Insights from Application Data

POC for Small Cell Planning

July 2018

Agenda

- What is Application Data?
- Overview of Application data
- Application/Tower Data Challenges/Limitations
- Application/Tower Comparison
- Application Advantage
- Use Cases
- Methodology
- Demo
- Next Steps

What is Application Data?

- Collected via Mobile Applications with user consent.
- Insightful when user's use "Location Services" with app
- Popular video streaming application give insights into quality of network

Due to the Telecommunications Act of 1996, telecom companies cannot collect certain user information, such as their location, without their consent. Obtaining consent is relatively easy for mobile applications due to the installation process and the ability to offer users opportunities to opt-in to geo-location based services.

Overview of App Data

- **Application data aggregated in two different ways:**
  - **Coordinate Bin**
    - Observations not linked directly to a tower, but instead individual lat/long bins
    - Available bin resolutions: 20m, 70m, 140m, 280m, 600m, 1km, 2km, or 5km
    - Bin summary statistics available for 35 day, 1 week, and 1 day time periods
  - **Cell Tower Bin**
    - Groups observations into bins by tower sector and provides summary statistics

- **Two themes of metrics:**
  - **RF Metrics**
    - Signal Metrics: LTE Signal Power, LTE Signal Strength, LTE RSRQ/RSRP
    - LTE Band Metrics
    - Roaming Metrics
  - **Video Resolution Metrics**
    - Application Throughput
    - Video Start Time
    - Video Resolution

Further Explanation of selected terms:

LTE Band Metrics: Application Data displays the share of traffic across LTE bands for a given operator's LTE network, share of traffic for a specific band and LTE Cell Sector IDs shown in Application Data are labeled with the band that sector operates on.

Application throughput: Network throughput measured at the application layer of the protocol stack. This is the number of useful information bits delivered by the network to the client application per unit of time. The amount of data considered excludes protocol overhead bits (e.g. TCP headers) as well as retransmitted data packets in lower layers of the network stack (e.g. RLC AM, TCP). Since Application throughput is an application specific measurement and excludes overhead bytes, this makes an excellent network quality metric to infer actual user experience in real world mobile applications.

Video Start Time: Length of time users wait for videos to start playing and whether this wait is consistent with a good user experience.

Video Resolution: Percentage of video observations that recorded application throughput greater than or equal to the bitrate recommended for a specified resolution.

How much data is available?

Estimate* of Data Availability (Nationwide): 35 day data

Aggregation	Resolution	Total Sample Size	%Nationwide Observations	% Missing Video Data	% Missing RF Data	Usable RF/Video Res Sample Size
Coordinate	20m	463249	11.75%	99.56%	5.79%	436276/76
	35m	831490	26.17%	99.31%	6.15%	780150/218
	70m	986989	43.30%	99.97%	3.63%	918487/631
	140m	**918155**	**60.87%**	**97.08%**	**7.20%**	**852040/1951**
	280m	724109	76.67%	88.84%	7.52%	669248/6244
	600m	493293	85.77%	85.04%	8.59%	450915/13004
	1km	306537	91.22%	92.43%	9.70%	276796/13455
	2km	175022	95.14%	81.88%	10.84%	153383/10279
	5km	86314	97.81%	72.04%	12.02%	75938/8516
Cell Sector		56597	85.20%	12.46%	1.65%	55663/51941

Time Frames Available: 1 day, 1 week, 35 days** (**default)

Data Recency: Application Vendor updates a minimum of 3-4 days after reporting period. Reporting period is Saturday to Friday.

*Estimate is based on data sets for April through May 2018

We used the 140-meter data set at the 35-day summary level for this exploration due to the volume of usable RF and Video Resolution data available at that resolution. One thing to note is that even though the 140m dataset has a smaller sample size than the 70m dataset, it represents a larger % of nationwide observations. 20m and 70m video resolution data are very sparse due to the application vendor's privacy filters.

As you can see, the most complete dataset available is the data aggregated by cell sector.

Application Data Limitations

Metrics	Application	LSR
Geo Bin Size	20m**-5km	50m
Cell Tower Info	✓	✓
RF Metrics	✓	✓
Video Resolution Metrics	✓	✗
Raw Data	✗	✓
Customer Lat/Long	✓	✗

- **Application Privacy Thresholds** prevent us from seeing areas with small number of observations. **As we increase the resolution, we lose information:
  - **No Video Resolution metrics** at high resolution due to privacy thresholds
  - Suspect we are seeing incomplete data due to the sparsity
  - Sparsity increases as we narrow time from 35 days to 1 day
  - Sparsity increases as we narrow coord bin from 5km to 20m
- **Application Vendor provides summary statistics only (due to**

LTE RSRQ 5th percentile	LTE RSRQ 10th percentile	LTE RSRQ 50th percentile	LTE RSRQ 90th percentile	LTE RSRQ 95th percentile
-8	-8	-9	-12	-14

The Application Data from the application vendor is heavily aggregated due to privacy. Also, if there are not enough observations within an aggregation “bin” that data is not included. The data is most interesting at a granular level, but also very sparse at a granular level.

Key points:
1. The smaller geographic bins where we do have data available, ie. 140m, do not allow us to tie observations to towers. However, we still can use to view quality of coverage.
2. Data is sparse at the more interesting resolutions: 1 day/20 meters
3. Application vendor provides only summary statistics. If we do not know the true distribution of variables due to aggregation, some assumptions for statistical analysis cannot be confirmed. Ie., “Is RSRQ normally distributed in an area?”
4. Since we do not have access to the raw data, we have no control over outliers.

The biggest advantages of using the application data:
1. Video Resolution Metrics
2. Customer lattitude and longitude

Known Tower Data Limitations

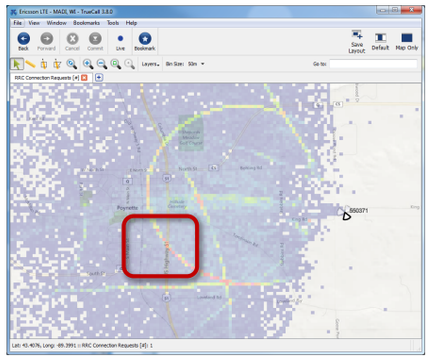
Metrics	Application	LSR
Geo Bin Size	20m-5km	50m**
Cell Tower Info	✓	✓
RF Metrics	✓	✓
Video Resolution Metrics	✓	✗
Raw Data	✗	✓
Customer Lat/Long	✓	✗

- **50m resolution is much smaller than the geolocation error of the latitude and longitude of the calls
- If care is not taken poor engineering decisions will be made under the assumption of little to no geolocation error
- **Distance from the tower to the caller is known, however the call falls on an arc around the tower, cannot be tied to an actual location.**

The main drawback to using the raw data from the cell towers is that we do not have an exact location of the caller. As we will see in a future slide, this inaccuracy makes it difficult to identify hot spots of poor coverage in populated areas.

Why do we know know the location of the caller? Distance from the tower to the caller is known, however the call falls on an arc around the tower, cannot be tied to an actual location. Also, the Telecommunications Act of 1996 prevents telecom companies from collecting user location from cell phone users.

Case Study – Poynette, WI
Where is the traffic?

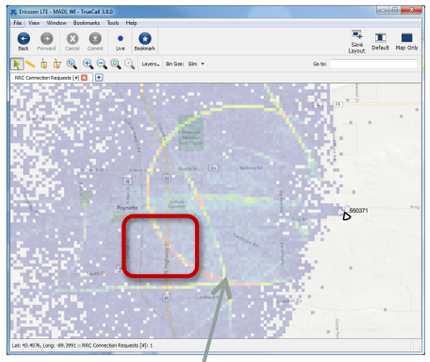


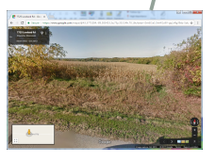
According to tower data
The traffic is evenly distributed on an arc

- In most calls all that is known about the location of the call is the cell it was on and the timing advance (distance from the cell)
- The arch is not real it is an artifact of the geolocation algorithm
- **Where on the arch is the traffic**

As an example of the poor location capabilities of the current raw data, we look at one call in a small Wisconsin town. We can see the arc on which the call falls and it is nearly the whole circle around the tower. Unfortunately, we cannot pinpoint the location of the caller in this instance. The caller may be calling from the golf club, the grocery store, the corn field, etc.

Case Study – Poynette, WI
Where is the traffic?





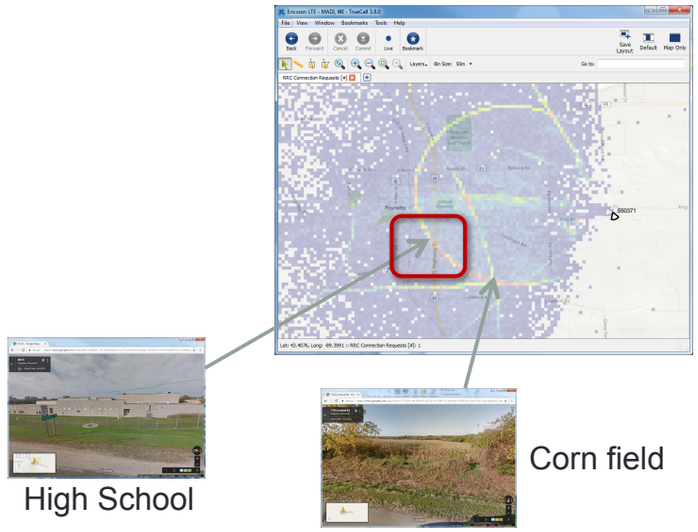
Corn field

According to tower data
The traffic is evenly distributed on an arc

- In most calls all that is known about the location of the call is the cell it was on and the timing advance (distance from the cell)
- The arch is not real it is an artifact of the geolocation algorithm
- **Where on the arch is the traffic**

Case Study – Poynette, WI

Where is the traffic?



According to tower data
The traffic is evenly distributed on an arc

- In most calls all that is known about the location of the call is the cell it was on and the timing advance (distance from the cell)
- The arch is not real it is an artifact of the geolocation algorithm
- **Where on the arch is the traffic**

Case Study – Poynette, WI

Where is the traffic?

According to tower data
The traffic is evenly distributed on an arc

- In most calls all that is known about the location of the call is the cell it was on and the timing advance (distance from the cell)
- The arch is not real it is an artifact of the geolocation algorithm
- **Where on the arch is the traffic**

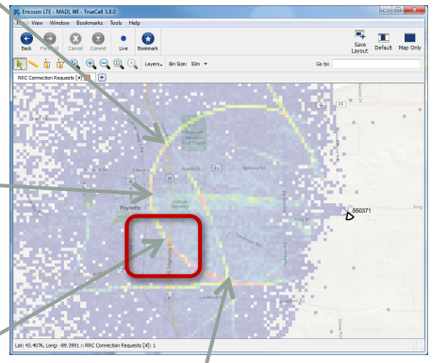
Case Study – Poynette, WI

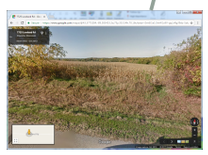
Where is the traffic?

Grocery Store

Residential Street

High School





Corn field

According to tower data
The traffic is evenly distributed on an arc

- In most calls all that is known about the location of the call is the cell it was on and the timing advance (distance from the cell)
- The arch is not real it is an artifact of the geolocation algorithm
- **Where on the arch is the traffic**

Case Study – Poynette, WI

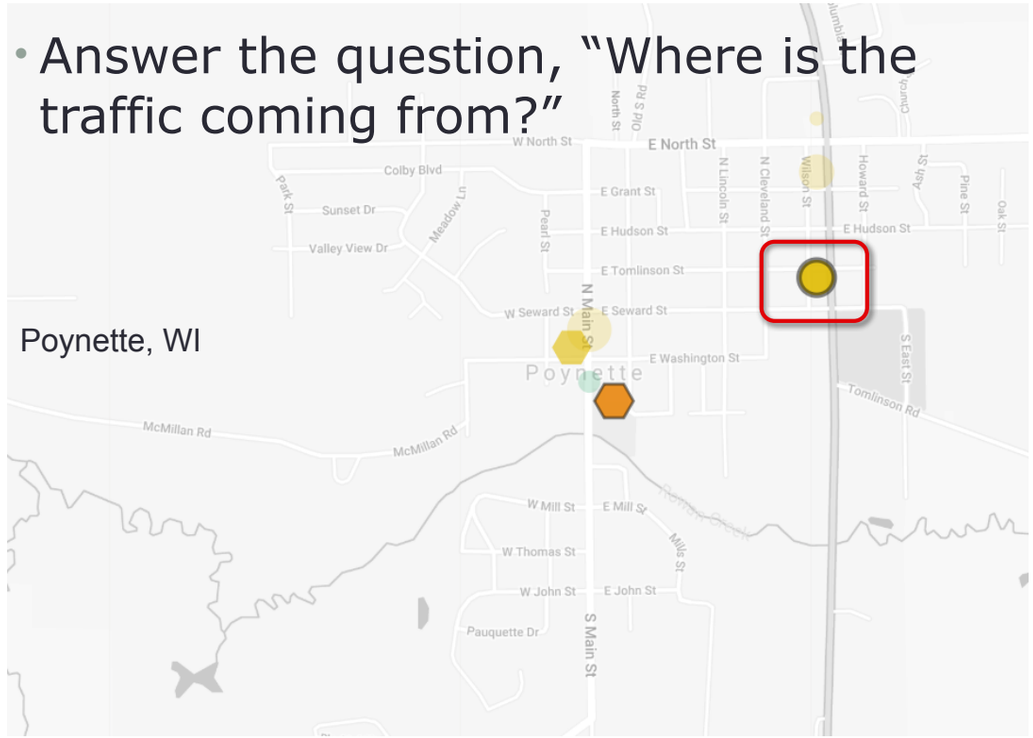
Where is the traffic?

According to tower data
The traffic is evenly distributed on an arc

- In most calls all that is known about the location of the call is the cell it was on and the timing advance (distance from the cell)
- The arch is not real it is an artifact of the geolocation algorithm
- **Where on the arch is the traffic**

Application Data Advantages

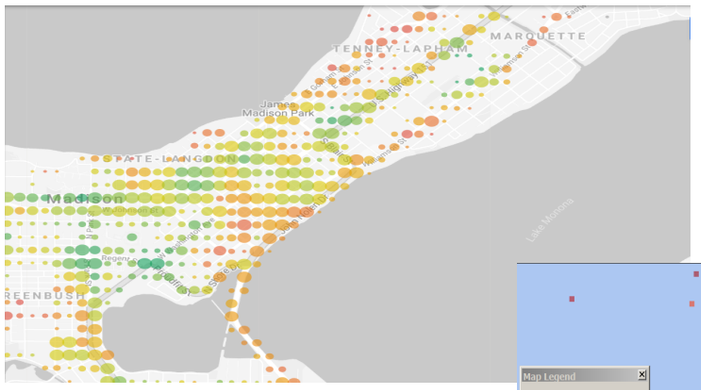
- Answer the question, “Where is the traffic coming from?”



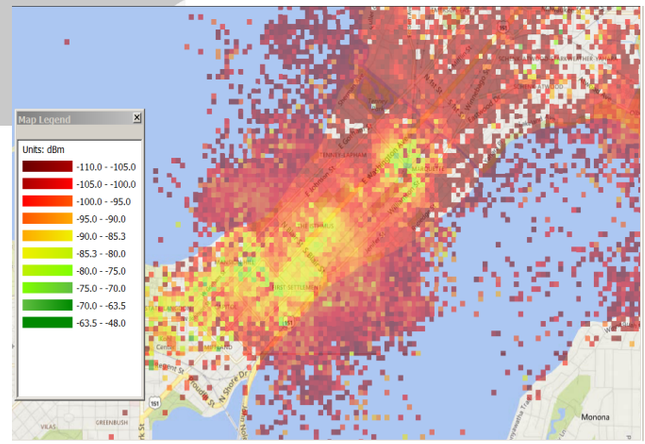
This slide shows a visualization using the data collected from the application for the same town as the previous slide: Poynette, WI.
One key take away from this slide is that we can see specific hot spots for towers in application data versus traffic arcs in tower data.

Here, we can see a 1 week view at 140 meter resolution for Poynette, Wisconsin. There is one cell ID selected and the corresponding traffic hotspots are visible. In a small town such as Poynette, WI (pop. 2528), data at more granular resolutions (70m, 20m, 1 day) are unavailable from the application vendor. So, we cannot pinpoint any specific calls or callers, we can see an aggregation of calls in the town over various periods of time.

App/Tower Comparison: RSRP Coverage Map



- 1 Week Data, June 2-8, 2018
- Same Cell Sector
- Same color scale



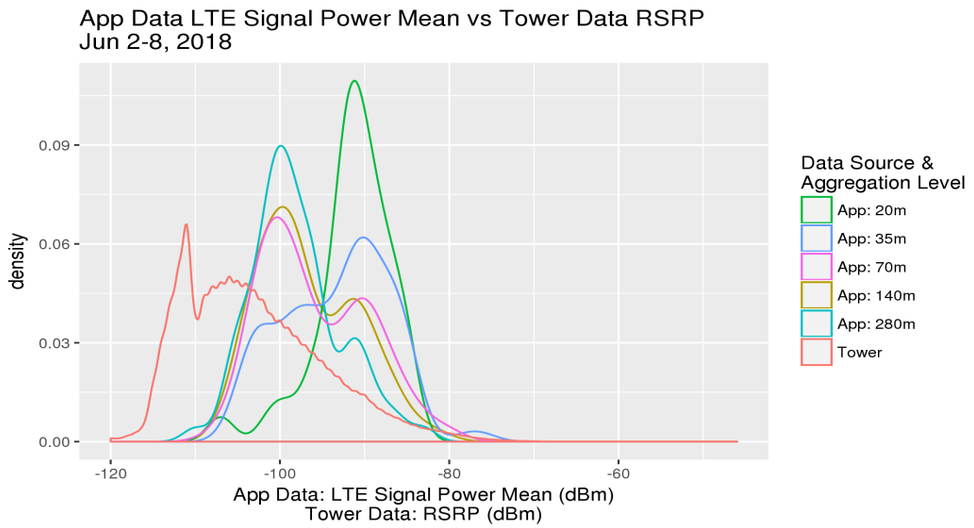
- Aggregated points in Application Data show more accurate centers of traffic
- Raw data collected via tower demonstrates inaccuracy of customer location estimates

We compared the application vendor's data dashboard (upper left) to the in-house raw tower data (bottom right). This comparison is for 1-week data, comparing the same tower, on the same week (June 2-8, 2018). The Application Data dashboard view is aggregated by cell sector at 140m bin size. The color scales have been modified to match.

An important note about the in-house raw tower data is that due to CPNI rules, the telecom company is unable to collect latitude and longitude values from its users. Rather, they have a general idea of where the user is located based on algorithms that determine the distance that the caller is to the towers. Callers on 3G networks are usually attached to multiple towers and the algorithm generates an arc on which the caller is located. The accuracy goes down on LTE networks since the called may only attached to one tower at a time and the location of the caller can only be placed on a circle around the tower.

The key takeaway from comparing the raw tower data to the application data is that the tower data plot demonstrates inaccuracy of data points using the arcs from the raw tower data. The area in madison pictured is surrounded by lakes on both sides. The raw tower data shows a significant number of calls being made from the lakes, whereas the application data has the users actual latitude and longitude and shows a lot of traffic along roadways and in metro areas.

App/Tower Comparison: RSRP Distribution



- Distribution Shapes are similar for 70m-280m; App data is shifted to the right of Tower data
- **Tower Data exposes a very different pattern of data than Application Data**

Here, we are comparing the distribution of RSRP values over 1 week period:

- Raw Tower Data in RED: distribution of RSRP for all 1.9 million calls made ending at cell
- Application Vendor Data in multicolor (1 curve for each distance bin)

The density distributions of 70m through 280m App data bins look similar in shape to LSR data, although they are shifted to the right. They are most likely better representations of the data. The shift to the right can be explained by the fact that cellular users will not be able to use data services at a very low value of RSRP. Therefore, you see the curves falling within the range that data services can be delivered.

Key Takeaway: Overall, the data is not the same. App vendor's collection methods, privacy filters and the way it defines things like "Signal Power Mean" create a disparity between our "ground truth" data from the towers and what we see in Application Data. Although different, the data still provides network insights.

App vs Tower Data Summary

Metrics	Application Data	Tower Data
Geo Bin Size	20m-5km	50m
Cell Tower Info	✓	✓
RF Metrics	✓	✓
Video Resolution Metrics	✓	✗
Raw Data	✗	✓
Customer Lat/Long	✓	✗

- App Data provides better location information for customers
- App Data provides useful insights into video resolution
- App Data **should not be considered ground truth data,** most useful as an "additional insight" rather than prescriptive tool
- App Data limitations affect ability to report "How Confident"
- Tower Data gives us raw data and all data

This slide summarizes the comparison between aggregated application data and the raw tower data. The main advantage of the application data is the user location and the video resolution information. However, we must take the application data with a grain of salt due to the privacy thresholds, the aggregations that we are given and the fact that we are looking at a segment of users as opposed to all users.

Use Case Analysis

Application Vendor Proposed Use Cases

- **Targeting generational network upgrades** to LTE and 5G
- **Root-cause assessment** of poor user experience
  - E.g., **capacity**, **configuration**, **coverage**, & **interference** problems
- **Mobility optimization** across LTE bands
- **Cell footprint optimization**
- **Small cell placement**

The application vendor claims that other telecom companies around the world are using the same data for the listed purposes.

MNO Proposed Use Cases

- **Use Case 1:** Predict and map data using RF metrics in order to visualize traffic density and quality. Visually inform potential problem sites or new cell sites.
- **Use Case 2:** Cell site planning, small cell. Identify "hot spots" of video traffic as candidates for small cells.
- **Use Case 3:** Predict "in-building" traffic via user data such as "LTE Power Percentile" and "RSRQ value"
- **Use Case 4:** RF model tuning using RSRP/RSRQ to refine RF propagation tools and traffic maps

These are the use cases that the Analytical Engineering team formulated.

Aggregated App Data might be able to help map RF metrics for comparison to current tools. We might be able to gain a better understanding of traffic density and quality which can inform future small-cell planning or upgrade to 5G. The data can potentially also provide a quick visual representation of problem areas, problem cell sites and potential new cell sites.

Deriving Insights: Use Case 1

Use Case Selection:

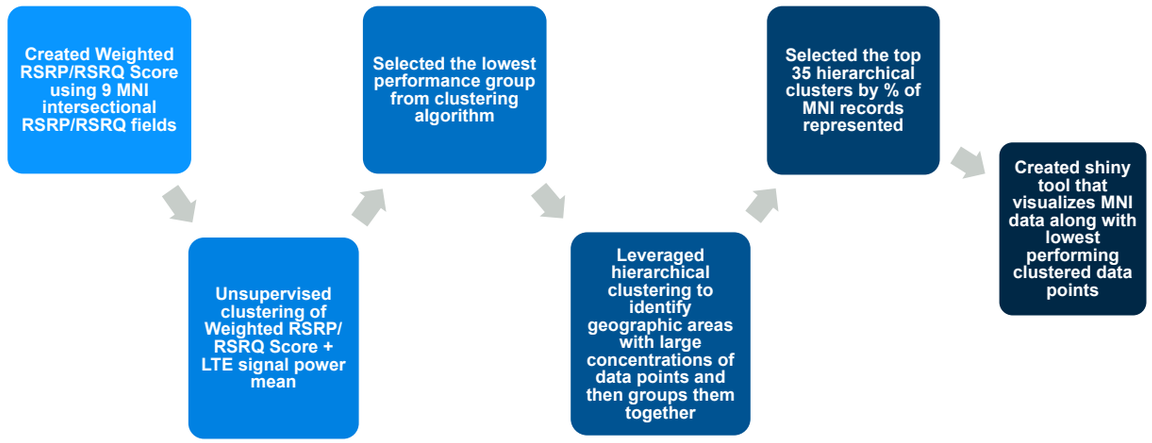
- Incomplete video resolution metrics lead us to focus on an **RF centric** use case
- Identified high level of **business value** around proving insights for small cell deployment
- Wanted to present insights from this data that would be **useful for a RF planner**

Benefits:

- Application level metrics by geographic location
- Identification of low performance areas custom to our network
- Push-button insights for RF engineers

Methodology

Methodology Steps



Methodology

Weighted RSRP/RSRQ Score

Application data displays the RSRP/RSRQ metric as an intersectional metric as follows:

		RSRQ Value		
		Poor	**OK**	**Good**
RSRP Value	**Poor**	Poor/Poor%	Poor/OK%	Poor/Good%
	OK	OK/Poor%	OK/OK%	OK/Good%
	Good	Good/Poor%	Good/OK%	Good/Good%

Good = > -89 dBm

For RSRQ:

- Poor = < -13 dB
- OK = -12 dB to -8 dB
- Good = > -7 dB

For RSRP:

- Poor = < -110 dBm
- OK = -109 dB to -90 dBm

Recall from a previous slide that one of the challenges with MNI data is that it is aggregated. So instead of raw RSRP/RSRQ value, we get 5th, 10th, 50th, 90th, 95th percentiles.

For our use case, we wanted to make it easier for anyone looking at the data to discern good data points from problem data points.

Further aggregation happens when the application vendor lumps RSRP/RSRQ into the intersectional metric you see here. For each row of data, whether is binned by cell ID or by distance, we get **9 separate columns representing the percent of traffic corresponding to the Poor, OK, Good measure defined by the application vendor.**

So we get percent of traffic that has poor RSRP and poor RSRQ in one column and so on for each of the 9 combinations. Poor RSRP is defined as <= -110 dBm, etc.

Methodology

Weighted RSRP/RSRQ Score

We reduced values in 9 fields down to 1 in order to more efficiently compare and analyze RSRP/RSRQ values between cells:

Poor/Poor	Poor/OK	Poor/ Good	OK/Poor	OK/OK	OK/Good	Good/ Poor	Good/ OK	Good/ Good
21.1	17.8	0	13.3	45.6	2.22	0	0	0

		RSRQ Value		
		Poor	OK	Good
RSRP Value	Poor	1	4	7
	OK	2	5	8
	Good	3	6	9

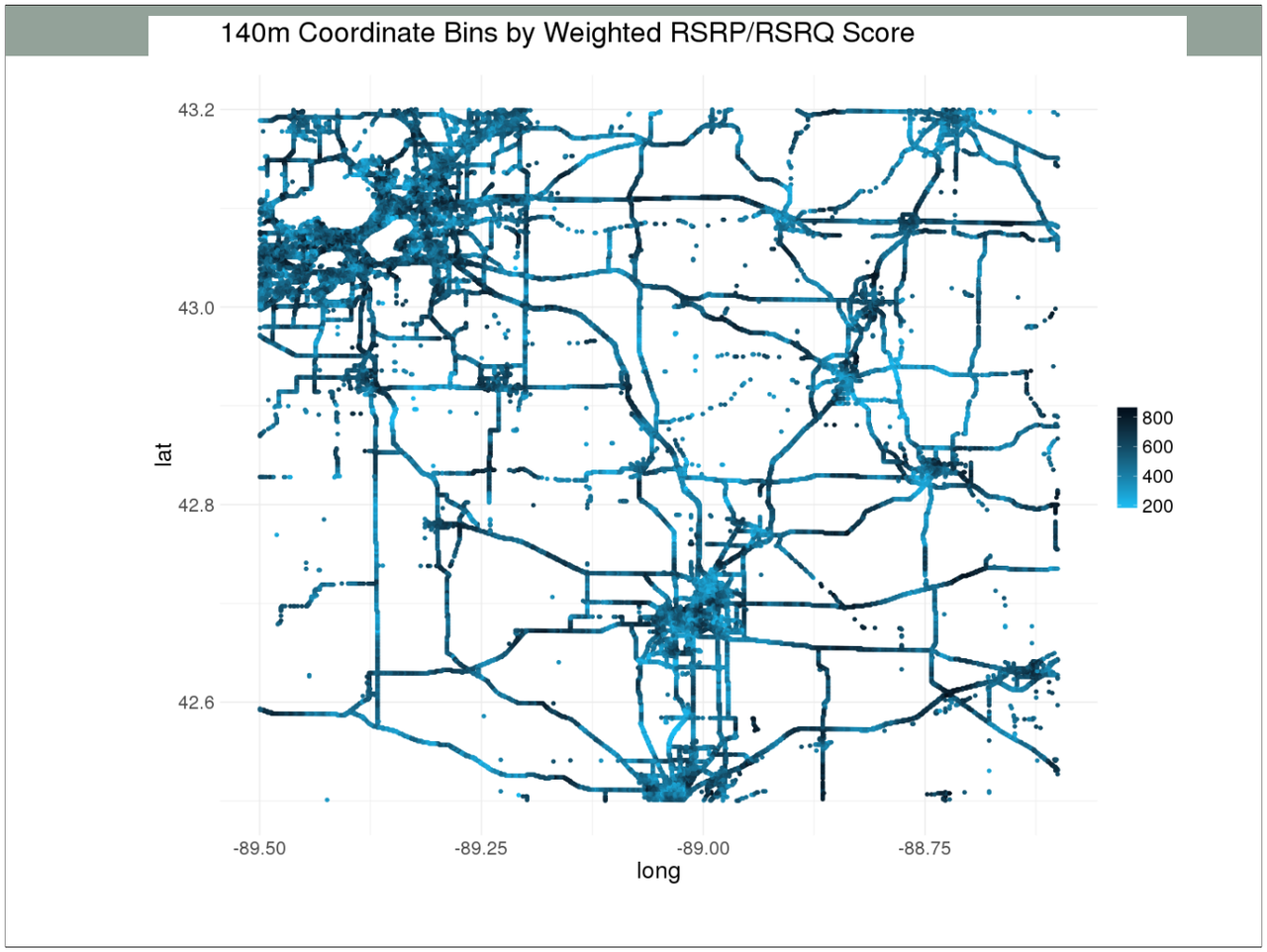
Example: We multiply each value by the weights and sum all 9 values to obtain the Combined RSRP RSRQ Score of 364.66.

		RSRQ Value (weights)		
		Poor	OK	Good
RSRP Value (weights)	Poor	21.1*1	17.8*4	0*7
	OK	13.3*2	45.6*5	2.22*8
	Good	0*3	0*6	0*9

Total: 364.66

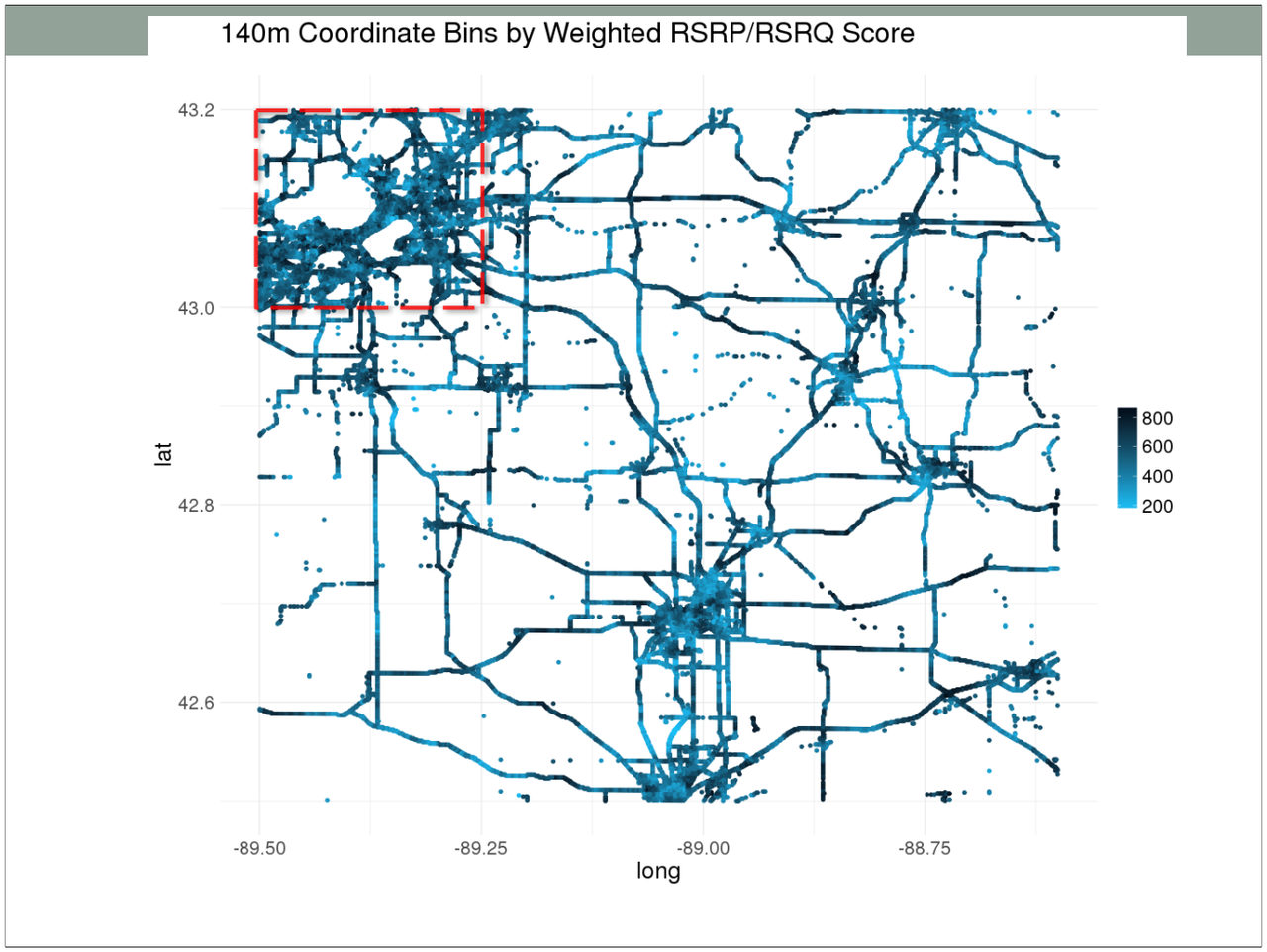
All 9 fields are important. This weighting method generates a score that we feel is both relevant and useful for our use cases.

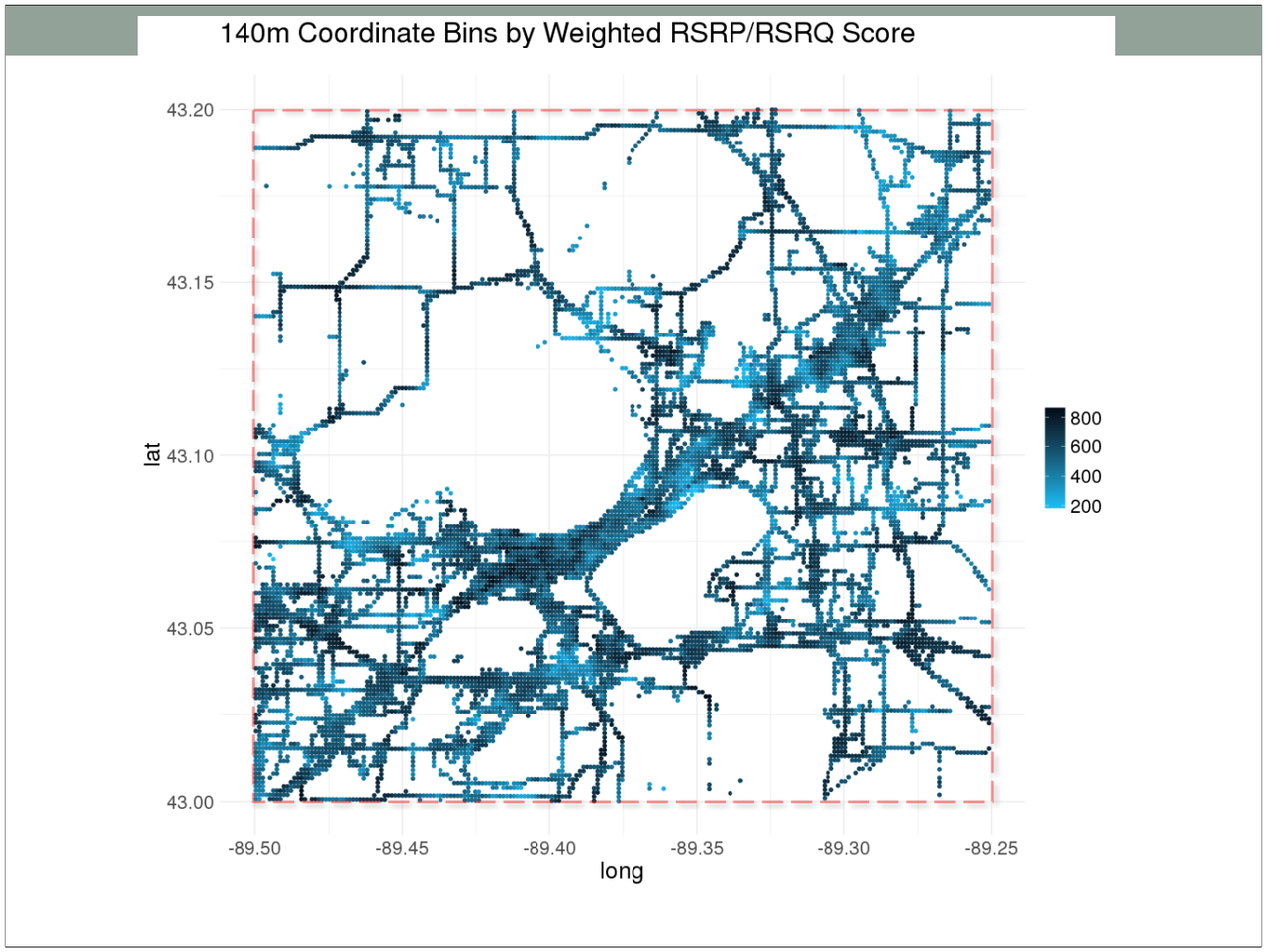
We are weighting RSRQ more than RSRP with this weighting system. Going forward, we recommend that the weighting scheme be correlated to a target variable using a data mining technique such as least squares regression.

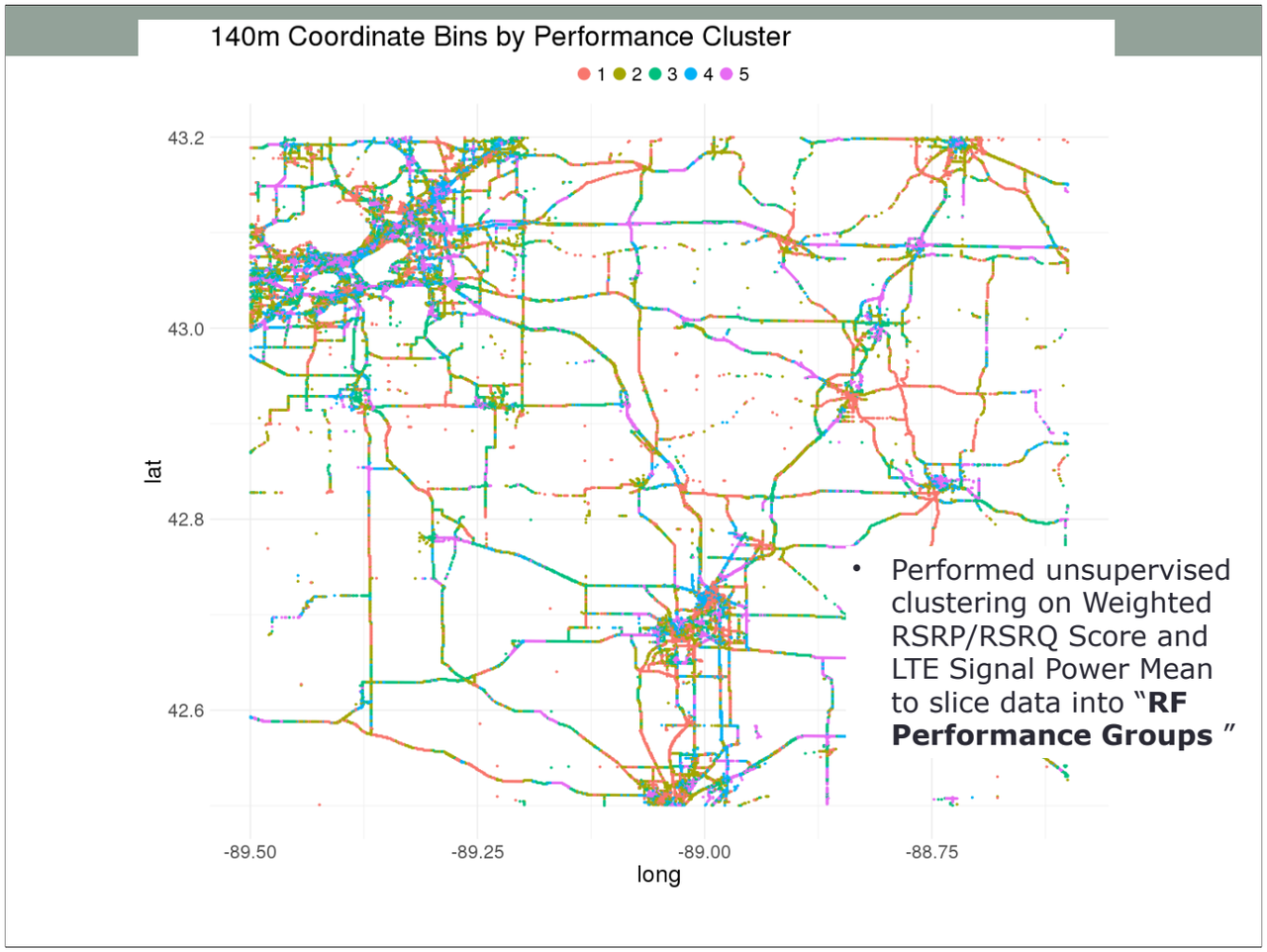


This plot shows 140m bin application data points in the Madison market colored by Weighted RSRP/RSRQ. Here, it's difficult to pinpoint areas with consistent performance across small enough geographic regions to action on.

With so many data points, we needed a way to parse through and filter for the ones that specifically sticks out as 'underperforming' in our use case.

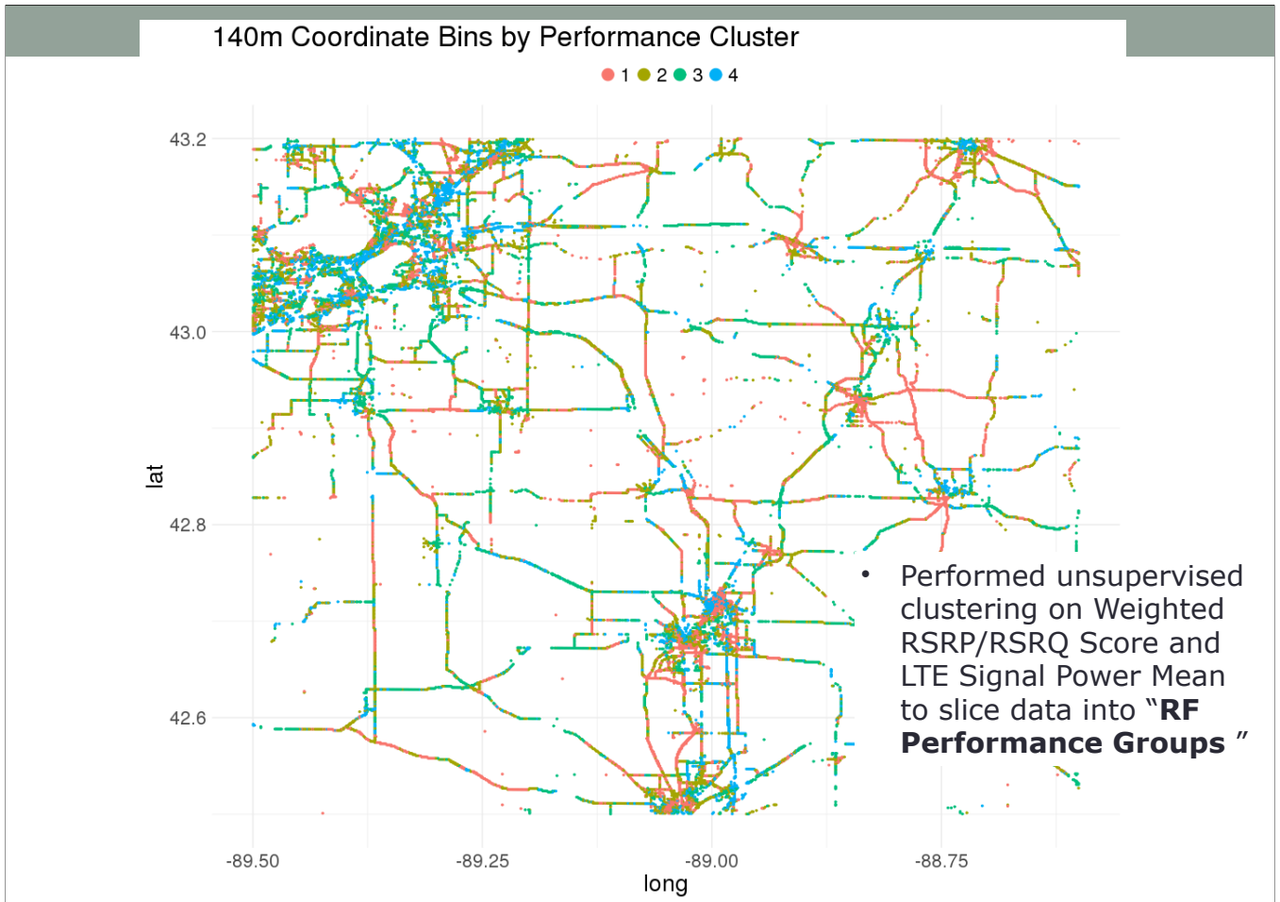

140m Coordinate Bins by Weighted RSRP/RSRQ Score
43.2
43.0
42.8
42.6
lat
-89.50
-89.25
-89.00
-88.75
long
800
600
400
200


140m Coordinate Bins by Weighted RSRP/RSRQ Score
43.20
43.15
43.10
43.05
43.00
lat
-89.50
-89.45
-89.40
-89.35
-89.30
-89.25
long
800
600
400
200

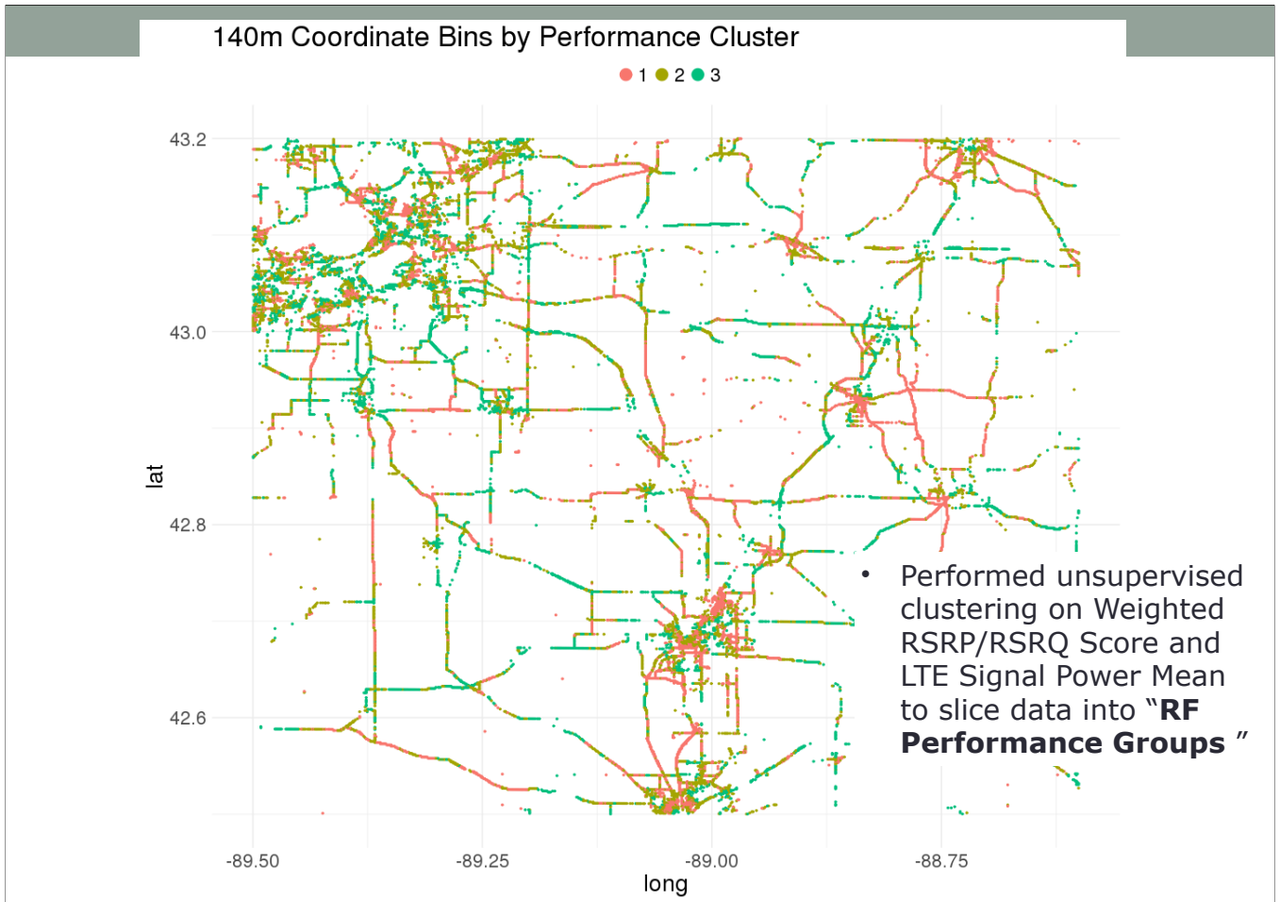


We used unsupervised k-means clustering on our weighted RSRP/RSRQ score along with LTE signal power mean to create 5 distinct clusters that we refer to as “RF performance groups.” The performance groups are numbered 1-5 with 1 being the “worst” and 5 being the “best”.

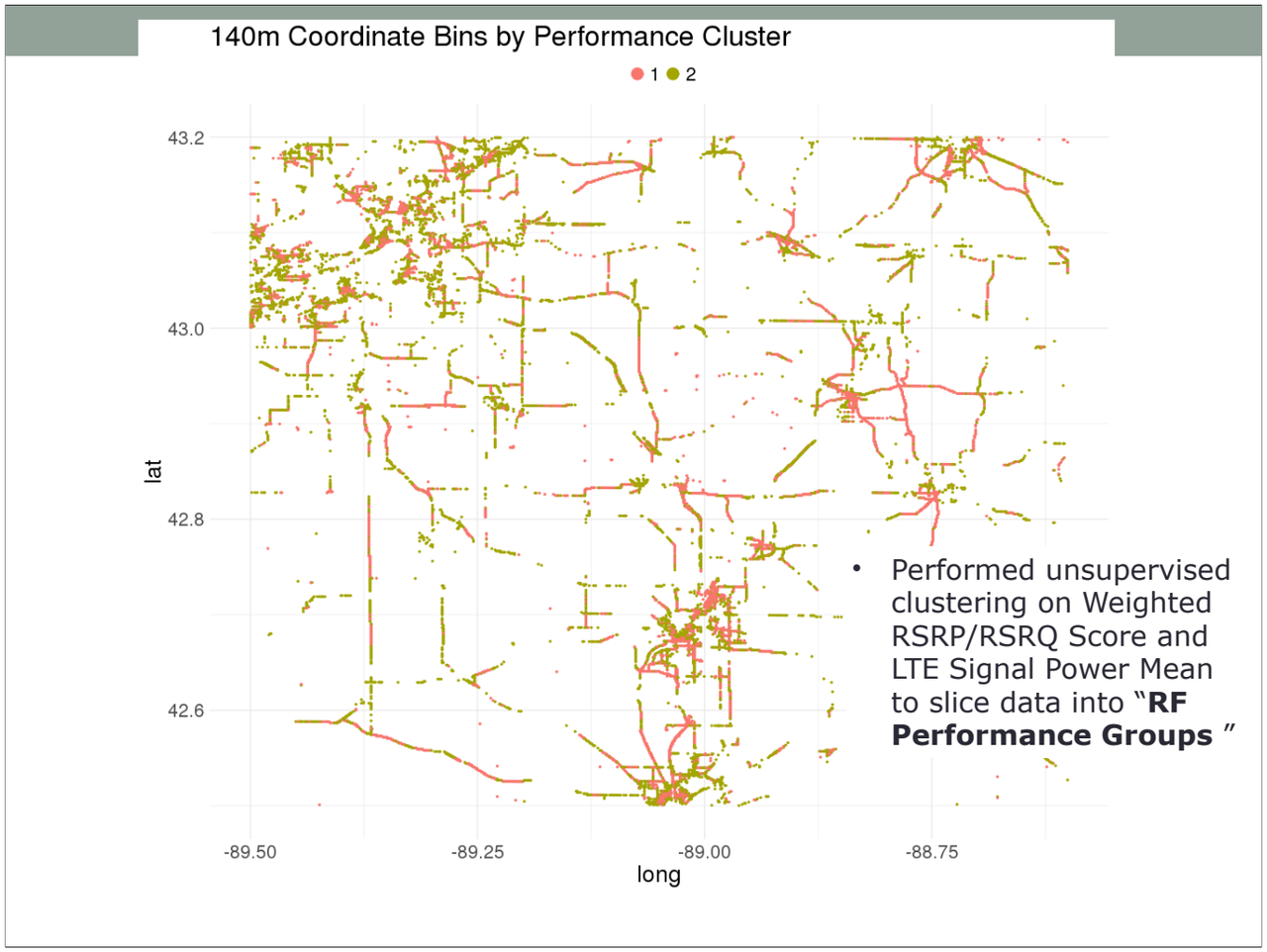
Unsupervised k-means clustering is a Data Mining method that fits within the “Automated Reasoning” aspect of the telecom company’s AI strategy.



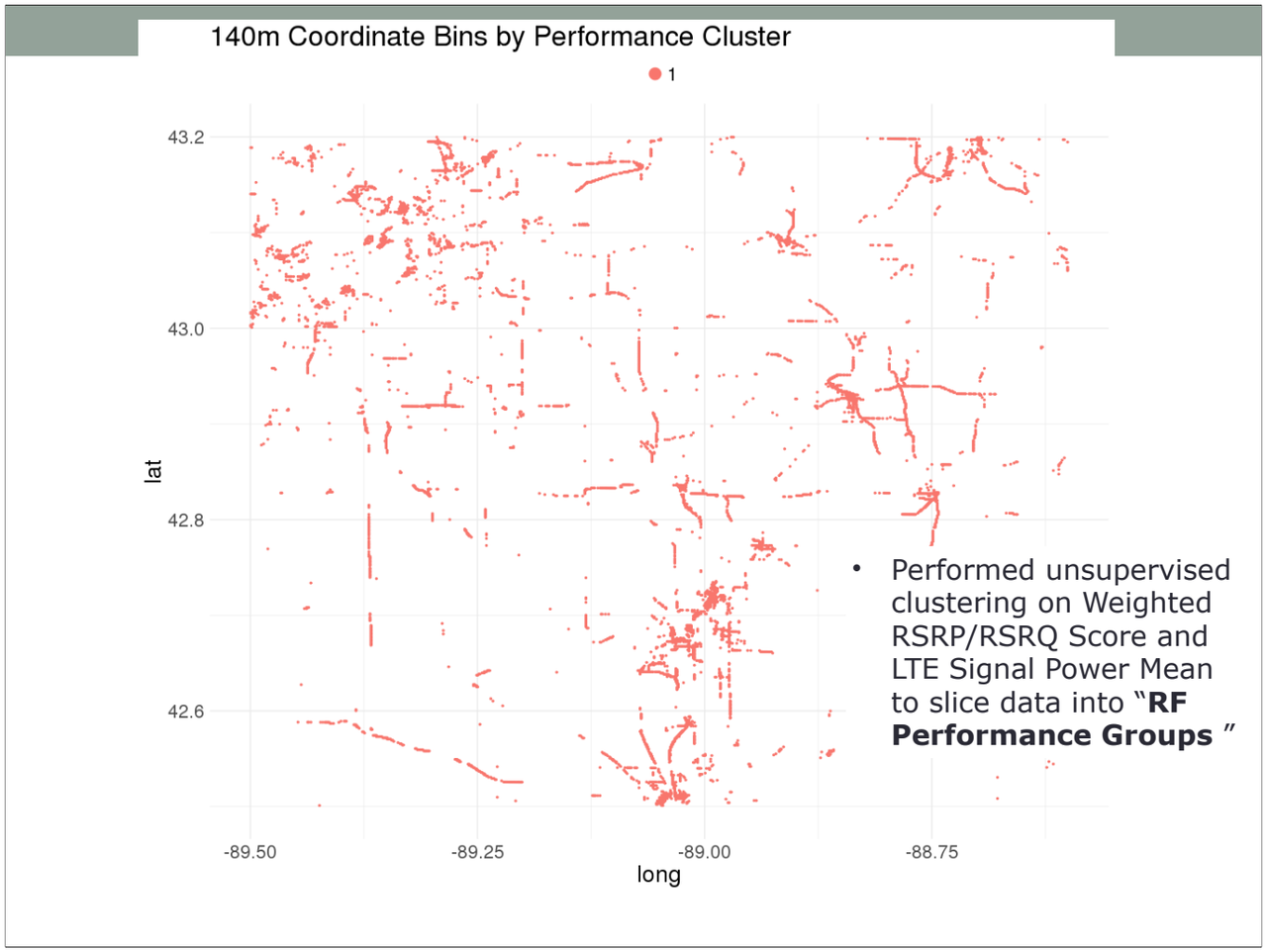
- Performed unsupervised clustering on Weighted RSRP/RSRQ Score and LTE Signal Power Mean to slice data into "**RF Performance Groups** "



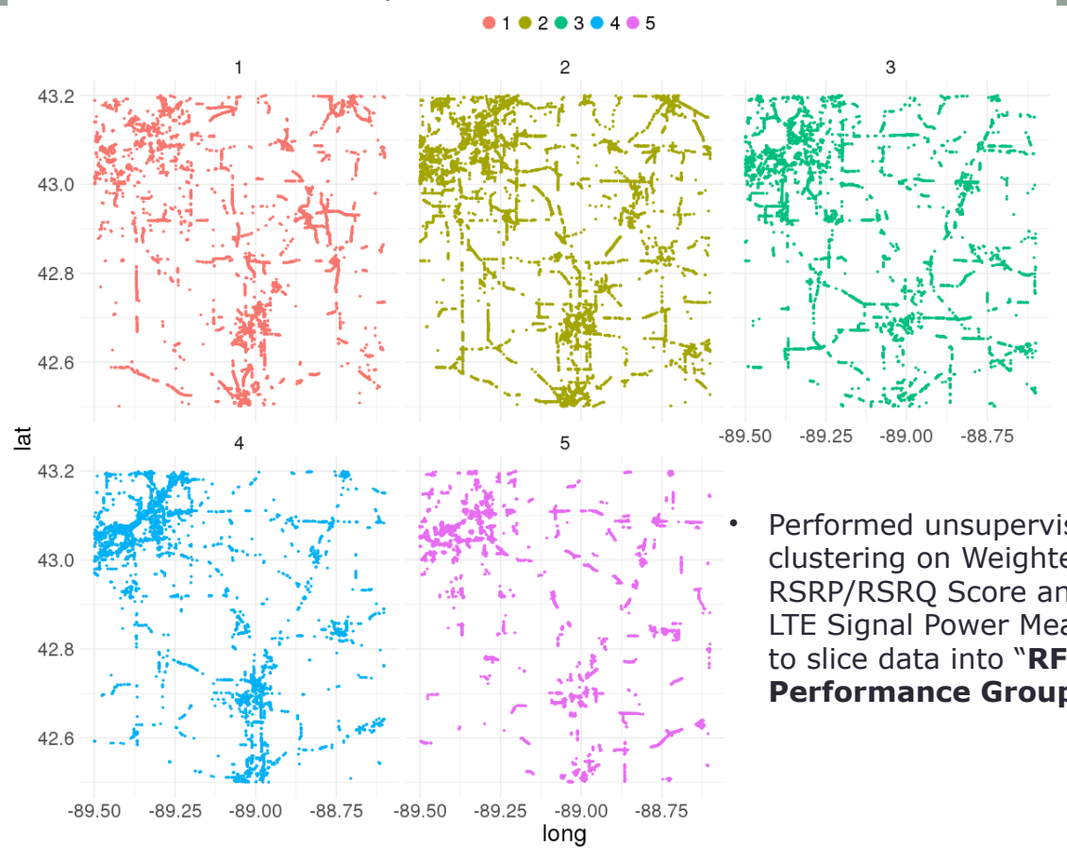
- Performed unsupervised clustering on Weighted RSRP/RSRQ Score and LTE Signal Power Mean to slice data into "**RF Performance Groups** "



- Performed unsupervised clustering on Weighted RSRP/RSRQ Score and LTE Signal Power Mean to slice data into "**RF Performance Groups** "



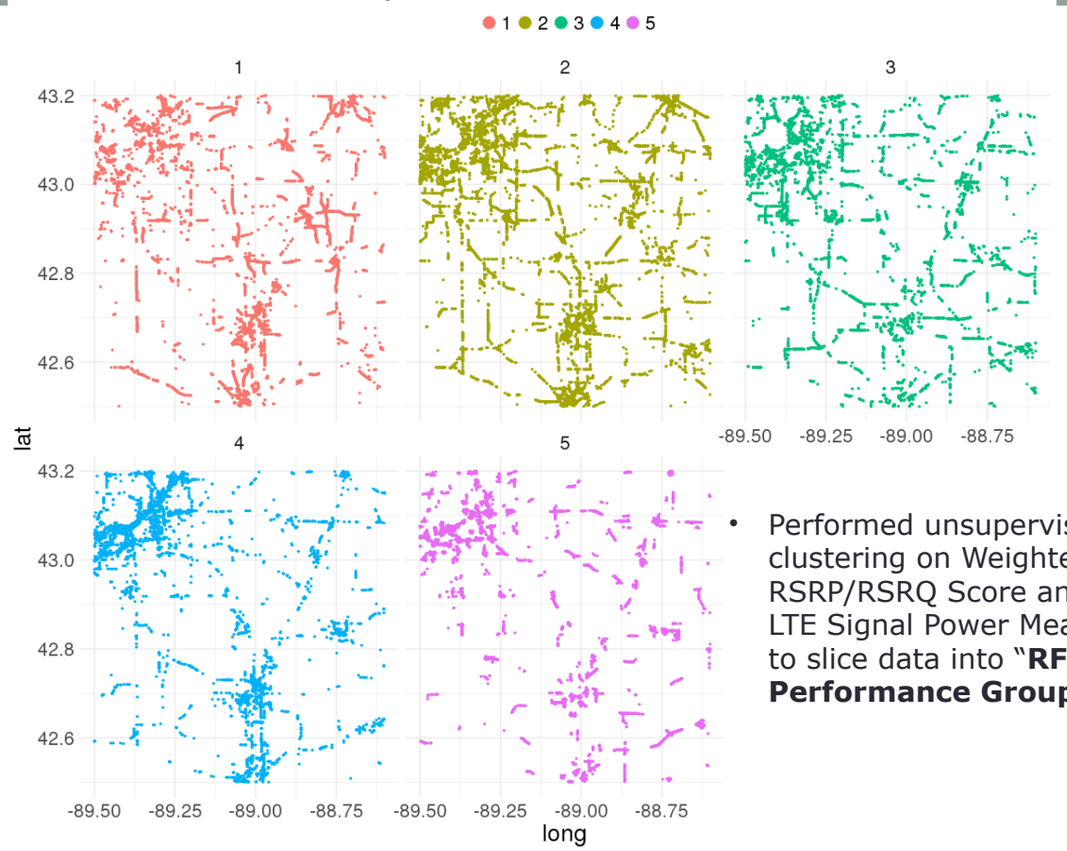
- Performed unsupervised clustering on Weighted RSRP/RSRQ Score and LTE Signal Power Mean to slice data into "**RF Performance Groups** "



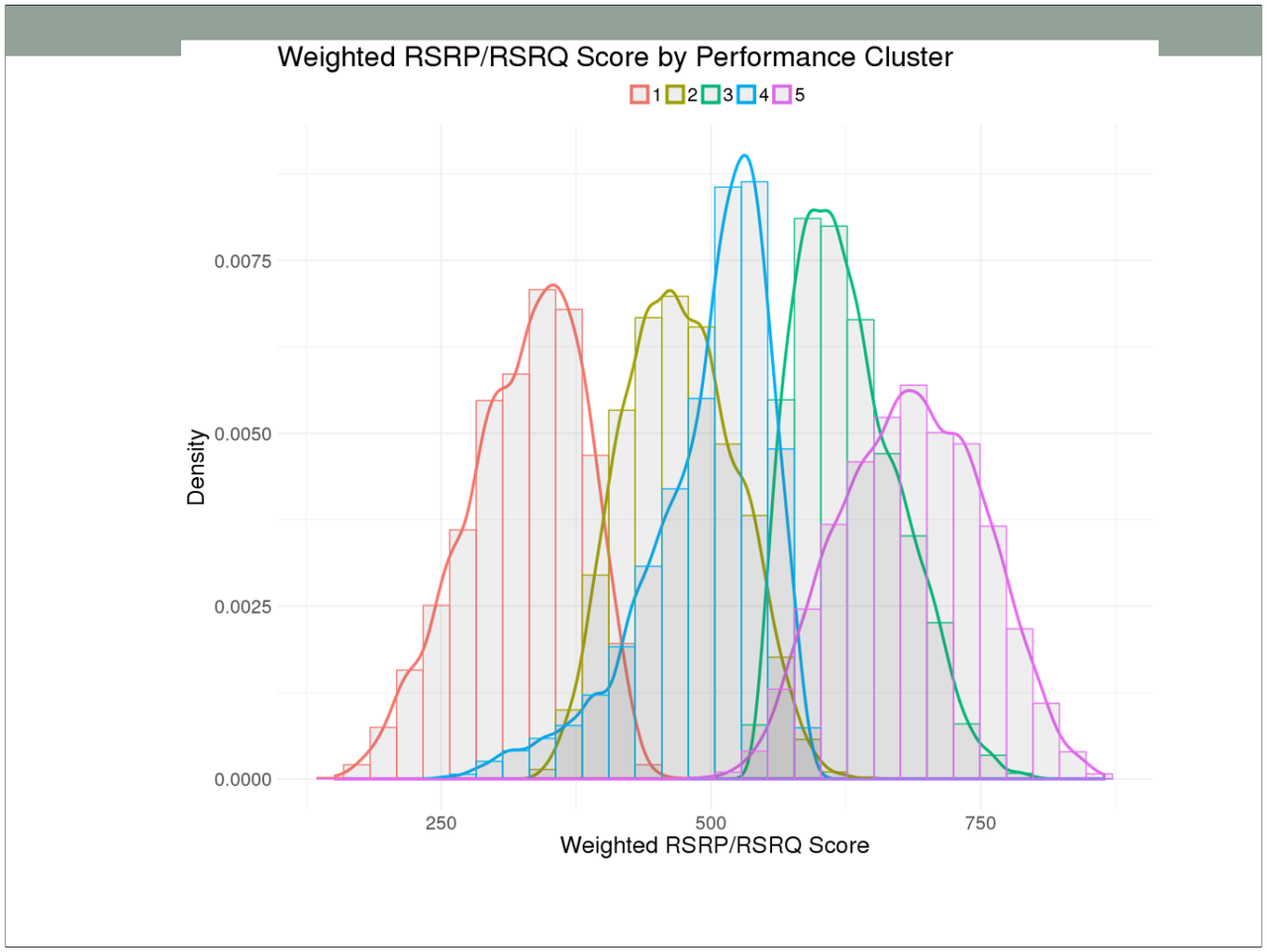
- Performed unsupervised clustering on Weighted RSRP/RSRQ Score and LTE Signal Power Mean to slice data into "**RF Performance Groups** "

- Performed unsupervised clustering on Weighted RSRP/RSRQ Score and LTE Signal Power Mean to slice data into “**RF Performance Groups** ”

Here we have separated each performance group. Visually, it is still hard to discern “problem areas”. In an upcoming slide we will address this issue by using hierarchical clustering.



- Performed unsupervised clustering on Weighted RSRP/RSRQ Score and LTE Signal Power Mean to slice data into "**RF Performance Groups** "



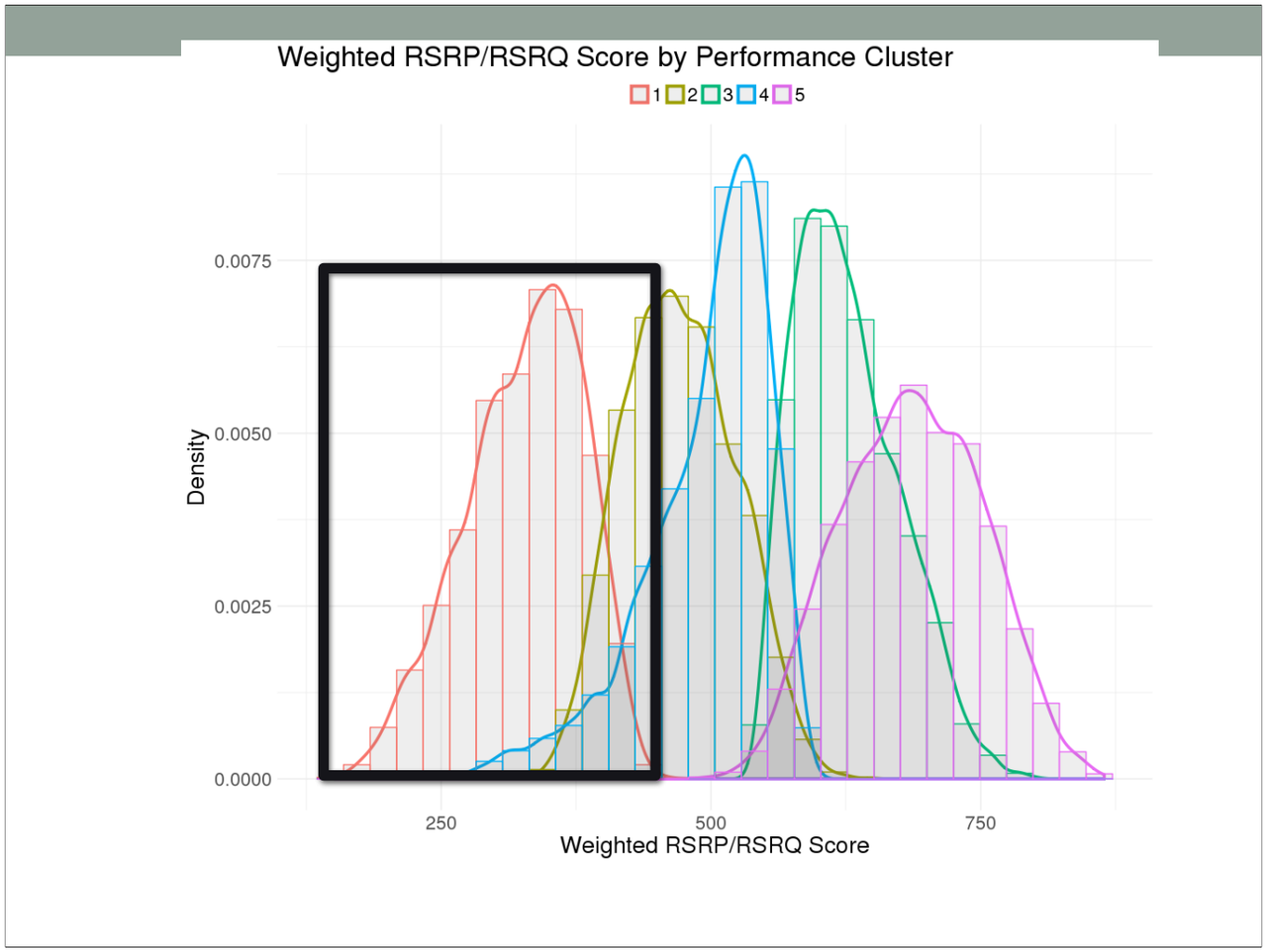
Each performance group (cluster) represents a differing level of *performance relative to every other cluster. In this example, cluster 1 (red) happens to be the lowest preforming cluster generated by this exercise. Here, the amount of overlap between clusters is overstated due to the absence of the second variable used for clustering, LTE signal power mean.

This allows us to filter our dataset for the lowest performance cluster (in this example red cluster 1) and use that as the basis for further modeling. In effect, we're only looking at the worst offending data points.

Summary table for Performance Clustering:

v_cluster	row_count_values_clust	avg_weighted_rsrp_rsrq_score	avg_observations	avg_LTE_signal_power_mean	
1	5742	326.5937		0.0000246	-102.46743
2	8122	471.0516		0.00002665	-97.86013
3	5479	626.6544		0.00002788	-90.73134
4	4794	496.7239		0.00004979	-87.78557
5	4990	686.3683		0.00005027	-78.605

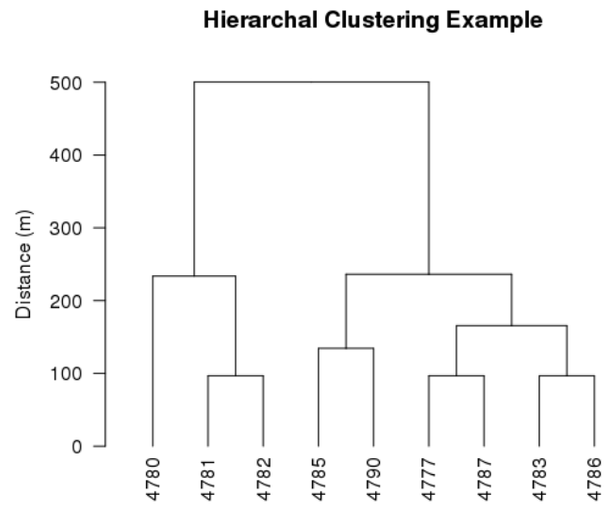
*Assuming our weighted RSRP/RSRQ score is a good proxy for RF performance


Weighted RSRP/RSRQ Score by Performance Cluster
1
2
3
4
5
Density
0.0075
0.0050
0.0025
0.0000
250
500
750
Weighted RSRP/RSRQ Score

Methodology

Hierarchical Clustering of Performance

- Selected lowest performance cluster (lowest average weighted RSRP/RSRP Score) and calculated distance matrix for all points
  - Performed hierarchical clustering based on distance matrix (in meters) of each 140m coord to every other 140m coord
    - Cut the "height" of each cluster tree allowing us to limit the maximum allowable distance between each 140m bin and



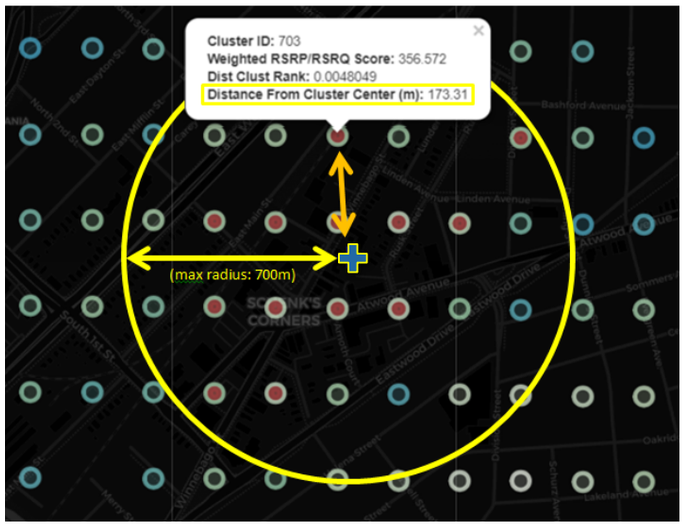
This plot shows an example of a hierarchical cluster based on distance (m). The points at the bottom represent individual 140m coordinates. In this example, 9 coordinate points were clustered together with a maximum distance from any individual point to the center of around 500m. We chose 500m because that represents the performance range of a small cell, thus selecting areas that could be served by construction of a small cell site. The algorithm looked through all the data, identified which data points were closest together, and then found the centrality for each grouped point.

If we look at the first hierarchical clustering example plot and start at the bottom far right, we see that data points 4786 and 4783 were determined close enough to group together by the algorithm at the lowest level of the hierarchy. The next pair over, (4787 and 4777), were also determined to be close enough together at the first level to be grouped. If we go up a level of the hierarchy, we see that the second pair (4787 and 4777) are close enough to the first pair (4786 and 4783) that they can be grouped together at around 150m. The first pair and the second pair, are however, not close enough distance wise to the third pair (4790, 4785) to be grouped at a height of 150m. If we cut our tree at 150m, the result in this example would be four separate clusters from right to left: #1(4786, 4783, 4787, 4777) , #2(4790, 4785) , #3(4782, 4781) , #4(4780). For our purpose in this example, we're cutting the height of our hierarchy at 500m, meaning all points here have been determined close enough to be grouped.

Methodology

Hierarchical Clustering of Performance

- Selected lowest performance cluster (lowest average weighted RSRP/RSRP Score) and calculated distance matrix for all points
  - Performed hierarchical clustering based on distance matrix (in meters) of each 140m coord to every other 140m coord
    - Cut the "height" of each cluster tree allowing us to limit the maximum allowable distance between each 140m bin and

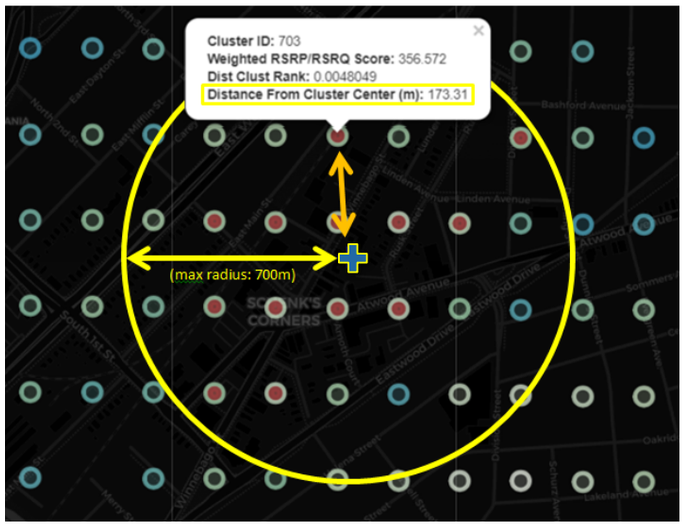


Methodology

Hierarchical Clustering of Performance

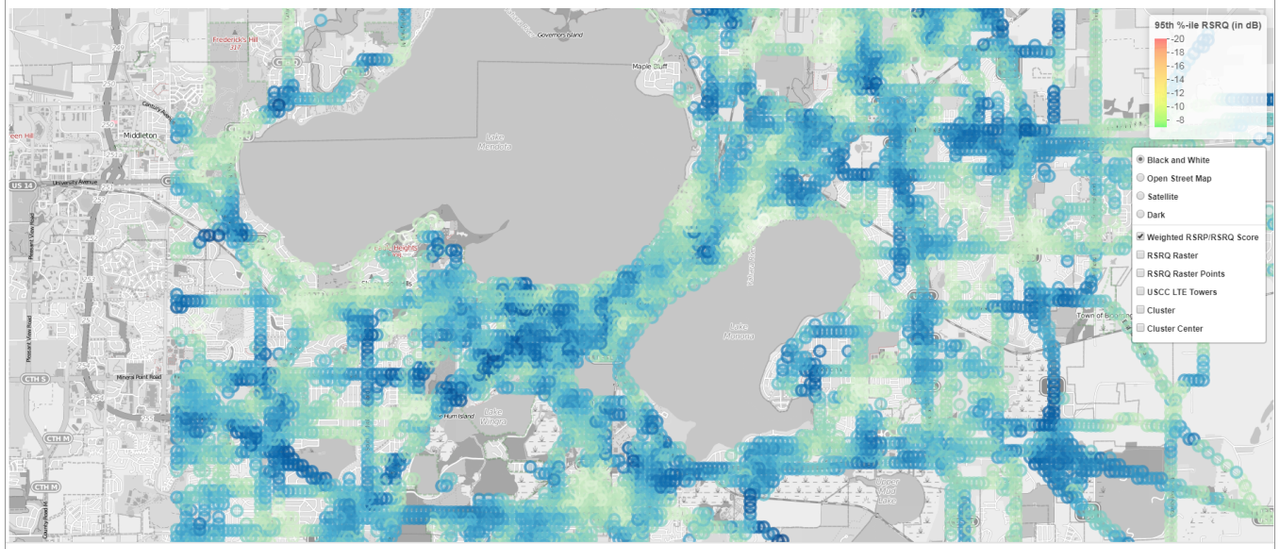
- Selected lowest performance cluster (lowest average weighted RSRP/ RSRP Score) and calculated distance matrix for all points

- Performed Hierarchical clustering based on distance matrix (in meters) of each 140m coord to every other 140m coord
  - Cut the "height" of each cluster tree allowing us to limit the maximum allowable distance between each 140m bin and derived optimal cluster center



Now we see what these points look like on the map. In this example, 9 coordinate points were clustered together with a maximum distance from any individual point to the center of around 500m. The algorithm looked through all the data, identified which data points were closest together, and then found the centrality for each grouped point (marked with the blue cross). Each blue circle represents a coordinate record in the application data. Each red point is determined by the unsupervised clustering algorithm to belong to performance group 1. The allowed max cluster radius was 700m – observed maximum radius was actually ~530m, resulting in 1,243 clusters generated in our proof of concept area of interest (Madison, WI). In our dashboard implementation, we use a field from the data set to determine which points represent the most number of observations. In other words, each bin, or circle on the graph, can represent a difference percentage of observations. We ranked our clusters by the percentage of observations they represent, thus our top ranked cluster has the highest percentage of observations and poor performance. Then we chose to highlight only the top 35 clusters on our dashboard. This choice helps Network Engineers to quickly focus in on the most underperforming areas with the highest traffic density. While we could have made this a modifiable option on the dashboard, we opted to hard code it into the app instead. This was due to recommendations by the Tools Architect to keep the dashboard simple with few user options to avoid non-standard ways of using the data presented.

Shiny Dashboard Demo

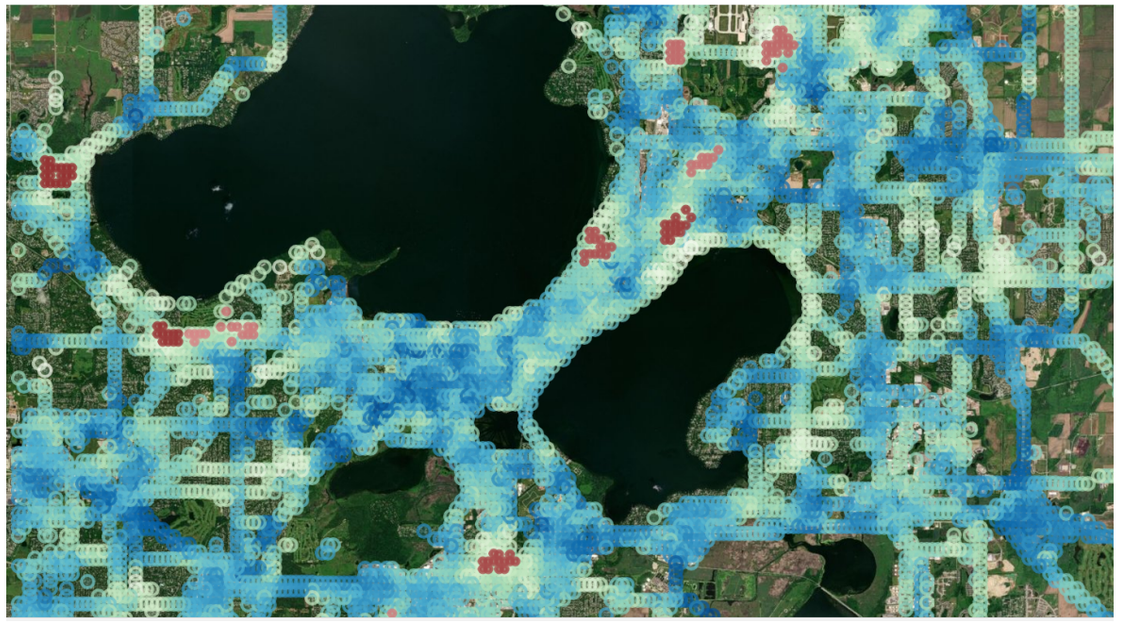


When we load the dashboard, we see the madison area with each aggregated observation plotted as a colored circle on the map. The darker blue circles are areas of good performance, while the lighter circles represent poor performance. The dashboard has a limited number of options by design. The options are available on the right-hand side of the map.

The map is generated using the leaflet package in R and features the ability to include multiple base layers. We have included 4 base layers in this dashboard: “Black and White”, “Open Street Map”, “Satellite”, and “Dark”. These layers can be changed by selecting a different radio button on the right. Here we have selected the “Black and White” radio button. Additionally, there are several layers that can be added based on what data we want to visualize. Multiple boxes can be checked at the same time and the options are “Weighted RSRP/RSRQ Score”, “RSRQ Raster”, “RSRQ Raster Points”, “USCC LTE Towers”, “Cluster” and “Cluster Center”.

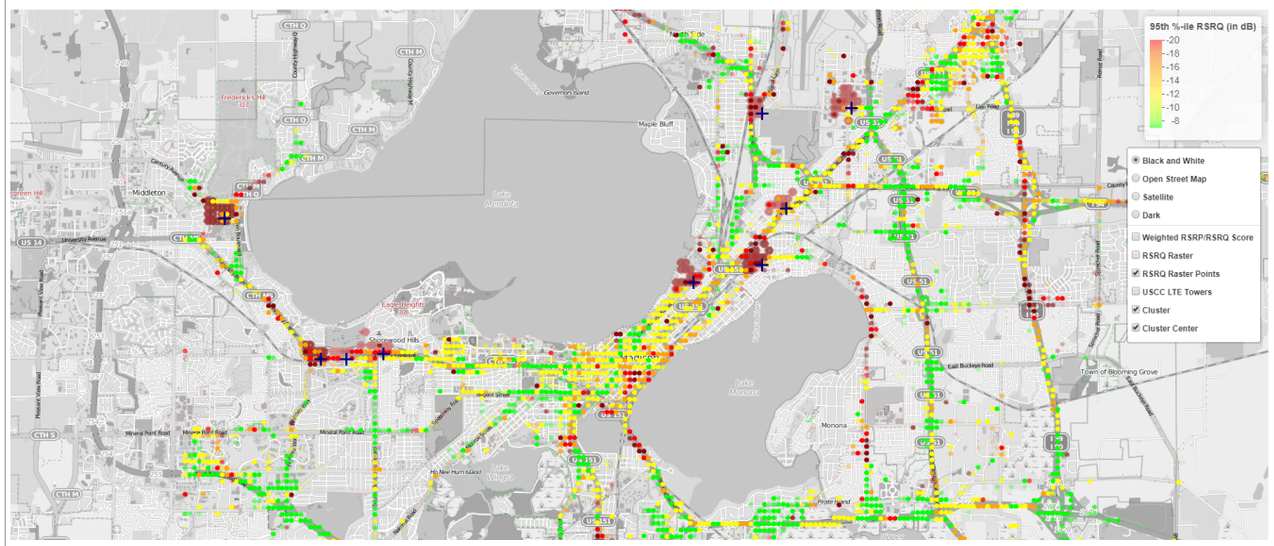
In this view, We have the “Black and White” map layer selected and we are visualizing the “Weighted RSRP/RSRQ Score” for each observation in the data set.

Shiny Dashboard Demo



Here, we have changed the map layer to the “Satellite” view and have added the clusters (red). The darkness of the clusters is based on the percentage of observations represented by cluster, so a darker red cluster represents a higher percentage of traffic. Hover over a data point will give more information about the underlying data of each observation.

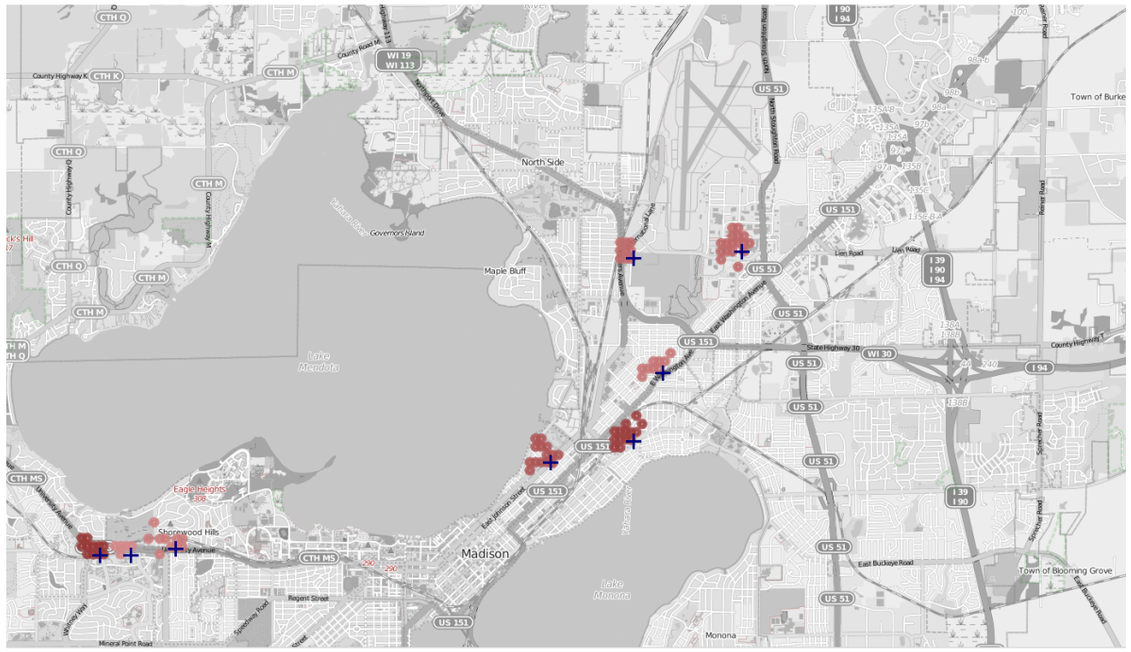
Shiny Dashboard Demo



Here, we have zoomed in and are back on the “Black and White” map layer. We have kept the clusters on the map and added the cluster centers, represented by the blue cross. Additionally, we turned off the weighted RSRP/RSRQ score and replaced it with the RSRQ raster points. These points represent only the 95th percentile value of all RSRQ values in that bin. In other words, for each 140m bin (data point), the 95th percentile value tells us that 95 percent of observations within that bin were at least X or better. We color the points based on the value of X. So if 95 percent of RSRQ values in the bin are between -10 and -8 dB or less, then we color it green. If 95 % are between -16 and -20 dB or less, then we color it red. We determined the color scale based on predestinated thresholds from Network Engineers.

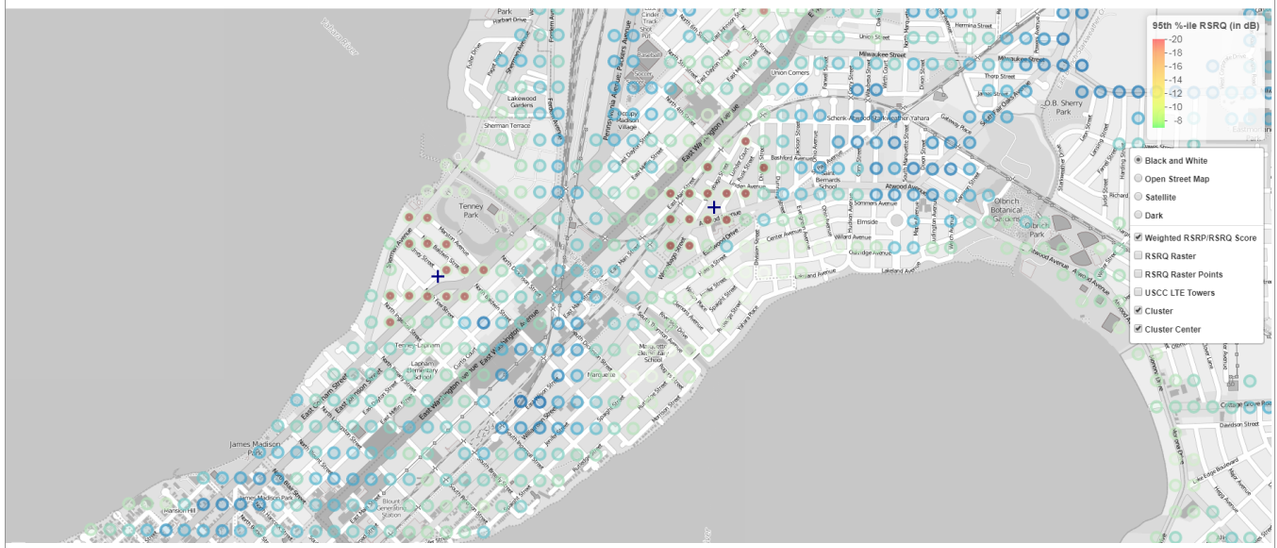
Looking at the data in this way lets us validate that our clusters are lining up with areas of red and orange “RSRQ Raster Points”, or poor areas with bad RSRQ values.

Shiny Dashboard Demo



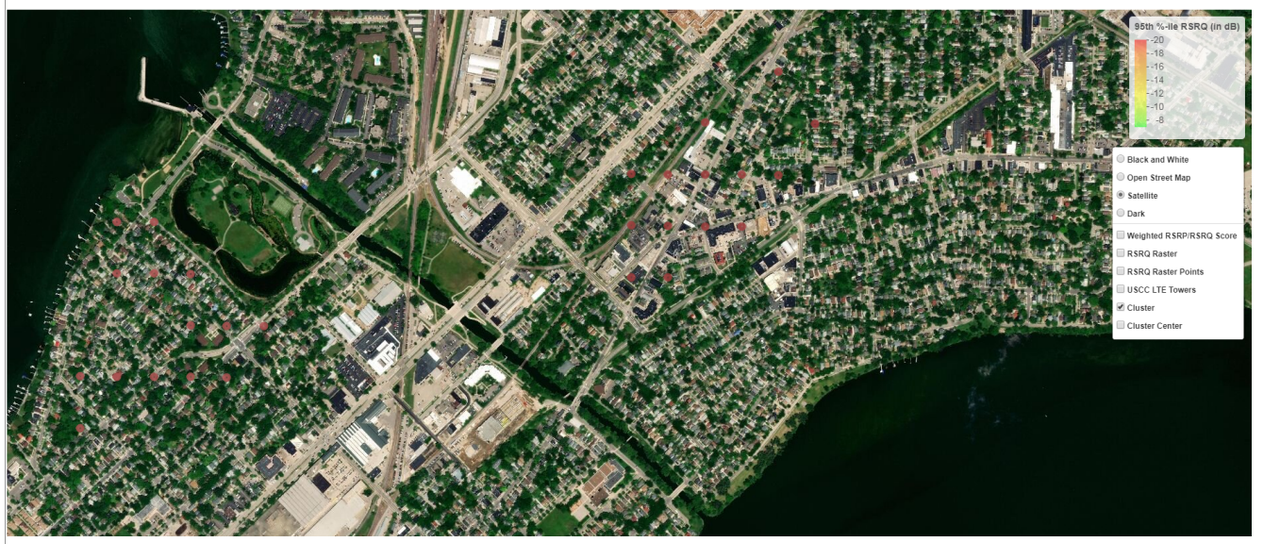
Turning off all the colored points leaves us with the clusters and their centers. This is an interesting high level view for engineers to begin trying to determine why these geographic areas are underperforming. Not shown here are the location of the cell towers serving this area. We used in-house data to add our cell towers to the map as an available layer option. This view could not be shown here for confidentiality reasons. The Madison, WI area is served by many towers. **Identifying hot spots of poor coverage with the precision shown here was not possible with the tools currently available to the company due to FCC's CPNI rules.**

Shiny Dashboard Demo



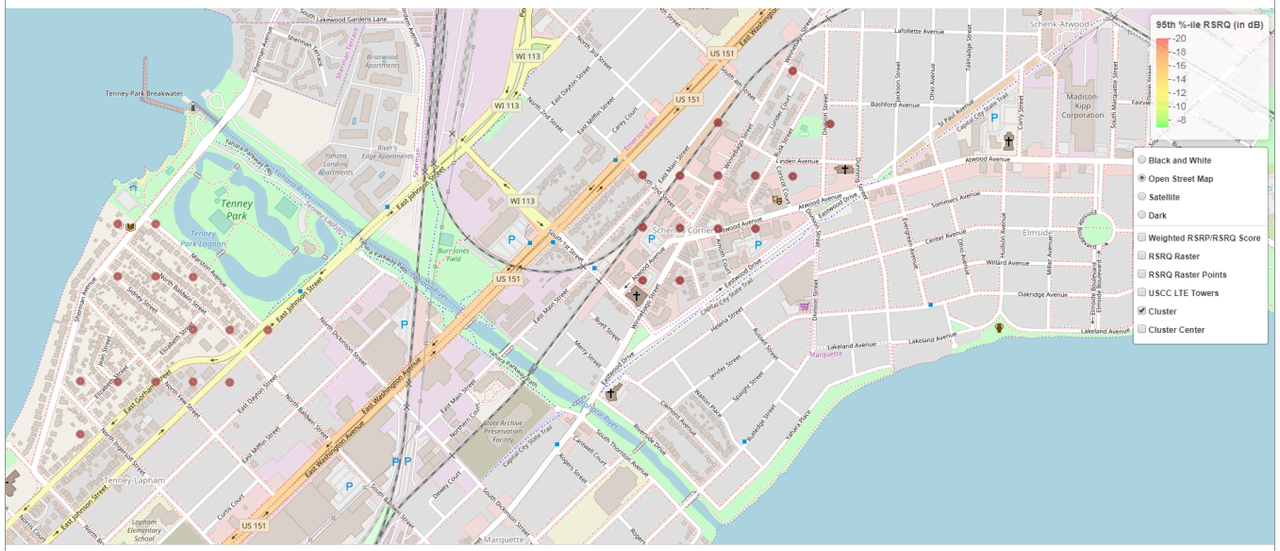
Zooming in to the data, we can start to see where the poor coverage is happening. It is not suprising that the area around the lakes is light green/blue (poor performance) due to the way the radio waves travel over water. Dealing with waterways in urban areas is a known challenge for cellular networks.

Shiny Dashboard Demo



Turning on the satellite map layer, we can see that there are a lot of buildings in the poor performance cluster.

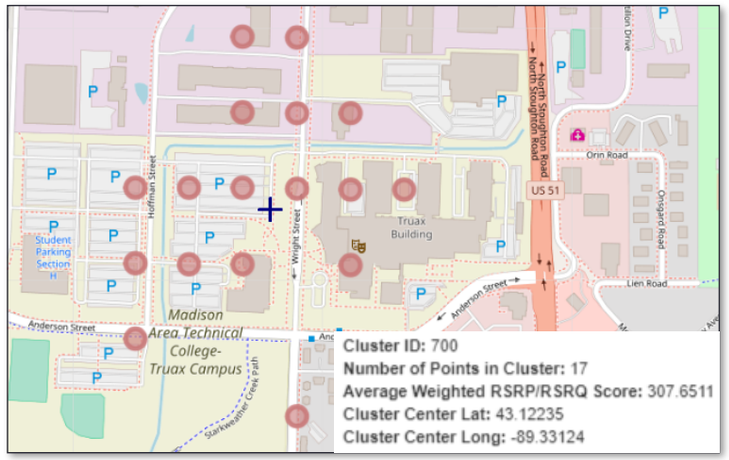
Shiny Dashboard Demo



Looking at the open street map now, we see a few parking garages in addition to the building. Further analysis could be done to confirm, but we expect that we might be seeing “in-building” traffic here. This is traffic from people using data services while indoors. Even though we may expect great coverage in the area and be located very close to a tower, the way the radio waves propagate through the urban materials of steel and concrete can result in lower than expected performance. Going forward, it would be interesting to see if clusters such as this were consistent from week to week, then determine how much traffic is actually coming from this area, and finally consider placing a small cell on a rooftop to serve the customers in this area. This is an example of the potential use of this dashboard application.

Summary: Use Case 1 POC

Using data from the Application Data tool, we created a visual product that **highlights geographic areas** containing a **high density of network traffic** along with **poor RF performance**.



These derived areas may represent optimal locations for future **small cell deployment**, but the general accuracy of these predicted areas is **limited** by the limitations outlined in the Application data.

Use Case 1 POC ~80% Completed

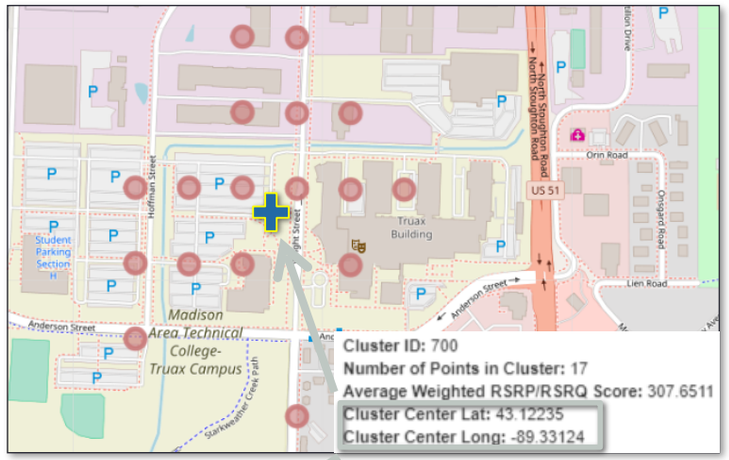
In the cluster pictured, we are again likely seeing a lot of in-building traffic from Madison Area Technical College. Further analysis of the cluster could reveal mean/median RSRQ/RSRP for the cluster, mean/median values over time and relative volume of traffic coming from this area.

Here we propose that we can have built-in “low performance” detection and with an automated optimization recommendation (the cluster center) from the Machine Learning Model giving us a potential small cell site.

Summary: Use Case 1 POC

Using data from the Application Data tool, we created a visual product that **highlights geographic areas** containing a **high density of network traffic** along with **poor RF performance**.

These derived areas may represent optimal locations for future **small cell deployment**, but the general accuracy of these predicted areas is **limited** by the limitations outlined in the Application data.



Cluster centers represent predicted optimal small cell locations.

Use Case 1 POC ~80% Completed

Predictive Analytics

- Currently have used three AI/ML methods to determine RF metric thresholds for video resolutions (see “Next Steps” slide):

			Measure of Performance		
Method	**Bin Size**	**time bin**	**AUC**		**rmd file**
Logistic Regression	5km	5wk	**0.7131**		MNI_classification.rmd
	280m		0.63		
	140m		0.63		
			Confidence	**Lift**	
Assoc Rules	20m	5wk	1	2.649351	MNI_Assoc_Rules.rmd
	35m		1	2.695025	
	70m		**0.8275862**	**2.397988**	
	140m		0.7191011	2.351113	
	280m		0.652439	2.34872	
	600m		0.6976744	2.514018	
	1km		0.8448276	2.717118	
	2km		0.9545455	2.714816	
	5km		1	2.518022	
Method	**Bin Size**	**time bin**	**Mean Squared Error**		**rmd file**
Random Forest	5km	5wk	**~.24**		MNI_Random_Forest.rmd

We looked briefly at using machine learning techniques to predict video resolution based on RF metrics. We tested Logistic Regression, Association Rules and Random Forest, the best we have so far is the Random Forest with a 24% error rate. We expect we can make this better with via tuning.

Next Steps

- Evaluate predictive power of RF metrics for Video performance

Resolution	Threshold RSRP/ RSRQ Values	%of coverage area currently supported
360p	TBD	TBD
480p	TBD	TBD
720p	TBD	TBD
>720p	TBD	TBD

- Explore relationship between Application RF performance metrics and time for each 140m coordinate bin
- Enhance clustering methodology so that problem areas can be assigned a relative priority
- Deploy shiny app so that, on demand, RF engineers can click a URL and get the information needed both visually and in csv form

We only explored 1 of 4 prospective use cases. We believe there may be predictive power in the RSRP/RSRQ score to predict expected video resolution capabilities in a geographic region. Something that the company does not currently posses, yet is very interested in, is a video resolution coverage map. Customers are no longer only concerned with their ability to make a call, they want access to all of their data services wherever they go. If we are able to use video resolution as a proxy for data service quality, then we may be able to generate a map that shows areas of good/bad data service. This coverage map would be very accurate if we could use the actual customer experience data gathered from the users application usage, such as we have been given here.

On the slide is an example of a table that we could potentially derive from the data. This would provide insights into the actual video coverage of our network and how to optimize for video resolution.

We also worked with a static data set in this proof of concept. More interesting trends and insights could be gained by looking at how the data changes over time.

Use Case 1 Proof of Concept is about 70-80% complete. If the engineering team deems it to be a useful tool, several Data Engineering tasks would need to be accomplished to make it a sustainable source of information:

1. Pull data via API
2. Create tables in Hadoop to store the data
3. Work with the automation team to improve the efficiency of the code
4. Clean up the User Interface
5. Improve the scoring algorithm and fine tune the clustering algorithms.

Questions?

Supplemental Slides

How much data is available?
1 week data

Estimate* of Data Availability (Nationwide)

Aggregation	Resolution	Time Frame	Total Sample Size	% Missing Video Data	% Missing RF Data	Usable RF/Video Sample Size
Coordinate	20m	1 week	11,642	99.86%	6.01%	10,936 / 0
	35m	1 week	56,471	99.90%	5.33%	53,442 / 1
	70m	1 week	130,686	99.80%	5.63%	123,290 / 5
	140m	1 week	188,401	99.39%	6.37%	176,357 / 19
	280m	1 week	200,050	97.98%	7.08%	185,845 / 180
	600m	1 week	170,464	95.18%	8.25%	156,383 / 1,022
	1km	1 week	123,039	92.37%	9.39%	111,471 / 2,951
	2km	1 week	80,449	90.17%	10.44%	72,037 / 3,623
	5km	1 week	46,721	86.06%	10.97%	41,594 / 3,105

Time Frames Available: 1 day, 1 week, 35 days** (**default)

Data Recency: Application updates a minimum of 3-4 days after reporting period. Reporting period is Saturday to Friday.

*Estimate is based on data sets for June 2 - 8, 2018

We included this slide to demonstrate the sparse amount of data available at more granular levels.

Deriving Insights: Use Case 1

- We explored the data and found that the provided video resolution metrics were incomplete, leading us to focus on an **RF centric use case**.
- Brainstormed internally and agreed that there would be a high level of **business value** around providing insights for **small cell deployment**.
- We iterated through various ways to present insights from this data that would be useful for an **RF planner**. We wanted to do more than replicate the application data dashboard.

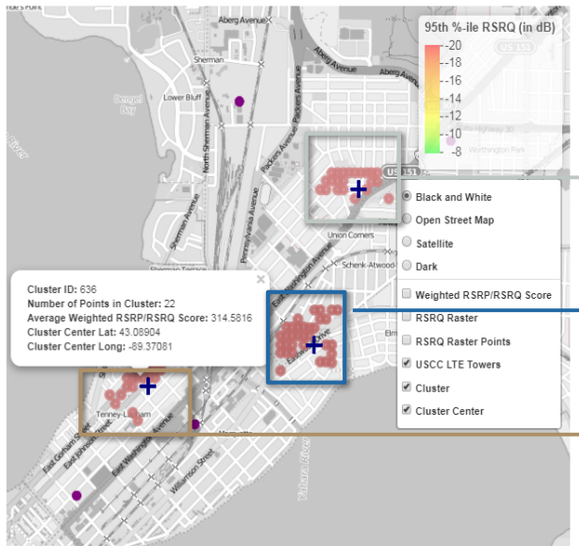
Additional Information regarding our choice to analyze Use Case 1:

Use Case Selection:
- Incomplete YouTube video resolution metrics lead us to focus on an **RF centric use case**
- Identified high level of business value around proving insights for **small cell deployment**
- Wanted to present insights from this data that would be useful for an **RF planner**

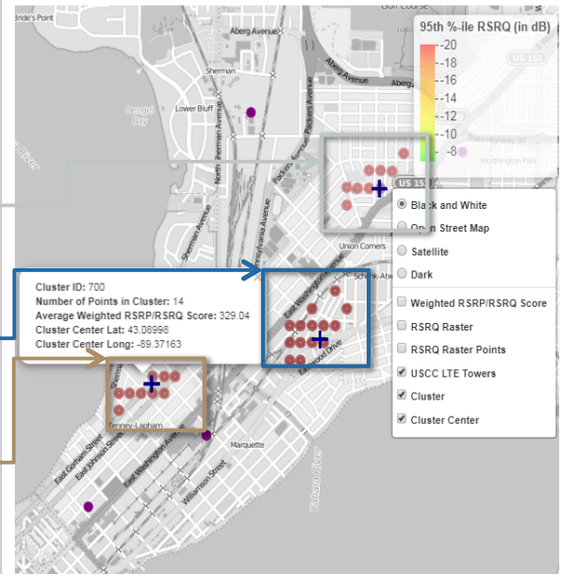
Benefits:
- Application level metrics by geographic location
- Identification of low performance areas custom to our network
- Push-button insights for RF engineers

Comparing Clustering:
70m bin versus 140m bin

MNI Project (Test App - MNI Resolution: 70m, Clustering Radius: 700m)



MNI Project Resolution: 140m, Clustering Radius: 700m



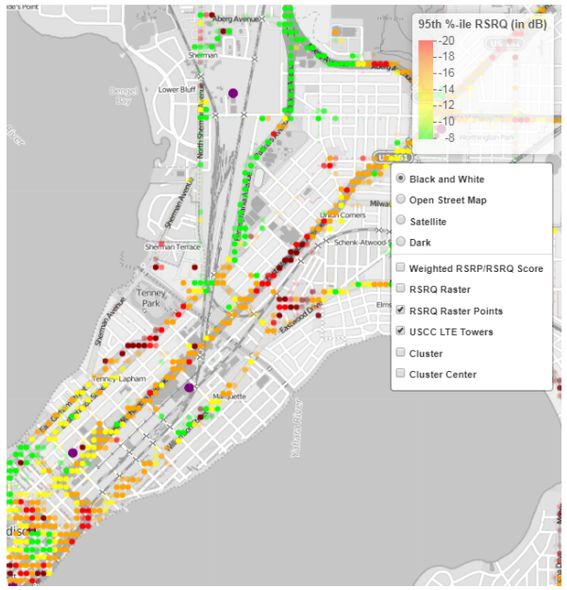
We iterated through several steps to compare and contrast the 70m dataset and the 140m data set. Ideally, we wanted to go with the smallest geographic bins.

Comparing the 70m Bin to the 140m bin, we can narrow in on the center of mass of the traffic. Also, clusters with higher traffic and poor performance will remain visible as we increase the resolution.

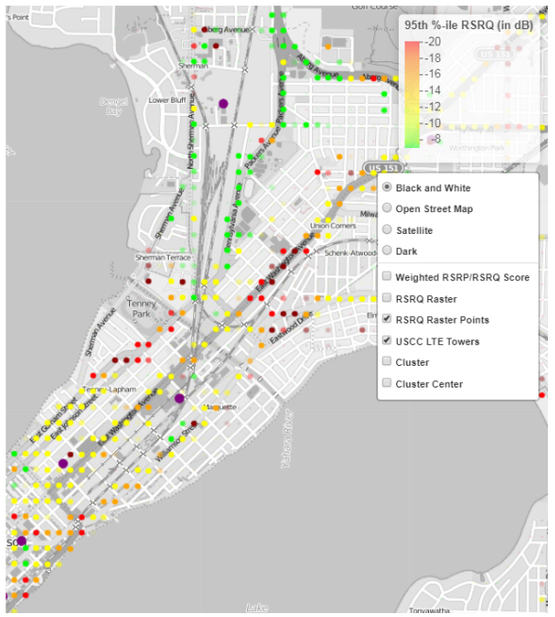
At each resolution, the cluster center is almost unchanged, however, **we lose a considerable amount of observations included in each aggregated bin below 140m.**

Comparing Data Availability: 70m bin versus 140m bin

MNI Project (Test App - MNI Resolution: 70m, Clustering Radius: 200m)



MNI Project Resolution: 140m, Clustering Radius: 700m



Visually, the comparison of Data Availability supports using 70m data since it looks like there are more data points. However, **we determined that the data points for 140m bins are "heavier", ie. They represent a larger % of observations (traffic) overall.** We can expect that as we "zoom out", the data set will represent a larger % of total traffic.

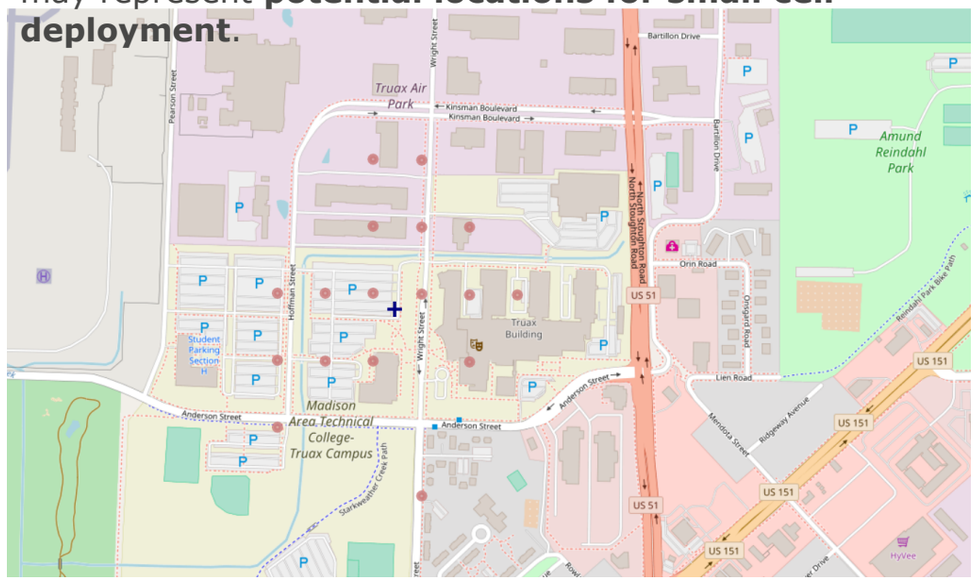
What it does

- Provides a mapping of the Madison Market, overlaid with data provided by the Application Vendor showing aggregate **RSRP/RSRQ** conditions per 140m lat/long cell.
- Clusters each 140m cell and estimates dense clusters that represent **poor RF conditions**.
- Provides the center point of these poor RF clusters, which may represent **potential locations for small cell deployment**.

Summary of the R Shiny App.

MNI Visualization

Provides the center point of these poor RF clusters, which may represent **potential locations for small cell deployment**.



In the cluster pictured, we are likely seeing a lot of in-building traffic from Madison Area Technical College. Further analysis of the cluster could reveal mean/median RSRQ/RSRP for the cluster, mean/median values over time and relative volume of traffic coming from this area.