement SDKs [36] integrated into

popular partnering apps, typically games, utilities or streaming apps. This way, they crowdsource measurements at large (city, country, or world-wide) scale and over long periods of time, but the measurements can be sparse in space (depending on end-user location) and time (measurements are collected infrequently so as to not drain user resources, such as battery or cellular data).

Either way, signal strength maps are expensive for both carriers (paying 100s of millions to third parties to collect data) and crowdsourcing companies (most of which use cloud services, thus collecting more measurements increases their operational cost). Yet, collecting a large number of measurements is necessary to obtain good coverage and accuracy for signal maps. Current technology and application trends, such as (i) 5G dense deployment of small cells and (ii) smart city/IoT monitoring and control at metropolitan scales, will only increase the need for accurate performance measurements [2, 11, 16], while data may be sparse, unavailable, or expensive to obtain in some locations, times and frequencies.

Our goal in this paper is to improve the tradeoff between cost (number of measurements) and quality (*i.e.*, coverage and accuracy) of signal maps via signal strength prediction from limited measurements. In general, there are two approaches in RSRP prediction: propagation models and data-driven approaches, reviewed in Section 5. Our approach falls in the second category and we employ a powerful machine learning framework that naturally incorporates multiple features and predictors. More specifically, we make the following contributions.

1. A rich RSRP prediction framework based on random forests (RFs). We develop a powerful machine learning framework based on random-forests (RFs), considering a rich set of features including, but not limited to, location, time, cell ID, device hardware, distance from the tower, frequency band, and outdoors/indoors location of the receiver, which all of them affect the wireless properties. To the best of our knowledge, this is the first time that location, time, device and network information are considered jointly for the problem of signal strength prediction. We assess the feature importance and we find cell ID, location, time and device type to be the most important. Moreover, this is the first time that RFs have been applied to the signal maps estimation problem. Prior work on data-driven prediction for signal maps was primarily based on geospatial interpolation techniques [6, 22, 29], which do not naturally extend beyond location, *i.e.*, (x, y) , features. Prior work on localization used a RFs model with only spatial coordinates as features [31]. We show that our RFs-based predictors can significantly improve the tradeoff between prediction error and number of measurements needed, compared to state-of-the-art data-driven predictors. They can achieve the lowest error of these baselines with 80% less measurements; or they can reduce the *RMSE* (root mean square error) by 13% for the same number of measurements. The absolute error can be reduced by up to 2dB – an improvement that can be crucial for determining the LTE/VoLTE performance in weak reception areas.



(a) Campus example cell x204: high density (0.66), (b) Campus: example cell x355: small density (0.12) more dispersed data (573). (c) NYC: Manhattan LTE TA, (d) NYC: zooming in Manhattan Midtown (Times Square) for some of the available cells.

Figure 1: LTE RSRP Map Examples from our datasets. (a)-(b): Campus dataset. Color indicates RSRP value. (c)-(d): NYC dataset. Data for a group of LTE cells in the Manhattan Midtown area. Different colors indicate different cell IDs.

2. Combining data and model based predictors. Although we show that RFs generally outperform propagation models at the cell or coarser granularity, we also observe the diversity of propagation models and data-driven predictors at finer scale. In areas where RSRP measurements are limited or inaccurate, our framework can naturally leverage propagation models to further improve RFs’ performance through ensemble learning. We demonstrate the potential benefits of the ensemble approach to harvest the diversity between model and RFs predictors and we provide guidelines on the use of stacking ensemble learning.

3. Real-world datasets. Our study leverages two types of real-world datasets: (i) a small but dense Campus dataset collected on a university campus; and (ii) several large but sparser NYC and LA datasets, provided by a mobile data analytics company. Examples are depicted in Fig. 1 and information about the datasets is provided in Table 3. We use these datasets to evaluate and contrast different methods and gain insights into tuning our framework. For example, cell ID is an important feature in areas with high cell density, which is encountered in urban areas such as Manhattan Midtown; in contrast, cell ID should be used to train cell-specific RFs in suburban areas. Furthermore, time features are important in cells with less dispersed measurements, *i.e.*, concentrated in fewer locations. To the best of our knowledge, the NYC and LA datasets are the largest used to date for RSRP or other RSS metric prediction. They contain 10.9 million LTE data points in areas of $300km^2$ and $1600km^2$ for NYC and LA respectively, instead of at most tens of km^2 in prior work [6, 13, 29]. This enables our paper to be the first to perform RSRP prediction in city-wide/metropolitan level, thus be applicable to crowdsourcing companies and operators.

The structure of the rest of the paper is as follows. Section 2 presents the prediction methods under comparison, including our random forests-based approach as well as baselines for comparisons. Section 3 presents the available signal map datasets. Section 4 provides evaluation results. Section 5 reviews related work. Section 6 concludes the paper.

2 RSRP PREDICTION

We start by defining the RSRP prediction problem. Then, we present state-of-the-art and our own prediction methods, summarized on Table 1. Broadly speaking, there are model-based and data-driven RSRP predictors. Our methodological contributions are (1) a data-driven (random forest-based) prediction framework, with a richer set of features than just the spatial features previously considered; and (2) a hybrid data and model-based approach. The evaluation of these methods based on the datasets is provided in Section 4.

2.1 Problem Statement

RSRP Definition. Although there are many KPIs related to received signal strength (RSS), including RSRP, RSRQ (reference signal received quality), RSSI (RSS Indicator), in this paper we focus specifically on reference signal received power (RSRP). This choice is both because RSRP is widely used for the assessment of LTE networks and as a case study that can potentially be applied to prediction of other RSS metrics.

3GPP [8] defines RSRP as the average over the power contributions of the resource elements that carry cell-specific reference signals within the considered frequency bandwidth (*e.g.*, 5 or 10 MHz wideband LTE channels). RSRP is typically reported in dBm by UEs (user equipment) for a 15 KHz sub-carrier. Basically, RSRP excludes interference and noise from other sectors, estimating more accurately the signal power of the serving cell. RSRP (jointly with RSRQ) measurements are mainly utilized by smartphones for cell selection, handover decisions, mobility measurements and power control.

LTE Cells vs. Tracking Areas. LTE networks use several identifiers including: the MCC (mobile country code), MNC (mobile network code), TAC (tracking area code) and the cell ID. The concatenation of the previous IDs results in the cell global identifier (CGI), which uniquely identifies a serving cell. We refer to this unique cell identifier as cell ID or *cID*. LTE also defines Tracking Areas (which we will refer to as LTE TA) by the concatenation of MCC, MNC and TAC, to describe a group of neighboring/overlapping cells, under common LTE management for a specific neighborhood or region.

The RSRP Prediction Problem. Our goal is to predict the RSRP value at a given location, time, and potentially considering additional contextual information (to be defined in Section 2.3), based on available measurement data either in the same cell *cID* or in the same LTE TA.

2.2 Model-Based Prediction: LDPL

There is a large literature on propagation models [5, 32, 38], some of which is reviewed in Section 5. They model the received signal strength given the location of receiver, transmitter and the propagation environment. As a representative baseline from the family of model-based predictors, we consider the Log Distance Path Loss (LDPL) propagation model, which is simple yet widely adopted in the literature [1, 30].

LDPL predicts the power (in dBm) at location \vec{l}_j at distance $\|\vec{l}_{BS} - \vec{l}_j\|_2$ from the transmitting basestation (BS) or cell tower, as a log-normal random variable (*i.e.*, normal in dBm) [1]:

$$P_{cID}^{(t)}(\vec{l}_j) = P_0^{(t)} - 10n_j \log_{10} \left(\frac{\|\vec{l}_{BS} - \vec{l}_j\|_2}{d_0} \right) + w_j^{(t)}. \quad (1)$$

Table 1: Overview of RSRP Prediction Methodologies evaluated in this paper. Methods proposed in this paper are marked in bold. Methods in regular font are prior art, evaluated as baselines for comparison. Methods in light gray font are reviewed but not implemented in this paper.

(1) Model Based (Radio Frequency Propagation Model)	1(a) LDPL: Eq. (1) (Log Distance Path Loss)	1(b) LDPL-knn (heterogeneous PLE)	1(c) WINNER I/II [5] and others: Ray Tracing [38], COST 231 [32].
(2) Geospatial Interpolation	2(a) OK: Ordinary Kriging [6, 29]	2(b) OKD:OK Detrending [6] (hybrid of model and data)	2(c) OKP: OK partitioning [6] (spatial heterogeneous)
Data Driven	3(a) $RF_{s_{x,y}}$ Spatial Features: $\mathbf{x} = (l^x, l^y)$	3(b) $RF_{s_{x,y,t}}$ Spatiotemporal: $\mathbf{x} = (l^x, l^y, d, h)$	3(c) $RF_{s_{all}}$ Full Features: $\mathbf{x} = (l^x, l^y, d, h, dev, \ \vec{l}_{BS} - \vec{l}_j\ _2, freq_{dl}, out)$ (and in some scenarios: cID)
Hybrid	4(4) Combine Model & Data Driven	4(a) Oracle Competitive Ensemble	4(b) Stacking Ensemble: Combine RFs with $LDPL - knn$

The most important parameter is n_j , *i.e.*, the path loss exponent (PLE), which has typical values between 2 and 6 and can be estimated (\hat{n}_j) from the data.¹ We consider two cases.

Homogeneous LDPL: Much of the literature assumes that the PLE n_j is the same across all locations. We can estimate it from Eq. (1) from the training data points.

Heterogeneous LDPL: Some recent work (*e.g.*, [1, 6]) considers that different PLE across locations. We considered several ways to partition the area into regions with different PLEs, and we present knn regression, where we estimate \hat{n}_j from its k nearest neighbors, weighted according to their Euclidean distance, which we refer to as “LDPL-knn”.

2.3 Proposed Data-Driven Prediction: RFs

In this paper, we apply a powerful machine learning framework: Random Forests (RFs) regression. RFs are an ensemble of multiple decision trees [4], which provides a good trade-off between bias and variance by exploiting the idea of bagging. RFs first build multiple decision trees based on sub-samples of the training data and splits between nodes using a random sample of features. For regression, the objective is to minimize the MSE at the terminal leaf. An RSRP value P can be modeled as follows given a set of features vector \mathbf{x} .

$$P|\mathbf{x} \sim \mathcal{N}(RF_{s_{\mu}}(\mathbf{x}), \sigma_{\mathbf{x}}^2) \quad (2)$$

where $RF_{s_{\mu}}(\mathbf{x})$, $RF_{s_{\sigma}}(\mathbf{x})$ are the mean and standard deviation respectively of the RSRP predictor, $\sigma_{\mathbf{x}}^2 = RF_{s_{\sigma}}(\mathbf{x}) + \sigma_{RF_s}^2$ and $\sigma_{RF_s}^2$ is the MSE of the predictor. Finally, the prediction is $\hat{P} = RF_{s_{\mu}}(\mathbf{x})$ since that is the MLE (maximum likelihood estimation) value that minimizes the MSE.

Random Forests are a well-known and successful machine learning model, which we apply for the first time to the RSRP prediction problem. The interesting question is then how to apply it and what are the important features depending on the scenario. For each measurement j in our data, we consider the following full set of features, available via the Android API:

$$\mathbf{x}_j^{\text{full}} = (l_j^x, l_j^y, d, h, cID, dev, out, \|\vec{l}_{BS} - \vec{l}_j\|_2, freq_{dl})$$

- **Location** $\vec{l}_j = (l_j^x, l_j^y)$. These are the spatial coordinates and the only ones considered by previous work on data-driven RSS prediction [6, 22] or in the context of localization [21, 31].

¹ $P_0^{(t)}$ is the received power at reference distance d_0 , calculated by the free-space path loss (Friis) transmission equation for the corresponding downlink frequency, gain and antenna directionality, and \vec{l}_{BS} the location of the transmitting antenna. The log-normal shadowing is modeled by $w_j^{(t)} \sim \mathcal{N}(0, \sigma_w^2(t))$ (in dB), with variance $\sigma_w^2(t)$ assumed independent across different locations. The cell, cID , affects several parameters in Eq. (1), including P_0 , w_j , the locations of transmitting (\vec{l}_{BS}) and receiving (\vec{l}_j) antennas.

- **Time features** $\mathbf{t}_j = (d, h)$, where d denotes the weekday and h the hour of the day that the measurement was collected. Using h as a feature implies stationarity in hour-timescales, which is reasonable for signal strength statistics.

- **The cell ID, cID .** This is a natural feature since RSRP is defined per serving cell. cID was defined in Section 2.1 as the cell global identifier, CGI (*i.e.*, the concatenation of MCC, MNC, TAC and LTE cell ID).

- **Device hardware type, dev .** This refers to the device model (*e.g.*, Samsung Galaxy9 or iPhone X) and *not* to device identifiers. We consider this feature for several reasons. First, there are different noise figures (NF), *i.e.*, electronic interference, and reception characteristics across different devices. Second, the RSRP calculation details differ across devices and manufacturers, since 3GPP just provides generic guidelines. Third, hardware manufacturing affects the mobile sensors output [12].

- **The downlink carrier frequency, $freq_{dl}$.** This is calculated by *EARFCN* (E-UTRA Absolute Radio Frequency Channel Number). We consider this feature because radio propagation and signal attenuation heavily depend on $freq_{dl}$.

- **$out \in \{0, 1\}$** is an approximate indicator of outdoors or indoors location, inferred from Android’s GPS velocity sensor.

- **Euclidean distance** $\|\vec{l}_{BS} - \vec{l}_j\|_2$, of the receiver at location l_j from the transmitting antenna BS (base station or cell tower).

Among the above features, the cell ID cID is particularly important, as it will be demonstrated in Section 4.2.1. It turns out that when there is a large number of measurements with the same cID , it is advantageous to train a separate RFs model per cID , using the remaining features:

$$\mathbf{x}_j^{cID} = (l_j^x, l_j^y, d, h, dev, out, \|\vec{l}_{BS} - \vec{l}_j\|_2, freq_{dl}).$$

When there are a few measurements per cID , then we can treat cID as one of the features in $\mathbf{x}_j^{\text{full}}$.

Why choose RFs for Data-Driven Prediction? We selected RFs regression because it is more a powerful and flexible data-driven prediction framework than the previous state-of-the art, *i.e.*, SpecSense [6], that used geospatial interpolation.

First, RFs can naturally incorporate all aforementioned features, while prior work considered only location features, since geospatial interpolation does not naturally extend to arbitrary features. We denote as $RF_{s_{x,y}}$, $RF_{s_{x,y,t}}$, $RF_{s_{all}}$ the RFs predictors with only spatial (l^x, l^y), spatial (l^x, l^y) and temporal (d, h), and all features, respectively. In Section 4, we assess feature importance in different datasets, using tools inherent to the RFs regression framework.

Second, RFs, by definition, partition the feature space with axis-parallel splits [23]. Examples of decision boundaries produced by $RF_{s_{x,y}}$ is depicted in Fig. 2. One can see the splits according to the spatial coordinates (lat, lng) and the produced areas agree with our knowledge of the placement and direction

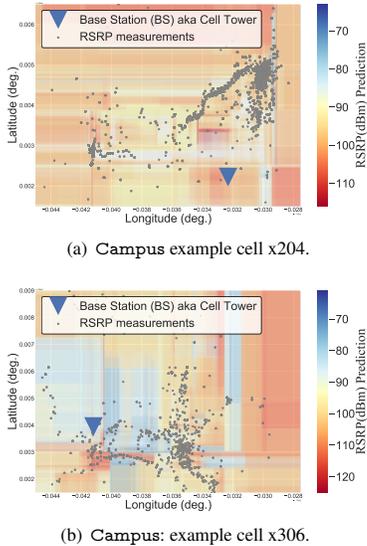


Figure 2: Examples of decision boundaries chosen by $\text{RFs}_{x,y}$. We can see that RFs can naturally identify spatially correlated measurements, *i.e.*, regions with similar wireless propagation characteristics.

of antennas on campus. These splits also happen in multiple dimensions if additional features are considered. Automatically identifying these regions with spatially correlated RSRP comes for free to RFs and is particularly important in the RSRP prediction problem because wireless propagation has different properties across neighborhoods [31]. In contrast, to address this spatial heterogeneity, [6] had to pre-process the data and partition them into disjoint areas with different interpolation parameters, which is a problem with its own extra complexity.

2.4 Baseline: Geospatial Interpolation

State-of-the-art approaches in data-driven RSS prediction [6, 22] have primarily relied on geospatial interpolation. However, this approach is inherently limited to use only spatial features (I^x, I^y) and does not naturally extend to additional dimensions and contextual information. The best representative of this family of predictors is ordinary kriging [22] and its variants [6], which are used as baselines for comparison in this paper.

Ordinary Kriging (OK): It predicts RSS at the testing location $\vec{l}_j = (I_j^x, I_j^y)$ as a weighted average of the K nearest measurements in the training set: $P_j = \sum_{i=1}^K w_i P_i$. The weights w_i are computed by solving a system of linear equations that correlate the test with the training data via the semivariogram function; more details can be found in [6].

Ordinary Kriging Partitioning (OKP): In [6], Voronoi-based partitioning is used to identify regions with the same PLE and apply a different OK model in each region. This is comparable to the heterogeneous propagation model, however it is impractical for city wide signal maps.

Ordinary Kriging Detrending (OKD) [6]: OK assumes spatial stationarity, which does not hold for RSS values. OKD incorporates a simplified version of LDPL in the prediction in order to address this issue [6]. This can be thought as a hybrid approach of data-driven (geospatial) and model-driven (LDPL). It is the best representative of the geospatial predictors and serves as our baseline for comparison in this paper.

2.5 Hybrid Model and Data-Based Prediction

Different predictors may capture different characteristics of RSRP depending on the location and training sample [9]. Fig. 9 and 10(a) present two example cells and demonstrate that there are locations where a modest model-based predictor (LDPL-knn) performs better than our best data-driven predictor (RFs_{all}), and vice versa. The intuition is that in areas with a large number of measurements, RFs outperform all methods, but in areas with limited measurements, a model-based predictor can help. We observe and demonstrate this diversity of data- and model-based predictors in our datasets, and we seek to harvest their diversity via ensemble learning, which comes natural to our machine learning framework. Ensemble learning benefits more from diversity than from individually strong predictors. Therefore, in order to handle the few locations with limited measurements, it suffices to complement our best data-driven predictor (RFs_{all}) with a reasonable model-based predictor (LDPL-knn), to get the best of both worlds.

Oracle Competitive Ensemble. The above intuition suggests that in order to harvest the diversity of model- and data-driven predictors (across locations and training samples), we should pick the best predictor (the one with the smallest variance) for a particular data point. This is essentially a competitive ensembling technique [10, 19], where only the weight of the best predictor is set to one and other weights are set to zero. A practical problem is that we only know the variance of the regressors on the training points. As a baseline for understanding the potential benefit of this approach, we consider the performance of an oracle predictor that has access to all data (test and training), computes the error the model-based (LDPL-knn) and data-driven (RFs_{all}) predictors and picks the best. Table 6 in Section 4 shows that the Oracle Competitive Ensemble indeed outperforms both and provides a lower bound to the error that any practical hybrid can achieve.

Stacking Ensemble. We considered several ensemble approaches to combine two base predictors: one model-based (LDPL) and one data-driven (RFs_{all}). We picked Stacking regression [3]: the predictions from the individual base regression models are used as meta-features to train a meta-regressor. We experimented with several options for the meta-regressor (*e.g.*, LASSO) and we selected RFs that are more shallow than the first-level predictor in order to avoid overfitting problems; more specifically, we use up to 5 levels and 100 ensemble trees.

3 DATA SETS

Table 2 summarizes the two datasets used in this paper. The first is a campus dataset and the second consists of two city-wide datasets from NYC and LA. Fig. 1 depicts some examples.

3.1 Campus dataset

Dataset Overview. We collected the first dataset on a university campus. This Campus dataset is relatively small: 180,000 data points, collected by seven Android devices that belong to graduate students, using 2 cellular providers. In terms of geographical area, it covers approximately 3km^2 , as the devices move between student housing, offices and other locations on campus. Some examples are depicted in Fig. 1(a)-(b).

Although small, this is a dense dataset, with multiple measurements over time on the same or nearby locations. The cells

Table 2: Overview of Signal Maps Datasets used in this study.

Dataset	Period	Areas	Type of Measurements	Characteristics	Source
Campus	02/10/17 - 06/18/17	Univ. Campus Area $\approx 3km^2$	LTE KPIs: RSRP, [RSRQ]. Context: GPS Location, timestamp, dev , cid . Features: $x = (l_j^x, l_j^y, d, h, dev, out, \ \bar{l}_{BS} - \bar{l}_j\ _2)$	No. Cells = 25 No. Meas $\approx 180K$ Density ($\frac{N}{m^2}$) Per Cell: 0.01 - 0.66 (Table 3) Overall Density: 0.06	Ourselves
NYC & LA	09/01/17 - 11/30/17	NYC Metropolitan Area $\approx 300km^2$	LTE KPIs: RSRP, [RSRQ, CQI]. Context: GPS Location, timestamp, dev , cid , EARFCN. Features: $x = (l_j^x, l_j^y, d, h, cid, dev, out, \ \bar{l}_{BS} - \bar{l}_j\ _2, freqat)$	No. Meas NYC $\approx 4.2M$ No. Cells NYC $\approx 88k$ Density NYC-all $\approx 0.014 \frac{N}{m^2}$	Mobile Analytics Company
		LA metropolitan Area $\approx 1600km^2$		No. Meas LA $\approx 6.7M$ No. Cells LA $\approx 111K$ Density LA-all $\approx 0.0042 \frac{N}{m^2}$	

in this dataset exhibit a range of characteristics: (i) the number of measurements N per cell varies from a few thousand up to 50 thousand; (ii) the measurement density (*i.e.*, $\frac{N}{sq\ m^2}$) also varies from 0.01 to 0.6; (iii) the measurements in some cells are concentrated in a few locations while in some others they are dispersed. These (number of measurements, density and dispersion) and other (mean and variance of RSRP) characteristics are reported for each cell of the **Campus** dataset in Table 4.

Data Collection. We developed a user-space app that uses the Android APIs to obtain radio layer and other information needed for RSRP prediction. Although the design of the monitoring system itself is challenging, it is out of this paper’s scope, space and blind review limitations. We only briefly describe the parts of data collection that are relevant to RSRP prediction.

On the device, we use the Android APIs to obtain LTE information: cellular RSRP, network carrier, radio access technology (RAT) to confirm that the network is LTE, and the relevant serving cell information cid as defined earlier. Each measurement is initiated by Android’s notifications/callbacks for network and location changes (*e.g.*, RSS or cell status change) and is also piggy-backed on location change notifications from other apps, in order to achieve a low energy footprint. Rich contextual information is also recorded at the time of the measurement, including: timestamp, device hardware type (dev) and location via the Google Location API, which offers both precision and low energy consumption. The measurements are saved locally in an SQLite database, converted to Javascript Object Notation (JSON) format and uploaded to MongoDB on our server. No personally identifiable information is used in this paper.

3.2 NYC and LA datasets

Dataset Overview. The second type consists of much larger datasets: 10.9M measurements in total, covering approx. $300km^2$ and $1600km^2$ in the metropolitan areas of NYC and LA, respectively, for a period of 3 months (Sep’17 - Nov’17). There are approx. 88,000 and 111,000 unique cell global identifiers (CGIs), in the NYC and LA, respectively.² Key characteristics are summarized in Table 2. An example of the NYC Midtown Manhattan neighborhood is depicted in Fig. 1(c)-(d).

To the best of our knowledge, these are the largest datasets used to date for LTE RSRP, or other RSS metrics prediction, in terms of any metric (number of measurements, geographical scale, number of cells etc). As such, they provide novel insight into the problem at a scale that is relevant to operators and crowdsourcing companies, which is orders of magnitude larger

²It should be noted that many of these cells are overlapping for extra capacity (*e.g.*, different frequency or different sectors by the same cell tower). Moreover, cellular providers share their infrastructure with virtual providers (*i.e.*, MNVOs) which usually have unique MNCs and subsequently create new CGIs.

than the scale previously considered in RSS prediction. Work in [22] uses up to 500 measurements for an area of $0.25km^2$, [6] uses 1500 locations samples from cellular networks for an area $\approx 15km^2$, a university campus area in [29] and ~ 1000 locations sampled at a $7km^2$ urban area in [13]. Work in [9] focuses on RSRQ (rather than RSRP) inference and considers 20,000 data points over approx. $20km^2$ in Edinburgh.

While these are large datasets, is important to note that they are also relatively sparse in space and time. Consider for example the density per cell (an average of approx. 300 measurements per cell) or per cell tower (approx. 500 measurements). The term cell tower refers to the physical location where several antennas are serving multiple different cells, usually indicated by common prefix in cid . There is also large heterogeneity across cells: we consider cells with more than 100 measurements and the maximum number of measurements per cell is in the order of 20,000. There is also sparsity in time: unlike the **Campus** dataset, there are no longer multiple measurements at different times for the same location. Additional data statistics for the data are omitted due to lack of space.

Data Collection. This dataset was collected by a major mobile crowdsourcing and data analytics company and shared with us. The company collects RSS and other KPIs through their measurement SDK, which is integrated into popular third party apps. They crowdsource from a large user base, but they also try to collect measurements infrequently so as to not burden each end-users, which explains the smaller overall density of the dataset compared to our **Campus** dataset, as it can be seen in Table 2. Each location data point is accompanied by rich network and contextual information, except for device or other personal identifiers (which are not stored, for privacy reasons). The details of the company’s collection methodology are both out of the scope of this paper and proprietary (to the company that collected and shared the dataset).

3.3 Common Description of Datasets

Data Format. For the purposes of RSRP prediction, we use the same subset of information from all datasets, *i.e.*, RSRP values and the corresponding features defined in Section 2.1. Measurements from both data sets are converted to GeoJSON format, which offers various advantages (lightweight JSON, compatibility with geospatial software, compact and intuitive representation of location information). A GeoJSON example with some of the KPIs fields (obfuscated) follows:

```

{
  "type": "Feature",
  "properties": {
    "timestamp": "2017-09-11T17:54:35EDT",
    "lteMeasurement": {
      "rsrp": -89,
      "rsrq": -20,
    }
  }
}

```

```

      "pci": 169, "earfcn": 9820},
"cell": { "ci": xxxxx710,
          "mnc": 410,
          "mcc": 310,
          "tac": xx22, "networkType": 4},
"device": {"manufacturer": "samsung",
           "model": "SM-G935P", "os": "android70"},
"locationMetaData": {"city": "New_York",
                     "accuracy": "x", "velocity": "x"}},
"geometry": {"type": "Point", "coords": [-73.9xx, 40.7xx]}}

```

Listing 1: GeoJSON example with LTE KPIs and location, in MongoDB (obfuscated for presentation).

Properties of Datasets. For each dataset, the following metrics describe characteristics that will turn out to affect RSRP prediction, as shown in Section 4.2.

- **Data Density:** Number of measurements per unit area ($\frac{N}{m^2}$).
- **Cells Density:** Number of unique cells (*cids*) per unit area, i.e., $\frac{|C|}{km^2}$. The higher it is, the more *cID* helps as a feature.
- **Dispersion:** In order to capture how concentrated or dispersed are the measurements in an area, we use the Spatial Distance Deviation (*SDD*) metric [20], defined as the standard deviation of the distance of measurement points from the center.

In the next section, we consider the datasets and perform prediction at different granularities: (i) per cell (*cID*) (ii) per Tracking Area (LTE TA). Examples of representative LTA TAs used in our evaluation, are summarized in Table 3.

OpenCellID. Both the LDPL and in the RFs predictors need the distance between the transmitting antenna and the receiver’s location (where RSRP is measured or predicted), $\|\vec{l}_{BS} - \vec{l}_j\|_2$. To that end, we lookup the location of the base-station, \vec{l}_{BS} , using the public APIs of a popular online crowd-sourced database `opencellid.org`. This is the only external information we need in addition to the main RSS datasets.

4 PERFORMANCE EVALUATION

We evaluate all predictors of Section 2 (both state-of-the-art and our own RFs-based ones) over the datasets of Section 3. Along the way, we provide insights into the prediction performance and into tuning the framework depending on the dataset.

4.1 Setup

4.1.1 RFs Setup. RFs require less tuning compared to prior-art techniques (e.g., estimating the parameters of the semi-variogram [6], lag [6] and spatiotemporal correlation matrices per environment [13]). The most important hyper-parameters for RFs are the number of decision trees (i.e., n_{trees}) and the maximum depth of each tree (i.e., max_{depth}). We used a grid search over the parameter values of the RFs estimator [28] in a small hold-out part of the data to select the best values. For the **Campus** dataset, we select $n_{trees} = 20$ and $max_{depth} = 20$ via 5-Fold Cross-Validation (CV); larger max_{depth} values could result in overfitting of RFs. For the **NYC** and **LA** datasets, we select $n_{trees} = 1000$ and $max_{depth} = 30$; more and deeper trees are required for larger datasets.

As discussed in Section 2.3, one important design choice is what granularity we choose to build our RFs models: per *cID* or per LTE TA (as defined in Section 2.1).

Training per *cID*: We can train a separate RFs model per cell (*cID*) using all features except *cID* (\mathbf{x}_j^{cID}). This is natural since RSRP is defined per serving cell (Section 2.1) but requires

Table 3: NYC and LA datasets: LTE TAs Examples.

	NYC (MNC-1) Manhattan Midtown	NYC (MNC-1) E. Brooklyn	LA (MNC-2) Southern
No. Measurements	≈ 63K	≈ 104K	≈ 20K
Area km^2	$1.8km^2$ (Fig. 1 (c-d))	$44.8 km^2$	$220 km^2$
Data Density $\frac{N}{m^2}$	≈ 0.035	≈ 0.002	≈ 0.0001
No. Cells $ C $	429	721	353
Cell Density $\frac{ C }{km^2}$	238.3	16.1	1.6

a large number of measurements per cell, which is the case in **Campus** dataset but not in **NYC** and **LA** datasets.

Training per LTE TA: Another option is to train one RFs model per Tracking Area (LTE TA), and use *cID* as one of the features in (\mathbf{x}_j^{full}). This is particularly useful in the **NYC** dataset, where there are less measurements for the same cell unit area, insufficient to train a model per *cID*. However, in urban areas, there is very high cells density in a region and data points from different cells in the same LTE TA can still be useful.

4.1.2 Baselines’ Setup. *LDPL methods.* For Eq. (1): we compute the distance from the base station using the online database from `opencellid.org`; breaking distance $d_0 = 1m$; *freqdl* is obtained from the *EARFCN* measurement. In addition, for *LDPL-knn*: we select empirically $k = 100$ neighbors for the **Campus** dataset and $k = 10\%$ of the training data-points in each cell for the **NYC** and **LA** datasets.

Geospatial Predictors. The number of neighbors was empirically set to $k = 10$. For geospatial methods, a larger k selected for *LDPL-knn* did not show any significant improvement, and it would result in much higher computational cost. An exponential fitting function of the semi-variogram function $\gamma(h)$ was selected [6]; the maximum lag (h) was set to 200m, as in [6], for the **Campus** and **NYC** environments, while it was set to 600m for the **LA** suburban environment. The approximated empirical semi-variogram $\hat{\gamma}(h)$ was calculated per 10m [6].

4.1.3 Splitting Data into Training and Testing. We select randomly 70% of the data as the training set $\{\mathbf{X}_{train}, \mathbf{P}_{train}\}$ and 30% of the data as the testing set $\{\mathbf{X}_{test}, \mathbf{P}_{test}\}$ for the problem of predicting missing RSRP values. The results are averaged over $S = 5$ independent random splits. These default choices are used unless otherwise stated. An exception is Fig. 6, where we vary the size of training set and we show that our RFs-based predictors degrade slower than baselines with decreasing training size.

4.1.4 Evaluation Metrics. We evaluate the performance of the predictors in terms of absolute error (RMSE) and Relative Improvement (ARI) as well as feature importance in RFs.

Root Mean Square Error (RMSE): If \hat{P} is an estimator for P , then $RMSE(\hat{P}) = \sqrt{MSE(\hat{P})} = \sqrt{E((P - \hat{P})^2)}$, in dB. We report *RMSE* for each predictor at different levels of granularity, namely: (i) per *cID* (ii) per LTE TA (in **NYC** and **LA**) or (iii) over the entire dataset (**Campus**).³

Absolute Relative Improvement (ARI): This captures the improvement of each predictor over the variance in the data:

³If we use RFs model per cell, denote \widehat{P}_j^{cID} the prediction for the measurement j , with the dedicated RFs model for that specific *cID*. Then for each cell, $MSE_{cID} = \frac{1}{N_{cID}} \sum_j^{N_{cID}} (P_j^{cID} - \widehat{P}_j^{cID})^2$ while for all data points $MSE_{all} = \frac{1}{N} \sum_j^N (P_j^{cID} - \widehat{P}_j^{cID})^2$.

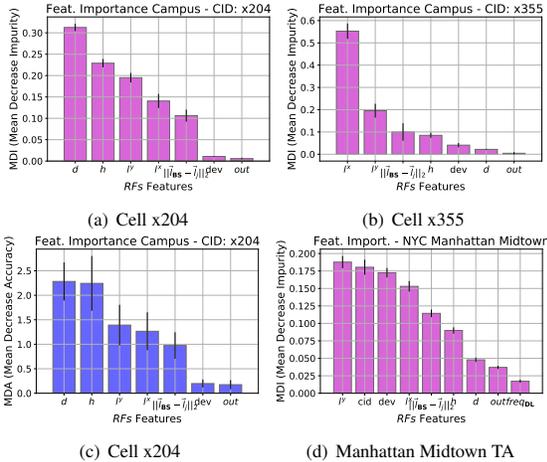
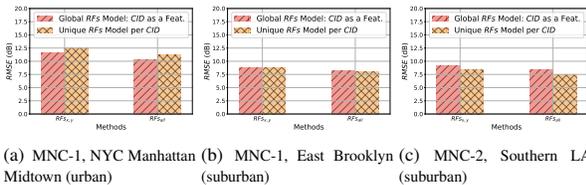


Figure 3: For Campus dataset (a), (b), (c): Feature Importance for two distinct cells (Rf's built per distinct cell). Cells' data are depicted in Fig. 1. For NYC dataset, (d) shows the MDI score for one LTE TA for MNC-1.



(a) MNC-1, NYC Manhattan (urban) (b) MNC-1, East Brooklyn (suburban) (c) MNC-2, Southern LA Midtown (suburban)

Figure 4: RMSE in NYC and LA datasets. This figure makes multiple comparisons: (1) urban vs suburban LTE TAs; (2) cID as feature vs. training a different Rf's model per cID ; (3) providers MNC-1 vs. MNC-2.

$ARI = 1 - \frac{1}{|C|} \sum_{i \in C} \frac{MSE_i}{Var_i}$, where $|C|$ is the number of the different cells in the dataset, and Var_i is cell i 's variance.

Mean Decrease Impurity (MDI), *a.k.a.* Gini Importance: This essentially captures how often a feature is used to perform splits in Rf's. It is defined as the total decrease in node impurity, weighted by the probability of reaching that node (approximated by the proportion of samples reaching that node), averaged over all trees in the ensemble [28].

Mean Decrease Accuracy (MDA), *a.k.a.* Permutation Importance: It measures the predictive power of each feature. The values of that feature are randomly permuted, and we measure the decrease in accuracy, when we predict with the remaining features and average over all trees in Rf's.

4.2 Results

4.2.1 Feature Importance.

a. Campus dataset: We train one Rf's model per cID for the set of features $\mathbf{x} = (l_j^x, l_j^y, d, h, \|\vec{l}_{BS} - \vec{l}_j\|_2, out, dev)$. We assess their importance w.r.t. MDI and MDA and representative results are shown on Fig. 3. We observe that, in cells with high data density and low dispersion, the most important are the time features (d, h) w.r.t. to both metrics. An example of such a cell is x204, which has $SDD = 325$, density=0.66 points/ m^2 and is depicted in Fig. 1(a). We see that (d, h) are the top features for this cell w.r.t. both MDI and MDA, as shown in Fig. 3(a) and Fig. 3(c), respectively. For the rest of the paper, we only report feature importance w.r.t. MDI. We also inspected the decision trees produced and these features are indeed being used at the higher levels of the decision trees.

Table 4: Campus dataset: Comparing Predictors per cell

cID	Cell Characteristics					RMSE (dB)						
	N	$\frac{N}{sq\ m^2}$	SDD	$\mathbb{E}[P]$	σ^2	LDPL-hom	LDPL-kNN	OK	OKD	Rf's x, y	Rf's x, y, t	Rf's all
x312	10140	0.015	941	-120.6	12.0	17.5	1.63	1.70	1.37	1.58	0.93	0.92
x914	3215	0.007	791	-94.5	96.3	13.3	3.47	3.59	2.28	3.43	1.71	1.67
x034	1564	0.010	441	-101.2	337.5	19.5	7.82	7.44	5.12	7.56	3.82	3.84
x901	16051	0.162	355	-107.9	82.3	8.9	4.60	4.72	3.04	4.54	1.73	1.66
x204	55566	0.666	325	-96.0	23.9	6.9	3.84	3.85	2.99	3.83	2.30	2.27
x922	3996	0.107	218	-102.7	29.5	5.6	3.1	3.16	2.01	3.10	1.92	1.82
x902	34193	0.187	481	-111.5	8.1	21.0	2.60	2.47	1.64	2.50	1.37	1.37
x470	7699	0.034	533	-107.3	16.9	24.8	3.64	2.73	1.87	2.78	1.26	1.26
x915	4733	0.042	376	-110.6	203.9	14.3	7.54	7.39	4.25	7.31	3.29	3.15
x808	12153	0.035	666	-105.1	7.7	4.40	2.41	2.42	1.60	2.34	1.75	1.59
x460	4077	0.040	361	-88.0	32.8	11.2	2.35	2.28	1.56	2.31	1.84	1.84
x306	4076	0.011	701	-99.2	133.3	18.3	4.85	4.30	2.80	3.94	3.1	3.06
x355	30084	0.116	573	-94.3	42.6	9.3	2.42	2.31	1.85	2.26	1.79	1.79

On the contrary, for more dispersed and less dense cells, such as cell x355 ($SDD = 573, 0.116N/m^2$, map in Fig. 1(b)), the location (l_j^x, l_j^y) is naturally the most important, as confirmed in Fig. 3(b). Feature importance for dev and out are close to zero, which is expected because of the small number of devices in the Campus dataset.

b. NYC and LA datasets: In this case, $freqdl$ is available and the datasets contain thousands of cells. We start with a Rf's model per LTE TA. As a representative example, we report the feature importance, in Fig. 3(d), for the LTE TA of a major mobile network carrier (MNC-1) located in NYC Midtown Manhattan and already depicted in Fig. 1(c)-1(d). The most important features turn out to be the spatial features (l_j^x, l_j^y) as well as the cell cID . This is because the data are sparser and the whole LTE TA is served by geographically adjacent or overlapping cells.

We also investigated whether we should train a separate Rf's per cID , or cID should be used as one of the features in a single Rf's. For a representative urban LTE TA (Manhattan Midtown), in Fig. 4(a) we calculate the RMSE for two cases: (i) when cID is used as a feature in a single Rf's per LTE TA and (ii) when a separate Rf's model is produced per cell. Interestingly, the prediction is better when cID is utilized as a feature. Given the sparsity of the data and the high overlap of the cells, Rf's benefit from the features of the additional measurements. Manhattan Midtown has a cells density of 238 per km^2 at it can be seen in Table 3: the cell size does not exceed the size of a few blocks or sometimes there are multiple cells within a skyscraper. On the contrary, for the suburban LA dataset, where the cells are not so densely deployed, a unique Rf's model per cell performs better than Rf's per LTE TA, as shown in Fig. 4(c). Likewise in the Campus (lower density than NYC) Rf's model per cID did better than using as a feature in a single Rf's model for the entire LTE TA. Similar results and findings were observed for the rest of cells and TAs, but are omitted due to space constraints. In general, Rf's trained per cID is usually a better option, but cID should be used as a feature in urban areas with high cells density.

4.2.2 Comparing RSRP Predictors. We compare the performance of the Rf's prediction framework against state-of-the-art geospatial interpolation techniques (OK and OKD) as well as model-driven techniques (LDPL-knn and LDPL-hom).

a. Campus dataset: Table 4 reports the RMSE for all predictors for each cell in the Campus dataset, and for the default 70-30% split. Fig. 5 compares all methods but calculating RMSE over the entire Campus dataset, instead of per cell. In both cases, we can see that our Rf's-based predictors

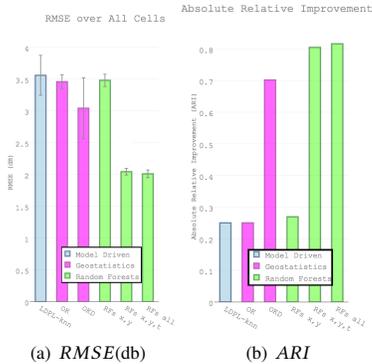


Figure 5: Comparison of all predictors over the entire Campus dataset (all points, all cells).

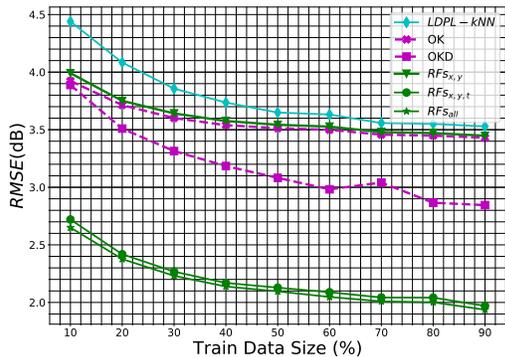
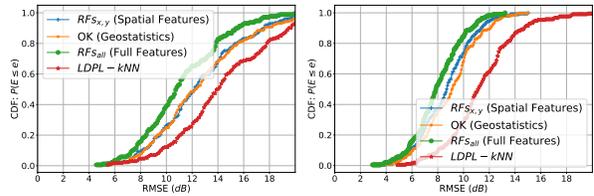


Figure 6: Campus dataset: $RMSE$ vs. Training Size. Our methodology (RFS with more than spatial features, *i.e.*, RFS _{x,y,t} , RFS_{all}) significantly improves the $RMSE$ -cost tradeoff: it can reduce $RMSE$ by 13% for the same number of measurements compared to state-of-the-art data-driven predictors (OKD); or it can achieve the lowest error possible by OKD (≈ 2.8 dB) with 10% instead of 90% (and 80% reduction) of the measurements.

outperform model (LDPL) and other data-driven (OK, OKD) predictors, as long as they use more features than just location.

Fig. 6 shows the $RMSE$ as a function of the training size (as % of all measurements in the dataset). First, the performance of OK and RFS _{x,y} is almost identical, as it can be seen for $RMSE$ over all measurements (Fig. 6 and Fig. 5) and $RMSE$ per cell (Table 4). This result can be explained by the fact that both predictors are essentially a weighted average of their nearby measurements, although they achieve that in a different way: OK finds the weights by solving an optimization problem while RFS _{x,y} uses multiple decision trees and data splits. Second and more important, considering additional features can significantly reduce the error. For the Campus dataset, when time features $t = (d, h)$ are added, RFS _{x,y,t} significantly outperforms OKD: it decreases $RMSE$ from 0.7 up to 1.2 dB. Alternatively, in terms of training size, RFS _{x,y,t} needs only 10% of the data for training, in order to achieve OKD’s lowest error (≈ 2.8 dB) with 90% of the measurement data for training. Our methodology achieves the lowest error of state-of-the-art geospatial predictors with 80% less measurements. The absolute relative improvement of RFS _{x,y,t} compared to OKD is 13%.

b. NYC and LA datasets: Fig. 7 shows the error for two different LTE TAs, namely for NYC Manhattan Midtown (urban) and for southern LA (suburban), where RFs have been trained per cID . CDFs of the error per cID of the same LTE TA are plotted for different predictors. Again, OK performance



(a) NYC Manhattan Midtown (b) LA Southern Suburb

Figure 7: NYC and LA datasets: CDFs for $RMSE$ per cID for two different LTE TAs, for the same major MNC carrier-1. RFS_{all} offer 2dB gain over the baselines for the 90th percentile.

is very close to RFS _{x,y} , because they both exploit spatial features. However, RFS_{all} with the rich set of features improves by approx. 2dB over the baselines for the 90th percentile, in both LTE TAs. Interestingly, the feature *dev* is now important, which is expected in this crowdsourced data, which has high heterogeneity of devices reporting RSRP.

There are multiple reasons why RFS_{all} outperform geospatial interpolation predictors. The mean and variance of RSRP depend on time and location and the complex propagation environment. RFs can easily capture these dimensions instead of modeling a priori every single aspect. For example, RFS _{x,y,t} predicts a time-varying value for the measurements at the same location in Fig. 1(a), while RFS _{x,y} or OK/OKD produce just a flat line over time. OK also relies on some assumptions (same mean over space, semivariogram depending only on the distance between two locations), which do not hold for RSRP. Even hybrid geospatial techniques (OKD) cannot naturally incorporate additional features (*e.g.*, time, device type, etc.).

4.2.3 Assessing location density and overfitting. In the Campus dataset, we observed that a significant fraction of the data comes from a few locations, *i.e.*, from participating grad students’ home and work. In other words, many data-points were reported from the same or nearby locations over time, which begs the question whether this leads to overfitting of RFs to those oversampled location. We investigated this question and found that our RFs predictors neither get a performance boost nor overfit. To that end, we utilize HDBSCAN, a state-of-the-art clustering algorithm, to identify very dense clusters of location that account for a significant number of measurements (cluster size 5% of the cell’s data). We refer to data from those locations as “dense”; we remove them and we refer to the remaining ones as “sparse-only” data. Fig. 8 reports the $RMSE$ of different methods when training and testing is based on (i) all-data, (ii) sparse-only data and (iii) sparse-only data with a 5% randomly sampled from the dense data. It can be clearly seen that our RFS _{x,y,t} and RFS_{all} have similar performance in all scenarios and consistently outperform baselines. OK and LDPL-knn’s errors slightly decrease in sparse data (OK cannot handle repeated locations) but are still higher than our error. Table 5 reports the error per cell for sparse-only data, and our proposed predictors outperform baselines in a cell-by-cell basis.

4.2.4 Lessons from different datasets. When possible, we already provided insights w.r.t. the characteristics of the datasets (*e.g.*, different density and dispersion) and their effect on feature selection and selection of training block of RFs (cID or LTE TA). We would like to further discuss the

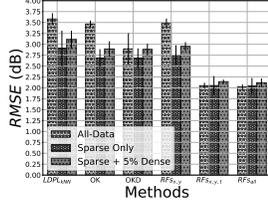


Figure 8: Our Approaches ($RFs_{x,y,t}$, RFs_{all}) outperform prior art in all the different sampling distributions we considered.

Table 5: Campus dataset: Comparing Predictors per cell, considering only sparse measurements (i.e., after removing measurements which create dense clusters).

cID	N	SDD	RMSE(dB)					
			LDPL kNN	OK	OKD	$RFs_{x,y}$	$RFs_{x,y,t}$	RFs_{all}
x312	4852	1240	1.66	1.49	1.46	1.62	1.06	0.91
x914	858	922	5.08	4.94	5.09	5.04	3.38	3.33
x034	514	532	6.94	6.52	6.59	6.6	5.52	5.25
x901	1549	218	3.07	2.79	2.86	2.9	1.9	1.97
x204	13099	535	2.53	2.48	2.46	2.57	1.93	1.93
x922	1927	309	3.62	3.66	3.56	3.66	2.13	2.17
x902	7589	245	2.45	2.06	1.89	2.08	0.92	0.92
x470	1357	431	3.72	0.75	1.51	0.79	0.48	0.52
x915	785	345	5.17	4.81	4.78	4.94	4.27	4.34
x808	5655	972	2.43	2.36	2.41	2.46	1.95	1.82
x460	1176	347	3.35	3.38	3.47	3.43	3.23	3.23
x306	1382	1131	5.84	5.15	5.13	5.34	4.14	4.3
x355	15356	790	2.68	2.54	2.53	2.58	2.04	2.03



Figure 9: Diversity of model and data predictors in our Campus dataset. The color indicates the magnitude of the difference between the model (LDPL-knn) and data (RFs) driven predictors. Yellow (lighter): LDPL-knn performs better, Black (darker): $RFs_{x,y,t}$ performs better. Color in the middle indicates mixed results.

effect of the Radio Frequency propagation environment in the prediction error. On the one hand, the Campus dataset has an average error of 2.2 dB while on the other hand the NYC LTE TA for Manhattan Midtown (see Table 3) has an average $RMSE$ of approximately 10dB (see Fig. 4(a)). The former is a sub-urban campus with very dense measurements in a small area, while the latter exhibits harsh wireless propagation conditions because of Manhattan skyscrapers, large number of people etc. It is interesting to note that, although the data density (number of measurements per m^2) is comparable (e.g., ≈ 0.035 for both NYC Midtown LTE TA and cell x808 in Campus), the Campus dataset has 180 thousand measurements for 13 cells, while LTE TA for NYC Midtown has approx. 63K measurements for 429 cells, thus less measurements per cell. For less harsh propagation environments such as Brooklyn (Fig. 4(b)) or suburban LA (Fig. 4(c)), the error decreases to approx 7.5dB, i.e., within the range of one signal bar (see Section 4.3).

4.2.5 Hybrid Model and RFs predictors. Although the data-driven approach generally outperforms model-based predictors, there are areas in our datasets (typically those with few measurements) where the reverse is true. An example is depicted in Fig. 9: it compares the test error of LDPL-knn with $RFs_{x,y,t}$ in different locations and demonstrates the diversity

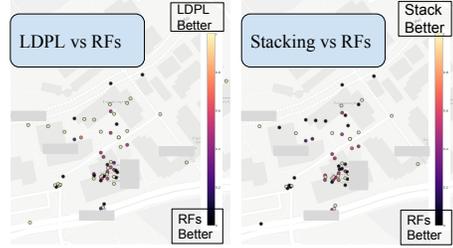


Figure 10: Campus dataset, zooming in cell x034 to understand diversity of predictors. The color of the dots indicates locations whether LDPL or Stacking Ensemble (light color) or RFs (dark color) is a better predictor. Left: LDPL-knn outperforms RFs in some locations (typically those with few measurements). Right: Stacking Ensemble manages to exploit that diversity to outperform RFs on those same locations.

of model and data-driven predictors in our datasets. As described in Section 2.5, a hybrid predictor should be able to harvest that diversity and achieve the best of both worlds.

The Oracle Competitive Ensemble, defined in Section 2.5, is the best one can hope for and puts a lower bound to the error that any hybrid predictor might hope to achieve. Table 6 compares the model-driven, data-driven and hybrid predictors for each cell of the Campus dataset. Although, at the cell level, $RFs_{x,y,t}$ performs better than LDPL-knn, the Oracle Competitive Ensemble is able to significantly outperform both. This is because, RFs does better in most but not all locations, when looking into finer granularity than the entire cID . Examples of zooming in two cells where predictors exhibit such diversity are shown in Fig. 9 and Fig. 10.

As expected, the practical Stacking Ensemble (also defined in Section 2.5) cannot achieve as low an error as Oracle Competitive Ensemble, but it performs better than either LDPL-knn or $RFs_{x,y,t}$. In Table 6, the last column (Gain) reports the reduction in error from $RFs_{x,y,t}$ to Stacking Ensemble and is small when reported across the entire cell. However, when we zoom in finer granularity, we can clearly see the reduction of error in a few locations. For example, Fig. 10 shows an example of relatively sparse and dispersed measurements for cell x034, where Stacking Ensemble outperforms $RFs_{x,y,t}$ for the majority of data points. The benefit of using the model can be seen outdoors and near buildings where fewer measurements were taken, which agrees with our intuition.

4.3 Discussion

Summary of Findings. We demonstrated the following facts. First, the RFs-based predictors outperform state-of-the art data-driven predictors in all scenarios, when more features beyond just location are considered. Second, RFs significantly outperform propagation models, except for a few areas with limited number of measurements, in which cases a hybrid predictor can combine the best of both worlds. Third, the most important features were primarily cID , location, time, device type. Finally, when RSRP prediction is desired for city-wide signal maps, we should train a separate RFs model per cell, when there is a large number of data points per cell, otherwise we should use cID as a feature. The latter is the case in large-scale crowdsourcing scenarios. Overall our RFs-based predictors offer superior performance (accurate RSRP predictions with less measurements/cost) and our methodological framework is general and extensible (can naturally be incorporating several features and ensemble methods).

Table 6: Campus dataset: Comparing the model-based (LDPL-knn), one of the best data-based (RFs_{x, y, t}), and hybrid (Oracle Competitive Ensemble, Stacking Ensemble) predictors, per cell.

cID	RMSE (dB)				
	LDPL kNN	RFs x, y, t	Oracle	Stack	Gain
x312	1.51	0.98	0.66	0.93	0.05
x914	3.49	1.96	1.66	1.93	0.03
x034	8.15	5.01	3.16	4.78	0.24
x901	4.72	1.86	1.5	1.82	0.04
x204	3.86	2.31	1.83	2.31	0.01
x922	3.22	2.16	1.73	2.13	0.03
x902	2.56	1.39	1.12	1.39	0.00
x470	3.77	1.27	0.94	1.26	0.01
x915	7.33	3.26	2.61	3.14	0.12
x808	2.41	1.83	1.42	1.82	0.01
x460	2.24	1.89	1.63	1.85	0.04
x306	5.15	3.53	2.43	3.51	0.02
x355	2.36	1.90	1.56	1.88	0.02

Importance of RSRP Prediction and Magnitude of Error. We argue that the reduction in prediction error (on the order of a couple of dBs, depending on the scenario) is significant. Absolute RSRP values, not only translate to network performance bars⁴ but also determine the performance of VoLTE. Work in [17] showed that the probability of occurrence of VoLTE problems is approx. 2 – 5% for RSRP ∈ [−105, −110), increases to 5 – 10% for [−115, −120) range and it goes up to more than 15% for RSRP ≤ −120 dBm. Since dropped or interrupted calls are the main customer complaint, a reduction of RSRP prediction error by a few dB can help accurately identify such regions (e.g., see RFs_{all} low error in cells x312, x902 in Table 4).

5 RELATED WORK

Signal Strength is a fundamental property of wireless networks, relevant in many application contexts. The most closely related work elaborated upon in Section 2 and used as baselines. Broadly speaking, signal strength prediction can be done through propagation models or through data-driven approaches.

Propagation models. Radio frequency propagation and path loss (equation-based) modeling has been extensively studied. Popular models of this family are the Hata model [34], the COST 231 [32] (e.g., Walfisch-Ikegami model), WINNER I/II [5] and the recent Ray tracing [38] which offers high accuracy. However, this family of models requires a detailed map of the environment (e.g., topology, street width, antennas’ height, no. of floors, sometimes buildings 3D maps), fine grained tuning and is computational expensive. A simple yet widely used [1, 6] propagation model is the Log Distance Path Loss (LDPL) model [30], which assumes shadowing following a log-normal distribution and path loss following logarithmic attenuation. We use the heterogeneous LDPL, i.e., a different path loss parameter per location, as the representative baseline of propagation models. We defer to Section 2.2 for details.

Data-Driven Predictors. A body of prior work, uses geospatial interpolation [6, 9, 22, 29] for RSS prediction, where RSS at a particular location is predicted by interpolating neighboring measurements. Work in [29] (WiMax data) and [6] (wideband RSSI) have developed methods which incorporate wireless propagation characteristics in geospatial models, namely Ordinary Kriging with Detrending (OKD) and regions partitioning (OKP, OKPD), already discussed in Section 2.4. As we described conclusively in this paper, geospatial predictors cannot naturally incorporate additional dimensions such as time,

⁴Although the mapping from RSRP values to signal bars differs across devices, typical RSRP values for iOS and Android devices are the following ranges: for 1-bar RSRP are below −115 dBm, for 2-bars RSRP is in [−114, 105] dBm, for 3-bars RSRP values ∈ [−104, −85] and finally for 4 signal bars (excellent reception) RSRP values are higher than −85 dBm.

frequency, hardware and network information, as our proposed work does. We defer to Section 2.4 and Section 4 for details.

Work in [13] uses Bayesian Compressive Sensing (BCS) to develop a framework for inference of missing signal strength values jointly with users incentives control. However, it requires fined tuned spatial and temporal correlation matrices per location, which is limiting for city-wide scale RSRP prediction. Similarly, a very detailed 3D map information from Light and Range Detection Data (LiDAR) is required for [7], which utilizes deep neural nets for RSRP prediction. Last but not least, prior signal map prediction work has been in much smaller geographical scale than ours: a 0.25km² in [22], a university campus in [29], a 7km² area in [13] and a 15km² area [6].

Localization. RSS modeling [14] is also important in the context of UE localization. Work in [31] develops UE localization algorithms based on UE Measurement Data (UMD), where the LTE RSRP likelihood is modeled via RFs, with training features only the measurements’ latitude and longitude. Similarly, [21] focuses on UE localization, and builds synopsis of RF coverage maps in order to facilitate the localization process.

Tools and datasets. There is a number of proprietary measurement tools that can collect cellular performance measurements from end-devices (e.g., [35], [27]). Tools in the research community include Mobilyzer [25] and Mobiperf App [15] with active and passive measurements, but do not include the cell-IDs and precise location information used in this paper. Moreover, datasets in the popular repository Crawdad [18] did not include large scale RSRP data. In order to obtain insights regarding RSRP, we developed the tool for collecting Campus dataset. The NYC and LA datasets used in this paper, were provided to us by a major mobile data analytics company (name omitted for anonymization purposes); it provides RSRP and several crucial features used in this paper, at metropolitan scale (see Section 3.2 for scale’s comparison with prior work).

Other. Work in [20] studies basestation localization and extensively analyzes crowdsourced signal strength data provided by `opencellid.org`. Finally, a category of work related to RSS measurements is spectrum monitoring [24, 40] and databases of cellular, DTV and radar bands [6, 39], for cognitive radio modeling [26] and spectrum sharing.

6 CONCLUSION

We developed a machine learning framework for cellular signal strength prediction, which is important for creating signal maps in a cost-efficient way, necessary for future 5G and IoT deployments. We used the powerful tools of random forests and ensemble learning, which we adapted in this context by evaluating different features and meta-features. We compared and evaluated different methods over different datasets. Some of the datasets under study are the largest used in this context and provide unique insight into city-wide signal map prediction.

Directions for future work include the following. On the methodological side, we plan to further explore ensemble learning and design methods specifically for this problem, in order to fully harvest the diversity of propagation and RFs-based predictors. On the application side, we plan to address problems related to RSRP prediction, including recommendations on where/when to collect the next measurements, understanding the impact of prediction error to applications (e.g., VoLTE) performance, and inference from signal strength maps to assist troubleshooting and deployment.

REFERENCES

- [1] Emmanouil Alimpertis, Nikos Fasarakis-Hilliard, and Aggelos Bletsas. 2014. Community RF sensing for source localization. *IEEE Wireless Comm. Letters* 3, 4 (2014), 393–396.
- [2] Tugce Bilen, Berk Canberk, and Kaushik R Chowdhury. 2017. Handover Management in Software-Defined Ultra-Dense 5G Networks. *IEEE Network* 31, 4 (2017), 49–55.
- [3] Leo Breiman. 1996. Stacked regressions. *Springer Machine learning* 24, 1 (1996), 49–64.
- [4] Leo Breiman. 2001. Random Forests". *Machine Learning* 45, 1 (Oct. 2001), 5–32.
- [5] Yvo de Jong Bultitude and Terhi Rautiainen. 2007. IST-4-027756 WINNER II D1. 1.2 V1. 2 WINNER II Channel Models. *EBITG, TUI, UOULU, CU/CRC, NOKIA, Tech. Rep., Tech. Rep* (2007).
- [6] Ayon Chakraborty, Md Shaifur Rahman, Himanshu Gupta, and Samir R Das. 2017. SpecSense: Crowdsensing for Efficient Querying of Spectrum Occupancy. In *Proc. of the IEEE INFOCOM*. Atlanta, Georgia, USA.
- [7] Rita Enami, Dinesh Rajan, and Joseph Camp. 2018. RAIK: Regional analysis with geodata and crowdsourcing to infer key performance indicators. (April 2018), 1–6.
- [8] ETSI. 2015. LTE, Evolved Universal Terrestrial Radio Access (E-UTRA), Physical layer Measurements (3GPP TS 36.214 version 12.2.0 Release 12. <http://www.etsi.org/>.
- [9] Mah Rukh Fida, Andra Lutu, Mahesh K. Marina, and Ozgu Alay. 2017. Zip-Weave: Towards efficient and reliable measurement based mobile coverage maps. In *Proc. of the IEEE INFOCOM*. Atlanta, Georgia, USA.
- [10] Yakov Frayman, Bernard F Rolfe, and Geoffrey I Webb. 2002. Solving Regression Problems Using Competitive Ensemble Models. *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence* (2002), 511–522.
- [11] Ana Gomez-Andrades, Raquel Barco, Pablo Munoz, and Inmaculada Serrano. 2017. Data analytics for diagnosing the RF condition in self-organizing networks. *IEEE Transactions on Mobile Computing* 16, 6 (2017), 1587–1600.
- [12] Andreas Grammenos, Cecilia Mascolo, and Jon Crowcroft. 2018. You Are Sensing, but Are You Biased?: A User Unaided Sensor Calibration Approach for Mobile Sensing. 2, 1 (2018), 11.
- [13] S. He and K. G. Shin. 2018. Steering Crowdsourced Signal Map Construction via Bayesian Compressive Sensing. In *Proc. of the IEEE INFOCOM*. Honolulu, HI, USA.
- [14] Tian He, Chengdu Huang, Brian M Blum, John A Stankovic, and Tarek Abdelzaher. 2003. Range-free localization schemes for large scale sensor networks. In *Proc. of the ACM MobiCom*. 81–95.
- [15] Junxian Huang, Cheng Chen, Yutong Pei, Zhaoguang Wang, Zhiyun Qian, Feng Qian, Birjodh Tiwana, Qiang Xu, Z Mao, Ming Zhang, et al. 2011. Mobiperf: Mobile network measurement system. *Technical Report. University of Michigan and Microsoft Research* (2011).
- [16] Ali Imran, Ahmed Zoha, and Adnan Abu-Dayya. 2014. Challenges in 5G: How to empower SON with big data for enabling 5G. *IEEE network* 28, 6 (2014), 27–33.
- [17] Yunhan Jack Jia, Qi Alfred Chen, Zhuoqing Morley Mao, Jie Hui, Kranthi Sontinei, Alex Yoon, Samson Kwong, and Kevin Lau. 2015. Performance characterization and call reliability diagnosis support for voice over LTE. In *Proc. of the ACM MobiCom*. 452–463.
- [18] David Kotz, Tristan Henderson, Ilya Abyzov, and Jihwang Yeo. 2009. CRAWDAD dataset dartmouth/campus (v. 2009-09-09). Downloaded from <https://crawdad.org/dartmouth/campus/20090909>. <https://doi.org/10.15783/C7F59T>
- [19] Ludmila I Kuncheva. 2002. Switching between selection and fusion in combining classifiers: An experiment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 32, 2 (2002), 146–156.
- [20] Zhijing Li, Ana Nika, Xinyi Zhang, Yanzi Zhu, Yuanshun Yao, Ben Y Zhao, and Haitao Zheng. 2017. Identifying value in crowdsourced wireless signal measurements. In *Proc. of the ACM World Wide Web Conference (WWW)*. 607–616.
- [21] Robert Margolies, Richard Becker, Simon Byers, Supratim Deb, Rittwik Jana, Simon Urbanek, and Chris Volinsky. 2017. Can you find me now? Evaluation of network-based localization in a 4G LTE network. In *Proc. of the IEEE INFOCOM*. IEEE, 1–9.
- [22] Massimiliano Molinari, Mah-Rukh Fida, Mahesh K Marina, and Antonio Pescape. 2015. Spatial Interpolation Based Cellular Coverage Prediction with Crowdsourced Measurements. In *Proc. of the ACM SIGCOMM Workshop on Crowdsourcing and Crowdsourcing of Big Internet Data (C2BID)*. London, UK, 33–38.
- [23] Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts.
- [24] Ana Nika, Zengbin Zhang, Xia Zhou, Ben Y Zhao, and Haitao Zheng. 2014. Towards commoditized real-time spectrum monitoring. In *Proceedings of the ACM workshop on Hot topics in wireless*. ACM, 25–30.
- [25] A. Nikraves, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao. 2015. Mobilyzer: An Open Platform for Controllable Mobile Network Measurements. In *Proc. of the ACM MobiSys*. 389–404.
- [26] Oluwaseyi Omotere, Lijun Qian, Riku Jantti, Miao Pan, and Zhu Han. 2017. Big RF Data Assisted Cognitive Radio Network Coexistence in 3.5 GHz Band. In *Proc. of the IEEE ICCCN*. 1–8.
- [27] Open Signal Inc. 2011. 3G and 4G LTE Cell Coverage Map. <http://www.opensignal.com>.
- [28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.
- [29] Caleb Phillips, Michael Ton, Douglas Sicker, and Dirk Grunwald. 2012. Practical radio environment mapping with geostatistics. *Proc. of the IEEE DYSpan* (Oct. 2012), 422–433.
- [30] T. Rappaport. 2001. *Wireless Communications: Principles and Practice* (2nd ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.
- [31] Avik Ray, Supratim Deb, and Pantelis Monogioudis. 2016. Localization of LTE measurement records with missing information. In *Proc. of the IEEE INFOCOM*. San Francisco, CA, USA, 1–9.
- [32] COST 231 Final Report. 1999. *Digital mobile radio towards future generation systems*. Technical Report.
- [33] Root Metrics Inc. [n. d.]. Metro RootScore Reports. <http://www.rootmetrics.com/us>.
- [34] Bernard Sklar. 1997. Rayleigh fading channels in mobile digital communication systems. I. Characterization. *IEEE Comm. Magazine* 35, 7 (1997), 90–100.
- [35] Speedtest by Ookla Inc. 2006. The Global Broadband Speed Test. <http://www.speedtest.net>.
- [36] Tutela Inc. 2011. Crowdsourced mobile data. <http://www.tutela.com>.
- [37] Jie Yang, Alexander Varshavsky, Hongbo Liu, Yingying Chen, and Marco Gruteser. 2010. Accuracy characterization of cell tower localization. In *Proc. of the ACM UbiComp*. 223–226.
- [38] Zhengqing Yun and Magdy F Iskander. 2015. Ray tracing for radio propagation modeling: Principles and applications. *IEEE Access* 3 (2015), 1089–1100.
- [39] Tan Zhang and Suman Banerjee. 2013. Inaccurate spectrum databases?: Public transit to its rescue!. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. ACM, 6.
- [40] Tan Zhang, Ashish Patro, Ning Leng, and Suman Banerjee. 2015. A wireless spectrum analyzer in your pocket. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*. ACM, 69–74.