RAIK: Regional Analysis with Geodata and Crowdsourcing to Infer Key Performance Indicators

Rita Enami, Dinesh Rajan, and Joseph Camp
Department of Electrical Engineering, Southern Methodist University

***Abstract*—Key Performance Indicators (KPIs) are important measures of the quality of service in cellular networks. There are multiple efforts by cellular carriers and 5G standardization to leverage the KPIs to minimize drive tests (MDT) and self-organize the network for optimal performance via user feedback. Such an approach accounts for user devices in the field of their operation according to their normal usage and circumvents a number of costs (e.g., manpower, equipment) traditionally covered by the carrier, either directly or through a third party. In this paper, we build a Regional Analysis to Infer KPIs (RAIK) framework to establish a relationship between geographical data and user data using crowdsourced measurements. To do so, we use a neural network and crowdsourced data obtained by user equipment (UE) to predict the KPIs in terms of the reference signal's received power (RSRP) and path loss estimation. Since these KPIs are a function of terrain type, we provide a two-layer coverage map by overlaying a performance layer on a 3-dimensional geographical map. As a result, we can efficiently use crowdsourced data (to not overextend user bandwidth and battery) and infer KPIs in areas where measurements have not or can not be performed. For example, we show that RAIK can use only geographical information to predict the KPIs in areas that lack signal quality data with a negligible mean squared error, a seven-fold reduction in error from state-of-the-art solutions.**

I. Introduction

Network operators use Key Performance Indicators (KPIs) to track network performance, including the received signal power, received signal quality, throughput, and delay. Historically, drive testing has been widely used by carriers and third party entities to collect a sufficient density of KPI data to accurately characterize the network performance. Despite providing detailed information at certain locations, this approach is costly in terms of manpower, time, and equipment. Even with the high costs associated with drive testing, carriers do not have access to some regions and often cannot anticipate the breadth of user devices, contexts, and functionalities with in-field operation. Further complicating the problem, drive testing may have to be repeated with changes to the physical environment, such as the construction of new buildings or highways, seasonal variations, or modifications to the spatial distribution of users in the network [1].

An alternative and less costly way to capture such KPIs is crowdsourcing, as outlined in the Minimization of Drive Test (MDT) effort of LTE release 10 in 3GPP TS 37.320 [2]. MDT allows carriers to monitor the in-situ network performance of end users to detect variations of the provided quality of service (QoS) to perform such actions as handover if the problem is confined to a single user or self-organization if the problem extends to one or more towers. For the latter problem, changes to the antennae configuration in terms of transmit power, tilt, or height can alleviate some issues while more persistent effects necessitate smaller cell deployment in detected network holes.

To make efficient use of the crowdsourced data (to preserve bandwidth and battery life of users), a natural extension of MDT is to interpolate the region's performance from discrete user locations using propagation models [3], [4] and coverage maps [5], [6]. Other studies have used crowdsourced data to measure network metrics (*e.g.*, [7]–[11]). However, none of these approaches directly considers geographical features of the environment in predicting the propagation characteristics and resulting KPIs. In this paper, we establish the relationship required by MDT efforts between geographical data and user-based data, from which we build a Regional Analysis to Infer KPIs (RAIK) framework. To do so, we predict the network coverage using neural networks alongside crowdsourced data collected by User Equipment (UEs) with an overlaid LiDAR dataset in that same region. RAIK is based on a feed-forward, back-propagation model, which employs multilayer perceptron (MLP) with the geographical features of a region to provide a KPI-based coverage map. To evaluate RAIK, we perform extensive in-field measurements from urban and suburban regions with diverse geographical features such as type, density, and height of the buildings and trees. RAIK forms a generalized framework that allows prediction of the KPIs in areas that have yet to receive crowdsourced channel quality measurements from users, relying solely on the terrain and clutter information of a given area.

Our work consists of the following three contributions:

- We introduce the Regional Analysis to Infer Key Performance Indicators (RAIK) framework, a learning structure that can create interconnected relationships between geographical information and KPIs.
- To understand the impact of tile size on the prediction results, we provide a coverage map based on the path loss exponent using crowdsourced data. We find in all the results that there is a tradeoff between the larger tile size having too much area which has distinct terrain and too small area without sufficient measurements. We show that this tenuous relationship is magnified in the downtown area due to the diversity from street to street.
- We consider the accuracy of predicting KPIs in areas in which the RAIK framework lacks any signal quality training, relying solely on the geographical features of the area. For sub-regions that are tested without prior training in that region, we find that the mean squared error (MSE) of the predicted path loss and the measured one (to test our prediction) can be as small as 0.01, which is a 7-fold reduction from state-of-the-art algorithms.

The remainder of the paper is organized as follows. In Section II, we present our framework, measurement set up, and path loss evaluation using crowdsourced data. In Section III, we present our prediction and in-field analysis of the

relationship between geographical features and our path loss prediction model using a neural network. We present related work in Section IV and conclude in Section V.

II. Regional Analysis to Infer KPIs (RAIK)

To construct a coverage map from a region of interest, we build a framework depicted in Fig. 1, which consists of the following steps. (*i.*) We first build an Android-based crowdsourcing infrastructure, which allows the widespread collection of in-field signal quality data coupled with the location of that user at the time of the measurement. (*ii.*) We then infer the propagation characteristics of a given region (regardless of the geographical features) by using the collected signal quality measurements through that area and a sliding square window of varying sizes. (*iii.*) Since the received signal attenuation is affected by foliage and buildings surrounding the user equipment (UE), we consider 3-dimensional geographical data from the region of interest. For this purpose, we use Light Detection and Ranging (LiDAR) data, which includes detailed information of buildings and foliage such as height and surface area (see Section. II-A for more details). (*iv.*) Lastly, the prediction function will receive the estimated channel characteristics and corresponding geographical features of an area to construct a two-layer map consisting of: (*a.*) the network performance information obtained by the UEs, and (*b.*) their corresponding location information overlaid on a map containing the foliage and buildings in the area.

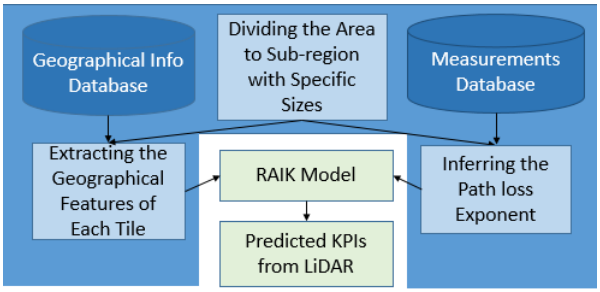


Fig. 1: Regional Analysis to Infer KPIs (RAIK) Framework.

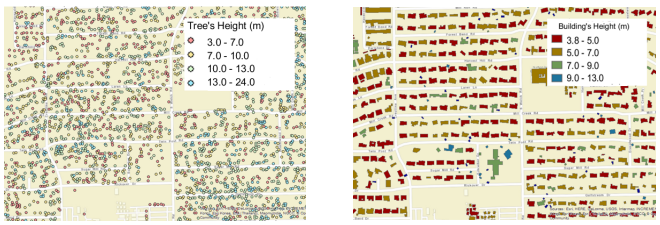
A. Data Acquisition Procedure for KPI Prediction

We now further describe the two data sets on which our RAIK model is based: (*a.*) received signal quality data collected by Android phones, and (*b.*) LiDAR, which describes the geographical features in the area. We consider these two data sets because the geographical features directly impact the received signal quality in a given region.

(a) Android-Based Crowdsourced Data. We have a crowdsourced dataset, which is built from voluntary participants that installed our publicly-available Android application (WiEye) to collect global radio measurements. To limit the power and bandwidth consumed by our app, signal quality from all visible cellular and WiFi base stations are recorded 10 times per day. We have a development version of our app that captures measurements at a frequency of 1 Hz, which we have used to emulate a more concentrated user base in relevant geographical regions in this paper. We specifically record received signal strength across all available technologies, GPS coordinates, Mobile Country and Network Codes, base station identification (CID, LAC), device identification, and velocity of the receiver (when locally collecting data).

We have acquired hundreds of millions of crowdsourced signal strength data points using WiEye. Locally, we collected an additional 10 million measurements with greater densities in three representative geographical regions in Dallas: downtown, single-family, and multi-family residential areas. We utilized obtained data from the downtown and single-family residential areas to train the model. Then, we used the multi-family residential area as a testing region, where the training from this area was not used. Generally, the density of the foliage in the single-family area is higher than the other two regions, the downtown area is mainly covered by tall buildings, and the multi-family area has a mixture of vegetation and moderately-sized buildings (*e.g.*, 2-3 stories).

(b) LiDAR-Based Geographical Features. To consider the vertical and horizontal footprint of trees and buildings, we use LiDAR (Light Detection and Ranging) data, which creates 3-dimensional point clouds of the Earth's surface. LiDAR employs a remote sensing method from airplanes or helicopters that transmits pulses of light to detect the distance from the earth. The laser sends these pulses and measures the time delay between the transmitted and the received pulse to calculate the elevation. LiDAR systems are equipped with a laser scanner that measures the angle of each transmitted pulse and the returned pulse from the surface, high precision clocks which record the time that the laser pulse leaves and returns to the scanner, an Inertial Navigation Measurement unit (IMU) to measure the angular orientation of the scanner relative to the ground (pitch, roll, yaw), a data storage and management system, and a GPS detector.

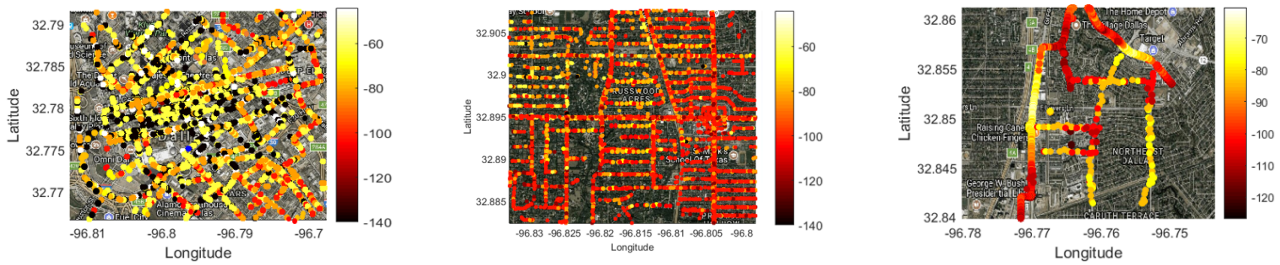


(a) Extracted Tree Data. (b) Extracted Building Data.

Fig. 2: 3-dimensional map from same region using LiDAR.

The LiDAR sampling rate is $400,000$ pulses per second, which creates millions of data points. Also, the accuracy of the collected points is about 15 cm vertically and 40 cm horizontally. Hence, LiDAR systems provide a high-resolution 3D geometric model for the earth, clutter, and foliage, with applicability across a broad range of fields such as geodesy, geometrics, archeology, geography, geology, geomorphology, seismology, forestry, atmospheric physics. [12]. Relevant to our work, we use LiDAR to represent a three-dimensional map of building and tree data in the three Dallas regions under test. Each record that corresponds to a tree in our 3-D map includes coordinates of the object, height, and area. We have the same information for buildings. Fig. 2 shows the detected trees and buildings in suburban region in Dallas. The background of each figure is from OpenStreetMaps to verify the accuracy of the LiDAR information from the same area.

KPI Metric for RAIK. From all the KPIs in the standard, we specifically target the Reference Signal's Received Power (RSRP) since: (*i.*) network providers seek to provide coverage over an area to deliver sufficient quality of service to customers, (*ii.*) a well-known relationship exists between the



(a) Downtown Region. (b) Single-Family Residential Region. (c) Multi-Family Residential Region.

Fig. 3: RSRP from Downtown (left), single-family residential (middle), and multi-family residential (right) regions.

received signal power and the throughput [13], and (*iii.*) UEs regularly measure the received signal power to keep track of visible base stations in case of the handover, even if the phone is idle. Thus, the battery consumption to measure RSRP is low and conducive to MDT efforts.

B. Propagation Over Three Representative Region Types

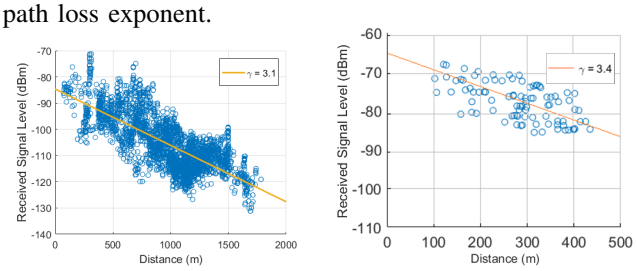
Large-scale fading refers to the average attenuation in a given environment to transmission through and around obstacles in an environment for a given distance [14]. There are three well-known types of models to predict large-scale fading: empirical, deterministic, and semi-deterministic. Empirical models such as [3] and [4] are based on measurements and use statistical properties. These models are widely-used because of their low computational complexity and simplicity. However,the accuracy of these models is not as high as deterministic models to estimate the channel characteristics. Deterministic models or geometrical models using the Geometrical Theory of Diffraction to predict the path loss. To consider the losses due to diffraction, detailed knowledge of the terrain is needed to calculate the signal strength [15]. Despite the accuracy of their model, their computational complexity is high and need detailed information about the region of interest. The last one, semi-deterministic models, are based on empirical and deterministic models [16]. In this study, we use an empirical approach since it is the type of modeling that could best leverage crowdsourced data.

Large-scale fading is a function of distance (d) between the transmitter and the receiver where γ is the path loss exponent. The path loss exponent typically varies between 2 in free space and 6 indoors, depending on the environmental type. Nominal values are in range of 2.7-3.5 in typical urban scenarios and between 3 to 5 in heavily shadowed urban environments [14]. The large-scale path loss for an arbitrary distance d_i between transmitter and receiver is defined according to:

$$L_p(d_i) = L_p(d_0) + 10nlog(d_i/d_0) \quad (1)$$

Here, n is the the path loss exponent, and $L_p(d_0)$ is the path loss at the reference distance d_0. To characterize propagation in a given region, we calculate the path loss exponent from mobile phone measurements, where a linear regression model is used to calculate the n. While mobile phones are not as precise as advanced drive testing equipment, we have shown that path loss is a recoverable parameter from UE signal quality measurements if the appropriate calibration is performed [17].

Fig. 3 depicts the collected RSRP in three representative geographical regions: downtown, single-family residential, and multi-family residential. In each region RSRP values are based on signals received from a single base station. The variation of the signal quality can be observed in each of the three regions. However, sudden changes on the received signal strength in the downtown area are more dramatic from street to street. In particular, we observe very strong signals adjacent to dead zones with respect to the RSRP. Since there are differing geographical features within each region, we calculate the path loss exponent (γ) obtained from received signal measurements taken by mobile phones in smaller sub-regions. To do so, we use a square window with an initial size of 200-m square. There is a tradeoff in the region size considered when considering the accuracy of the RAIK framework. If the sub-region considered is too large, the variation in the geographical features present imprecision in the inferred path loss. If the sub-region considered is too small, the amount of signal quality measurements is insufficient to infer a precise path loss exponent.



(a) Entire downtown region. (b) 200-m square sub-region.

Fig. 4: Path loss exponent (γ) calculated over entire downtown region (left) versus a 200 m x 200 m sub-region (right).

For example, Fig. 4 shows that the calculated γ from the RSRP values over the entire region (3.1) is different from the one calculated from the RSRP of a 200-m square sub-region (3.4). Hence, we will use a filter over each region with sizes of 100-m, 200-m, and 300-m square in Section III to understand both the role of these window sizes and the resulting path loss exponent (γ) in each region and sub-region.

III. Prediction and In-Field Evaluation of KPIs

We now train and evaluate the RAIK framework with our signal quality measurements from the three region types discussed in Section II-B. To do so, we first consider a performance metric and the impact of choosing different sizes of sub-regions (*i.e.*, tile size) over which to compute those predicted metrics. We then consider homogeneous training and

testing, where the neural network is trained and tested in the same region for the downtown and single-family residential region. Lastly, we consider heterogeneous testing where we use the training from these aforementioned region types but test on a different region type: multi-family residential.

A. MultiLayer Perceptron Components Used in RAIK

Neural network algorithms have been widely applied to predict the channel propagation in wireless networks [18]–[21]. In our study, we use a multilayer perceptron (MLP) artificial neural network introduced in [22], and [23]. MLP performs the Levenberg-Marquardt back-propagation algorithm as a supervised learning technique for training the network [24]. MLP consists of three layers: input, output and hidden layers. A neural network in its general form is described as:

$$Z = f(\Sigma w_{i,j}^T.x_i + b) \tag{2}$$

where x_i is the input vector. Each node in a layer is connected to the nodes in the next layer with certain weights $w_{i,j}$. In neural network algorithms, the goal is finding the best selection of weights as the inputs' coefficients such that the difference between the predicted values and the target values are minimized. Here, $w_{i,j}^T$ is the transpose vector of the selected weights by the model associated to the inputs (x_i) and b is the bias vector.

$$w^T = w_1, w_2, .., w_n \tag{3}$$

We employ the sigmoid function [25], which is easily differentiable with respect to the network parameters, and this plays an important role in training of the neural network. It is expressed as:

$$S(x) = \frac{1}{1 + exp(-x)} \tag{4}$$

B. Training and Performance Metrics of RAIK

The model's performance highly depends on the selected features for the model's input and their accuracy. The selected input features are defined as: (*a.*) distance between the transmitter and the receiver, (*b.*) percentage of the area covered by buildings (*i.e.*, foot print) [26], trees (*i.e.*, canopy or crown), and free space (*i.e.*, unoccupied by trees or buildings), (*c.*) number of buildings and trees, (*d.*) average height of the buildings and trees, and (*e.*) standard deviation of the heights of the buildings and trees. All the input parameters have been provided from a 3-dimensional LiDAR database. The model's output is the path loss exponent acquired from radio measurements in each sub-region. To increase the efficiency of the model, all features are normalized to fall in a range of [0, 1].

Performance Metrics for Evaluating the RAIK Model. To evaluate the model, we apply first-order statistical metrics including the minimum and maximum difference between observed and predicted values and the standard deviation of the errors, denoted as e_{min}, e_{max}, and e_{σ}, respectively. The mean error, e_{mean}, shows the average error across all records, which indicates whether there is a systematic bias (a stronger tendency to overestimate or underestimate) in the model. We employ the linear correlation, r, between the predicted, γ_i, and actual values, $\hat{\gamma}_i$. This metric varies between -1.0 and +1.0. Linear correlation returns +1 if there is a perfect positive correlation between the two input variables, -1.0 if there is a perfect negative correlation, and 0.0 if there is no correlation. Finally, the performance of the model is evaluated by calculating the mean squared error (MSE), defined as:

$$min \frac{i}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{5}$$

Here, y_i and $\hat{y}_i$ are the desired output and the obtained output calculated by the neural network, respectively [27].

The Impact of Tile Size on RAIK Performance. We now consider the influence of tile or sub-region size on the estimated path loss and the accuracy of the prediction model. To this end, we train the model on data obtained from tiles with three different sizes: 100-m square, 200-m square, and 300-m square. Table I shows the comparison between the performance metrics of RAIK for tiles with the three different sizes in tile's side length. It is shown that by increasing the tile size to a value larger or smaller than 200-m square, the statistical metrics of the performance decrease in both regions, gradually. Of particular note, the tile size choice most affects the downtown area due to the diversity in height, area, and density of trees and buildings. In both region types, we see that the selection of too large or too small tile sizes impedes the ability of the model to capture the correlation between geographical characteristic changes and the resulting channel propagation, making the KPI prediction noisy. Since the 200-m square tile size showed the best performance in terms of MSE, e_{min}, e_{max}, e_{mean}, and e_{σ} across region types, we will use this size for the remainder of the paper to study various other issues related to KPI prediction.

TABLE I: Impact of tile size on RAIK performance.

Region	Single-family			Downtown		
Tile's Side (m)	100	200	300	100	200	300
MSE	**.03**	**.02**	**.02**	**0.07**	**0.02**	**0.05**
e_{min}	-.61	-.13	-.4	-.7	-.32	-.43
e_{max}	.38	.21	.33	.-43	.23	27
e_{ave}	.04	.02	.03	.1	.04	.07
σ_{err}	.21	.17	.16	.32	.23	.22
r	.78	.97	.92	.63	.90	.86

TABLE II: Impact of Measurement Number in Propagation Prediction.

Region	Meas #	Performance Metric					
		MSE	e_{min}	e_{max}	e_{mean}	σ_{err}	r
Single Family	A	**.02**	-.17	.27	.03	.21	.85
	B	**.01**	-.18	.24	.02	.2	.93
	C	**.01**	-.13	.2	.02	.13	.95
Down-twon	A	**.05**	-.42	.3	.08	.32	.77
	B	**.03**	-.21	.24	.06	.23	.84
	C	**.02**	-.15	.23	.02	.17	.87

Impact of the Number of Measurements on RAIK Prediction. We now study the impact of the number of measurements in a given 200-m square tile on the accuracy of the KPI prediction using the RAIK framework. For this purpose, we consider tiles that have different ranges of measurements: 600-800, 800-1000, and 1000-1200 depicted as 'A', 'B' and 'C' respectively. A tile has to be within the range to be considered for training the RAIK framework. Table II shows the prediction accuracy for the single-family residential and downtown regions when such an approach is taken for the training. We observe that there is a trade-off that occurs. Increasing the minimum number of measurements forces tiles

to be not considered, having fewer records for input into RAIK. On the other hand, increasing the minimum number allows for better accuracy of the channel characteristics for those tiles that *are* considered. The latter effect can be seen by the increase in the correlation coefficient as the minimum number of measurements required for each tile is raised. Overall, we see a net RAIK prediction benefit to increasing the measurement requirement for both regions considered for these ranges of measurement number.

Homogeneous Training and Testing But in Adjacent Sub-Regions. Next, we study RAIK performance when training and testing in the same region type, which we refer to as homogeneous training and testing. However, the training (70%) and testing (30%) data come from differing sub-regions in the same region type. As depicted in Fig. 5a, the downtown sub-region used for testing is adjacent to the sub-region used for training. This is emulating a crowdsourcing context in which a carrier lacks measurements from users in a certain area but has other users providing data nearby. Fig. 5b shows the results from this particular training/testing data combination. We show that the absolute error between the actual γ and the predicted one is extremely bounded (0.18). In fact, the absolute error of the majority of the testing sub-region is below 0.1.

(a) Adjacent sub-region. (b) Absolute Error.

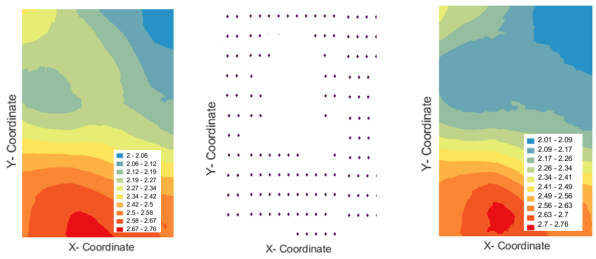
Fig. 5: Prediction in downtown adjacent sub-region.

Table III depicts the results of the testing phase of homogeneous approach. In particular, it can be seen that the model shows a better performance with collected data from the single-family region. This can be explained because the variation of the buildings' height and terrain type in the single-family region is more self-similar as compared with the downtown area. To evaluate RAIK in the context of the most

TABLE III: RAIK with Homogeneous Training and Testing.

Region	**MSE**	e_{min}	e_{max}	e_{mean}	σ_{err}	r
Single Family	**.01**	-.17	.19	.02	.13	.92
Down-town	**.02**	-.2	.24	.09	.21	.89

relevant related works in predicting γ, we compared it with the Kriging algorithm, which is a common approach to address the spatial propagation prediction [28]. To predict the lost data in a region, Kriging employs regression of the surrounding values of that region by assigning weights to these values to capture the spatial correlation of field of interest. Many studies have used Kriging to estimate the path loss [5], [29], [30]. For example, an empirical Okumura-Hata model with Inverse Distance Weighting (IDW) and Kriging has been evaluated in prior work [30]. They have shown that the approach with Kriging shows an improved performance versus just Okumura-Hata model and IDW. Fig. 6a shows the Kriging calculated path loss when all signal quality measurements are used. As we just performed for a downtown region, we introduce a measurement hole, emulating a lack of crowdsourced measurements, shown in Fig. 6b. The dots denote the available points, and the lack of dots denotes lack of signal quality measurements. Fig. 6c shows the Kriging prediction results with these signal quality measurements removed and find the MSE to be 0.07. We perform the same analysis on this region with RAIK and find the MSE to be 0.01. The use of geographical data to predict KPIs reduced the error seven-fold.



(a) Actual Path Loss. (b) Hole in Sub-Region. (c) Kriging Prediction.

Fig. 6: Adjacent sub-region analysis with Kriging Algorithm.

Heterogeneous Training and Testing. Since the aim is providing a generalized RAIK model to predict the channel characteristics based on geographical information alone, we consider the obtained input-output pairs from different base stations located in single-family residential and downtown environments as our data for training. We then test in an entirely new region (multi-family residential) but also test in the two regions that were used for training. There are two issues to evaluate here. First, up to this point, the training and testing has only taken place in the same region. Here, we have two different regions composing the training set, which may confuse the model. Second, we would like to evaluate how well the RAIK framework can predict in region types that have received no training.

Table IV shows the RAIK performance across the three regions. We observe that both the single-family residential region and downtown region perform slightly worse from their respective homogeneous training and testing performance described in Table III. In particular, we observe that the variation of error for single-family and downtown increase from .13 and .21 to .21 and .27, respectively. Although, the MSE for single-family increased from .01 to .03 and downtown increased from .02 to .04, the model still shows an acceptable performance in comparison with the homogeneous model. Furthermore, the performance of the multi-family region that contributed no training data to the RAIK model is 0.03. The variation of error has increased slightly in comparison with two well-trained regions, but it is still comparable. In summary, while slight improvements in RAIK KPI prediction can be achieved when training data is available from adjacent regions of similar geographical features, RAIK can still predict KPIs across regions with distinct geographical features.

TABLE IV: RAIK with Heterogeneous Training and Testing.

Region	**MSE**	e_{min}	e_{max}	e_{mean}	σ_{err}	r
Single Family	**.02**	-.22	.24	.05	.21	.90
Down-town	**.03**	-.27	.3	.07	.27	.87
Multi Family	**.03**	-.3	.33	.09	.3	.82

IV. Related Work

Many different propagation models have been used to predict the coverage area of a network, such as Okumura-Hata [3], [4] and the Longley-Rice irregular terrain model [31]. In these models, one must collect radio signal measurements from a specific region to be able to calibrate the model for that region to find the appropriate constants. The other method to predict propagation coverage over an area is utilizing geostatistical modeling techniques, where the measurements are collected strategically and different interpolation techniques are applied to predict the propagation model of the uncovered locations. For example, a radio environment map of 2.5 GHz WiMax utilized geostatistical modeling and interpolation [5]. Still another work proposed a modified version of Kriging algorithm to reduce the computational complexity of the spatial interpolation to produce the coverage map [6]. In contrast, we specifically target a relationship between the signal quality of a network at a given location and the geographical features in that area to predict the KPIs of that region and regions that lack accessibility or crowdsourced measurements.

V. Conclusion

In this paper, we used the knowledge of geographical features of a region to extend crowdsourced measurements such as those within the MDT effort of the 3GPP standard to predict the KPIs in that region. To do so, we built an Android-based crowdsourcing infrastructure and performed in-field measurements to create a high density of UE measurements in three representative region types: downtown, single-family residential, and multi-family residential. With these RSRP measurements, we studied the relationship between the size of smaller, square sub-regions under consideration with regards to the calculated path loss exponent, showing the tradeoff of too large and too small sub-regions. We then used LiDAR data to extract tree and building data to build a Regional Analysis to Infer Key Performance Indicators (RAIK) framework, which created a relationship between these geographical features and the received signal level in different sub-regions. Using the RAIK framework, we showed that KPIs can be predicted with very low error in areas that lack access or users to produce crowdsourced measurements. We believe that this work will serve as a fundamental step in extending the reach of MDT measurements taken by carriers and thereby reduce the load on users and their devices.

Acknowledgement

This work was in part supported by NSF grants: CNS-1150215, and CNS-1526269. Also, we would like to thank Textron System for their support in this measurement campaign.

References

[1] S. Yi, S. Chun, Y. Lee, S. Park, and S. Jung, *Radio Protocols for LTE and LTE-advanced*. John Wiley & Sons, 2012.

[2] 3GPP, *ETSI TS 137 320 "Radio measurement collection for Minimization of Drive Tests (MDT)"*, 2011.

[3] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE transactions on Vehicular Technology*, vol. 29, no. 3, pp. 317–325, 1980.

[4] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, "Field strength and its variability in vhf and uhf land-mobile radio service," *Rev. Elec. Commun. Lab*, vol. 16, no. 9, pp. 825–73, 1968.

[5] C. Phillips, M. Ton, D. Sicker, and D. Grunwald, "Practical radio environment mapping with geostatistics," in *Proc. of IEEE DySPAN*, 2012.

[6] H. Braham, S. B. Jemaa, B. Sayrac, G. Fort, and E. Moulines, "Low complexity spatial interpolation for cellular coverage analysis," in *Proc. of IEEE WiOpt*, 2014.

[7] S. Sonntag, J. Manner, and L. Schulte, "Netradar-measuring the wireless world," in *Proc. of IEEE WiOpt*, 2013.

[8] A. Nikravesh, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao, "Mobilyzer: An open platform for controllable mobile network measurements," in *Proc. of ACM MobiSys*, 2015.

[9] J. Yoon, S. Sen, J. Hare, and S. Banerjee, "Wiscape: A framework for measuring the performance of wide-area wireless networks," *IEEE Transactions on Mobile Computing*, vol. 14, no. 8, pp. 1751–1764, 2015.

[10] Mobiperf, "Mobiperf, measuring network performance on mobile platforms," 2013.

[11] S. Rosen, S.-j. Lee, J. Lee, P. Congdon, Z. M. Mao, and K. Burden, "Mcnet: Crowdsourcing wireless performance measurements through the eyes of mobile devices," *IEEE Communications Magazine*, vol. 52, no. 10, pp. 86–91, 2014.

[12] J. B. Campbell and R. H. Wynne, *Introduction to remote sensing*. Guilford Press, 2011.

[13] A. Nikravesh, D. R. Choffnes, E. Katz-Bassett, Z. M. Mao, and M. Welsh, "Mobile network performance from user devices: A longitudinal, multidimensional analysis.," in *PAM*, vol. 14, pp. 12–22, Springer, 2014.

[14] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*, vol. 2. Prentice Hall, 1996.

[15] F. Ikegami and S. Yoshida, "Analysis of multipath propagation structure in urban mobile radio environments," *IEEE transactions on Antennas and Propagation*, vol. 28, no. 4, pp. 531–537, 1980.

[16] G. K. Chan, "Propagation and coverage prediction for cellular radio systems," *IEEE transactions on vehicular technology*, vol. 40, no. 4, pp. 665–670, 1991.

[17] R. Enami, Y. Shi, D. Rajan, and J. Camp, "Pre-crowdsourcing: Predicting wireless propagation with phone-based channel quality measurements," in *ACM MSWiM (to appear)*, 2017.

[18] T. Balandier, A. Caminada, V. Lemoine, and F. Alexandre, "170 mhz field strength prediction in urban environment using neural nets," in *Proc. of IEEE PIMRC*, 1995.

[19] R. Fraile and N. Cardona, "Macrocellular coverage prediction for all ranges of antenna height using neural networks," in *Proc. of IEEE ICUPC*, 1998.

[20] P.-R. Chang and W.-H. Yang, "Environment-adaptation mobile radio propagation prediction using radial basis function neural networks," *IEEE transactions on vehicular technology*, vol. 46, no. 1, pp. 155–160, 1997.

[21] I. Popescu, A. Kanstas, E. Angelou, L. Nafornita, and P. Constantinou, "Applications of generalized rbf-nn for path loss prediction," in *Proc. of IEEE PIMRC*, 2002.

[22] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[23] S. Haykin, "Neural networks: a comprehensive foundation prentice-hall upper saddle river," *NJ MATH Google Scholar*, 1999.

[24] M. H. Hassoun, *Fundamentals of artificial neural networks*. MIT press, 1995.

[25] R. Rojas, *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.

[26] A. Neskovic and N. Neskovic, "Microcell electric field strength prediction model based upon artificial neural networks," *AEU-International Journal of Electronics and Communications*, vol. 64, no. 8, pp. 733–738, 2010.

[27] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[28] M. Oliver and R. Webster, "A tutorial guide to geostatistics: Computing and modelling variograms and kriging," *Catena*, vol. 113, pp. 56–69, 2014.

[29] A. Konak, "Estimating path loss in wireless local area networks using ordinary kriging," in *Proceedings of the Winter Simulation Conference*, pp. 2888–2896, Winter Simulation Conference, 2010.

[30] S. Kolyaie and M. Yaghooti, "Evaluation of geostatistical analysis capability in wireless signal propagation modeling," in *Proc. 11th International Conference on GeoComputation*, 2011.

[31] G. A. Hufford, A. G. Longley, W. A. Kissick, *et al.*, *A guide to the use of the ITS irregular terrain model in the area prediction mode*. US Department of Commerce, National Telecommunications and Information Administration, 1982.