

**Examining the Digital Divide in the Pacific Northwest: An Empirical Look at Internet
Quality on the County Level**

Matthew Sandell

Data Analytics Senior Capstone Fall Quarter, Eastern Washington University

DSCI 481: ML-Data Science Fundamentals

Dr. Alexandros Paparas

December 11, 2024

Introduction

Internet access is extremely important and influential both economically and socially. Increasing availability and quality while reducing cost positively impacts local and state economies while connecting individuals with the myriad of resources the internet has to offer. While affordable, quality internet is crucial, not all regions perform the same. To learn more about these disparities, research can be conducted to determine what factors into quality internet, and how local and state officials can encourage improvement in this area.

The scope of the project is restricted to the northwestern states of the continental United States – Washington, Oregon, and Idaho. The level of granularity is dictated by the availability of quality, reputable statistics relating to internet usage. Due to these constraints, data for this project is collected on the county level, as this balances accessibility with local relevance. By combining several different sources from government resources, a reliable and useful dataset can be created with metrics for all 119 counties in the three states studied.

By addressing aspects such as quality, availability, and digital literacy, we can create a singular index variable for each county. This synthesized index variable reflects the current performance of each county and provides vital information on current infrastructure levels. This information is useful at the county level, to help local administrators better serve their respective communities. It can also help at the state level – if the data indicates that a particular state is lagging in providing quality internet to its population, state officials can invest in more effective broad infrastructure methods to improve the quality of life of individuals in the state. This benefits everyone from the individual and family level all the way up to the national level, ranging from better job opportunities and increased social connectivity to a more robust and productive economy. Additionally, the index variable has uses for individuals as well – internet quality and access may be an important factor when determining a region to move to, especially for those who work primarily remotely/online.

Due to the nature of creating an index variable, there are inherent obstacles that must be overcome prior to and during the analysis. Chiefly, the dimensionality of the data poses an issue. Over a dozen initial variables must be reduced and weighted into a single variable. To achieve an accurate weighting based on the level of impact each variable has, we turn to machine learning methods such as factor analysis and principal component analysis.

Multicollinearity also poses a problem. Because variables are collated from several disparate sources, there is the possibility of highly correlated variables leading to multicollinearity and compromising the integrity of the data. Correlation tests must be conducted and adjustments made to the dataset if variables are highly correlated. Correlation tests work in tandem with sampling adequacy tests to identify high correlations before the weighting of the variables and creation of the index.

Because the index was created mainly to support policymakers and decision-making at the county and state levels, the analysis needs to generate interpretable insights regarding regional differences in performances, not just rankings of counties. We use the index data to see how regions are performing, comparing descriptors such as urban and rural. Geographic trends are an important part of identifying how infrastructure can be improved. Some variables impact given counties and/or regions more heavily than others. When paired with clustering methods such as k-means and hierarchical, the index data can show not just how well a county’s internet is, but also the reasons behind its performance. Underserved counties can be identified on each variable, highlighting where improvement can be made, ranging from increased funding to more efficient infrastructure.

The Data

Before the index variable itself is created, we must identify, locate, and combine relevant data to form a single initial dataset. Government websites provided the entirety of the dataset, split between federal sites such as the Census Bureau and state sites such as the Washington Secretary of State. Because of this, we can assume the integrity and veracity of the data.

| Variable | Measurement Method | Source |
|-----------------------|--|---|
| County | Text-based identifier of county | All sources contain county identifier |
| State | Text-based identifier of state | All sources contain state identifier |
| Desktop_or_Laptop | Percentage (0%-100%) of county population with a desktop and/or laptop | U.S. Census Bureau |
| Smartphone | Percentage (0%-100%) of county population with a smartphone | U.S. Census Bureau |
| Tablet_or_Other | Percentage (0%-100%) of county population with a tablet or other device | U.S. Census Bureau |
| Per_Capita_Income | Numeric value in the thousands representing per capita personal income in the given county | Bureau of Economic Analysis |
| Less_than_High_School | Percentage (0%-100%) of county population aged 18 and older with any level of schooling less than graduating high school | U.S. Census Bureau |
| High_School_Graduate | Percentage (0%-100%) of county population aged 18 and older with high school diploma/equivalency as highest level of education | U.S. Census Bureau |
| Bachelor_Up | Percentage (0%-100%) of county population aged 18 and older with a bachelor's degree or higher | U.S. Census Bureau |
| Provider_Count | Numeric count of unique internet providers in each county | Federal Communications Commission |
| Branches_per_1000 | Numeric average of public libraries per 1000 county residents | Washington Secretary of State, Idaho Commission for Libraries, Oregon State Library, U.S. Census Bureau |
| Download_Speed | Percentage (0% - 100%) of county population with access to highest available download speed | Federal Communications Commission |
| Upload_Speed | Percentage (0% - 100%) of county population with access to highest available upload speed | Federal Communications Commission |
| Rural_Access | Percentage (0% - 100%) of county population in rural census blocks with broadband internet access | Federal Communications Commission |
| Broadband_Access | Percentage (0% - 100%) of county population with access to broadband internet | Federal Communications Commission |
| Worked_from_Home | Percentage (0% - 100%) of county labor force that worked from home | U.S. Census Bureau |

Variables were chosen for their relevance to internet access in several ways:

Device Ownership: Desktop or Laptop, Smartphone, and Tablet or Other.

Education: Less Than High School, High School Graduate, and Bachelor Up.

Internet Speed: Download Speed and Upload Speed.

Internet Access: Rural Access and Broadband Access.

Wealth and Affordability: Per Capita Income and Provider Count are indicators of economic conditions, with more providers typically associated with lower prices, and wealthier regions able to afford better internet service.

Public Internet Use: Branches Per 1000 indicates the reliance on public internet.

Work From Home Rates: Work from Home rates were chosen as a signal of internet quality, as regions with bad internet tend to have lower work from home rates.

By having each variable relate to at least one of these topics, we ensure that the dataset is comprehensive in its treatment of what factors influence general internet quality.

Most variables came cleaned and ready directly from the source. Others required some preprocessing. Some education variables needed to be aggregated, as the initial data was split into smaller pieces. Because of this, these variables are combined data of several variables in the original Census Bureau dataset.

Additionally, the Branches Per 1000 variable is a composite of two datasets – population of each county and number of public libraries. The resulting variable is the number of public libraries per 1,000 people in each county.

If outliers were present, we would run the analysis with two datasets – one with the outliers and one without. This would allow us to see which model is more accurate by controlling for influential values. Because none of the variables used had any missing values or outliers that needed to be corrected, instead the next step was to join the dataset together into a united dataset with each row representing a county and each column representing one of the 14 variables used.

Note that initially an additional education variable was used, but due to high correlations between two variables revealed by a Bartlett's Correlation test, it was dropped. Doing this shifted our measure of sampling adequacy from 0.5 to 0.69, calculated using a Kaiser-Meyer-Olkin (KMO) test, well above the threshold of 0.6 required to assume adequate sampling.

Methodology

Before an analysis of the counties can be performed, the variables must be weighted to create the index variable itself. Because this project requires dimension reductionality – reducing the data from 14 variables to 1, we have two principal options: factor analysis (FA) and principal component analysis (PCA) are both viable options for

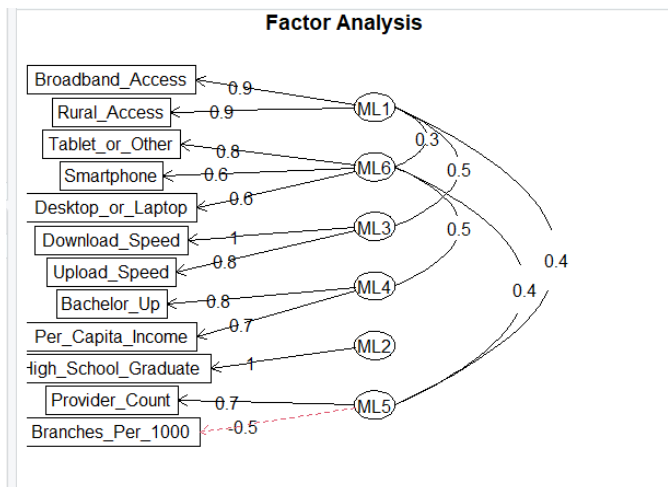
creating the index variable. While separate processes, both methods group similar variables and weight them accordingly.

We choose to conduct both methods, and then compare the results to see which is better, based on how interpretable the results are in the context of the data.

Factor analysis was conducted by using chi square and Mahalanobi's Distance to find outlying observations. We run the FA process for the whole dataset and the subset excluding outliers. This controls for undue influence that the outliers potentially exert on the overall dataset.

Bartlett's Correlation and KMO tests were used to remove highly correlated variables and test for sampling adequacy. Once we have adequate data, we use a scree plot and parallel test to determine the optimal number of factors to group the variables into. Because the numbers are different for each method, we tested the model using 2, 3, 4, 5, and 6 factors.

An acceptable model occurs when we have a Tucker Lewis value above 0.9 and RMSEA and RMSR values below 0.08. For each factor count, we drop variables that load highly into 2 factors and rerun the model to find the optimal loadings for each number of factors. For both the full dataset and excluding outliers, loadings are muddled and unintuitive. Some factors have only one variable, and the models require dropping too many variables from the data.



The best distribution of variables is shown above, sorted into 6 factors. While statistically acceptable, because we cannot find a balanced, interpretable model for the data using factor analysis, we conclude that this method is not ideal for our specific dataset and situation. Instead, we turn to PCA for creation of the index variable.

There were two forms of principal component analysis (PCA) that we used with this dataset. First, we conducted robust sparse PCA, which forces clearer loadings than standard PCA. Interpreting the results of sparse PCA relates to principal component scores. These are found by running the standardized data to find loadings of each variable on each principal component (PC). The signs of some loadings are swapped if it results in more interpretable loadings. To confirm the model adequately fits the data, the percentage of variance explained (PVE) is utilized. A PVE value above 0.7 is ideal, while higher values are more desirable. The final score for each observation is calculated by using the eigenvalues and principal component scores for each loading to create a singular number for each observation. In this case, each county has its own numeric score. This process is done three separate times. Each iteration uses a different number of PCs – first 4, then 5, and finally 6. Essentially, this changes the distribution of how variables are grouped together. The goal is to find a model with clear, interpretable loadings and a sufficiently high PVE.

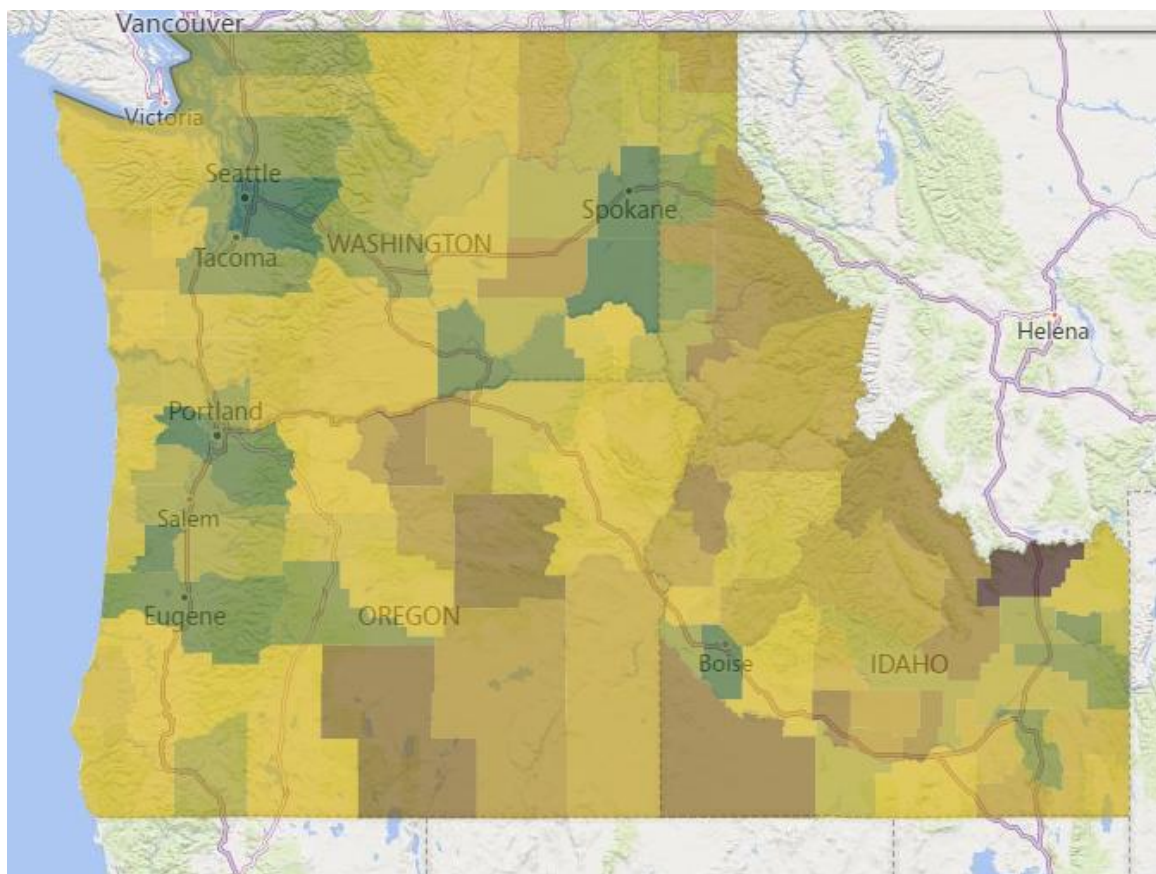
The results of sparse PCA for this data have a PVE of 0.766, indicating the model adequately explains variance in the dataset. Below are the variable loadings into each of the 4 principal components.

| | PC1 | PC2 | PC3 | PC4 |
|-----------------------|--------------|--------------|----------|-----------|
| Desktop_or_Laptop | 0.009912724 | 0.159112192 | 0.000000 | 0.000000 |
| Smartphone | 0.603524123 | -0.021575624 | 0.000000 | 0.000000 |
| Tablet_or_Other | 0.137101385 | 0.000000000 | 0.000000 | 0.000000 |
| Per_Capita_Income | 0.000000000 | 0.000000000 | 0.000000 | 0.000000 |
| Less_than_High_School | -0.006235886 | -0.627252908 | 0.000000 | 0.000000 |
| High_School_Graduate | 0.000000000 | 0.000000000 | 1.001544 | 0.000000 |
| Bachelor_Up | 0.048201373 | 0.744612147 | 0.000000 | 0.000000 |
| Provider_Count | 0.773771527 | -0.031777812 | 0.000000 | 0.000000 |
| Branches_Per_1000 | 0.000000000 | 0.000000000 | 0.000000 | 0.000000 |
| Download_Speed | 0.000000000 | 0.000000000 | 0.000000 | 0.6330421 |
| Upload_Speed | 0.000000000 | 0.000000000 | 0.000000 | 0.7687604 |
| Rural_Access | 0.000000000 | 0.000000000 | 0.000000 | 0.000000 |
| Broadband_Access | 0.000000000 | 0.000000000 | 0.000000 | 0.000000 |
| Worked_from_Home | -0.033782680 | 0.002908676 | 0.000000 | 0.000000 |

Note that as we increase principal components, PVE always increases. However, using 5 or 6 principal components only marginally changes the loadings. Because the change is minimal, we err on the side of as few PCs as can accurately be used. The data indicates that the optimal number of components is 4. Also take note that certain variables are not plotted into any principal component. We conclude that these variables are not that important to the creation of the index variable.

Additionally, High School Graduate is alone in PC3. We conclude that this variable is sufficiently important on its own to warrant a separate principal component. Overall, the loadings make sense – similar variables are grouped and correlated as expected. For example, Download and Upload Speed are both together in PC4, while PC2 contains a negative loading for Less Than High School and a positive loading for Bachelor and Up. Under the assumption that higher levels of education and quality internet are correlated, you would expect to see that poor education has a negative relationship and great education has a positive relationship with the final index score. Because the results are statistically adequate with a PVE of 0.766 and are largely interpretable and intuitive, we can safely conclude that robust sparse principal component analysis is a valid option for weighting the given variables.

As a visual confirmation, we put the index variable scores into PowerBI to generate a gradient map showing internet across the geographical regions of Washington, Idaho, and Oregon states.



Here we see high-performing counties in dark green and poorly performing counties in dark yellow/brown. Light yellow or green indicates mediocre or average internet. Dark green counties are largely metropolitan or highly populated areas. The worst-performing

counties are mostly extremely rural/underdeveloped counties. Note also the radiating effect, noticeable around Portland and Seattle. Counties surrounding these large cities also perform well; the performances ebb worse as they move farther from population cores and toward rural areas. The Eugene and Spokane areas are also examples of that.

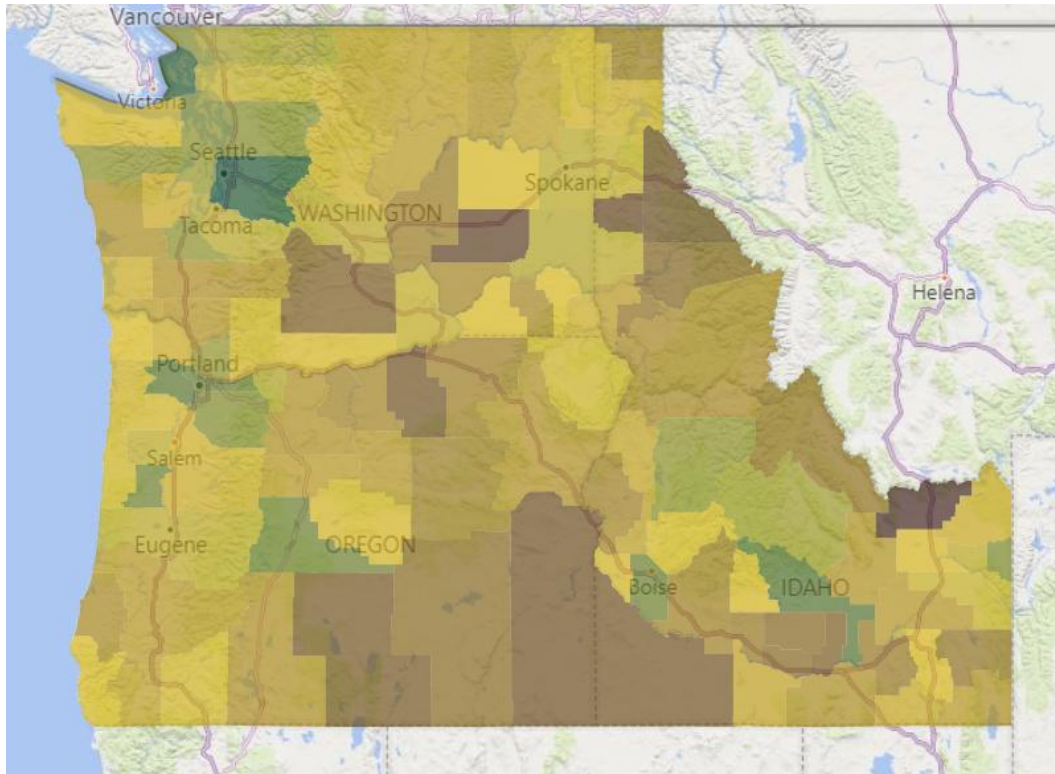
We also conduct standard PCA, which is more flexible in how the loadings are computed. Sparsity is not enforced in loadings, meaning variables can load more heavily across several PCs and there is largely a lack of perfect 0.00 loadings.

The process begins the same way as robust sparse, calculating covariance, eigenvalues, and eigenvectors. These values are used to create the loadings. However, the initial loadings apply to a number of principal components equal to initial variables, in this case 14. To apply the dimension reductionality aspect of PCA and find the ideal number of principal components to use, we calculate PVE and create a scree plot. The result indicates that 5 is the ideal number, so we start there. We convert just the first 5 principal components into matrices. This allows us to multiply each principal component by its accompanying PVE. The resulting sum is the total score for each observation.

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|-----------------------|-------------|-------------|-------------|-------------|-------------|
| Desktop_or_Laptop | 0.36794555 | -0.28348761 | 0.10318540 | 0.02321241 | 0.11310830 |
| Smartphone | 0.35461913 | 0.08832078 | 0.05724252 | -0.29150688 | -0.05245919 |
| Tablet_or_Other | 0.34748860 | -0.15793292 | 0.20854061 | -0.09187554 | -0.05938998 |
| Per_Capita_Income | 0.21156058 | -0.12345880 | -0.48077711 | -0.13001231 | -0.25612052 |
| Less_than_High_School | -0.20651596 | 0.42146086 | -0.17610919 | -0.27593110 | -0.06691200 |
| High_School_Graduate | -0.03681305 | -0.22876587 | 0.56590351 | 0.30044969 | 0.14814663 |
| Bachelor_Up | 0.35401931 | -0.21873140 | -0.29666965 | 0.08704424 | -0.08041741 |
| Provider_Count | 0.26269129 | 0.15707010 | 0.19178570 | -0.22929984 | -0.33849839 |
| Branches_Per_1000 | -0.27909636 | -0.06706839 | -0.33200950 | 0.25713549 | 0.22800900 |
| Download_Speed | 0.22176990 | 0.33472766 | -0.01085982 | 0.46749644 | -0.24391386 |
| Upload_Speed | 0.12675604 | 0.32719090 | -0.01304425 | 0.56957907 | -0.29733319 |
| Rural_Access | 0.25184856 | 0.35030979 | -0.06225828 | 0.00130708 | 0.58772769 |
| Broadband_Access | 0.29829945 | 0.35887946 | 0.02060308 | -0.03998067 | 0.41668202 |
| Worked_From_Home | 0.19864717 | -0.30670138 | -0.34722529 | 0.22778376 | 0.22117878 |

The loadings indicate that device ownership variables are grouped together. Additionally, Rural and Broadband Access are together, as are Download and Upload speeds. Education rates are also somewhat together. This satisfies expectations that “like terms” go together and make sense in the context of the data.

Because the results are interpretable and insightful, we again turn to PowerBI to visually show the distribution of scores and confirm our results.



Another map is generated, using the standard PCA scores instead of robust sparse. Again, dark green counties represent great internet, with dark brown counties having the worst internet.

Both forms of PCA give similar results and maps. However, the results for sparse PCA tend to be more accurate in terms of expectations – it's presumed that metropolitan areas/highly-populated regions would have better internet. Both graphs support this, but standard PCA is less accurate in that it misses certain cities and includes rural regions that score highly in one variable. An example is the heavily rural Blaine County, Idaho, which ranks as the richest per capita income county in the dataset, heavily skewing its final ranking in standard PCA. Instead, sparse PCA takes this into account but is more balanced in its value assignments. This is because sparse PCA is better at handling outliers and large ranges of values.

Comparison with factor analysis is largely inappropriate/unnecessary, as factor analysis has been shown not to work well for this dataset. The loadings for sparse PCA are more easily interpreted than in standard PCA and sparse PCA handles outlying values better. The PowerBI maps show sparse to be more accurate in fitting the values to expectations as compared to standard PCA. Because of this, we chose to continue with the robust sparse scores.

Now the first goal of the project has been achieved – the 14 original variables have been weighted and combined to form a singular index score that represents the generalized performance of each county’s internet, accounting for how influential each variable is.

Next, we utilize cluster analysis to gain more actionable insights into what these scores represent, and group counties based on shared characteristics. Cluster analysis is useful because it groups similar counties together. Instead of evaluating 119 counties directly, we can instead interpret three to five clusters of counties, each one composed of counties sufficiently close to each other in variables associated with that cluster. Each county is assigned to its highest cluster, and we interpret what this means in the context of the dataset.

To create the clusters, we choose between two methods – k-means and hierarchical. K-means analysis requires the number of clusters before the analysis, while hierarchical helps us find the best number of clusters. Because of this and other computational differences, we utilize both methods and choose the one that works best for the data.

The clustering method that gave the best results was k-means. We use elbow, silhouette, and gap statistic tests to indicate the optimal number of clusters to use. Each method gave a different number, so we compare the loadings for each number of clusters to determine the most interpretable/intuitive number of loadings. Between 2, 3, and 5 clusters, we find 3 clusters to be the most interpretable and intuitive.

| Cluster | Desktop_or_Laptop | Smartphone | Tablet_or_Other | Per_Capita_Income | Less_than_High_School | High_School_Graduate | Bachelor_Up | Provider_Count | Branches_Per_1000 | Download_Speed | Upload_Speed | Rural_Access | Broadband_Access | Worked_from_Home |
|---------|-------------------|--------------|-----------------|-------------------|-----------------------|----------------------|--------------|----------------|-------------------|----------------|--------------|--------------|------------------|------------------|
| 1 | -0.46695643 | -0.141437844 | -0.395105272 | -0.350207553 | 0.487068674 | -0.149260025 | -0.546039037 | -0.049557726 | -0.058086358 | 0.235587723 | 0.313869889 | 0.383355845 | 0.400296887 | -0.471317211 |
| 2 | 0.979666636 | 0.910951529 | 0.933366511 | 0.592584096 | -0.629082325 | -0.032280743 | 1.050695993 | 0.603781694 | -0.530381292 | 0.389007515 | 0.090281819 | 0.383725069 | 0.503148094 | 0.578576135 |
| 3 | -0.300955537 | -1.109921128 | -0.586620159 | -0.153974574 | -0.090182655 | 0.378730015 | -0.437799867 | -0.832872444 | 0.955184803 | -1.125144713 | -0.631353392 | -1.441993968 | -1.665564176 | 0.134019095 |

Shown above are the k-means loadings for 3 clusters. We exported the k-means results for each variable by cluster into Excel, where conditional formatting is applied to highlight positive and negative values as a visual aid. The higher number loading for each variable means that cluster performs better at the given variable.

Cluster 1 performs well at Less Than High School. We can conclude this cluster represents counties with relatively poor education, because this is a negative variable – higher rates indicate more people not completing high school.

Cluster 2 performs well at device ownership variables, Income, Bachelor and Up, and Provider Count. This cluster represents wealthy and educated counties with many providers.

Cluster 3 appears to perform well at High School Graduate and Library Branches. However, when sorting the k-means data by High School Graduate rates, we do not see

cluster 3 performing well at the top, so we ignore this variable in our analysis. This cluster represents counties that have high reliance on public internet.

Note that Download Speed, Upload Speed, Rural Access, and Broadband Access perform well in both clusters 1 and 2. Because we can't distinguish these clusters based on these variables, we ignore them for high performances. Additionally, the same thing happens in clusters 2 and 3 for Work From Home rates. Judging by the number of "positive" variable performances, cluster 2 appears to be the most well-rounded cluster in terms of overall internet quality. Clusters 1 and 3 appear to perform the best at the fewest variables and are associated with negative variables.

Here, we see somewhat uneven distributions of counties into clusters:

Cluster 1 contains 55 of the 119 observations.

Cluster 2 has 39 counties.

Cluster 3 has the remaining 25 observations.

This would indicate that a large portion of counties have mediocre internet, in the sense that they are grouped in the middle cluster for general performance. Cluster 2 contains the "best" counties, notably many of the metropolitan areas. Cluster 1 represents about half the counties, those with relatively poor education. Cluster 3 covers counties with high reliance on public internet.

The 3-cluster analysis above was chosen over other cluster counts because running 2 clusters results in binary clustering. This means positive variables are correlated together in one cluster and negative variables are correlated in the other. While this can show which counties are good and bad in a general sense, it does not reveal any additional useful information. Having 5 clusters spreads the variables too thin, resulting in many variables loading highly onto numerous clusters or no clusters at all. The clusters that contain variables are muddled and less intuitive compared to 2 or 3 cluster groupings.

Because these clusters fit the data poorly, we can conclude that using two or five clusters is not optimal. This is because the goal of cluster analysis and k-means is to group observations based on shared characteristics. Five clusters group observations poorly. Using two clusters works somewhat well but is too vague in that it creates only two types of counties – "good" and "bad". We therefore conclude that in terms of k-means analysis, using three clusters is optimal, as we can withdraw intuitive and interpretable insights from the groupings. It is also useful to have three categories, as this covers the range of possible evaluations without splitting the observations into overly-specific clusters.

Next, we conduct another round of k-means, this time using the principal components as latent variables – this creates different variable values to cluster. If the numbering is more consistent on each variable than k-means of the original variables, we can look more into the observations in each PC cluster and learn more, but first we must determine the best clustering method.

To that end, we conduct hierarchical clustering. There are two forms of hierarchical clustering – agglomerative and divisive. We chose to use the agglomerative, or “bottom-up” approach because it runs four separate clustering methods and allows us to pick the best one, whereas divisive runs only one method.

As with k-means, the first step is determining the optimal number of clusters. To do this we run three tests: within-cluster sum of squares (WSS), silhouette, and gap stat. We used 3 clusters because each test agreed on that number. Agglomerative clustering gives us 4 coefficients, one for each of the four clustering methods. We used the Ward method because it had the highest coefficient.

We broke the observations into three clusters: then we’re essentially conducting k-means again, albeit with a different method of forming the clusters. We calculate the k-means of the cluster/variable pairings and export them into Excel to check if they are somewhat intuitive.

This process is repeated by once again using the principal components from robust sparse PCA as latent variables to create differing cluster groups.

At this point, we have four different columns, each listing cluster numbers for each county in the dataset. Cluster numbers for seven counties in the dataset are shown below.

| Cluster K-Means PC | Cluster Hier PC | Cluster K-Means Var | Cluster Hier Var |
|--------------------|-----------------|---------------------|------------------|
| 3 | 1 | 2 | 1 |
| 1 | 2 | 3 | 2 |
| 4 | 1 | 2 | 1 |
| 2 | 1 | 3 | 2 |
| 2 | 1 | 3 | 2 |
| 3 | 1 | 1 | 3 |
| 1 | 3 | 2 | 1 |

Because each method clusters in a different manner, we must determine which method is optimal – that is, that clusters the most accurately. To do this, we check visually by sorting each variable high-to-low. This shows us whether each method is consistent in how it clusters – counties that are good in each variable should be in the same cluster, if that method is valid. For our data, it appears that k-means is a more consistent clustering

method compared to hierarchical, but it is unclear whether the original variables or the PCs as latent variables cluster better. To find out, we conducted three tests for quality of clustering: Davies-Bouldin, Dunn Index, and silhouette.

K-means test results for initial variables are shown on the left and PCs as latent variables are shown on the right:

```
> print(davies_bouldin_km_var) > print(davies_bouldin_km_pc)
$davies_bouldin                $davies_bouldin
[1] 2.7915                      [1] 9.578152

> print(dunn_index_km)         > print(dunn_index_pc_km)
$dunn                          $dunn
[1] 4.261814e-05              [1] 4.81207e-05

> print(mean_silhouette_km)   > print(mean_silhouette_km_pc)
[1] -0.01021946              [1] -0.117919
```

Davies-Bouldin measures cluster similarity. A lower score is ideal, as this indicates that clusters are not similar. The above data indicates that the original variables cluster better by Davies-Bouldin. The Dunn Index measures separation between clusters as well as compactness. A higher score is preferable, which we see here for PC clustering. The silhouette score also measures compactness and separation of clusters. A higher score is ideal, seen here for the original variables.

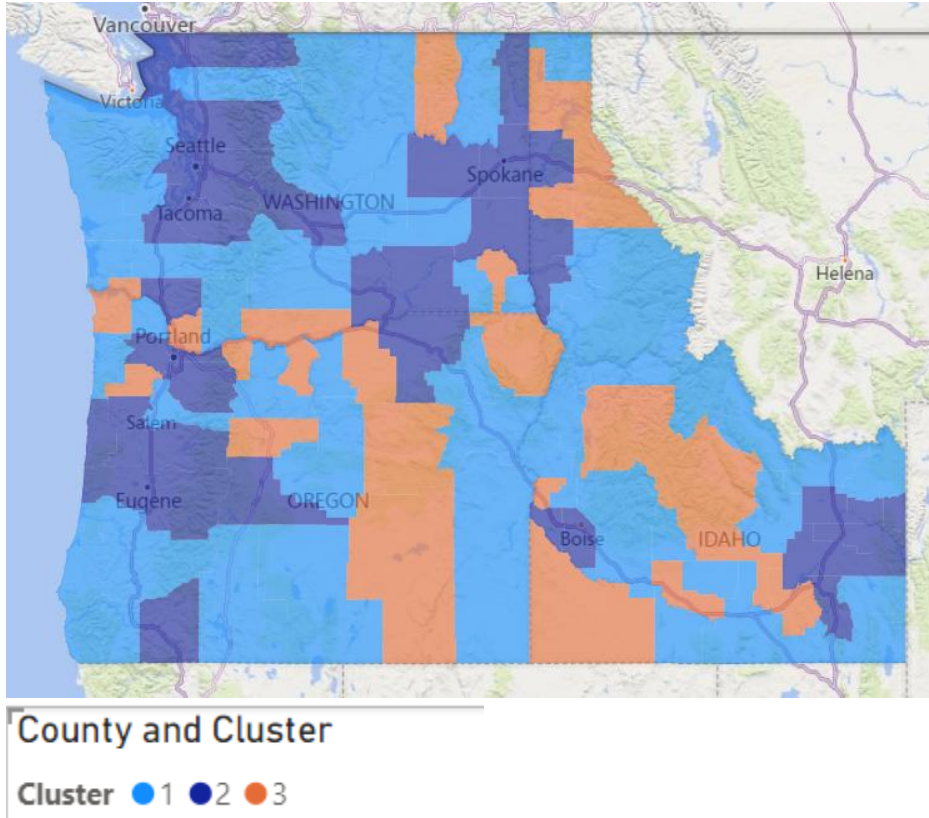
The majority of the methods favor using the original variables for k-means over the PCs. We conclude that this is the optimal method of clustering the dataset.

Results

Below, each county is listed with its respective index score, sorted high to low.

| County | State | Index Score | County | State | Index Score | County | State | Index Score |
|-------------|-------|-------------|--------------|-------|-------------|--------------|-------|-------------|
| King | WA | 1.5122 | Skagit | WA | 0.2616 | Yakima | WA | -0.2726 |
| Washington | OR | 1.3258 | Franklin | WA | 0.2524 | Wahkiakum | WA | -0.2764 |
| Whitman | WA | 1.2733 | Twin Falls | ID | 0.2457 | Skamania | WA | -0.2985 |
| Spokane | WA | 1.2347 | Bonner | ID | 0.2416 | Coos | OR | -0.3192 |
| Ada | ID | 1.1934 | Clatsop | OR | 0.1922 | Grays Harbor | WA | -0.3363 |
| Benton | OR | 1.1616 | Bingham | ID | 0.1830 | Power | ID | -0.3708 |
| Clackamas | OR | 1.0661 | Douglas | WA | 0.1779 | Franklin | ID | -0.3969 |
| Multnomah | OR | 0.9947 | Pend Oreille | WA | 0.1722 | Camas | ID | -0.3990 |
| Benton | WA | 0.9882 | Columbia | OR | 0.1401 | Caribou | ID | -0.4385 |
| Lane | OR | 0.9600 | Lincoln | OR | 0.1065 | Malheur | OR | -0.4871 |
| Madison | ID | 0.9189 | Umatilla | OR | 0.0971 | Boise | ID | -0.5154 |
| Kootenai | ID | 0.8326 | Cowlitz | WA | 0.0339 | Lincoln | ID | -0.5425 |
| Snohomish | WA | 0.8209 | Jefferson | OR | 0.0329 | Oneida | ID | -0.5689 |
| Pierce | WA | 0.7995 | Valley | ID | 0.0058 | Washington | ID | -0.5719 |
| Latah | ID | 0.7990 | Grant | WA | 0.0050 | Ferry | WA | -0.5754 |
| Walla Walla | WA | 0.7909 | Fremont | ID | -0.0034 | Idaho | ID | -0.5838 |
| Bonneville | ID | 0.7621 | Jefferson | WA | -0.0083 | Custer | ID | -0.5859 |
| Thurston | WA | 0.7572 | Columbia | WA | -0.0215 | Jerome | ID | -0.5932 |
| Deschutes | OR | 0.7570 | Lewis | WA | -0.0428 | Sherman | OR | -0.6604 |
| Kitsap | WA | 0.7522 | Crook | OR | -0.0445 | Wheeler | OR | -0.6881 |
| Bannock | ID | 0.7498 | Tillamook | OR | -0.0893 | Harney | OR | -0.6961 |
| Whatcom | WA | 0.6684 | Elmore | ID | -0.1069 | Boundary | ID | -0.7389 |
| Kittitas | WA | 0.6385 | Pacific | WA | -0.1086 | Morrow | OR | -0.7723 |
| Marion | OR | 0.6043 | Hood River | OR | -0.1154 | Adams | WA | -0.7894 |
| Clark | WA | 0.5999 | Douglas | OR | -0.1367 | Minidoka | ID | -0.8253 |
| Canyon | ID | 0.5066 | Wasco | OR | -0.1389 | Benewah | ID | -0.8508 |
| Jefferson | ID | 0.4815 | Union | OR | -0.1498 | Shoshone | ID | -0.8794 |
| San Juan | WA | 0.4786 | Garfield | WA | -0.1610 | Butte | ID | -0.8990 |
| Polk | OR | 0.4750 | Payette | ID | -0.1743 | Clearwater | ID | -0.9039 |
| Island | WA | 0.4189 | Cassia | ID | -0.1743 | Lemhi | ID | -0.9319 |
| Yamhill | OR | 0.4110 | Klamath | OR | -0.1841 | Gooding | ID | -0.9852 |
| Nez Perce | ID | 0.4009 | Josephine | OR | -0.1881 | Adams | ID | -0.9932 |
| Linn | OR | 0.3778 | Mason | WA | -0.1935 | Gilliam | OR | -1.0030 |
| Chelan | WA | 0.3591 | Baker | OR | -0.2214 | Lewis | ID | -1.0305 |
| Lincoln | WA | 0.3405 | Curry | OR | -0.2223 | Owyhee | ID | -1.0651 |
| Jackson | OR | 0.3358 | Gem | ID | -0.2367 | Grant | OR | -1.1182 |
| Stevens | WA | 0.3174 | Clallam | WA | -0.2368 | Lake | OR | -1.1456 |
| Blaine | ID | 0.3026 | Wallowa | OR | -0.2414 | Clark | ID | -1.7980 |
| Asotin | WA | 0.2856 | Klickitat | WA | -0.2425 | | | |
| Teton | ID | 0.2819 | Bear Lake | ID | -0.2629 | | | |
| | | | Okanogan | WA | -0.2703 | | | |

Having a full list of scores is useful for checking a specific county, but the results in table form are not interpretable in the aggregate. To help better understand the composition of each cluster, we utilize PowerBI filled maps to show the geographic distribution of clusters.

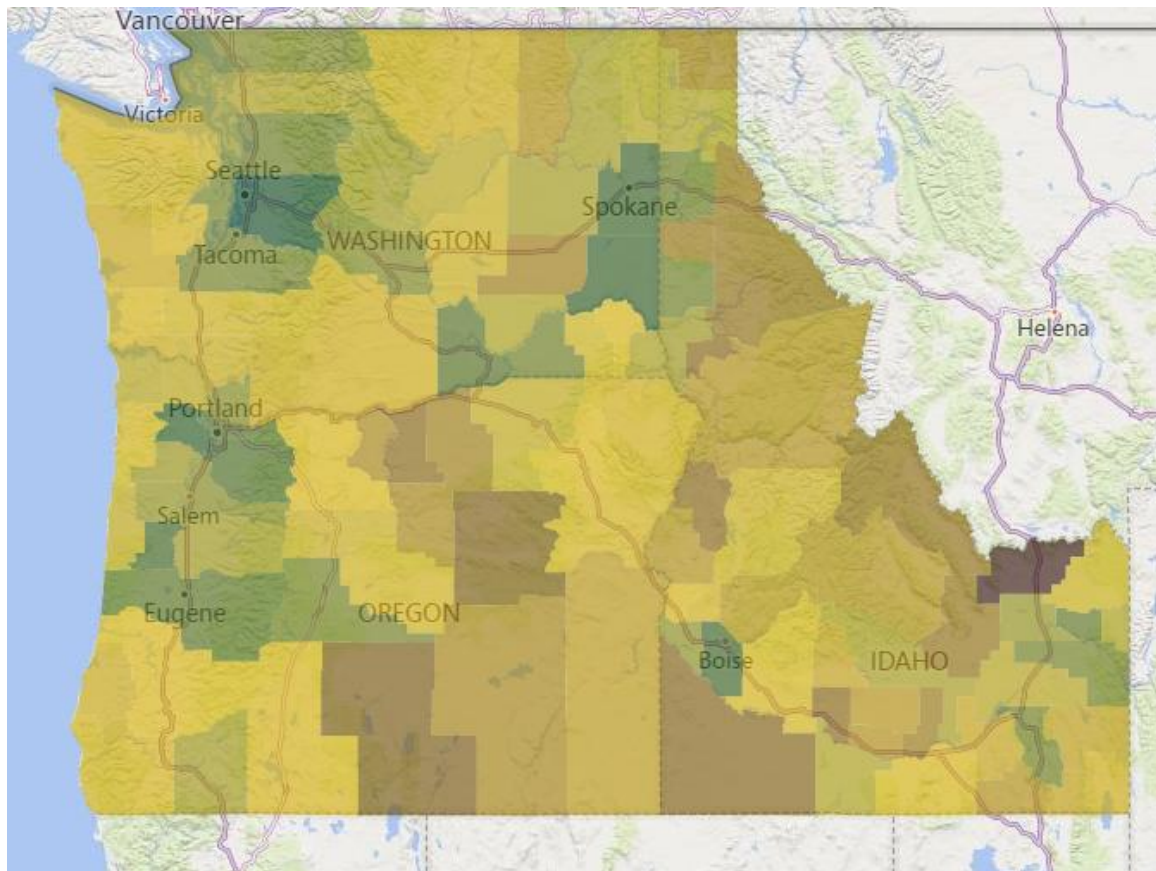


Purple counties, such as those containing Seattle, Portland, Spokane, and Boise belong in cluster 2, representing the counties with the best internet. Based on the cluster analysis we know this cluster represents wealthy and educated counties with many providers and devices. Orange counties are those in cluster 3, with high reliance on public internet infrastructure. These counties appear to be mainly found in rural regions. Similar results are found for light blue cluster 1, those with large populations that didn't graduate high school that we determined to represent middle-of-the-road counties. Occupying about half the dataset, we see these counties distributed largely as a buffer between population cores from cluster 2 and rural areas in cluster 3. Note that while some counties in cluster 2 are directly adjacent to cluster 3 counties, due to the varying distribution of population inside of each county, we assume there is a gap between rural and urban regions.

Overall, the map confirms the notion that highly populated/metropolitan areas experience the highest quality internet, while extremely rural/low-populated regions

experience the worst internet. Further, we grouped different counties by similar characteristics by sorting them into clusters.

Comparison of the clustering analysis map with the robust sparse PCA index score map largely confirms the conclusions of the research. Dark green counties, representing those with the highest internet scores, overlap extremely well with counties occupying cluster 2.



Putting these maps together paints a compelling picture – the index map highlights regions with great and poor internet performances. The clustering map shows us that those counties are grouped together based on shared characteristics. That is, the performance of a county is not random, but rather a result of specific variable performances. Cluster 2 clearly represents counties that perform high in positive variables, which is confirmed by the counties in this cluster occupying almost all the top index scores. At the bottom end of the index score spectrum, the worst-performing counties are split between clusters 1 and 3. This indicates neither cluster is significantly worse than the other. This is supported by cluster analysis, which showed that both clusters perform well in negative variables. Poor education and high reliance on public internet are unique factors, not directly competing. Intuitively, this shows that these counties don't perform uniformly – differing reasons

account for poor performances. Some counties see poor internet linked to low education, while others see high reliance on public internet.

Conclusion

While the internet is universally available and necessary across the Pacific Northwest, the quality varies greatly across counties, for different reasons. We leveraged variables relating to personal device ownership, education, internet speed, internet access, wealth and affordability, public internet use, and work from home rates to create a comprehensive dataset for each county. We used robust sparse principal component analysis to create an index score for each county in Washington, Idaho, and Oregon that measures relative internet performances between counties. K-means cluster analysis of the variables was also used, revealing that metropolitan areas generally have the best internet. The data indicates this is due to high rates of personal device ownership, education, income, and internet provider options. Counties with lower index scores tend to have poor education rates and reliance on public internet infrastructure. The insights we gained in this project can inform public policy and decision-making centered around reasons for excellent and poor internet conditions.

As education rates are shown to influence internet performance, it would be advisable to focus on increasing high school graduation and higher education rates. By prioritizing improving education, counties can increase digital literacy, creating a ripple effect of a literate population using the internet more effectively. The effects of this could include increased social cohesion, greater community, and increased economic conditions.

Additionally, reliance on public internet is a significant factor in counties that struggle in the index. Digital device ownership is heavily linked to great internet performance for counties. Working to increase personal ownership of laptops, tablets, and smartphones through subsidized/rebated devices or device loan programs could be a great step towards improving internet conditions.

While great effort was made to ensure the integrity of the dataset, weighting and clustering processes, and conclusions, additional work could be conducted to confirm the results of this research. Namely, constructing similar indices with differing but still relevant variables and methods by independent parties could confirm this research or indicate other factors that could contribute to internet quality. Additionally, because the data used is regional, applicability outside of the Pacific Northwest region is unknown. Differing parts of the country may have unique regional reasons for internet performances.