Adjusted KNN Model in Estimating User Density in Small Areas with Poor Signal Strength

Rong Duan , Guang-qin Ma
rduan,gma @research.att.com
AT&T Labs
200 S Laurel Ave
Middletown, NJ 07748

Abstract
Localized user density estimation is fundamental in many fields such as urban planning, traffic engineering, disease control, location based marketing and telecomm capacity planning. Modern mobility technologies provide the capability for measuring the localized user density dynamically and precisely. However, this is only limited to the areas that have good signal strength. It is a challenge to accurately estimate user density for areas with poor signal strength. However, user density can be estimated from other big data collected by telecommunication providers from different sources. This paper is a case study leveraging big data for developing a business solution. Exploratory Data Analysis (EDA) is applied to quantify the good signal vs bad signal, and a group of important variables that are highly related to user density are selected. An adjusted K-Nearest-Neighbor is applied to infer bad coverage user densities from the good coverage areas. Instead of predefining the K, different percentile measurements are provided to increase the robustness in business decision.

.

Keywords
Business Analytics, Human mobility, KNN, Signal Strength, Variable Selection

1. Introduction

With the dramatic growth in mobile data traffic, mobile service providers are constantly upgrading their technology and infrastructure. The macrocell technologies have evolved from 2G, 3G to 4G/LTE in the past decade with significant improvements in speed, capacity and coverage. However, with the limitations of radio signal penetration, signal strength and coverage might still not be good enough to satisfy traffic demand in high traffic volume areas, hard-to-reach sites or indoor locations. Furthermore, other factors constrain the further implementation of macrocell technologies. For example, tower zoning and permitting regulations, limited licensed wireless spectrum, radio interference among different macrocells and the high cost. As a complementary technology, flexible and low cost smallcell is proposed to enhance localized capacity and coverage. Smallcell technology is developed to separate the licensed macrocell spectrum into smaller chunks and reuse it to improve the small area signal strength. Distributed Antenna System (DAS), Pico Cell, femtocell are different smallcell technologies that are used to improve the signal strength inside buildings or small outdoor areas to extend the signal coverage. Market analysts from Infonetics Research forecasted a $2.7 billion smallcell market by 2017[1]. The major mobile service providers have started to install smallcell services. AT&T has

announced the smallcell deployment in Disney parks [2] and Verizon has stated its smallcell deployment into their 4G LTE network [3]. To identify the smallcell locations is the first critical step in the success of the smallcell deployment, and localized user density estimation is the foundation for this step. The user density is defined as the unique number of mobile devices in a bin. The challenges for estimating the smallcell user density are the dynamics of human mobility and the small coverage area. A typical macrocell could cover an area with a radius of a few tens of kilometers, but a smallcell, by design, only covers 10 meters to 1 or 2 kilometers [4]. With the relative larger coverage area, macrocell user density can be roughly estimated by population density and business density, but this method isn't appropriate for smallcells due to human mobility. As illustrated in Fig. 1, the small red dots are the user locations and the large black dot is the macrocell antenna facing Miami International airport and Dolphin ExpressWay. The macrocell users are not distributed uniformly. The major highway and airport terminals have more users than the lagoon, minor roads and airport runway. User locations are highly related to the location type.

Figure (1) User distribution for one macrocell antenna. The big black dot is the macrocell antenna location and the small red dots are the user locations. The major highway and airport terminals have more users than the lagoon, minor roads and airport runway. User locations are highly related to the location type.

Fig. 2 shows the scatterplot of user density vs population density for an area, where the user density is the actual mobile device user density for the bin and population density is the residential density and employee density for the bin. (a) All bins in the area, (b) are the bins with at least -75db received signal code power (RSCP) and (c) are the bins with at least -100db RSCP. It can be seen, user density for each bin can't be linearly estimated from population

density directly. The reason to choose the bins with different RSCP is to eliminate the possible missing user count issue when signal strength is poor. We will discuss how to quantify the poor signal strength in Section 2.1. Population density is a static statistics, it only represents people's residential or working locations. It is not able to characterize the mobile device usage locations. As shown in Fig. 1, the highway area has high-density mobile device users, but low residential or employee density.

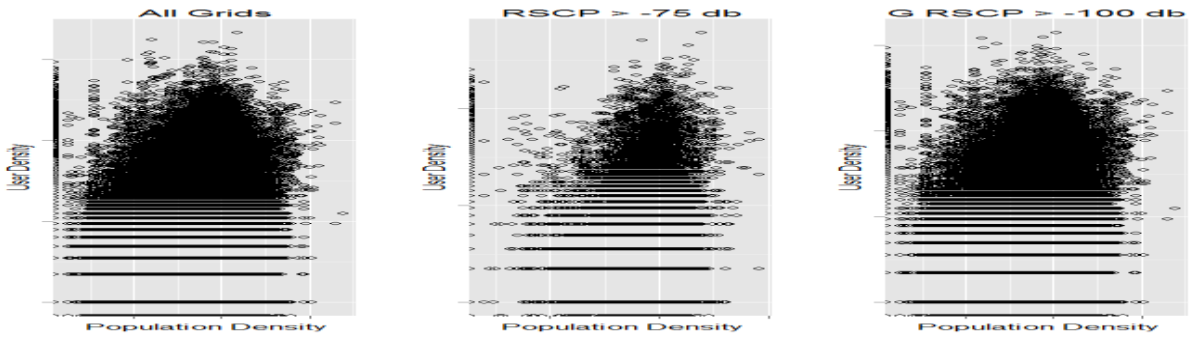


Figure (2) User density vs. Population density scatter plots for the bins with different signal strength. Each circle represents one bin. Both x-axis and y-axis are log scale.

With the development of position technologies, mobile devices can be tracked precisely. Cell Tower Triangulation (CTT) and Assisted GPS (AGPS) are two positioning technologies used by mobile service providers. The Cell Tower Triangulation (CTT) method calculates the mobile device position base on the time difference that the signal travels among different cell towers and the distance among the cell towers. AGPS utilizes the established Global Navigation Satellite System information. It records the satellite positions in an assistance server, and integrates it with cell tower information to enhance the position performance. These two methods depend on the signal strength between the mobile device and the base stations. They perform well only in the good radio signal coverage areas. The user density will be underestimated for poor coverage areas due to the missing connectivity between the mobile devices and the base stations.

Mobile service providers have collected a lot of information related to capacity planning and network dimensioning, which includes geographic, demographic and firmographic information related to the area. In this paper, we propose a big data approach to estimate the user density in poor coverage area from good coverage areas. First, we quantify the poor signal strength threshold and identify the variables that are highly related to user density. Then for each small bin with poor signal strength, we apply the adjusted K-nearest neighbors model to infer the user density from the good coverage candidate bins.

This paper is organized as follows: Section 2 is data exploration to quantify the poor signal strength and user density related variable selection. Section 3 introduces the system and the proposed method. Section 4 evaluates the performance of the proposed method based on real world application. A conclusion and future work are given in Section 5.

2. Data and Model Exploration

To optimize capacity planning and monitor network performance, mobile service providers have collected a large amount of nationwide data related to each geographic area. The geographic area can be split into 100m X 100m bins, and it can be further drilled down to 25m X 25m for metropolitan areas if needed. These data can be categorized into four major groups: performance measurements, position information, location information and usage information. No individual customer information is used in the study – all information is anonymous and characterized at bin level. Performance measurements include RSCP, dropped call count, successful call count; Position information consists of latitude, longitude, Zip Code, county, state, etc.; Location information includes the location type(suburban or urban, indoor or outdoor), population density, employee density, etc., geographic, demographic and firmographic information; Usage information includes the number of users and the data volume they used in the area. All the candidate variables and the descriptions are described in Table (1). A region that covers more than 400k bins is used as example in this study and all the scales are removed from the plot to mask the business information. Each variable is normalized by Median Absolute Deviation (MAD) to handle the outliers.

Group	Name	Description
Position	Latitude, Longitude	Bin bottom left corner Latitude and Longitude
	Zip Code,County,State	Bin Zip Code, County, State
Location	Area Type	Urban, Suburban, Rural, Venue site Indicator
	Usage Type	Indoor, Outdoor, mixed Indicator
	Terrain Clutter Type	Bin terrain type
	House Unit Density	House unit density for the bin from Census
	Population Density	Population density for the bin from Census
	Employee Density	Employee density for the bin from Duns & Bradstreet
	Business Density	Business density for the bin from Duns & Bradstreet
	Business Type	Business type from Standard Industrial Classification(SIC) code
	Nearest Business Type	The closest business type
	Nearest Business Distance	The distance between the bin and the closest business
	Nearest Road ID, County, State	The closest road ID
	Nearest Road Type	The closest road type defined in Census TIGER
	Nearest Road Distance	The distance between the bin and the closest road
Performance	RSCP Median	The median of the RSCP for all mobile devices in the bin
	RSCP Median 850	The median of the RSCP for 850 band mobile devices in the bin
	RSCP Median 1900	The median of the RSCP for 1900 band mobile devices in the bin
	Dropped Call	Dropped Call Count in the bin

	Successful Call	Successful Call Count in the bin
Usage	Unique User Count	Unique number of mobile device in the bin
	Voice/Data Call Count	Number of Voice Call and Data Call count
	Voice Erlang	Voice Usage
	Data Uplink/Downlink Volume	Data Usage

Table(1) Candidate Variables and Definitions

2.1 Quantitative definition of good signal strength vs. poor signal strength

RSCP is a signal strength quality index in UMTS wireless network. Its range is from -25dbm to -120dbm, the higher the better. RSCP impacts the wireless service quality. The call might be dropped if RSCP is not strong enough, but the user density is not impacted under this situation since the user has been counted in the area. But when the RSCP is too weak to establish a call, the user density will be underestimated. To quantify the RSCP threshold that influences the user density measurement, we explore the relationship between user density and signal strength. In Fig. 3, each black point represents one bin, X-axis and Y-axis are the user density and RSCP measurement for each bin respectively. The straight blue line is the RSCP equal to -100dB. In Fig. 3, we see for the bins with RSCP $\leq$ -100db, the user density is obviously smaller than the bins with RSCP >-100db, and decreasing as RSCP is decreases below -100dB. This phenomena can be interpreted as user density is underestimated when RSCP $\leq$ -100db. Since the bin size is fixed in this study, we will use user density and user count interchangeable here. According to the above observation, the threshold for segmenting good and bad signal strength can be defined as -100db. For good single strength area, the user density is correctly measured, and for the bad signal coverage area, the user density is underestimated. The objective is to estimate the user density for the bad signal strength area base on those good signal strength candidate bins.

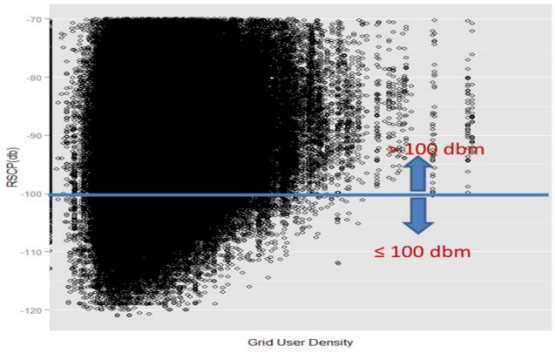


Figure (3). User density relation with signal strength. Each black circle represents one 100m X 100m bin. Y-axis is RSCP and X-axis is user density. The user density is obviously smaller while RSCP <-100db.

2.2 Model and Variable Selection

After quantifying the signal strength threshold, we still need to identify the variables that impact the user density, since the irrelevant variables will decrease the performance due to the ***curse of dimensionality***. Regression tree and exhaustive linear models are used to explore the relationship between user density and other variables in bins with good signal strength. As the regression tree model result showed in Fig. 4, Nearest Road Type, Nearest Road Distance, Longitude, Employee density and Terrain Type are identified as important variables that impact user density. But the model result also divulges that the tree model is not good enough to estimate user density directly. There are totally 459,068 bins in the collected data and the average user count is 380 in each bin. 81% of bins are categorized in the same group with 251 average user count in each bin. The reason for the poor performance is the complicated data structure. The data is the mixture of discrete and continues variables, and the number of values for each discrete variable is very different. For example, Usage Type has 3 distinct values, Area type has 4, Nearest Road Type has 18 and Terrain Type has 40 distinct values. Fig. 5 illustrates the distribution of Nearest Road Type from the observed data. 80% has the road type as local, rural road and City Street. The complicated data structure and unbalanced data distribution trigger the biased variable selection using the tree model [4] only.

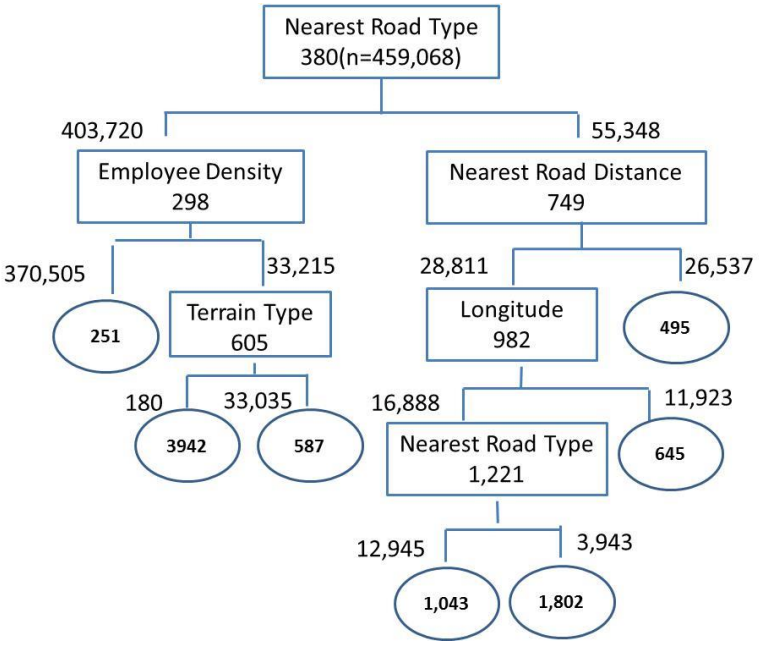


Figure (4). Tree model for variable selection.

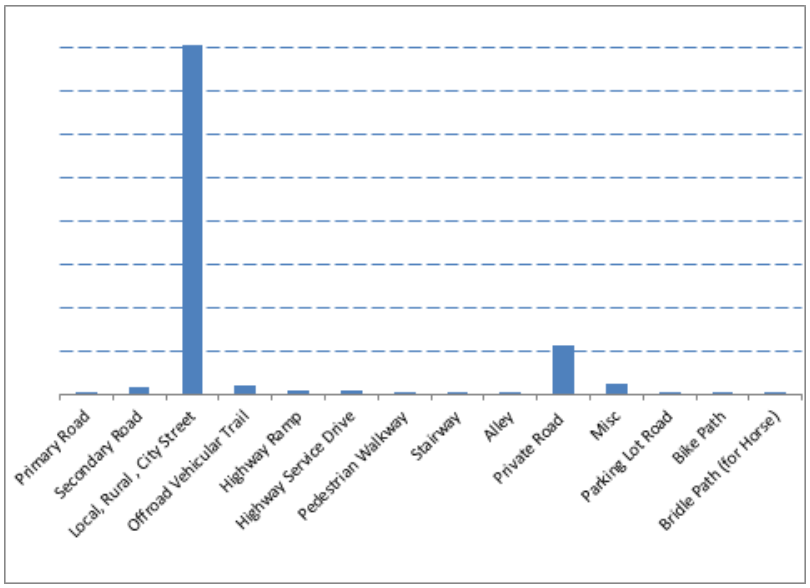


Figure (5) Nearest Road Type Distribution

To complement tree model, exhaustive linear regression is applied in variable selection. Table (2) shows the important variables selected by linear regression model. Besides the variables selected by the tree model, Usage Type, Area Type, Nearest Business Distance, Latitude, Population density, Business density and Residency density are significant variables in the linear model. As with the tree model, the linear model selects significant variables, but the linear relation is not sufficient in estimating user density, the adjusted R-square is only 0.15 for linear regression.

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	152.4	66.4	2.295	0.0218
GRID_LAT	0.0	0.0	-14.28	< 2e-16
GRID_LONG	0.0	0.0	21.584	< 2e-16
MOBILITY_RATIO	19.2	10.1	1.895	0.0582
INDOOR_RATIO	-11.7	8.0	-1.473	0.1408
TERRAIN_CLUTTER_TYPE	-3.3	0.1	-24.66	< 2e-16
RESIDENCE_DENSITY_SQKM	0.0	0.0	14.809	< 2e-16
POPULATION_DENSITY_SQKM	0.0	0.0	-15.351	< 2e-16
BUSINESS_DENSITY_SQKM	0.4	0.0	97.482	< 2e-16
EMPLOYEE_DENSITY_SQKM	0.0	0.0	24.041	< 2e-16
NEAREST_ROAD_TYPES1200	-538.0	8.9	-60.417	< 2e-16
NEAREST_ROAD_TYPES1400	-883.9	8.7	-101.87	< 2e-16
NEAREST_ROAD_TYPES1500	-834.3	91.4	-9.131	< 2e-16
NEAREST_ROAD_TYPES1630	-270.8	12.8	-21.211	< 2e-16
NEAREST_ROAD_TYPES1640	-535.8	28.2	-19.01	< 2e-16
NEAREST_ROAD_TYPES1710	-400.1	70.6	-5.67	1.43E-08
NEAREST_ROAD_TYPES1730	-825.1	83.9	-9.832	< 2e-16
NEAREST_ROAD_TYPES1740	-848.9	20.9	-40.639	< 2e-16
NEAREST_ROAD_TYPES1750	-820.6	27.2	-30.173	< 2e-16
NEAREST_ROAD_TYPES1780	-680.6	36.6	-18.607	< 2e-16
NEAREST_ROAD_TYPES1820	-1193.0	630.4	-1.892	0.0585
NEAREST_ROAD_DISTANCE_M	-1.0	0.0	-30.132	< 2e-16
NEAREST_BUSINESS_DISTANCE_M	-0.4	0.0	-17.443	< 2e-16
AREA_TYPESUBURBAN	-78.3	3.5	-22.548	< 2e-16
AREA_TYPEURBAN	-20.1	3.9	-5.102	3.36E-07
AREA_TYPEVENUE	445.9	13.4	33.174	< 2e-16
USAGE_TYPEMIXED	51.5	3.5	14.907	< 2e-16
USAGE_TYPEOUTDOOR	127.2	5.5	23.113	< 2e-16

Table(2). Linear model for variable selection

Neither the tree model nor the linear model is good enough in estimating the user density independently due to the complicated data structure and the nonlinear relationship between the variables. But the variables selected from both models provide the importance of the variables in interpreting the user density. We propose the nonparametric adjusted K-nearest neighbor (KNN) model to estimate user density based on the important variables selected from both tree and linear regression models as listed in Table (3).

Continuous Variables	**Categorical Variables**
1.Latitute	9.Usage Type
2.Longitude	10.Area Type
3.Nearest Road Distance	11. Terrain Type
4.Nearest Business Distance	12. Nearest Road Type
5.Employee Density	
6.Population Density	
7. Residency Density	
8. Business Density	

Table (3). Selected variables that related to user density.

3. **Approach**

K-nearest neighbors (KNN) is a basic instance learning method in machine learning. The key idea is to store all training samples $<X_i, f(X_i)>$ in a candidate pool. For any given query X_q, take the average of the ***K*** nearest neighbors,

$$\hat{f}(X_q) = \frac{1}{K}\sum_{i=1}^{k} f(X_i) \quad if\ f(X_i) \in R \qquad (1)$$

Where $X_i = \{x_{i1}, x_{i2}, \ldots x_{in}\}$ is n-dimension vector and the neighborhood is defined by the Euclidean distance: $d_{ij} = \sqrt{\sum_{t=1}^{n}(x_{it} - x_{jt})^2}$. In this study, X_i are the variables that were identified in Table 3 for the *i*th bin and $f(X_i)$ is the user density for the bin. The training samples are the information from good signal strength bins and the query X_q is the information from the poor signal strength bin. For each poor signal strength bin, we would like to estimate $\hat{f}(X_q)$ from its ***K*** nearest neighbors good signal bins**,** where ***K*** is a pre-defined constant. Capacity planning is a dynamic decision making procedure with different design strategies. To increase the robustness for engineers to optimize the design, we provide the minimum, average and maximum for different percentile of the nearest neighbor to replace the average of ***K*** nearest neighbor in Eq.(1).

$$\hat{f}(X_q) = F(\ percentile(f(X_1), f(X_2), \ldots, f(X_k))\ \ if\ f(X_i) \in R) \qquad (2)$$

Where F(.) represents minimum, average and maximum function.

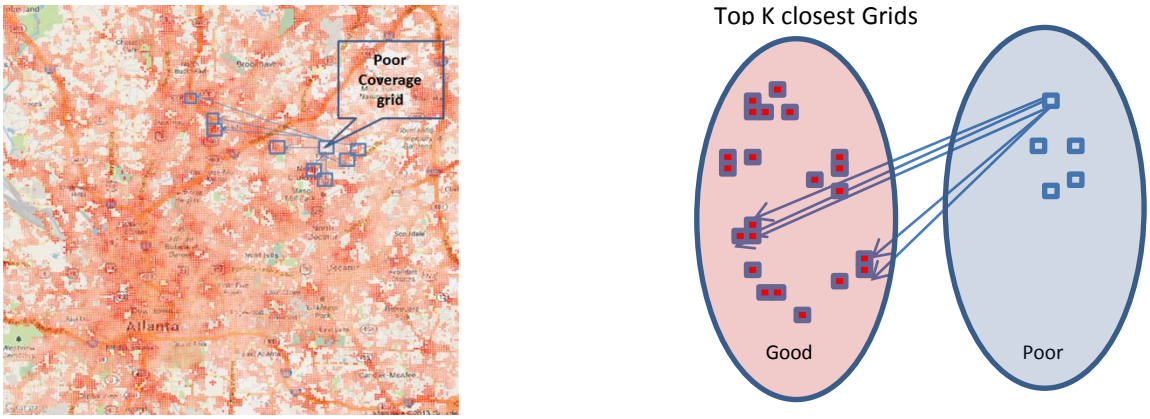


Figure (6) Approach Illustration. For each of the poor signal bin, find K closest bins from the pool of good signal bins. K is defined by different percentile.

4. Experiment

We apply the adjusted KNN model for the whole dataset. For each signal strenght RSCP $\leq$ -100 dB bin, find the ***K*** closest bins from all the bins with RSCP > -100db, and ***K*** is defined by different percentile. Each distinct value of the categorical variables is considered a subgroup. And the distance calculation only occurs among the continues variables in the same subgroup. The system flow chart is illustrated in Fig.7.

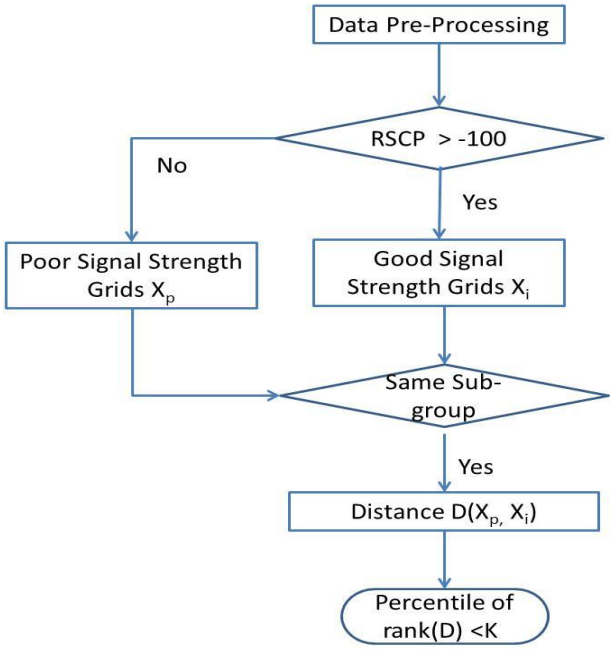


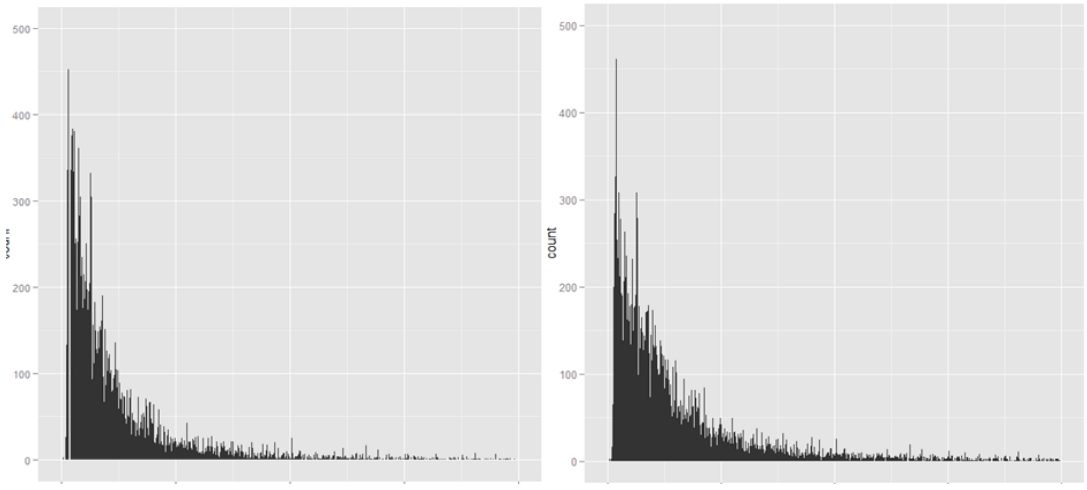
Figure (7). System Flow Chart

The two neighbour bins as shown in Fig.8, share the same value for location type information. The Area Type, Usage Type and Terrain Type are urban, indoor and Residential with trees respectively. And the RSCP are -116.3 dB and -116.9 dB. The original user density for bin 1 is 0 and bin 2 is 137. After applying the adjusted KNN model, the minimum, mean and max value user density estimation for the different percentiles is listed in Fig. 8.

#Sub	K=10	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Min	220	381	532	601	746	779	779	854	1047	1263	1366
Avg	307	531	743	838	1041	1087	1087	1191	1461	1762	1906
Max	1005	1739	2432	2744	3407	3558	3558	3899	4784	5769	6241

Figure (8) Two neighbor bins example. Above is the geolocation for the two bins and the bottom is the value when K is 10 for different percentiles.

In Figure 9, for the whole dataset, we compare the user density for poor signal coverage area before and after applying the adjusted KNN model. The underestimated user density on the left is augmented on the right by inference from the good signal coverage information.



Original Projected

Figure (9). Average user density for poor signal coverage bins before and after applying 10% KNN.

5. Conclusion

This paper illustrates a case study in exploiting big data to provide business solutions. EDA method is applied to quantify the poor signal strength threshold that impacts user density estimation. Regression tree and exhausted linear regression model are used to choose the variables that relate to user density. An adjusted KNN model is employed to estimate the user density for poor signal strength bins from neighboring good signal strength bins. The estimated user density for poor signal bins is boosted after inference from the good signal bins.

Reference:

1. S. Téral and R. Webb; Smallcell Report http://www.infonetics.com/cgp/login.asp?id=656 March 28, 2013
2. M. Reardon; http://news.cnet.com/8301-1035_3-57594983-94/at-t-uses-small-cells-to-improve-service-in-disney-parks/ *CNET news, Jul 23, 2013*
3. K. White; Smallcells: Small, but Valuable Addition to 4G LTE network http://www.verizonwireless.com/news/article/2013/05/4G-LTE-network-small-cells.html, May 2013
4. Game changing economics for Smallcell Deployment http://www.amdocs.com/About/media-room/Documents/Whitepaper/game-changing-economics-for-small-cell-deployment.pdf Amdocs OSS, Oct, 2013
5. W.-Y. Loh; Regression Trees with Unbiased Variables Selection and Interaction Detection. Statistica Sinica 12(2002), Page 361-386.