Mobile Internet Quality Estimation using Self-Tuning Kernel Regression

Hanyang Jiang [1]   Henry Shaowu Yuchi [1]   Elizabeth Belding [2]   Ellen Zegura [3]   Yao Xie [1]

Abstract

Modeling and estimation for spatial data are ubiquitous in real life, frequently appearing in weather forecasting, pollution detection, and agriculture. Spatial data analysis often involves processing datasets of enormous scale. In this work, we focus on large-scale internet-quality open datasets from Ookla. We look into estimating mobile (cellular) internet quality at the scale of a state in the United States. In particular, we aim to conduct estimation based on highly *imbalanced* data: Most of the samples are concentrated in limited areas, while very few are available in the rest, posing significant challenges to modeling efforts. We propose a new adaptive kernel regression approach that employs self-tuning kernels to alleviate the adverse effects of data imbalance in this problem. Through comparative experimentation on two distinct mobile network measurement datasets, we demonstrate that the proposed self-tuning kernel regression method produces more accurate predictions, with the potential to be applied in other applications.

1. Introduction

Since the invention of cellular-network-based wireless internet access more than two decades ago, the mobile internet has become an integral component of daily life, enabling a myriad of activities across diverse spheres, including communication, entertainment, commerce, and education. While urban areas generally boast robust mobile internet connectivity due to extensive infrastructure, rural areas often experience compromised connectivity (Vogels, 2021). This disparity is largely due to the lack of sufficient cell tow-

---

*Equal contribution [1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA [2]Department of Computer Science, University of California Santa Barbara, California, USA [3]School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia, USA. Correspondence to: Yao Xie <yao.xie@isye.gatech.edu>.

ers and infrastructural limitations, driven by geographical dispersion, lower population density, and reduced economic incentives for investment in these regions. Precisely finding these areas can lead to better planning of networks (Niu et al., 2021).

Identifying these underserved or unserved areas with poor connectivity is vital for telecommunication companies and policymakers (Mangla et al., 2021). By targeting these regions, service providers can strategically enhance infrastructure, improving coverage and service quality. On a broader scale, comprehensive connectivity mapping can inform policy-making, ensuring digital inclusivity and narrowing the digital divide between urban and rural areas. Improving mobile internet connectivity in rural areas also holds the potential for significant socio-economic development, contributing to regional growth and inclusivity. Thus, recognizing and addressing regions with poor mobile internet connectivity is a crucial step toward promoting digital equity and advancing societal progress. In the United States, the Broadband Equity, Access, and Deployment (BEAD) program will allocate substantial resources for expansion in the next few years, guided in part by available data regarding spatial quality (National Telecommunications and Information Administration, 2022).

In this work, we examine the question of predicting mobile internet quality based on an in-depth analysis using one of the largest open datasets, namely the Ookla dataset (Ookla, 2022), which captures mobile internet connectivity measurements worldwide. Users conduct speed tests on their internet connections using Ookla's tools. When a user initiates a speed test, data such as the user's IP address, the Internet Service Provider's (ISP) identity, and the connection's speed (both download and upload measured connection bandwidth) are logged. However, like many real-world datasets, it presents certain challenges. It is characterized by a relatively high variation of data and a conspicuous absence of measurements in specific areas.

To approach these issues, we develop a novel self-tuning bandwidth kernel regression method. The hallmark of our method is its flexibility in the spatial imbalance of data. In contrast to conventional models, our approach is designed to adaptively determine the kernel regression bandwidth for a location. More specifically, we use larger kernel regres-

sion bandwidth where the data gets sparser. This adaptive capability enables our model to more effectively handle the varying data density and spatial distribution inherent in our dataset.

In addition, we are cognizant of the computational resources required to process large quantities of data. To mitigate this issue, we incorporate techniques specifically designed to reduce the computational cost of our model, making it more efficient for real-world deployment. The robustness and effectiveness of our proposed method are substantiated by a comprehensive comparison with two baseline models: Gaussian Process regression (Rasmussen et al., 2006) and the basic kernel regression (Parzen, 1962). This comparative analysis is carried out across two distinct datasets, offering further insight into the performance and adaptability of our method.

In the domain of data-driven mobile internet quality prediction, there are two principal approaches currently: geospatial interpolation (Molinari et al., 2015; Riihijarvi & Mahonen, 2018; Tripkovic et al., 2021; Eller et al., 2021; Chakraborty et al., 2017) and supervised machine learning (Alimpertis et al., 2019; Rozenblit et al., 2018; Eller et al., 2022). Geospatial interpolation employs sparse measurements, treated as anchoring points, to generate dense signal-strength maps in their immediate vicinity. Kriging is used in these works (Molinari et al., 2015; Riihijarvi & Mahonen, 2018; Eller et al., 2021). (Molinari et al., 2015) compares the performances of Kriging and points out that regular Kriging is fairly robust. (Eller et al., 2021) develops a propagation-aware Gaussian process regression considering factors like the block of buildings. While these techniques are conducive to estimating performance in existing networks, they offer limited applicability in network planning as adjustments can only be studied post-deployment. Furthermore, barring a few exceptions, these methodologies do not inherently account for the effect of the propagation environment.

On the other hand, supervised machine learning methods have also been explored (Enami et al., 2018; Wu et al., 2020; Rozenblit et al., 2018; Alimpertis et al., 2019). Some techniques (Alimpertis et al., 2019) primarily focus on creating local performance maps, utilizing absolute coordinates or cell identifiers. However, these approaches lack cross-area generalization capabilities. There exist alternative supervised strategies (Shoewu et al., 2018; Jo et al., 2020) that restrict themselves to area-independent features, like the distance between the user equipment and base stations. Nevertheless, in numerous cases, any observed error reductions predominantly arise from fitting the distribution of a specific measurement campaign rather than enhancing the efficacy of traditional methods at a macroscopic level.

In this study, we aim to create an estimation of mobile internet quality that covers an entire state, not just urban areas as many previous studies have done. Some works including (Adarsh et al., 2021) realize the importance of this field, but they focus more on the data collection and analysis part instead of developing methods. The lack of data from rural areas makes this a tough challenge, setting our work apart from existing approaches. We also show that our method performs much better than Kriging which is commonly used in geospatial interpolation work. The results of our work can guide active sampling in underserved areas, helping to improve our study further.

The rest of the paper is organized as follows. We first discuss the literature relevant to cellular network connectivity prediction. Then we give a detailed description of the datasets used in the paper. After that, we introduce the methods used in this paper. Finally, we show the numerical experiments on the two different datasets.

2. Data Description

The dataset utilized in our study is sourced from the open datasets provided by Ookla. The data gathered by Ookla from 2019 to 2022 encapsulates the performance metrics of mobile internet connections for a multitude of users worldwide. Key variables in this dataset include geographical coordinates (longitude and latitude), mean download speed (MB/s), mean upload speed (MB/s), count of tests conducted in each area (aggregated for user privacy into $600m^2$ grid blocks), the number of distinct devices utilized for testing, and a comprehensive score assessing the connection speed, among other variables.

The following provides a brief description of some of these variables:

- Location: Denotes the location of the center of a $600m^2$ grid block where the users measure the connection speed.

- Download speed: Represents the measured rate of data transfer from the server to the user's device. This is a critical metric as it affects the speed of web page loading and file downloading, and hence is quite noticeable to the user. Download speeds are reported as an average in the grid location.

- Upload speed: Corresponds to the measured data transfer rate from the user's device to the server. This metric is crucial for tasks like video calls, cloud file uploads, and user-generated live streaming. Upload speeds are reported as an average in the grid location.

- Number of tests at each location: This refers to the number of times tests were conducted in a given grid block. More tests tend to indicate a more reliable
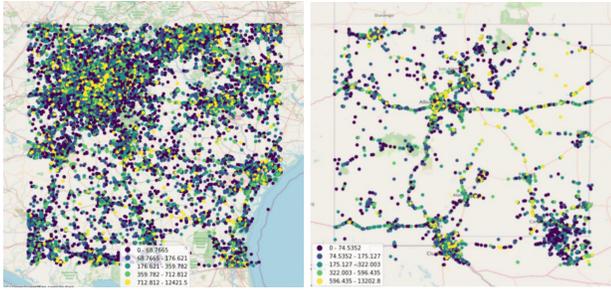
Figure 1. Mobile connection score of Georgia and New Mexico, where the data in Georgia is very dense, and in New Mexico is sparse.
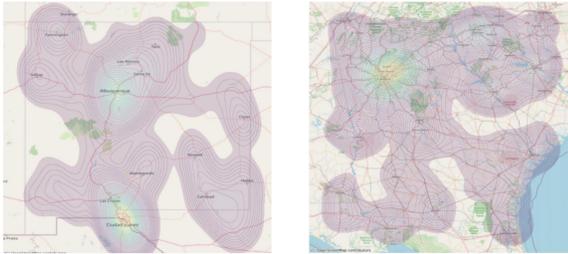
Figure 2. Kernel density estimation of Georgia and New Mexico dataset.

measurement score. However, in many locations, only a single test is conducted.

- Number of devices used for testing: This refers to the variety of devices used for testing in a certain grid block, as the performance can vary across different mobile phones and mobile providers.

- Score: This variable combines both the download and upload speeds to provide a holistic view of the connection bandwidth performance.

We have mobile connection data from the states of Georgia in the United States Southeast and New Mexico in the United States Southwest. The geographic scatter plot of the connection scores for each dataset is shown in Figure 1. There are 28,587 data points in the Georgia dataset. The mean score across grid blocks is 460.067, the min score is 0.056, and the max score is 12421.537. The standard deviation of the score is 616.3. The frequency of scores is shown in Figure 3. There are 7,579 data points in the New Mexico dataset. The mean score across grid blocks is 388.128, the min score is 0.174, and the max score is 13202.754. The standard deviation of the score is 504.5. The frequency of scores is shown in Figure 4.
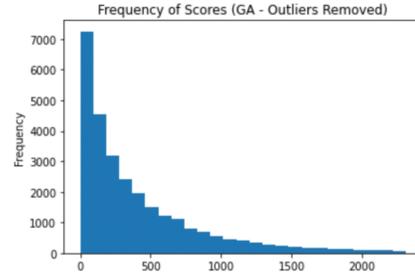
Figure 3. The frequency of scores in Georgia dataset.
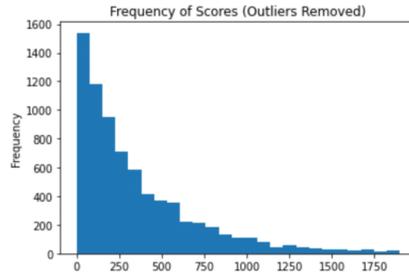
Figure 4. The frequency of scores in New Mexico dataset.

3. Methodology

We first employ standard kernel density estimation (Rosenblatt, 1956) to generate a heatmap depicting the mobile connection quality across two distinct geographical areas, namely Georgia and New Mexico. These heatmaps, as displayed in Figure 2, offer an illustrative representation of connection quality, enabling us to visually assess the spatial distribution of mobile connectivity in these regions.

However, upon inspection of the resulting heatmaps, it becomes evident that the standard kernel density estimation approach does not provide satisfactory coverage in certain rural areas. There are locations where the recorded measurements are notably sparse, resulting in large regions with no substantial estimations. Given that our study has a particular interest in understanding and improving the mobile connection quality in these rural and underserved areas, this observation underscores a significant limitation of the simple kernel density estimation method.

The apparent lack of estimations in the rural regions does not necessarily indicate an absence of mobile connectivity; rather, it highlights the deficiency of data in these areas, potentially due to their sparse population or challenging geographical conditions. It is precisely these areas where there is a need for robust and accurate analytical methods. Given that conventional kernel density estimation falls short in this regard, our observations demonstrate the necessity for adopting more sophisticated, data-driven techniques.

The more advanced methodologies should ideally be able

to compensate for the sparse data in rural areas and provide reliable estimations, thereby offering a more comprehensive view of the mobile connectivity landscape. Moreover, this situation also points towards the need for further data collection efforts in such under-represented regions, which can help in refining our analysis and augmenting the efficacy of our predictive models.

3.1. Gaussian Process Regression (Kriging)

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. This powerful Bayesian methodology can be used for various tasks such as regression, classification, and optimization. Gaussian processes are completely determined by their mean and covariance functions, and they provide a probabilistic, non-parametric approach to modeling data, meaning they can flexibly model complex relationships without relying on a fixed functional form.

Denote the output of the Gaussian process by $f(\mathbf{x})$, where $\mathbf{x}$ is the input. We can then express a Gaussian process as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \qquad (1)$$

where $m(\mathbf{x})$ is the mean function, and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function. They determine the properties of the functions drawn from the GP.

Given a set of n observations $\mathbf{y} = [y_1, \ldots, y_n]^T$ at locations $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, we can express the joint distribution of these observations as follows:

$$\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(m(\mathbf{X}), K(\mathbf{X}, \mathbf{X})). \qquad (2)$$

Here, $m(\mathbf{X})$ is a vector where the i-th entry is $m(\mathbf{x}_i)$, and $K(\mathbf{X}, \mathbf{X})$ is a covariance matrix where the entry at the i-th row and j-th column is $k(\mathbf{x}_i, \mathbf{x}_j)$.

Given the observed data, we can make predictions for new inputs $\mathbf{x}_*$ using the posterior predictive distribution:

$$f(\mathbf{x}_* \mid \mathbf{y}, \mathbf{X}, \mathbf{x}_*) \sim \mathcal{N}\left(\bar{f}_*, \text{cov}(f_*)\right), \qquad (3)$$

where $\bar{f}_*$ is the predictive mean and $cov(f_*)$ is the predictive covariance.

The Gaussian process offers several advantages, such as providing a measure of uncertainty (via the predictive variance) in addition to point estimates and the flexibility of specifying different covariance functions. However, they also have some drawbacks. Its computational complexity scales cubically with the number of observations, which makes it difficult to tackle large datasets.

In our problem, we denote the inputs and response by $x \in \mathbb{R}^2$ and $y \in \mathbb{R}$, respectively. we utilize the kernel

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp(-(\mathbf{x} - \mathbf{x}')^T \text{diag}(\theta_1, \theta_2)(\mathbf{x} - \mathbf{x}')), \qquad (4)$$

where θ_1, θ_2 and σ are parameters to be trained through maximum likelihood. Once the model parameters are trained, we can use them for prediction.

3.2. Kernel Regression

Kernel regression is a non-parametric technique to estimate the conditional expectation of a random variable. The objective is to find a non-linear relationship between the input variable and the corresponding output. Kernel regression employs kernel functions, which allow it to capture more complex patterns than linear regression.

For a given dataset with inputs $\mathbf{x}$ and corresponding outputs $\mathbf{y}$, the kernel regression estimate $\hat{y}$ at a new input point x is given by:

$$\hat{y}(x) = \frac{\sum_{i=1}^n K_h(\|x - x_i\|) y_i}{\sum_{i=1}^n K_h(\|x - x_i\|)}. \qquad (5)$$

Here, $K_h(u) = \frac{1}{h} K(\frac{u}{h})$ is a kernel function, and h is a bandwidth parameter. The kernel function $K(\cdot)$ is often chosen to be a Gaussian kernel, though other choices are also possible. The bandwidth parameter h controls the width of the kernel, and hence the smoothness of the estimated function.

In the Gaussian case, the kernel function $K(\cdot)$ is defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right). \qquad (6)$$

Kernel regression estimates the conditional mean function without imposing a parametric form for the functional relationship between predictors and the outcome variable. It allows for flexible, data-driven model specification.

The bandwidth h is a crucial parameter in kernel regression. If h is too small, the estimate will be very rough, capturing too much noise in the data (overfitting). If h is too large, the estimate will be too smooth, not capturing important patterns in the data (underfitting). The choice of an appropriate h often involves cross-validation or some other form of bandwidth selection strategy.

3.3. Self-tuning Bandwidth Kernel Regression

Kernel regression, by its standard definition, utilizes a constant bandwidth h that is applied uniformly across all data points. However, this one-size-fits-all approach might not be the most suitable in scenarios where the data exhibits spatial imbalances, as is the case in our study.

In regions where data points are sparse, the lack of neighboring points could potentially lead to unrepresentative averages and, consequently, inaccurate predictions. To overcome this, we propose the use of a larger bandwidth in such
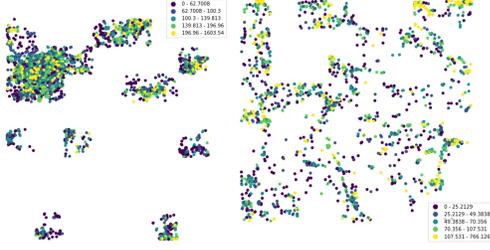
Figure 5. Visualization of dense and sparse regions in the Georgia dataset.
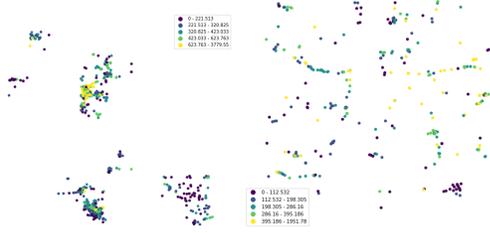
Figure 6. Visualization of dense and sparse regions in the New Mexico dataset.

areas, thereby encompassing more points for computation and increasing the representativeness of the estimates.

To this end, we propose the self-tuning bandwidth in the kernel regression framework (STBKR). The self-tuning bandwidth mechanism adapts the bandwidth for each point based on its surrounding density of points. The formula for our self-tuning bandwidth kernel regression can be expressed as follows:

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} K_{h(x)}(\|x - x_i\|)y_i}{\sum_{i=1}^{n} K_{h(x)}(\|x - x_i\|)}. \quad (7)$$

In this formula, we let $h(x) = cR_k(x)^2$. Here, c is a parameter that is determined by cross-validation, and $R_k(x)$ denotes the Euclidean distance from a given data point x to its k-th nearest neighbor. Notably, in areas where data is sparsely distributed, the distance $R_k(x)$ will be larger, thereby leading to an increased bandwidth $h(x)$.

This strategy of self-tuning bandwidth effectively addresses the issue of spatial imbalance in the dataset. By allowing the bandwidth to adapt based on the local data density, it ensures a more representative sampling of neighbors, leading to more accurate and reliable predictions. Furthermore, the choice of c through cross-validation aids in avoiding overfitting or underfitting, further strengthening the robustness of our regression model.

3.4. Data Preprocessing and Efficient Computation

The dataset under consideration presents a challenge due to its high variation and the fact that only a single measurement is taken at approximately half of the locations. This inherent variability in the data leads to the need for a variance reduction technique, specifically, the application of a k-nearest neighbor averaging method.

The k-nearest neighbor approach enables us to approximate the value at a given point x by utilizing the mean of its k nearest neighbors. Denote the k nearest neighbors to a given point x as $x_{n(x)_1}, x_{n(x)_2}, \cdots, x_{n(x)_k}$. The initial value y at point x is substituted by the average of these neighboring points as follows:

$$y_{\text{avg}}(x) = \frac{1}{k}(y_{n(x)_1} + y_{n(x)_2} + \cdots + y_{n(x)_k}). \quad (8)$$

This averaging process inherently smoothens the dataset, helping to alleviate the issue of high variance.

However, a secondary challenge presents itself due to the size of the dataset. The computation of kernel regression predictions across the entire test set can take up to several hours, a prohibitively long duration for many applications. To mitigate this computational cost while retaining the prediction accuracy of our model, we resort to another variant of the k-nearest neighbor method, this time employing a weighted sum instead of summing over the entire training set as depicted in Equation (5).

With the incorporation of a kernel function $K_h(\cdot)$ that introduces a weighting based on the distance between the points, the predictive function $\hat{y}(x)$ is calculated as follows:

$$\hat{y}(x) = \frac{\sum_{i=1}^{k} K_h(\|x - x_{n(x)_i}\|)y_{n(x)_i}}{\sum_{i=1}^{k} K_h(\|x - x_{n(x)_i}\|)}. \quad (9)$$

By adopting this approach, we can preserve the accuracy of the algorithm while dramatically reducing the computational time. This modification is instrumental in making the kernel regression computationally feasible for large datasets, thus striking a desirable balance between computational efficiency and predictive accuracy.

4. Numerical Experiments

4.1. Experiment Setting

We initiate our comparative study by partitioning the data into training and test sets with a split of 80% and 20% respectively. In our comparison, we feature our proposed self-tuning bandwidth kernel regression method (STBKR) and compare it with two baseline methods: The Gaussian process regression (GP) and fixed bandwidth kernel regression (FBKR). However, Gaussian process regression presents a

Table 1. Errors in dense areas of Georgia dataset.

	GP	FBKR	STBKR
MAE	245.62	83.03	81.6
MSE	144105.22	20746.26	19594.86
MNE	5524.22	2334.2	2248

Table 2. Errors in sparse areas of Georgia dataset.

	GP	FBKR	STBKR
MAE	178.69	102.1	102.9
MSE	73267.2	47111.7	37534.83
MNE	2312.76	2399.2	1942.5

Table 3. Parameters for dense and sparse areas in Georgia dataset.

	FBKR	STBKR
DENSE	$k = 5, c = 0.01$	$k = 10, c = 0.01$
SPARSE	$k = 5, c = 0.05$	$k = 5, c = 0.05$

Table 4. Parameters for dense and sparse areas in New Mexico dataset.

	FBKR	STBKR
DENSE	$k = 5, c = 0.005$	$k = 5, c = 0.05$
SPARSE	$k = 5, c = 0.075$	$k = 5, c = 0.02$

computational challenge when dealing with large training data due to its requirement to compute the inverse of a massive matrix. We address this problem by sparsifying the training data through downsampling, followed by applying the Gaussian process to the thinned-out data.

To optimize the parameters h, k, and c for the other two kernel regression methods, we implement a 5-fold cross-validation on the training set. The loss functions we apply in this study are Mean Absolute Error (MAE), Mean Squared Error (MSE), and Maximum Norm Error (MNE).

Furthermore, we segment our dataset into dense and sparse regions. To achieve this, we first divide the map into a 15×15 uniform grid. Subsequently, we count the data points within each region. Regions with a count higher than the average are classified as dense, while the remaining ones are classified as sparse.

Regarding the kernel regression methods, we tune the bandwidth magnitude c and the number of neighbors k separately in the dense and sparse regions. The motivation behind this approach is the reduced quantity of points in the sparse regions, which necessitates a larger bandwidth to include a comparable number of neighbors. This tailored treatment enhances the adaptability of our methods to the spatial distribution of the data, promoting better overall performance.

Besides predicting the connection speed, we also tried a simpler task which is to predict whether an area's connection is served, underserved, or unserved. This distinction is useful in policy-making, as it has been used to prioritize the order of broadband network expansion.

- Served: Download speed $\geq$ 100Mbps and upload speed $\geq$ 20Mbps.

- Underserved: Download speed $\geq$ 25Mbps and upload speed $\geq$ 3Mbps and is not in served category.

- Unserved: The rest cases.

4.2. Prediction

In this experiment, we want to predict the score of each location in the test set accurately. First, we use 5-fold cross-validation on the training set to choose the best parameter c and k in dense regions and sparse regions for both datasets. The results are shown in Table 3 and 4.

For the Gaussian process, we do downsampling on training sets to reduce computational costs. To utilize the training data more efficiently and reduce the uncertainty in unmeasured areas, we split the map into a 15×15 uniform grid and select training data uniformly in each grid. This makes the training data spread more uniformly and thus greatly reduces the uncertainty around the training set.

Then we compare the three methods separately in dense areas and sparse areas of the Georgia and the New Mexico dataset.

Upon analysis of the experimental results, it becomes apparent that the Gaussian Process (GP) regression underperforms in regions of higher data density compared to those characterized by sparsity. This deficiency may be attributed to the employed downsampling technique, which uses a uniform distribution of training data across the spatial domain, thus failing to sufficiently represent areas of high data concentration. Furthermore, GP demonstrates much worse performances in comparison to kernel regression methods across both region types. This can be ascribed to GP's inherent limitation in handling large volumes of training data. Despite the allocation of 80% of the total data to the training set, the downsampling process results in a mere 10% of the data being used for training GP. In stark contrast, kernel regression methods exhibit the capacity to effectively utilize the entirety of the training set, leading to superior performance.

In a comparison of the standard kernel regression method

Table 5. Errors in dense areas of New Mexico dataset.

	GP	FBKR	STBKR
MAE	172.31	53.13	54.55
MSE	80417.66	9224.37	8950.31
MNE	3038.29	1055.09	1049.89

Table 6. Errors in sparse areas of New Mexico dataset.

	GP	FBKR	STBKR
MAE	115.3	77.34	77.55
MSE	33754.41	31870.79	24624.49
MNE	1905.78	1172.39	1057.99

and the proposed self-tuning kernel regression, our method consistently manifests superior performance in both regions across the datasets in question. While the Mean Absolute Error (MAE) remains relatively consistent between the two kernel regression methods, our self-tuning method exhibits a marked decrease in Mean Squared Error (MSE) and Maximum Norm Error (l_∞ error). This indicates that our method offers enhanced predictive accuracy in those locations deemed more challenging, aligning with the primary objectives of our research.

4.3. Classification

In this section, we classify locations into three types: served, underserved, and unserved regions. For the Georgia dataset, we have 48% served, 49% underserved, and 3% unserved areas. In the New Mexico dataset, 21% are served, 60% are underserved, and 19% are unserved. This shows significant differences in mobile connectivity between these two states.

We use the three methods discussed previously to predict both download and upload speeds. Each point in the test set is then classified based on these predictions. The final accuracy of this classification is provided below.

From the table 7 and 8, we can see that our proposed method still generates a more accurate classification in all cases except the sparse region in the Georgia dataset. However, the difference is very small in that case, which does not obscure the overall advantage of our method.

5. Conclusion

In this work, we focus on the cellular network quality estimation based on Ookla open dataset. We proposed a novel self-tuning bandwidth kernel regression method (STBKR) for the specific problem on two different datasets. The difficulty of the problem comes from data imbalance and high
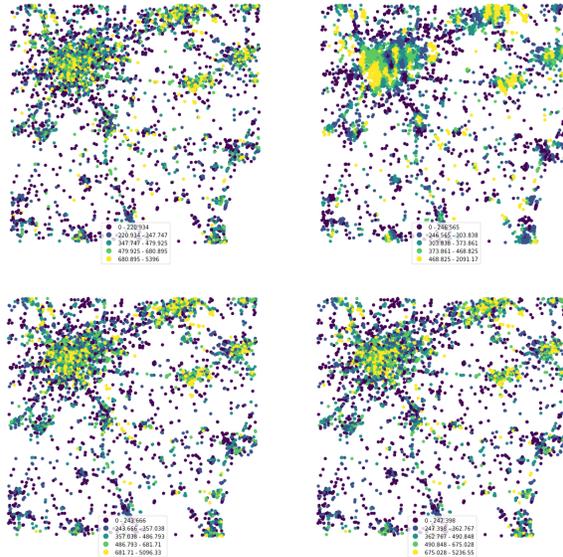
Figure 7. Top-left is the original test data after data averaging. Top-right is the prediction of GP. Bottom-left is the prediction of kernel regression. Bottom-right is the prediction of self-tuning kernel regression.

Table 7. Classification accuracy on Georgia dataset.

	GP	FBKR	STBKR
DENSE	52%	86.6%	88.3%
SPARSE	67.4%	86.2%	85.9%

variation. We compared our method with two established approaches: the Gaussian process (GP) and the basic kernel regression. Our findings indicate that GP struggles with the dataset, especially in dense regions, whereas both kernel regression methods perform impressively due to their ability to employ the entire training set effectively.

Our method consistently shows superior performance over the other two baseline methods, particularly at locations with challenging conditions. This is further validated in our classification experiment, which seeks to categorize various locations based on connectivity conditions. Despite substantial differences in mobile connectivity across the datasets, our proposed method demonstrates superior classification accuracy.

To summarize, our self-tuning bandwidth kernel regression method offers a promising alternative for dealing with large datasets, delivering high predictive accuracy and effective classification performance across diverse regions. Its superiority over the Gaussian process regression and fixed bandwidth kernel regression methods substantiates its potential as a valuable tool in data analysis.
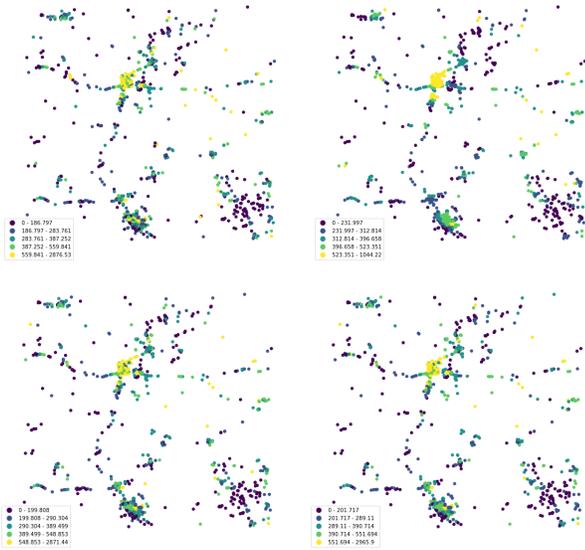
Figure 8. Top-left is the original test data after data averaging. Top-right is the prediction of GP. Bottom-left is the prediction of kernel regression. Bottom-right is the prediction of self-tuning kernel regression.

Table 8. Classification accuracy on New Mexico dataset.

	GP	FBKR	STBKR
DENSE	27.4%	88.7%	91.3%
SPARSE	20.8%	90.9%	93.3%

References

Adarsh, V., Nekrasov, M., Paul, U., Mangla, T., Gupta, A., Vigil-Hayes, M., Zegura, E., and Belding, E. Coverage is not binary: Quantifying mobile broadband quality in urban, rural, and tribal contexts. In *2021 International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–9. IEEE, 2021.

Alimpertis, E., Markopoulou, A., Butts, C., and Psounis, K. City-wide signal strength maps: Prediction with random forests. In *The World Wide Web Conference*, pp. 2536–2542, 2019.

Chakraborty, A., Rahman, M. S., Gupta, H., and Das, S. R. Specsense: Crowdsensing for efficient querying of spectrum occupancy. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9. IEEE, 2017.

Eller, L., Svoboda, P., and Rupp, M. Propagation-aware gaussian process regression for signal-strength interpolation along street canyons. In *2021 IEEE 93rd Vehicu-lar Technology Conference (VTC2021-Spring)*, pp. 1–7. IEEE, 2021.

Eller, L., Svoboda, P., and Rupp, M. A deep learning network planner: Propagation modeling using real-world measurements and a 3d city model. *IEEE Access*, 10: 122182–122196, 2022.

Enami, R., Rajan, D., and Camp, J. Raik: Regional analysis with geodata and crowdsourcing to infer key performance indicators. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6. IEEE, 2018.

Jo, H.-S., Park, C., Lee, E., Choi, H. K., and Park, J. Path loss prediction based on machine learning techniques: Principal component analysis, artificial neural network, and gaussian process. *Sensors*, 20(7):1927, 2020.

Mangla, T., Showalter, E., Adarsh, V., Jones, K., Vigil-Hayes, M., Belding, E., and Zegura, E. A tale of three datasets: Towards characterizing mobile broadband access in the united states. *arXiv preprint arXiv:2102.07288*, 2021.

Molinari, M., Fida, M.-R., Marina, M. K., and Pescape, A. Spatial interpolation based cellular coverage prediction with crowdsourced measurements. In *Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data*, pp. 33–38, 2015.

National Telecommunications and Information Administration. Broadband equity, access and deployment (bead) program, 2022. URL https://broadbandusa.ntia.doc.gov/funding-programs/broadband-equity-access-and-deployment-bead-progr Accessed: 2023-06-02.

Niu, Z., Zhou, S., and Crespi, N. Greening 6g: New horizons. *Shaping Future 6G Networks: Needs, Impacts, and Technologies*, pp. 39–53, 2021.

Ookla. Global broadband performance metrics: Q1. http://www.speedtest.net/global-index, 2022. Accessed:2022-09-20.

Parzen, E. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3): 1065–1076, 1962.

Rasmussen, C. E., Williams, C. K., et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

Riihijarvi, J. and Mahonen, P. Machine learning for performance prediction in mobile cellular networks. *IEEE Computational Intelligence Magazine*, 13(1):51–60, 2018.

Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pp. 832–837, 1956.

Rozenblit, O., Haddad, Y., Mirsky, Y., and Azoulay, R. Machine learning methods for sir prediction in cellular networks. *Physical Communication*, 31:239–253, 2018.

Shoewu, O. O., Adedoyin, M. A., Akinyemi, L. A., and Oborkhale, L. I. Fuzzy-logic based path loss models for metropolitan environment. In *2018 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 71–76. IEEE, 2018.

Tripkovic, S., Svoboda, P., Raida, V., and Rupp, M. Cluster density in crowdsourced mobile network measurements. In *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pp. 1–7. IEEE, 2021.

Vogels, E. A. Some digital divides persist between rural, urban and suburban america. 2021.

Wu, L., He, D., Ai, B., Wang, J., Qi, H., Guan, K., and Zhong, Z. Artificial neural network based path loss prediction for wireless communication network. *IEEE access*, 8:199523–199538, 2020.