City-Wide Signal Strength Maps: Prediction with Random Forests

Emmanouil Alimpertis
UC Irvine
ealimper@uci.edu

Athina Markopoulou
UC Irvine
athina@uci.edu

Carter T. Butts
UC Irvine
buttsc@uci.edu

Konstantinos Psounis
USC
kpsounis@usc.edu

ABSTRACT

Signal strength maps are of great importance to cellular providers for network planning and operation, however they are expensive to obtain and possibly limited or inaccurate in some locations. In this paper, we develop a 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 other features. 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, *i.e.,* requiring 80% less measurements for the same prediction accuracy, or reduces the relative error by 17% for the same number of measurements. Second, we leverage two types of real-world LTE RSRP datasets to evaluate 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.

CCS CONCEPTS

• **Networks → Network measurement**; **Network performance analysis**.

KEYWORDS

Signal Strength Maps; LTE; RSRP; RSS; Random Forests; Prediction

ACM Reference Format:
Emmanouil Alimpertis, Athina Markopoulou, Carter T. Butts, and Konstantinos Psounis. 2019. City-Wide Signal Strength Maps: Prediction with Random Forests. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313726

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. 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 [17] (*e.g.,* throughput, delay, frequency band, location, time) associated with the measurement. Signal maps consist of a large number of measurements of KPIs and are of crucial importance to cellular operators, for network management, maintenance, upgrades, operations and troubleshooting [19].

Although cellular providers can collect measurements on the network edge themselves (*e.g.,* via wardriving [31]), they increasingly choose to outsource the data collection 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. Mobile analytics companies (*e.g.,* OpenSignal [22], RootMetrics [28], Tutela [30] *etc.*) crowdsource measurements directly from end-user devices, via standalone mobile apps [22], or measurement SDKs [30] integrated into partnering apps. Thus large scale signal maps collection is achieved, 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 millions dollars to third parties) and crowdsourcing companies (most of which use cloud services, thus collecting more measurements increases their operational cost). Moreover, trends, such as (i) 5G dense deployment of small cells and (ii) smart city and IoT deployments, will only increase the need for accurate performance measurements [4, 9, 14, 16], while data may be sparse, unavailable, or expensive to obtain.

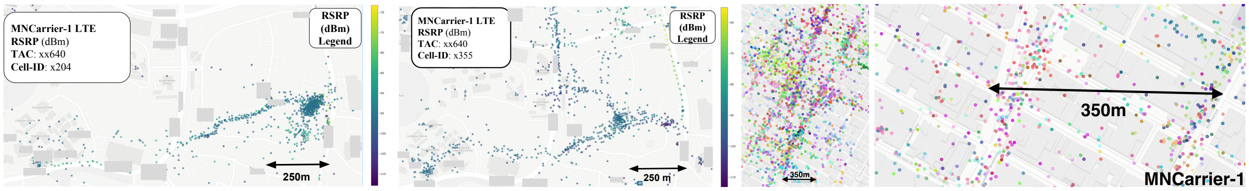
Our goal in this paper is to improve the tradeoff between cost (number of measurements) and quality (*i.e.,* coverage and accuracy) of signal maps via signal strength prediction from limited measurements. In general, there are two approaches in RSRP prediction: propagation models and data-driven approaches.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

1. RFs **RSRP prediction framework.** 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

Acknowledgments: This work was supported by NSF Awards 1649372, 1526736, 1444060; as well as by Networked Systems and CPCC at UC Irvine and the Broadcom Foundation Fellowship program.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313726



(a) Campus example cell x204: high density (0.66), low dispersion (325).

(b) Campus: example cell x355: small density (0.12) more dispersed data (573).

(c) NYC: Manhattan LTE TA

(d) NYC: zooming in Manhattan Midtown (Time 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.**

techniques [8, 20, 24], which do not naturally extend beyond location features. 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 17% for the same number of measurements.

2. 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 2. 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 signal strength) prediction, in terms of any metric (number of measurements, geographical scale, number of cells *etc.*). 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 and tens of thousands of measurements in [13] or smaller scale in [8, 15, 20, 24].

The structure of the rest of the paper is as follows. Section 2 presents our random forests-based prediction approach as well as baselines and prior work for comparisons. Section 3 presents the available signal map datasets. Section 4 provides evaluation results. Section 5 concludes the paper.

2 RSRP PREDICTION

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), defined by 3GPP in [12]. This choice is both because RSRP is widely used for various operations in LTE (*e.g.,* cell selection, handover decisions [10], network quality assessment *etc.*) and as a case study that can potentially be applied to prediction of other RSS metrics. Typically, RSRP is reported in dBm by UEs (user equipment) as the average power over several, narrow-band, control channels.

LTE Cells vs. Tracking Areas. A serving LTE cell is uniquely identified by the CGI (cell global identifier) which is the concatenation of the following identifiers: the MCC (mobile country code), MNC (mobile network code), TAC (tracking area code) and the cell ID. We abbreviate and refer to CGI 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 cells, under common LTE management for a specific area.

The RSRP Prediction Problem. Our goal is to predict the RSRP value at a given location, time, and potentially considering additional contextual information (Section 2.3), based on available measurements either in the same cell *cID* or in the same LTE TA. Table 1 summarizes our prediction methods and related work.

2.2 Model-Based Prediction: LDPL

As a representative baseline from the family of model-based predictors, we consider the Log Distance Path Loss (LDPL) propagation model [25], which is simple yet widely adopted in the literature (*e.g.,* [2], [8]). LDPL predicts the power (in dBm) at location $\vec{l}_j$ at distance $||\vec{l}_{\text{BS}} - \vec{l}_j||_2$ from the transmitting basestation (BS *a.k.a.* cell tower), as a log-normal random variable (*i.e.,* normal in dBm) [2]:

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

[1]We consider two cases regarding path loss exponent (PLE) n_j.

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: Recent work (*e.g.,* [2, 8]) considers 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 $\widehat{n_j}$ 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

We apply Random Forests (RFs) regression. RFs are an ensemble of multiple decision trees [5], which provides a good trade-off between

[1] $P_0^{(t)}$ is the received power at reference distance d_0 (typically $1m$), calculated by the free-space path loss (Friis) transmission equation for the corresponding downlink frequency, gain and antenna directionality, and $\vec{l}_{\text{BS}}$ the location of the transmitting antenna. n_j, *i.e.,* PLE, has typical values between 2 and 6. The log-normal shadowing is modeled by $w_j^{(t)} \sim \mathcal{N}(0, \sigma_j^2(t))$ (in dB), with variance $\sigma_j^2(t)$ assumed independent across different locations. The cell, cID, affects several parameters in Eq. (1), including P_0, w_j, the locations of transmitting ($\vec{l}_{\text{BS}}$) and receiving ($\vec{l}_j$) antennas.

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 [6], COST 231 [27], Ray Tracing [32], Hata Model [29] *etc.*	Limitations: Requires environment's info *e.g.,* topology, street width, BSs height, 3D map.
Data Driven	(2) Geospatial Interpolation[8]	2(a) OK: Ordinary Kriging	2-(b) OKD:OK Detrending (hybrid of model and data)	2-(c) OKP: OK partitioning (spatial heterogeneous)	(4) Other Data-Driven: 4(a) Bayesian Compressive Sensing [15]. Limitation: no arbitrary features. 4(b) Deep Neural Nets [11] Limitation: Needs 3D Map/LiDAR Data.
	(3) Random Forests (RFs)	3(a) $RFs_{x,y}$ [26] Spatial Features: $\mathbf{x} = (l^x, l^y)$	**3-(b)** $RFs_{x,y,t}$ **Spatiotemporal:** $\mathbf{x} = (l^x, l^y, d, h)$	**3(c)** RFs_{all} **Full Feats: (and in some scenarios:** cid**)** $\mathbf{x} = (l^x, l^y, d, h, dev, \|\|\vec{l}_{BS} - \vec{l}_j\|\|_2, freq_{dl}, out)$	

bias and variance by exploiting the idea of bagging [5]. An RSRP value P can be modeled as follows given a set of features vector $\mathbf{x}$.

$$P|\mathbf{x} \sim \mathcal{N}(RFs_{\mu}(\mathbf{x}), \sigma^2_{\mathbf{x}}) \tag{2}$$

where $RFs_{\mu}(\mathbf{x})$, $RFs_{\sigma}(\mathbf{x})$ are the mean and standard deviation respectively of the RSRP predictor, $\sigma^2_{\mathbf{x}} = RFs_{\sigma}(\mathbf{x}) + \sigma^2_{RFs}$ and σ^2_{RFs} is the MSE of the predictor. The prediction $\widehat{P} = RFs_{\mu}(\mathbf{x})$ is the MLE (maximum likelihood estimation) value minimizing the MSE.

For each measurement j in our data, we consider the following full set of features, available via the Android API:

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

• *Location* $\vec{l}_j = (l^x_j, l^y_j)$. These are the spatial coordinates and the only ones considered by previous work on data-driven RSS prediction [8, 20] or in the context of localization [19, 26].

• *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, *i.e.,* the CGI defined in Section 2.1.

• *Device hardware* type, *dev*. This refers to the device model (*e.g.,* Galaxy9 or iPhone X) and *not* to device identifiers. This feature captures (i) the different noise figures (NF), *i.e.,* electronic interference, and reception characteristics across different devices and (ii) the RSRP calculation may differ across devices and manufacturers [12].

• *The downlink carrier frequency, $freq_{dl}$.* Radio propagation and signal attenuation heavily depend on frequency calculated by *EARFCN* (E-UTRA Absolute Radio Frequency Channel Number).

• *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 $\vec{l}_j$ 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 shown in Section 4.2.1. 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:

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

When there are a few measurements per *cID*, then we can treat *cID* as one of the features in $\mathbf{x}^{\mathbf{full}}_{\mathbf{j}}$. We denote as $\mathrm{RFs}_{x,y}$, $\mathrm{RFs}_{x,y,t}$, RFs_{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.

Why RFs for Data-Driven Prediction? First, RFs can naturally incorporate all aforementioned features, since geospatial interpolation [8, 24] does not naturally extend to arbitrary features. Second, RFs, partition the feature space with axis-parallel splits [21].

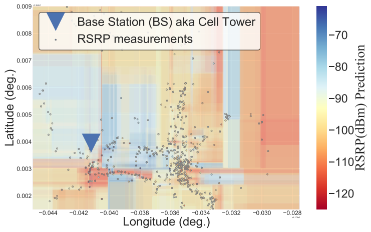


Figure 2: Example of decision boundaries chosen by $\mathrm{RFs}_{x,y}$ for Campus cell x306. We can see that RFs can naturally identify spatially correlated measurements, *i.e.,* regions with similar wireless propagation characteristics.

An example of decision boundaries produced by $\mathrm{RFs}_{x,y}$ 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 of antennas on campus. Automatically identifying these regions with spatially (and temporal) correlated RSRP comes for free to RFs and is particularly important in RSRP prediction because wireless propagation has different properties across neighborhoods [26]. In contrast, prior art (OKP, [8]) requires additional preprocessing for addressing this spatial heterogeneity.

2.4 Baseline: Geospatial Interpolation

State-of-the-art approaches in data-driven RSS prediction [8, 13, 20, 24] have primarily relied on geospatial interpolation, which however is inherently limited to only spatial features (l^x, l^y). The best representative of this family of predictors is ordinary kriging (OK) [20] and its variants [8], which are used as baselines for comparison in this paper.

Ordinary Kriging (OK): It predicts RSS at the testing location $\vec{l}_j = (l^x_j, l^y_j)$ 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 $\gamma(h)$ [8].

Ordinary Kriging Partitioning (OKP) [8]: Voronoi-based partitioning is used to identify regions with the same PLE and apply a distinct OK model in each region. It is comparable to the heterogeneous LDPL, yet is impractical for city wide signal maps.

Ordinary Kriging Detrending (OKD) [8, 24]: OK assumes spatial stationarity, which does not hold for RSRP. OKD incorporates a version of LDPL in the prediction in order to address this issue [8]. This can be thought as a hybrid approach of geospatial and LDPL and serves as our baseline for comparison in this paper.

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 and Table 3 summarizes representative LTA TAs examples used in our evaluation.

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 $\approxeq 3km^2$	LTE KPIs: RSRP, [RSRQ]. Context: GPS Location, timestamp, dev, cid. Features: $\mathbf{x} = \left(l_j^x, l_j^y, d, h, dev, out, \|\|\vec{l}_{\text{BS}} - \vec{l}_j\|\|_2\right)$	No. Cells = 25, No. Meas $\approxeq$ 180K Per Cell Density: 0.01 - 0.66 (Table 3) Overall Density: 0.06	Ourselves [3]
NYC & LA	09/01/17-11/30/17	NYC Metropolitan Area $\approxeq 300km^2$	LTE KPIs: RSRP, [RSRQ, CQI]. Context: GPS Location, timestamp, dev, cid. EARFCN. Features:$\mathbf{x} = \left(l_j^x, l_j^y, d, h, cid, dev, out, \|\|\vec{l}_{\text{BS}} - \vec{l}_j\|\|_2, freq_{dl}\right)$	No. Meas NYC $\approxeq$ 4.2M, No. Cells NYC $\approxeq 88k$ Density NYC-all $\approxeq 0.014\ N/m^2$	Mobile Analytics Company
		LA metropolitan Area $\approxeq 1600km^2$		No. Meas LA $\approxeq$ 6.7M, No. Cells LA $\approxeq$ 111K Density LA-all $\approxeq 0.0042\ N/m^2$	

3.1 Campus dataset

Dataset Overview. We collected the first dataset on UC Irvine campus. This Campus dataset is relatively small: 180,000 data points, collected by seven different users, using 2 cellular providers. In terms of geographical area, it covers approximately $3km^2$, as the devices move between locations (*e.g.,* housing, office *etc.*) on campus. Some examples are depicted in Fig. 1(a)-(b). Although small, the cells in this dataset exhibit a range of characteristics (reported in Table 4) regarding the (i) number of measurements, (ii) mean and variance of RSRP and (iii) dispersion and density (*e.g.,* multiple measurements over time on the same or nearby locations).

Data Collection. We developed a user-space app [1, 3] 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, we defer to [3] due to space limitations. No personally identifiable information is collected.

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). Key characteristics are summarized in Table 2. An example of the NYC Midtown Manhattan neighborhood is depicted in Fig. 1(c)-(d). While these are large datasets, they are also relatively sparse in space and heterogeneous. For example, the density is now approx. 300 measurements per cell on average and up to the order of 20 thousands max. We only consider cells with more than 100 measurements. There is also sparsity in time: unlike the Campus dataset, there are no longer multiple measurements at different times for the same location.

Data Collection. This dataset was collected by a major mobile analytics crowdsourcing company and shared with us. They collect measurements from a large user base infrequently so as to not burden each end-users, which explains the smaller overall density of the dataset, as shown in Table 2. Each location data point is accompanied by rich network and contextual information, except for device or other personal identifiers.

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. We utilize GeoJSON format to represent and process our data, which offers several practical advantages.

Properties of Datasets. For each dataset, the following metrics describe characteristics that affect RSRP prediction.

- *Data Density:* Number of measurements per unit area (N/m^2).

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	$\approxeq$ 63K	$\approxeq$ 104K	$\approxeq$ 20K
Area km^2	$1.8km^2$ (Fig. 1 (c-d))	44.8 km^2	220 km^2
Data Density N/m^2	$\approxeq$ 0.035	$\approxeq$ 0.002	$\approxeq$ 0.0001
No. Cells $\|C\|$	429	721	353
Cell Density $\|C\|/km^2$	238.3	16.1	1.6

- *Cells Density:* Number of unique cells (*cids*) per unit area, *i.e.,* $|C|/km^2$. 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 [18], defined as the standard deviation of the distance of measurement points from the center.

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), $||\vec{l}_{\text{BS}} - \vec{l}_j||_2$. To that end, we lookup the location of the BS, $\vec{l}_{\text{BS}}$, using the public APIs of a popular online crowdsourced database opencellid.org.

4 PERFORMANCE EVALUATION

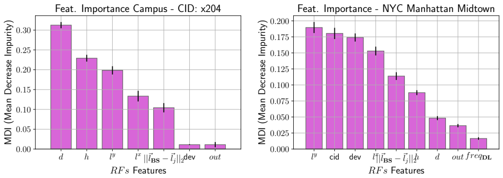
4.1 Setup

4.1.1 RFs Setup. 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 [23] 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 overfit RFs. For the NYC and LA datasets, we select $n_{trees} = 1000$ and $\max_{depth} = 30$; more and deeper trees are needed for larger datasets. One important design choice is what granularity we choose to build our RFs models: per *cID* or per LTE TA (defined in Section 2.1).

•Training per cID: We can train a separate RFs model per cell using all features except *cID* ($\mathbf{x_j^{-cID}}$), since RSRP is measured per serving cell, but requires many measurements per cell. This 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 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 per cell unit area, insufficient to train a model per *cID*. In urban areas, there is very high cells density in a region and data points from different cells in the same LTE TA can be useful.

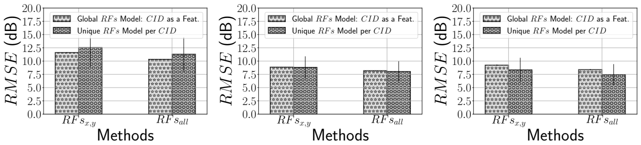
4.1.2 Baselines' Setup. For LDPL-knn: we select empirically k = 100 neighbors for the Campus dataset and k = 10% of the training data points in each cell for the NYC and LA datasets. For *Geospatial Predictors,* the number of neighbors was empirically set to k = 10; a larger k did not show significant improvement. An exponential fitting function of the semivariogram function $\gamma(h)$



(a) Cell x204

(b) Manhattan Midtown TA

Figure 3: For Campus dataset **(a) : Feature Importance for cell x204 (**RFs **built per distinct cell). Cells' data are depicted in Fig. 1. For** NYC dataset**, (b) shows the** MDI **score for one LTE TA for MNC-1.**



(a) MNC-1, NYC Manhattan Midtown (urban)

(b) MNC-1, East Brooklyn (suburban)

(c) MNC-2, Southern LA (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** RFs **model per** cID**; (3) providers MNC-1 vs. MNC-2.**

Table 4: Campus dataset**: Comparing Predictors per cell**

	Cell Characteristics					RMSE (dB)						
cID	N	$\frac{N}{\text{sq } m^2}$	SDD	$\mathbb{E}[P]$	σ^2	LDPL hom	LDPL kNN	OK	OKD	RFs x,y	RFs x,y,t	RFs 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**

was selected [8]; the maximum lag (h) was set to 200m, as in [8], for the Campus and NYC environments, while it was set to 600m for the LA suburban environment. The empirical semivariogram $\widehat{\gamma(h)}$ was calculated per $10m$ [8].

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. 5, 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 Absolute Relative Improvement (ARI) as well as feature importance in RFs.

Root Mean Square Error (RMSE): If $\widehat{P}$ is an estimator for P, then $RMSE(\widehat{P}) = \sqrt{MSE(\widehat{P})} = \sqrt{E((P - \widehat{P})^2)}$, 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), (iii) over all data (Campus).

Absolute Relative Improvement (ARI): This captures the improvement of each predictor over the variance in the data: $ARI = 1 - (1/|\mathrm{C}|)\sum_{i \in \mathrm{C}}(MSE_i/Var_i)$, where $|\mathrm{C}|$ 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: It captures how often a feature is used to perform splits in RFs. It is defined as the total decrease in node impurity, weighted by the probability of reaching that node (approx. by the proportion of samples reaching that node), averaged over all trees in the ensemble [23].

4.2 Results

4.2.1 Feature Importance.

a. Campus dataset*:* We train one RFs model per cID for the set of features $\mathbf{x} = (l_j^x, l_j^y, d, h, ||\vec{l}_{\mathrm{BS}} - \vec{l}_j||_2, out, dev)$. Results w.r.t. MDI 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 MDI. An example of such a cell is x204 (depicted in Fig. 1(a)), which has $SDD = 325$, density=0.66 points/m^2 and its MDI is shown in Fig. 3(a). Feature importance for dev and out are close to zero, which is expected because of the small number of devices in the Campus dataset. For more dispersed and less dense cells, such as cell x355 ($SDD = 573, 0.116N/m^2$, map in Fig. 1(b)), the location features (l_j^x, l_j^y) have higher MDI; results are omitted.

b. NYC and LA datasets*:* In this case, $freq_{dl}$ is available and the datasets contain thousands of cells. We start with a RFs model per LTE TA. As a representative example, we report the feature importance, in Fig. 3(b), for the LTE TA of a major mobile network carrier (MNCarrier-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 and dev. 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 RFs per cID, or cID should be used as one of the features in a single RFs. 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 RFs per LTE TA and (ii) when a separate RFs 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, RFs benefit from the features of the additional measurements. Manhattan Midtown has a cells density of 238 per km^2 (see 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 RFs model per cell performs better than RFs per LTE TA, as shown in Fig. 4(c). In the Campus dataset (lower density than NYC), the RFs model per cID did better than using as a feature in a single RFs 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, RFs 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 against baselines, both geospatial interpolation (OK and OKD) and model-driven (LDPL-knn and LDPL-hom). *a.* Campus dataset*:* Table 4 reports the $RMSE$ for all predictors for each cell in the Campus dataset, for the default 70-30% split. Fig. 7 compares all methods,

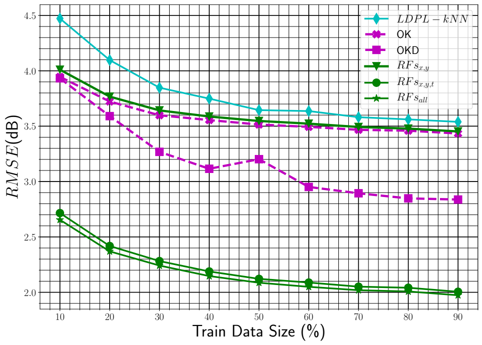


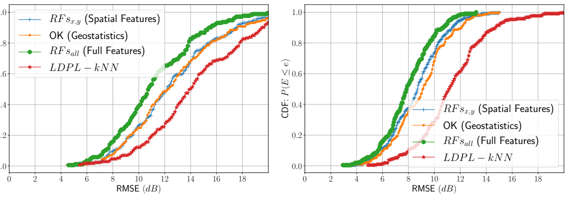
Figure 5: Campus dataset: *RMSE* **vs. Training Size. Our methodology (RFs with more than spatial features, *i.e.,* $\text{RFs}_{x,y,t}$, RFs_{all}) significantly improves the RMSE-cost tradeoff: it can reduce *RMSE* by 17% for the same number of measurements compared to state-of-the-art data-driven predictors(OKD); *or* it can achieve the lowest error possible by OKD (≃ 2.8dB) with 10% instead of 90% (and 80% reduction) of the measurements.**

calculating *RMSE* over the entire Campus dataset, instead of per cell. We can see that our RFs-based predictors ($\text{RFs}_{x,y,t}$, RFs_{all}) outperform model (LDPL) and other data-driven (OK, OKD) predictors.

Fig. 5 shows the *RMSE* as a function of the training size (as % of all measurements in the dataset). First, the performance of OK and $\text{RFs}_{x,y}$ is almost identical, as it can be seen for *RMSE* over all measurements (Fig. 5 and Fig. 7) 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 $\text{RFs}_{x,y}$ uses multiple decision trees and data splits. Second and more importantly, considering additional features can significantly reduce the error. For the Campus dataset, when time features $\mathbf{t} = (d, h)$ are added, $\text{RFs}_{x,y,t}$ significantly outperforms OKD: it decreases *RMSE* from 0.7 up to 1.2 dB. Alternatively, in terms of training size, $\text{RFs}_{x,y,t}$ needs only 10% of the data for training, in order to achieve OKD's lowest error (≃ 2.8dB) 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 $\text{RFs}_{x,y,t}$ compared to OKD is 17%, shown in Fig. 7(b).

b. NYC and LA datasets*:* Fig. 6 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 is very close to $\text{RFs}_{x,y}$, 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 important, which is expected since this data, has heterogeneous 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, $\text{RFs}_{x,y,t}$ predicts a time-varying value for the measurements at the same location in Fig. 1(a), while $\text{RFs}_{x,y}$ 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.*). Finally, RFs significantly outperform propagation models, except for a few areas with limited number of measurements.



(a) NYC Manhattan Midtown

(b) LA Southern Suburb

Figure 6: NYC and LA datasets**: CDFs for *RMSE* per *cID* for two different LTE TAs, for the same major MNCarrier-1.** RFs_{all} **offer 2dB gain over the baselines for the 90th percentile.**

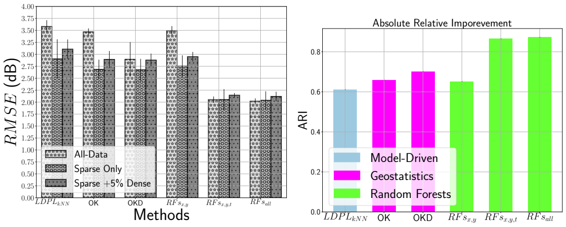


Figure 7: Comparison of all predictors over the entire Campus dataset**. Left (a)** $RMSE(dB)$ **under various scenarios, Right (b) ARI over all data points. Our Approaches ($\text{RFs}_{x,y,t}$, RFs_{all}) outperform prior art in all scenarios.**

4.2.3 *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 grad students' home and work, which begs the question whether this leads to overfitting. We investigated this question and found that our RFs predictors neither get a performance boost nor overfit. To that end, we utilize HDBSCAN [7], a state-of-the-art clustering algorithm, to identify very dense (spatially) clusters 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. 7(a) 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 $\text{RFs}_{x,y,t}$ and RFs_{all} have similar performance in all scenarios and consistently outperform baselines (similarly in cell-by-cell basis; results omitted). Please note that OK and LDPL-knn's errors slightly decrease for "sparse-only"; OK cannot handle repeated locations and LDPL-knn may overfit.

5 CONCLUSION

We developed a machine learning framework for cellular signal strength prediction, which is important for creating signal maps in a cost-efficient way, crucial for future 5G and IoT deployments. We used the powerful tool of random forests, which we adapted in this context by evaluating different features. The datasets under study are the largest used in this context and provide unique insight into city-wide signal map prediction. Future work includes a hybrid ML-propagation model to harvest the diversity of both worlds.

REFERENCES

[1] A. Shubaa, A. Le, E. Alimpertis, M. Gjoka, A. Markopoulou. 2016. AntMonitor: System and Applications. *arXiv:1611.04268* (Nov. 2016).

[2] E. Alimpertis, N. Fasarakis-Hilliard, and A. Bletsas. 2014. Community RF sensing for source localization. *IEEE Wireless Comm. Letters* 3, 4 (2014), 393–396.

[3] E. Alimpertis and A. Markopoulou. 2017. A system for crowdsourcing passive mobile network measurements. In *14th USENIX NSDI'17, Posters Sessions*. USENIX Association, Boston, Massachusetts, USA.

[4] T. Bilen, B. Canberk, and K. R. Chowdhury. 2017. Handover Management in Software-Defined Ultra-Dense 5G Networks. *IEEE Network* 31, 4 (2017), 49–55.

[5] L. Breiman. 2001. Random Forests". *Machine Learning* 45, 1 (Oct. 2001), 5–32.

[6] Y. J. Bultitude and T. Rautiainen. 2007. *IST-4-027756 WINNER II D1. 1.2 V1. 2 WINNER II Channel Models.* Technical Report.

[7] R. JGB Campello, D. Moulavi, and J. Sander. 2013. Density-based clustering based on hierarchical density estimates. In *PAKDD '13*. Springer, 160–172.

[8] A. Chakraborty, Md S. Rahman, H. Gupta, and S. R Das. [n. d.]. SpecSense: Crowdsensing for Efficient Querying of Spectrum Occupancy. In *Proc. of the IEEE INFOCOM '17*. 1–9.

[9] B. Cici, E. Alimpertis, A. Ihler, and A. Markopoulou. 2016. Cell-to-cell activity prediction for smart cities. In *Proc. of the IEEE INFOCOM WKSHPS '16*. IEEE, 903–908.

[10] H. Deng, C. Peng, A. Fida, J. Meng, and Y C. Hu. 2018. Mobility Support in Cellular Networks: A Measurement Study on Its Configurations and Implications. In *Proc. of the ACM IMC '18*. ACM, 147–160.

[11] R. Enami, D. Rajan, and J. Camp. 2018. RAIK: Regional analysis with geodata and crowdsourcing to infer key performance indicators. In *Proc. of the IEEE WCNC '18*. 1–6.

[12] 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/`.

[13] M. R. Fida, A. Lutu, M. K. Marina, and O. Alay. 2017. ZipWeave: Towards efficient and reliable measurement based mobile coverage maps. In *Proc. of the IEEE INFOCOM '17*. IEEE.

[14] A. Gomez-Andrades, R. Barco, P. Munoz, and I. Serrano. 2017. Data analytics for diagnosing the RF condition in self-organizing networks. *IEEE Transactions on Mobile Computing* 16, 6 (2017), 1587–1600.

[15] S. He and K. G. Shin. 2018. Steering Crowdsourced Signal Map Construction via Bayesian Compressive Sensing. In *Proc. of the IEEE INFOCOM '18*. IEEE, 1016–1024.

[16] A. Imran, A. Zoha, and A. Abu-Dayya. 2014. Challenges in 5G: How to empower SON with big data for enabling 5G. *IEEE network* 28, 6 (2014), 27–33.

[17] J. Johansson, W. A Hapsari, Sean Kelley, and G. Bodog. 2012. Minimization of drive tests in 3GPP release 11. *IEEE Comm. Magazine* 50, 11 (2012).

[18] Z. Li, A. Nika, X. Zhang, Y. Zhu, Y. Yao, B. Y Zhao, and H. Zheng. 2017. Identifying value in crowdsourced wireless signal measurements. In *Proc. of the ACM WWW '17*. ACM, 607–616.

[19] R. Margolies, Richard Becker, S. Byers, S. Deb, R. Jana, S. Urbanek, and C. Volinsky. 2017. Can you find me now? Evaluation of network-based localization in a 4G LTE network. In *Proc. of the IEEE INFOCOM '17*. 1–9.

[20] M. Molinari, M R. Fida, M. K Marina, and A. Pescape. 2015. Spatial Interpolation Based Cellular Coverage Prediction with Crowdsourced Measurements. In *Proc. of the ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big Internet Data (C2BID)*. ACM, 33–38.

[21] K. P. Murphy. 2012. *Machine Learning: A Probabilistic Prespective.* The MIT Press, Cambridge, Massachusetts.

[22] Open Signal Inc. 2011. 3G and 4G LTE Cell Coverage Map. `http://www.opensignal.com`.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct. (2011), 2825–2830.

[24] C. Phillips, M. Ton, D. Sicker, and D. Grunwald. 2012. Practical radio environment mapping with geostatistics. *Proc. of the IEEE DYSPAN '12* (Oct. 2012), 422–433.

[25] T. Rappaport. 2001. *Wireless Communications: Principles and Practice* (2nd ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.

[26] A. Ray, S. Deb, and P. Monogioudis. 2016. Localization of LTE measurement records with missing information. In *Proc. of the IEEE INFOCOM '16*.

[27] COST 231 Final Report. 1999. *Digital mobile radio towards future generation systems.* Technical Report.

[28] Root Metrics Inc. 2018. Metro RootScore Reports. `http://www.rootmetrics.com/us`.

[29] Bernard Sklar. 1997. Rayleigh fading channels in mobile digital communication systems. I. Characterization. *IEEE Comm. Magazine* 35, 7 (1997), 90–100.

[30] Tutela Inc. 2011. Crowdsourced mobile data. `http://www.tutela.com`.

[31] J. Yang, A. Varshavsky, H. Liu, Y. Chen, and M. Gruteser. 2010. Accuracy characterization of cell tower localization. In *Proc. of the ACM UbiComp '10*. ACM, 223–226.

[32] Z. Yun and M. F. Iskander. 2015. Ray tracing for radio propagation modeling: Principles and applications. *IEEE Access* 3 (2015), 1089–1100.