Investigation of Optimal Reimbursement Policy for Wifi Service by Public Libraries

Vinod Bakthavachalam

August 28, 2014

Data Overview

The purpose of this analysis is to investigate what the optimal reimbursement policy should be for public libraries that provide wifi service to their customers. The FCC currently has three policies under proposal: (1) a reimbursement rate based on size of the library only, (2) a rate based on number of users of the wifi, and (3) a rate based on a combination of the two. Specifically, this analysis seeks to determine whether any of the proposed reimbursement systems would differentially affect rural, town, suburban, and city libraries through examining whether the trends for size and/or users differ across library type.

The data comes from the Institute of Museum and Library Services and is publicly available[1]. Survey results of public libraries in 2012, the most recent year available, consist of three files: one with state level aggregated totals of libraries within the area, data on each unique library system, and data on all branches/locations for all library systems. For this analysis the last two files provided the data needed. The first file with data on each unique branch contained variables that served as an estimate of the number of users of the wifi service (PITUSR), a proxy for the cost of providing that service (OTHOEXP), longitude (LONGITUD) and latitude (LATITUDE) of the building, and a location indicator of the library (LOCALE).

Estimated users was annualized and included only those using public computers within the library, so it did not account for personal devices, making it likely an underestimate of true usage levels. Cost of providing the service included items in addition to the cost for just providing wifi, so it was an overestimate. Despite these limitations the two variables were used as if they were unbiased measures. Interpretation of the results must therefore be taken with caution. The location indicator separated libraries into four main areas: city, suburb, town, and rural, and then three subareas within each area. A new indicator based on this variable was created where the subareas were combined within each category to give the four main areas. All observations were therefore sorted into four groups to tackle the issue of interest.

The variable for size (SQ_FEET) was in the second data set with observations for all branches/locations within a library system. This data set was much larger than the first one (17586 vs. 9305). Hence, some matching needed to be done in order to obtain one dataset with both the size and number of users. Both datasets contained a variable called FSCSKEY that was a unique identifier assigned to each library in the sample by the IMLS. Different branches/locations of the same library system had the same identifier. To merge the datasets the maximum size of all branches for each library system was taken and then matched to the first dataset. This yielded a combined dataset that had 9305 observations, a one to one correspondence with the original set. In this way all variables necessary for the analysis were placed into a single data set. Finally, observations with either negative size, negative number of users, or a location outside a US state/DC (such as Guam or Puerto Rico) were excluded, leaving 9024 libraries to be analyzed. The breakdown across the four groups was:

```
##   CITY  RURAL SUBURB   TOWN
##    484   4047   2298   2195
```

The most prevalent group was rural libraries while there were around half as many suburban and town libraries. City libraries were the least common group.

Summary of Results

Several different methods of analysis were undertaken to examine the issue of whether the proposed reimbursement methods would differentially affect the four types of libraries and examine if the trend of users and/or size across libraries could be

[1] See: http://www.imls.gov/research/pls_data_files.aspx

determined, which would allow a formulaic cap to be constructed. Each method sought to examine the question in a slightly different way to check the robustness of the conclusions as well as see if additional insight could be obtained.

As a first pass a scatterplot of size by users was created to examine visually whether there was any clustering by library type on these two variables (Figure 1). All libraries were plotted and points were color coded by type, so that clusters if there would be visually available. The hope was this plot would pick up any trends/groupings. The plot gave some evidence of clustering, but it was overshadowed by libraries that had very large size and/or user levels. From the plot it appeared that the majority of these points were city libraries with some suburban ones as well.

To examine these more formally, histograms and empirical CDFs were constructed to look at the distributions of each variable (Figure 2). Both had long right tails, expected from the scatterplot, and additional analysis revealed that 295 libraries or over 60% of libraries with size or user levels in the top 5 percent were city libraries and represented the majority of the user base (216661201 users which is over 63% of the total). Further analyses excluded the top 5 percent of each variable to better examine trends among the average library, but conclusions from here on should be tempered as they do not include the majority of city libraries.

The scatterplot above was reconstructed (Figure 3). With extreme values removed, the plot now revealed groupings of library type. Rural were the smallest, followed by town, then suburban, and lastly city had the largest values on the two variables. Hypothesis tests and boxplots confirm this trend (Figure 4). Even in users per square foot city libraries have significantly larger values than the other types. The four groups are significantly different in both user and size levels and therefore no single cap is appropriate. Rather a cap that is a function of either size, users, or both is needed.

In fact a plot of the median values across type revealed an almost linear relationship between size and user levels and library type (Figure 3). A straight line fit through rural, town, and suburban libraries. The relationship changed depending on whether city libraries were included or not. This suggests that a policy that works for rural, town, and suburban libraries would not be suitable for city libraries as they are much different than the other types, keeping in mind that the majority of city libraries have extreme size and user levels this is especially true.

Geospatial plots were then constructed that overlaid the distribution of libraries across size and users on the US Map, plotting the longitude and latitude of each library (Figures 5 and 6). The idea here was to examine the clustering by location in another way and also confirm the validity of the location indicator constructed. The plots revealed similar trends as the scatterplot as larger libraries with a bigger user base were concentrated around major cities and in more urban areas. Again, city libraries seem to be systematically titled towards bigger buildings that house more users of wifi.

Next, a k-means algorithm with four clusters was carried out to examine more formally the correlation between size and user levels and library type (Figures 7 and 8). If there was a true trend then the clusters from the algorithm should match the clusters of library type fairly closely. The algorithm produced similar results as rural libraries appeared smallest on both variables followed by town and suburban libraries and then city libraries were in the cluster with the largest size and user levels. This formalizes the visual results in the scatterplot and also suggests that within libraries there is more variation across user levels than size levels. This raises the question of whether user levels can predict size, reducing the dimension of the data and allowing the FCC to make a cap solely based on user levels. Regression results of size on user levels do suggest this is possible as the percent of the variance of size explained by users is over 50%.

Finally, because the goal is to reimburse libraries for the cost they incur in providing wifi, a proxy for this cost was regressed on size, user levels, library type, and interaction terms. Three models were constructed to match the three types of caps in discussion: just using user levels, just size levels, and both size and user levels as regression predictors. The results reveal that size is the worst predictor of cost but the most similar across different library types as many interactions between size and library type are not significant. The best predictor of cost is both size and users, but many coefficients involving size are not significantly different from zero, so the model reduces to one that is very similar to using just users. Many interaction terms of users with type are significant, suggesting that the effect of user levels on cost is different for each library type.

Overall, a combination of size and users appears to be a better predictor of cost than just size or users alone based on the regression and clustering results. Hence a cap based on a combination of both size and users is likely the most appropriate while a cap based on just size seems to be the least appropriate. Still, any policy proposal will differentially affect city libraries as they are significantly different from the other types. These libraries also actually account for the majority of the reported user base (again around 63% of total users). It appears that potentially two separate policies might be better with one for rural, town, and suburban libraries and another for city libraries. More work should be done to obtain estimates of the true cost of providing wifi service and examining whether the conclusions here hold.

Analysis

As a first pass, a two dimensional scatterplot was constructed with size on the y-axis and users on the x-axis. Points were color coded by library type. If there was a different trend in size and/or users for each library type, then clusters should

be visible. Figure 1 shows two main things. First, there are a lot of extreme values and most of these these extreme values are city libraries (blue points far from origin). There are some suburban outliers as well (yellow points). Second, there does appear to be some clustering. Many green points (rural) are close to one another near the origin. There also appear to be groupings of red (town) and yellow points (suburban) near the origin as well as blue points (city) slightly farther out.

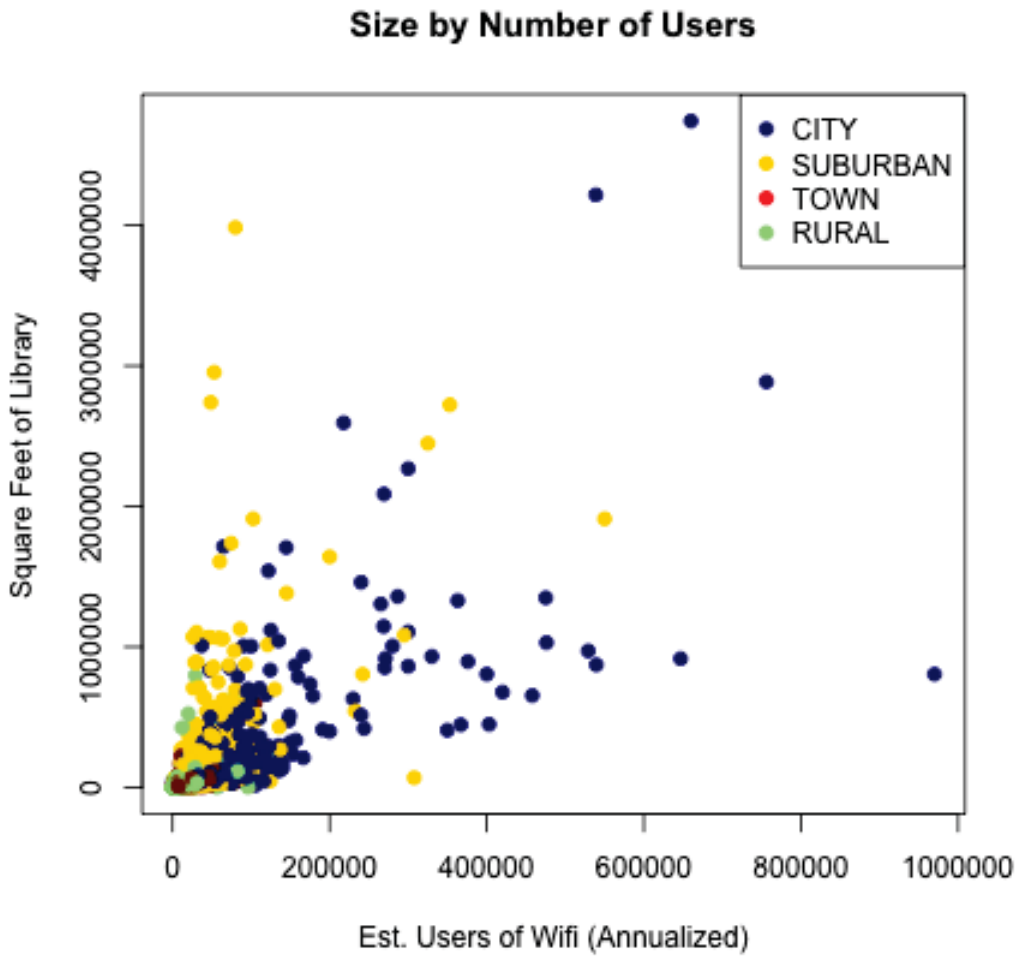


Figure 1: *The scatterplot is of size against users with points color coded by library type. The plot shows that many city and some suburban libraries have large values that obscure trends in the majority of libraries.*

To further investigate the first issue of large outliers, histograms and empirical CDF plots for size and users were constructed (see Figure 2). Note that for both histograms the max value is much greater than the 95th percentile, and the histogram is highly skewed right with a long tail. The empirical CDF plot confirms this as much of the probability mass is concentrated within a short window of values. Because of this fact observations above the 95th percentile were excluded to remove extreme outliers and obtain a better sense of the distributions. The hope was to visualize the clump of points closer to the origin in the scatterplot above and get a better idea of the average library. However, it is important to note the breakdown of type in the outliers. Out of libraries in the top five percentile of either size or users, 295 are city libraries which are 60.9504 percent of all city libraries. By contrast the next most prevalent type in the outliers are suburban libraries, but only 11.8364 percent of suburban libraries are actually outliers. Also, any analysis after removing outliers actually misses the majority of the user base as all outliers represent 63.6904 percent of total reported users of wifi. Hence, any conclusions from using the restricted data set will be missing the majority of city libraries and internet users, but will hold for the majority of other library types. Still, a user based cap on these average libraries only will actually miss the majority of the wifi user base.

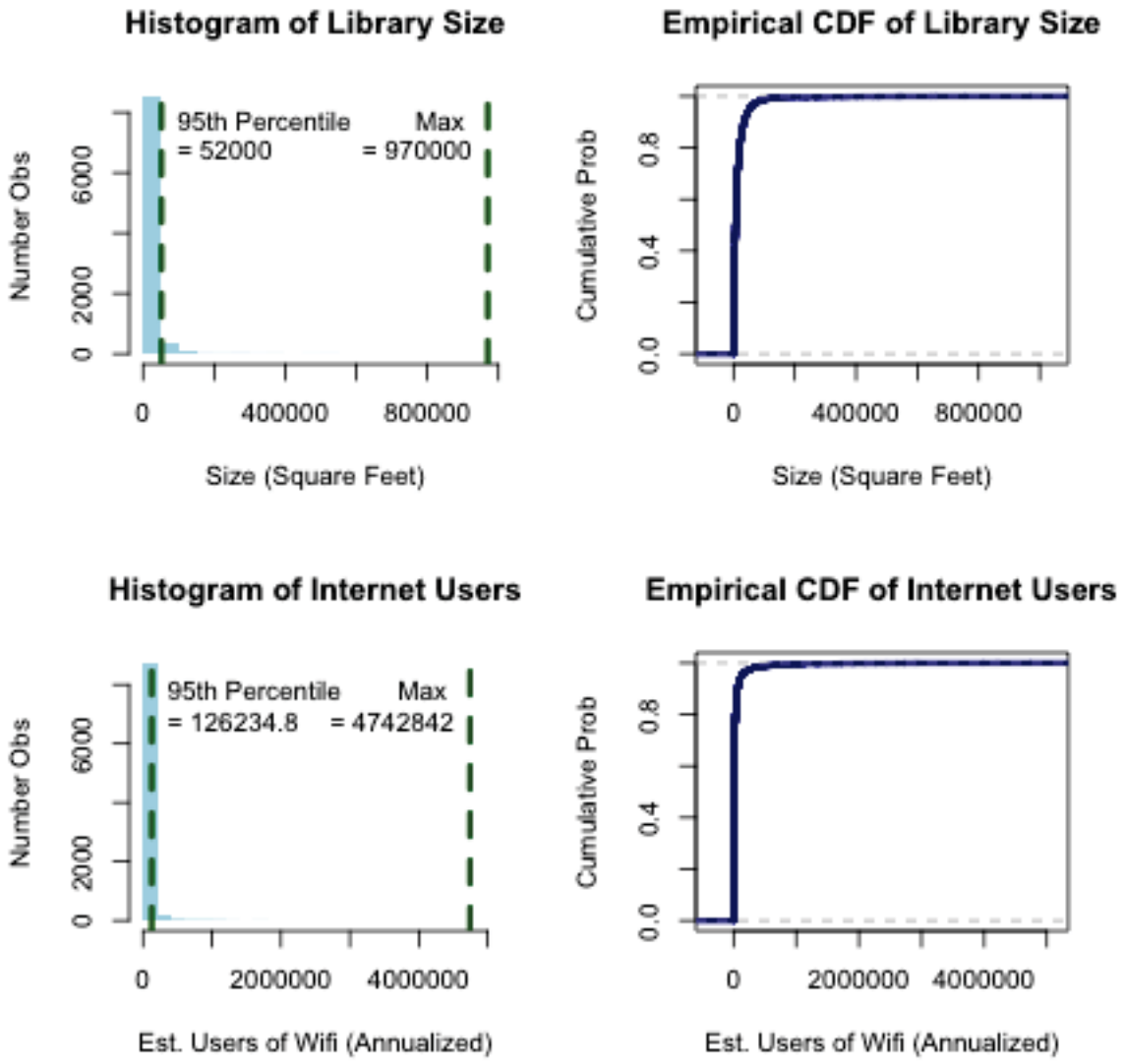


Figure 2: *The left hand plots are histograms of the data while the right hand plots are the empirical CDF. Both demonstrate that the distributions of both size and users are highly skewed right with large positive values. Note from Figure 1 it is clear that a majority of these libraries are city libraries.*

After removing observations with either a large size or large number of users, there were 8388 libraries left. The two dimensional scatterplot was recreated with the restricted sample (see Figure 3). It illustrates a less pronounced clustering effect. Rural, town, and suburban libraries do seem similar to one another as many points are close together. However, rural libraries appear smallest in terms of size and users on average compared to town and suburban libraries, and suburban ones seem larger than town ones. City libraries are very spread out but do appear to be the largest and have the most users, although there are much fewer of them than the other three, so atypical values have more influence on the average value. The plot of the median values for each library type in the two variables confirms this trend. One thing of note is that the relationship between size and users for suburb, town, and rural libraries is very collinear as the regression line through the median values for these three types shows. City libraries are very different from this relationship and including them dramatically changes the regression line. It appears that suburb, town, and rural libraries are more similar to one another than city libraries.

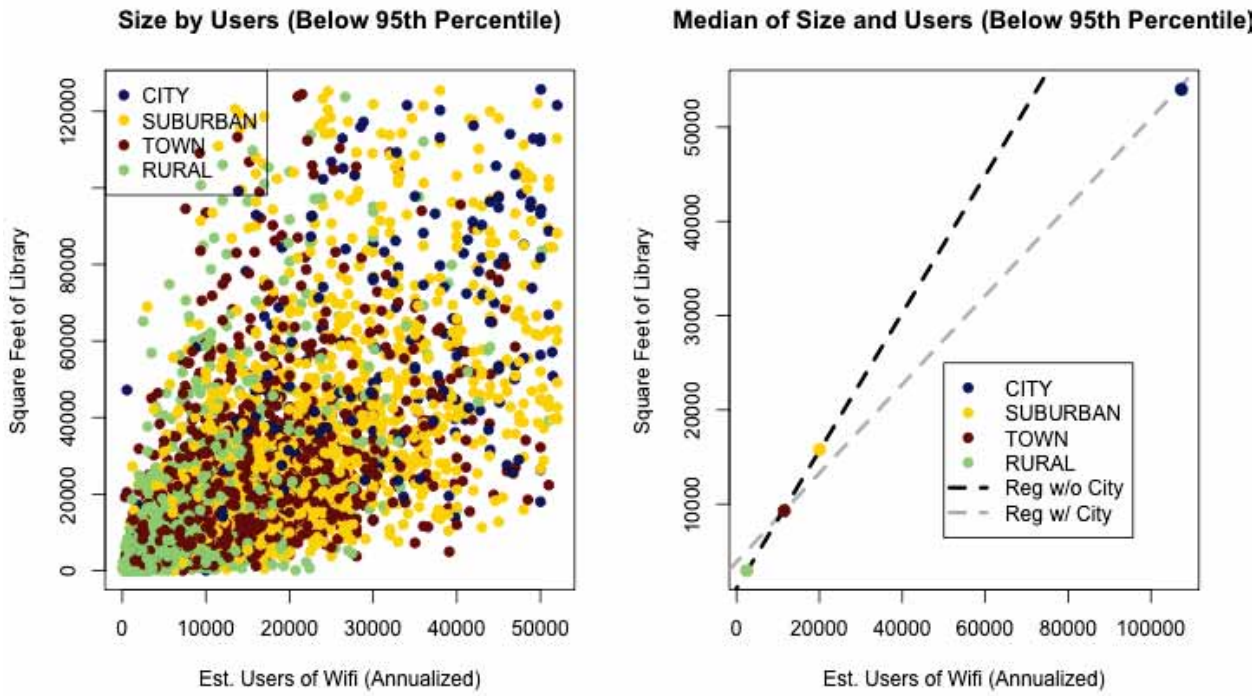


Figure 3: *The left panel shows the restricted scatter plot. The groupings are now clearer after removing outliers. Rural libraries are the smallest in both variables, followed by town, then suburb, and lastly city libraries have the largest values. The right hand panel plots the median of size and users by library types. The trend across type in median values confirms the pattern in the scatterplot to the left. The dotted lines represent the fitted regression lines with and without city libraries. Clearly including city libraries or not changes the functional relationship between size and users.*

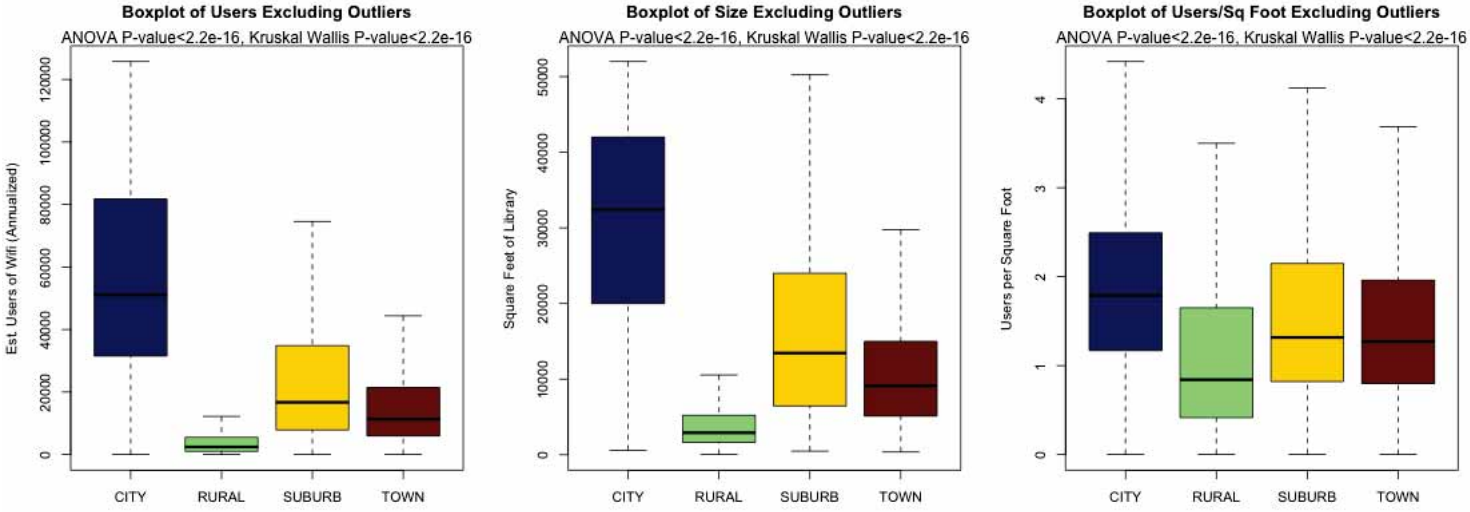


Figure 4: *Side by side box plots across library type for users and size confirms the trends seen in Figure 3. Hypothesis tests confirm that these average levels are significantly different from one another as a group and when comparing them individually. This suggests that all four library types do separate somewhat into four groups on the basis of size and users (at least for the average library in each group). The rightmost plot shows users per square foot and even there city libraries are significantly larger than rural, suburban, or town libraries (significant p-value). Here suburban and town are more similar though. This further asserts that city libraries are systematically different.*

The difference in values and spread can be examined more formally by looking at boxplots of size and users split up by library type (see Figure 4). Hypothesis tests were also carried out for each variable with the null hypothesis assuming that

the average values were the same in all populations. Both an ANOVA (assuming normally distributed data) and a Kruskal Wallis (no assumption of normality) test were carried out due to the high skewness of the data as seen in the histograms. As expected, regardless of which test was used, the null hypothesis was rejected. Libraries do appear to be systematically different in terms of average size and users of wifi across the four types (even if users per square feet is used the results are similar). In addition, the boxplots also illustrate the trend in spread seen in the scatterplot as city libraries have the highest variance in size and users, far more than the other three types. Rural libraries are the least variable across size and users. Although the spreads are clearly different, pairwise comparisons of the average values of both users and size were done to individually test whether the four types were significantly different from one another using Tukey's test for multiple comparisons, which assumes that the variances in the four populations are homogeneous. This assumption is not too bad for the most important comparison of suburb to town libraries as they have the most similar values. All pairwise hypothesis test had p-values$< \alpha = 0.05$, so all null hypotheses were rejected at the traditional significance level. It appears that overall these four library types differ significantly from each other in terms of size and user levels.

In order to look at this another way, geospatial plots were constructed of the libraries with one focusing on users and the other on size (see Figures 5 and 6). Here the sample was restricted to those observations within the continental US (i.e. excluding HI and AK). Points using the longitude and latitude of each library were overlayed on a US map. The color of the points was based on a gradient scale of number of users/size with purple/yellow indicating low users and green/dark red indicating high users. The plot demonstrates clearly that there is correlation between library location and users/size. Areas near major cities/high population areas have darker circles (either green or dark red), indicating larger libraries and more users of wifi. In the middle of the country, both plots have points with lighter colors than on the coasts. The main trends in these plots support the conclusions from the scatterplot above and confirm the location indicator created was a valid measurement.

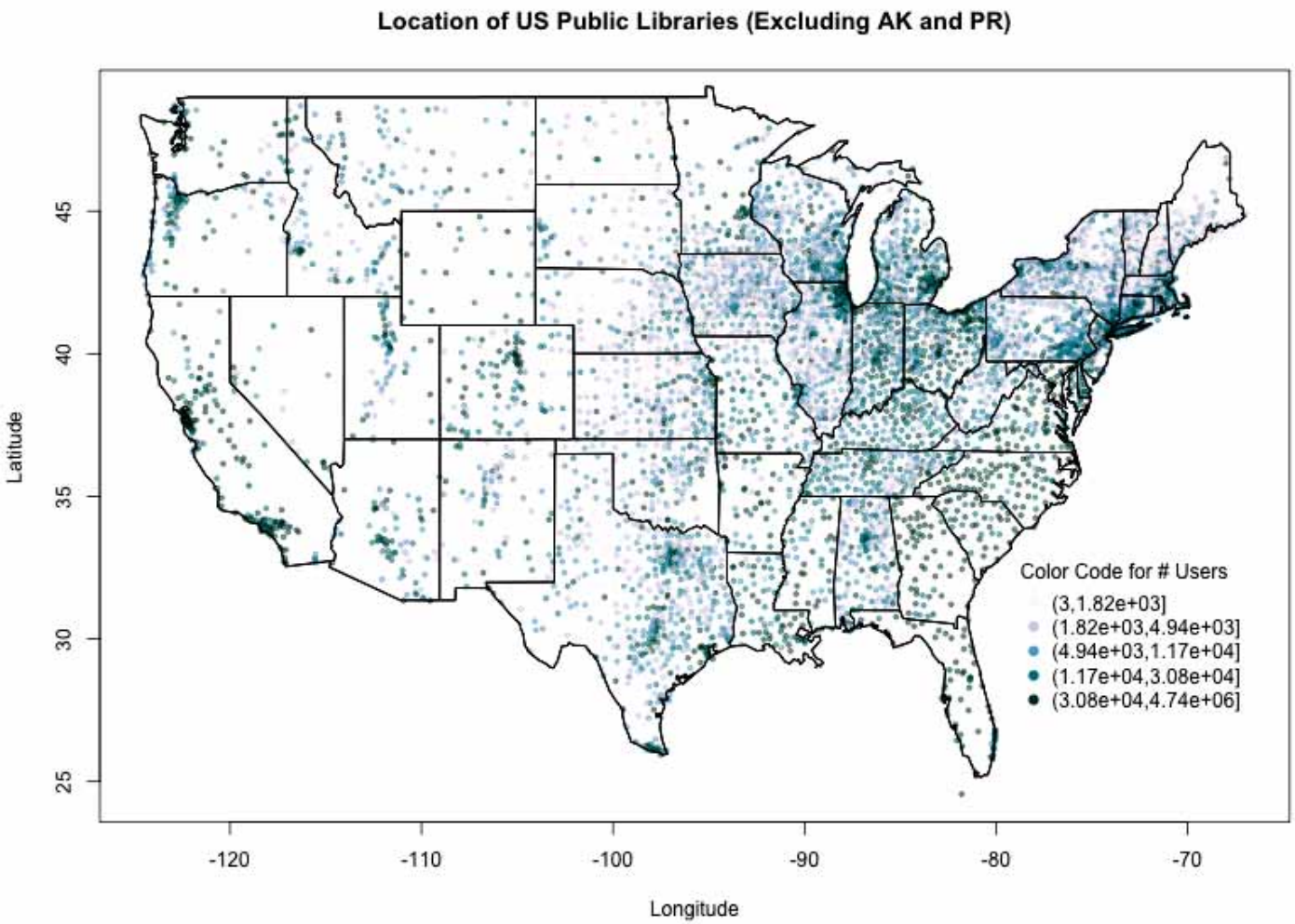


Figure 5: *This plot represents the location of libraries with points color coded by user levels. Light purple points represent low users and green points represent high users. The plot makes it clear that libraries with high user levels are near major cities/more urban US areas. This confirms the scatterplot analysis and supports the validity of the location indicator constructed that sorted libraries into the four types.*

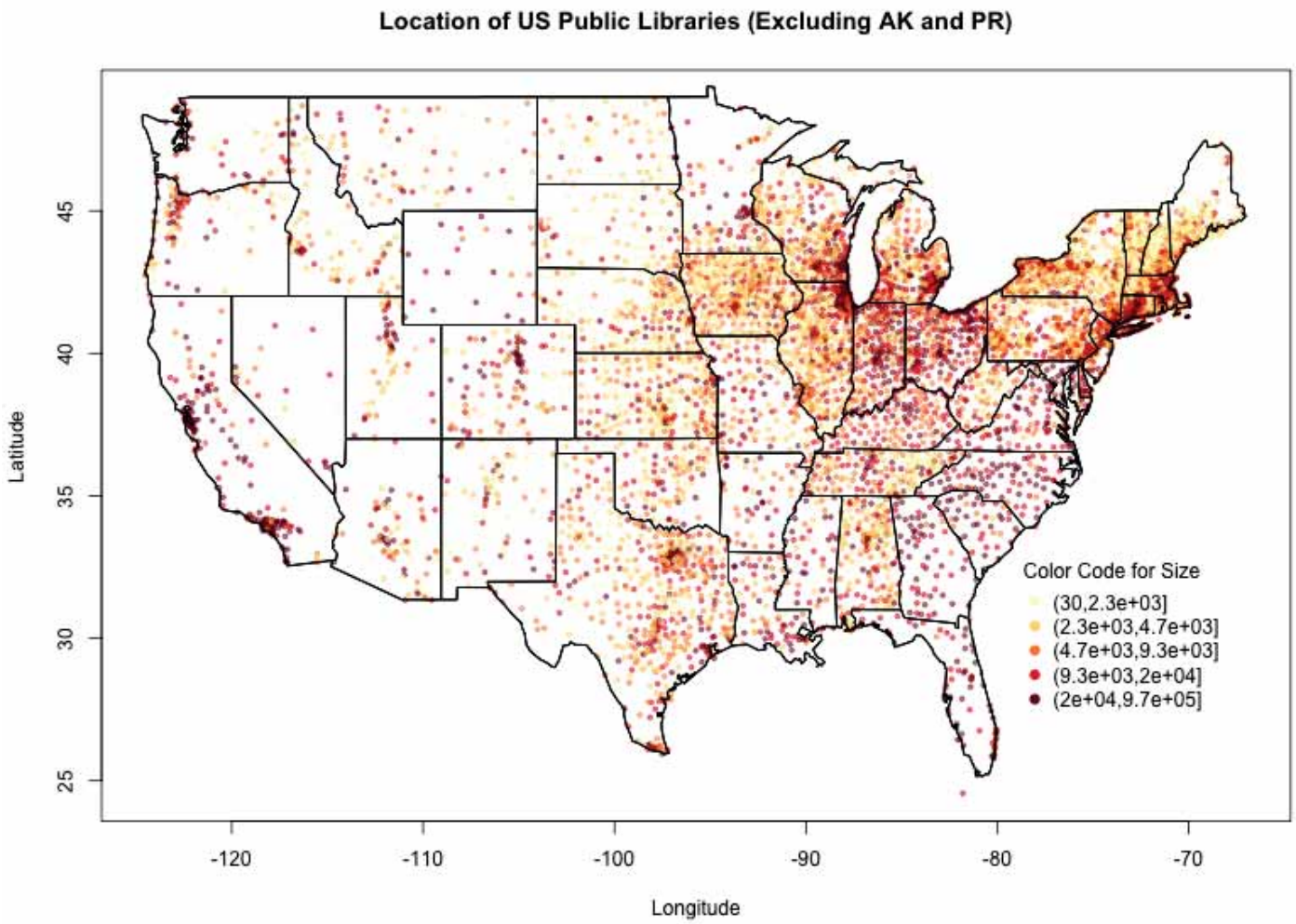


Figure 6: *This plot represents the location of libraries with points color coded by size levels. Light yellow points represent low users and dark red points represent large libraries. The plot makes it clear that large libraries are near major cities/more urban US areas. This confirms the scatterplot analysis and supports the validity of the location indicator constructed that sorted libraries into the four types.*

To formalize the cluster analysis a k-means clustering algorithm was carried out with four clusters expected based on the size and user variables. The plot below contains the results of the clustering assignments compared to the true labeling of city, rural, suburban, and town (see Figure 7). Shapes of points denote cluster assignment by the algorithm: square, triangle, circle, and diamond. Color denotes the actual type of a library: blue (city), suburban (yellow), red (town), green (rural). If there is a true clustering of library type by size and users, then the assignment by k-means should be similar to that by type. The results visually demonstrate that a majority of the rural libraries (green points) are classified into a single cluster by the algorithm, again supporting the idea that rural libraries are systematically different from the others. There does not appear to be a definitive pattern of town, suburb, and city across clusters.

However, further examination of the cluster composition makes the distributions clearer (see Figure 8). The breakdown of the amount of each library type across clusters is:

```
##                 1        2       3       4
##   CITY   0.40741 0.359788 0.06349 0.16931
##   RURAL  0.02465 0.007968 0.89766 0.06972
##   SUBURB 0.19891 0.072557 0.36476 0.36377
##   TOWN   0.09597 0.021790 0.51646 0.36579
```

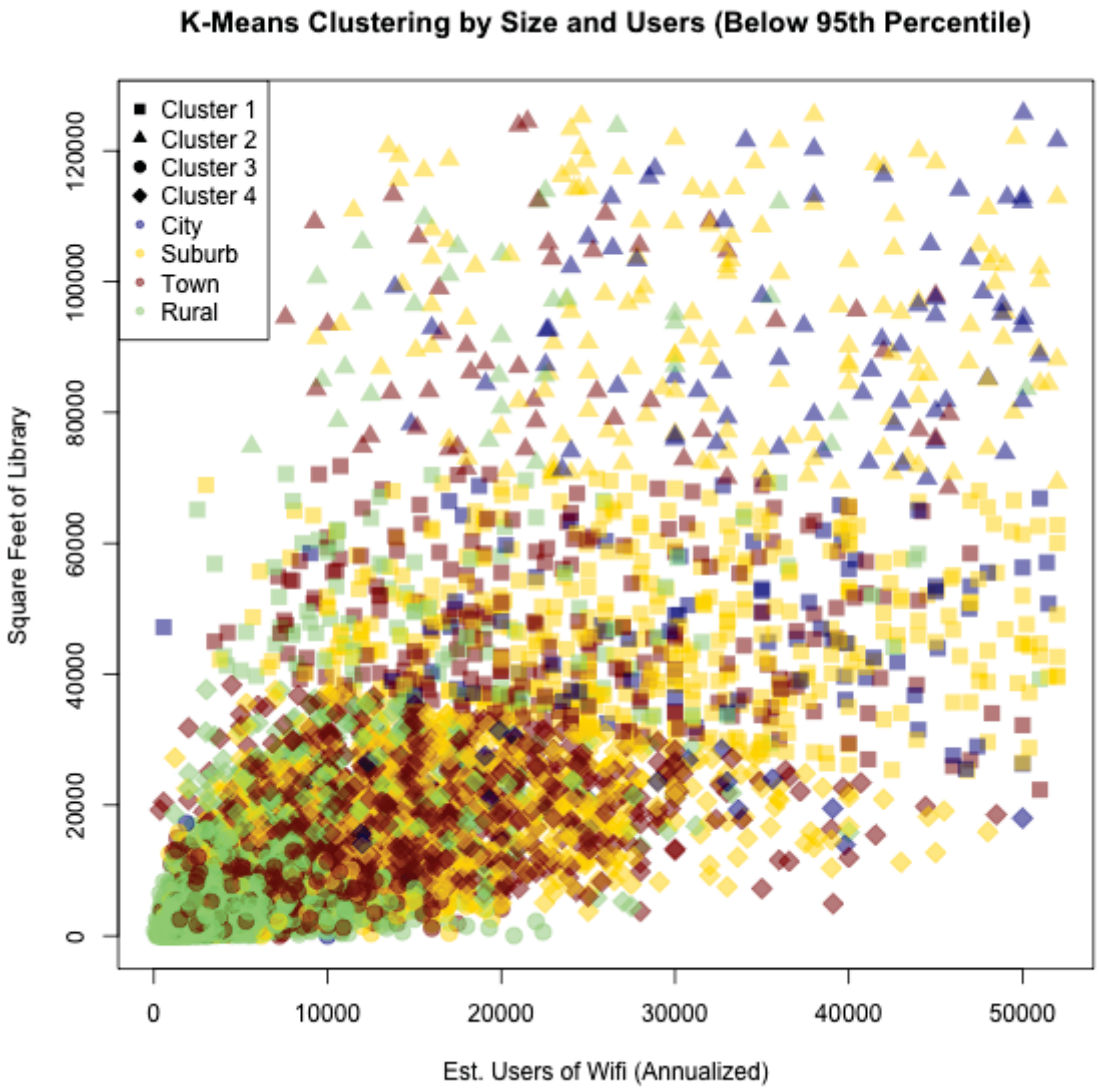


Figure 7: *This plot shows the results of k-means clustering with four clusters specified. Color represents true labeling of library type and shape denotes cluster assignment. There is some similarity between cluster assignment and library type, confirming the trends seen in the scatterplot and boxplot analyses. Rural libraries tilt lower on size and users followed by town, suburban, and city libraries. This is confirmed by Figure 8.*

The majority of rural libraries are in cluster 3 with the smallest size and user levels, which matches the findings of other analyses. A large percentage of town points are in clusters 3 and 4 and most suburb libraries are also in clusters 3 and 4. City libraries are mostly in the two largest clusters as well. Overall, the results here match the trend observed earlier where city libraries are higher in both size and users than the other types while rural and suburban are similar and rural libraries are smallest in both variables.

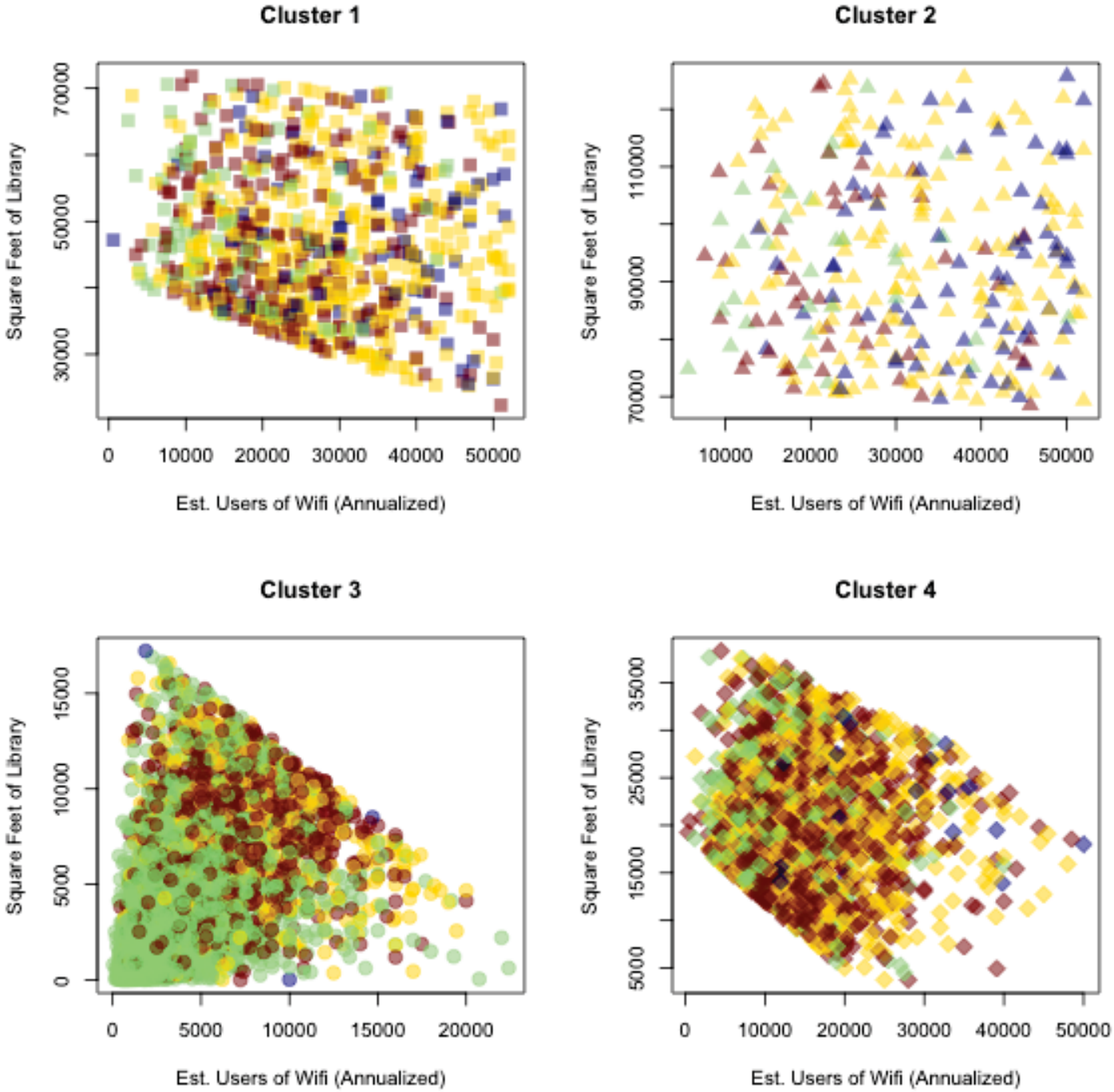


Figure 8: *These panels represent the four individual clusters from the k-means algorithm. It is clear that there is grouping of specific library types within clusters. The majority of rural libraries are in cluster 3. Suburban and town libraries are between clusters 3 and 4. City libraries are concentrated in clusters 1 and 2 (those farthest from the origin). Again this supports the conclusions of previous analysis and provides further evidence that library types are different from another (especially with respect to city libraries). See table above for cluster breakdown across library types.*

The cluster analysis also shows that the majority of the points appear more variable in the direction of wifi users than size. This raises a question of whether the total variation can be reduced to just the variation in users. One way to assess this is to see how well users can be predicted by size. As the geospatial plot showed, larger libraries tend to have more users, which are concentrated near big urban centers, so the two variables are correlated ($r = 0.7254$). Users should then be able to predict size of a library fairly well. The results of the regression are below.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	4401.3576	95.2695	46.20	0.0000
Users	0.3616	0.0037	96.50	0.0000

The coefficient of users is significant, and the $R^2 = 0.5262$. See the appendix for diagnostic plots. The fit could easily be increased using nonlinear regression techniques or adding polynomial terms as the scatterplot showed a somewhat nonlinear shape. The central point is that most of the variation of urban vs. rural is in users and that even the majority of the variation in size can be predicted by users, allowing the dimension of the data to be reduced for the average library. Therefore, a cap based on just on just users may be more appropriate as user levels capture a lot of the variation in the data and also can predict typical library size levels as well.

Now, since the analysis is supposed to involve a reimbursement for the cost of wifi service to libraries, it makes sense to examine the effect of these two variables and library type on the distribution of cost. Keep in mind though that the measure of cost used is a biased measure of the true cost for providing wifi.

Three separate regressions were run to see which combination of variables could best explain the variation in cost: (1) using size as the predictor, (2) using number of users as the predictor, and (3) using both as predictors. Each regression included indicator variables for library type and interaction terms between library type and the predictors. The third regression also included a coefficient for the interaction between size and users. These indicator variables and interaction terms were useful in examining whether the effect of size and/or users on cost was different for each library type. The results of each regression follow. See the appendix for diagnostic plots on each regression.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	119521.9511	33902.4048	3.53	0.0004
Rural	-119525.0607	34806.1425	-3.43	0.0006
Suburb	-123926.5902	34676.0654	-3.57	0.0004
Town	-109042.6567	34717.8543	-3.14	0.0017
Size (Sq Ft)	9.2404	1.0021	9.22	0.0000
RuralxSize	2.1495	1.2467	1.72	0.0847
SuburbxSize	3.7032	1.0584	3.50	0.0005
TownxSize	-1.4012	1.1144	-1.26	0.2087

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	105651.6216	26849.2260	3.93	0.0001
Rural	-79129.1017	27521.3310	-2.88	0.0041
Suburb	-59604.7525	27476.2547	-2.17	0.0301
Town	-75227.4181	27494.2599	-2.74	0.0062
Users	5.2162	0.4085	12.77	0.0000
RuralxUsers	-0.4346	0.4854	-0.90	0.3706
SuburbxUsers	1.2902	0.4373	2.95	0.0032
TownxUsers	-1.0284	0.4675	-2.20	0.0279

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-48515.7031	51804.3521	-0.94	0.3491
Rural	48521.1634	52709.0296	0.92	0.3573
Suburb	65768.7363	52660.9261	1.25	0.2118
Town	43533.7655	52781.0856	0.82	0.4095
Size	6.8270	1.7752	3.85	0.0001
Users	5.9527	1.0345	5.75	0.0000
RuralxSize	-1.4915	2.0947	-0.71	0.4765
SuburbxSize	-1.9286	1.8599	-1.04	0.2998
TownxSize	-2.2984	1.9316	-1.19	0.2342
RuralxUsers	-2.6131	1.1492	-2.27	0.0230
SuburbxUsers	-3.4140	1.0940	-3.12	0.0018
TownxUsers	-2.8861	1.1367	-2.54	0.0111
SizexUsers	-0.0001	0.0000	-1.73	0.0844
RuralxSizexUsers	0.0001	0.0000	1.64	0.1015
SuburbxSizexUsers	0.0001	0.0000	3.96	0.0001
TownxSizexUsers	0.0000	0.0000	1.31	0.1887

For the first two regression models, using users as the predictor is clearly better than size. The R^2 of the model with users is 0.4047 compared with 0.3596 for size, so users explains more of the variability in cost. Focusing on the better model, cost is increasing in users. The coefficients of rural, suburb, and town are significant, but the interaction terms suburbxusers and townxussers are as well. It appears that all libraries have different average costs than one another and that users affects costs of wifi differently for suburb and town libraries. In the model with size, suburbxsize is significant, but the other interaction terms are not. As expected from the cluster analysis, different library types are more similar in size to one another. It appears the association of size with cost is more consistent across library types. The model with both users and size explains the most of the variability in cost even after controlling for the additional predictors. Model 3 has an Adj. R^2 of 0.4554 compared with 0.4039 in model 2.

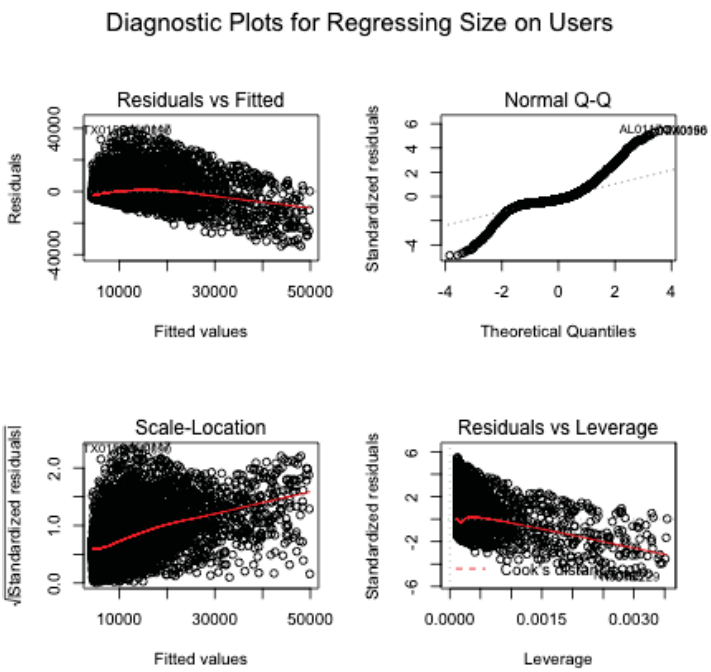
Within model 3, after adjusting for the effects of size and users, the average costs of different library types does not appear significant. This suggests that size and users capture most of the variation in cost of wifi among libraries. Increased users and size increases the observed cost. Interaction terms between type and user levels are significant, suggesting the effect of users on cost is different across library types like in model 2. Most of the interaction terms with size are not significant, suggesting the effect of size on cost is similar across library types. Even though model 3 has a higher Adj. R^2 than model 2, they are similar, and it appears that using just users has almost as much predictability of cost as using both size and users. Most of the coefficients from size in model 3 are not significant anyway, so the reduced model is very similar to model 2. This matches the results seen when predicting size from users and from the geospatial analysis where a lot of the variability was captured by users. It is worth noting, however, that the F-test for whether the two models are identical rejects the null hypothesis, suggesting that model 3 is indeed better than model 2 in terms of explaining cost.

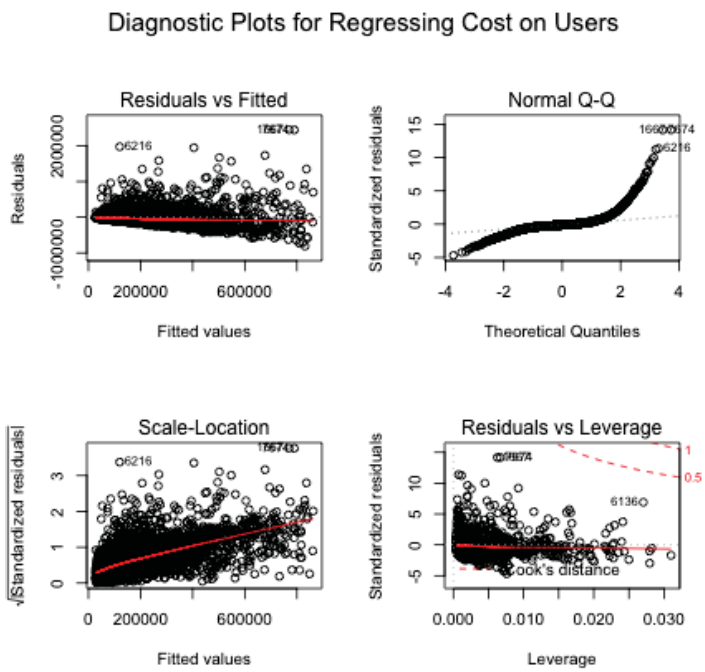
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	5201	142798619434310.84				
2	5209	156550977498249.22	-8	-13752358063938.38	62.61	0.0000

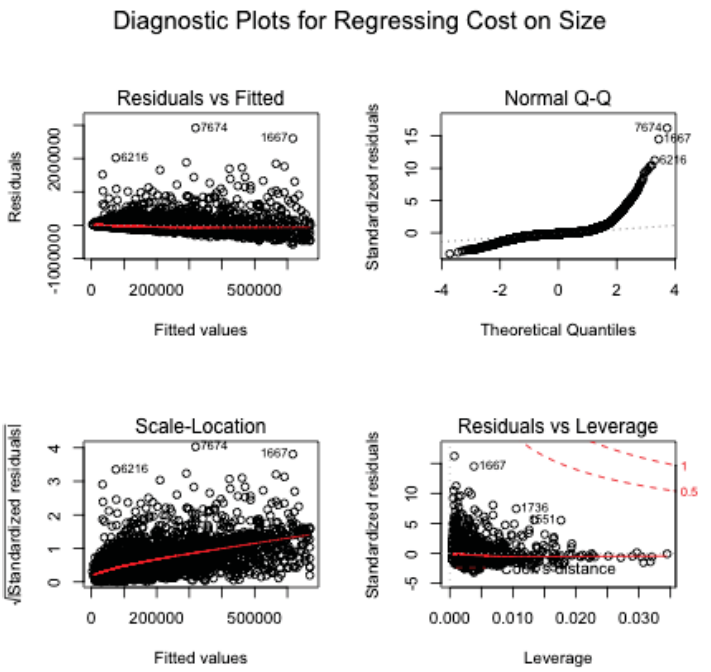
Overall, the results suggest that depending on the method of reimbursement used, it may differentially affect suburb, town, rural, and city libraries. Rural libraries are systematically smaller and have less users followed by town, suburb, and city libraries. City libraries have the highest variance across the two variables and constitute the majority of the outliers. For the majority of the data, there is more variation in user levels and much of the variation of size can be explained by user levels as well.

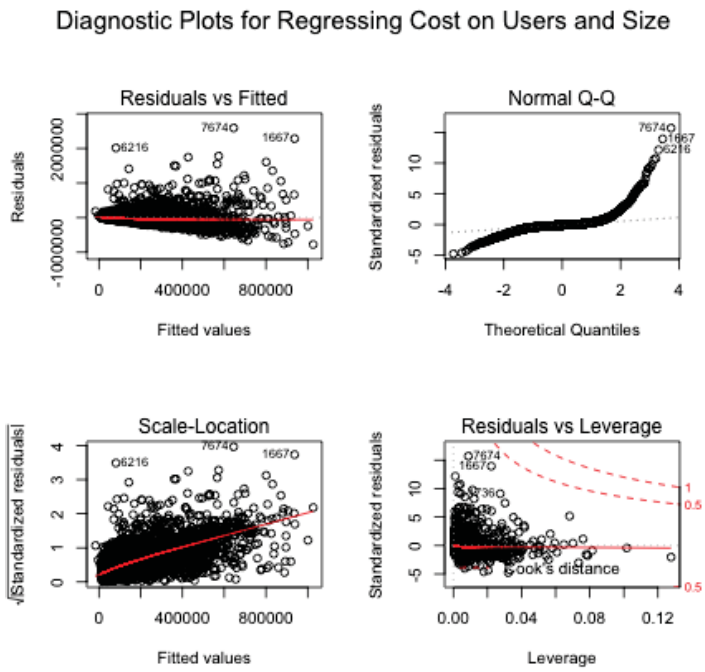
Predicting costs of providing the wifi service do not materially change the results as again user levels explain almost as much variability as both size and users and interaction terms with library type and user levels appear significant while those with size are not. Therefore, when deciding how to set the proposed cap, it is clear that the four library types will be affected differently by a cap based on just users or a combination of both size and users (mainly due to variation across user levels). In fact any proposal will need to account for the fact that city libraries are vastly different than the other three types. They probably need to be treated differently to get an optimal reimbursement policy. Therefore it is likely that two separate policies are needed with one for rural, town, and suburban libraries and a separate one for city libraries that constitute the majority of the extreme values. These could be based on user levels which capture most of the variation across library type and also explain a lot of the spread in size levels or be constructed using a combination of both size and user levels to best explain the variation in cost. Additional studies need to be done to obtain different estimates of the true cost of providing wifi service and examining whether the conclusions here hold when using other measures.

Appendix


Diagnostic Plots for Regressing Size on Users
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Diagnostic Plots for Regressing Cost on Users
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Diagnostic Plots for Regressing Cost on Size
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Diagnostic Plots for Regressing Cost on Users and Size
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage