SMU Data Science Review

Volume 6
Number 1 *Spring 2022*

Article 2

2022

Adjusting Community Survey Data Benchmarks for External Factors

Allen Miller
Southern Methodist University, allenmiller@smu.edu

Nicole M. Norelli
Southern Methodist University, nnorelli@smu.edu

Robert Slater
Southern Methodist University, rslater@smu.edu

Mingyang N. Yu
Southern Methodist University, nyu@smu.edu

Adjusting Community Survey Data Benchmarks for External Factors

Allen Miller[1], Nicole M. Norelli[1], Mingyang Nick Yu[1], Robert Slater
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
{allenmiller, nnorelli, nyu, rslater} @smu.edu

Abstract. Using U.S. resident survey data from the National Community Survey in combination with public data from the U.S. Census and additional sources, a Voting Regressor Model was developed to establish fair benchmark values for city performance. These benchmarks were adjusted for characteristics the city cannot easily influence that contribute to confidence in local government, such as population size, demographics, and income. This adjustment allows for a more meaningful comparison and interpretation of survey results among individual cities. Methods explored for the benchmark adjustment included cluster analysis, anomaly detection, and a variety of regression techniques, including random forest, ridge, decision tree, support vector, gradient boosting, KNN, and ensembles. The final models used ensemble regression methods to predict trust in government and identify important features and cluster analysis to assign similar cities to clusters for comparison. The voting regression model predictions were compared to the actual raw scores, and cities that scored significantly above and below predictions were identified. These overperformers and underperformers may have additional factors not accounted for within the model contributing to their score.

1 Introduction

As municipal-level leaders are encouraged to make more data-driven decisions (Sawicki & Craig, 1996), particularly regarding the impact of local government spending on resident quality of life, objective evaluation of city performance is essential. Many factors influence the quality of life metrics. Some cities begin from a much more advantageous position due to factors municipal-level leaders cannot directly control, such as superior economic and natural resources, which can be reflected in resident satisfaction. If adjustments can be made for some of these factors, a more accurate evaluation of city performance can emerge. This will allow municipal-level leaders to focus their efforts on meaningful change and expend resources on projects that will improve their residents' satisfaction and quality of life.

Polco is a U.S. company that provides an online public input platform for cities to collect resident data. Local governments use the platform to gather input from their residents and then use that data to inform decision-making. This study combined The National Research Center at Polco's proprietary survey data with public datasets such as the U.S. Census and local government spending data to adjust resident satisfaction

benchmarks based on factors municipal-level leaders cannot easily influence. This research used the National Community Survey (NCS), a long-standing, well-established community survey that has been administered throughout the United States (Miller & Kobayashi, 2000) to develop a mathematical model for establishing benchmarks. This model will allow for more accurate comparisons between cities, leading to better data-driven decisions by local government officials. In other words, this research aimed to use The National Research Center at Polco's National Community Survey data in combination with public datasets to build a predictive model for resident satisfaction benchmark adjustments in various city types.

While the NCS gathers data regarding various domains, such as overall economic health, infrastructure, safety, and natural environment, this study focused on an overall measure of confidence in the local government. Recent NCS data from the previous five years was cleaned, and all the cities within this recent subset were identified. A survey of relevant literature indicated an association between certain characteristics, such as economic prosperity, education, health, and quality of life measures (Diener, 2013; Lawless & Lucas, 2011; Rentfrow et al., 2007). This study explored using data regarding these characteristics to adjust benchmark expectations for cities. Data from the U.S. Census, specifically the American Community Survey (ACS), provided demographic and population information about the identified cities. This data was incorporated, and specific characteristics, such as city size, education levels, poverty levels, and employment levels, were explored.

The five cities from the NCS with the highest confidence in local government ratings were all mid-sized cities with less than 100,000 residents. These five cities administered their surveys between 2017 and 2021. Three of the five were from 2020 and 2021 during the global pandemic. Four of the five cities were in the Midwest region of the United States. All five cities had a lower-than-average percent of households below the poverty level, and all the top five cities had an above-average median household income.

The five cities with the lowest confidence in local government ratings had a much wider population size range, from below 10,000 to above 100,000 residents. The surveys were administered between 2018 and 2020. They were also spread out among the geographical regions of the United States. While the percentage of households below the poverty level varied among the five cities, the median household income was lower than average for three of the five.

Average education for high school, bachelor's degree, and master's degree is higher for the five cities with the highest confidence in local government scores compared to the five cities with the lowest confidence. On average, 54% of citizens in the top five cities hold bachelor's degrees or higher, compared to 34% in the cities with the lowest confidence in local government. The average difference for master's degrees or higher is 15% compared to 9%. Median household income is 36% higher for cities with high confidence, while per capita income is 42% higher on average. There is a 236% increase in the average poverty rate for the five cities with low confidence scores compared to those with high confidence scores.

Current literature regarding resident satisfaction measures does not provide an established method for benchmark adjustments. Raw satisfaction scores can be difficult to interpret. For instance, scoring higher in an overall feeling of safety than the overall quality of the natural environment does not necessarily indicate a problem with the

city's natural environment. The most common method for providing context to the scores is to compare them to a national average score on the same or similar question (Miller & Kobayashi, 2000). However, the limitation of this strategy is that not all cities start from the same position. Creating a model that adjusts for factors that cannot be easily influenced would provide a more accurate assessment of performance in each survey area. Research within the citizen survey, resident satisfaction, and related domains use a variety of theoretical and methodological approaches. This study utilized predictive Regression and random forest models, cluster analysis, and anomaly detection to explore various methods and approaches to benchmark adjustments.

2 Literature Review

Community survey data and its utilization within the public policy is a broad topic. The theoretical foundation of this study's model was informed by literature across several fields. This literature included examining survey data methods and limitations, theoretical approaches to satisfaction measures, the relationship between objectively measured characteristics and community satisfaction scores, survey usage at the local government level, and analysis methods used with survey data in previous studies.

2.1 Survey Methods

Surveys are the required instrument when the government or private entities want to quantify public opinions or sentiments. While surveys allow for gathering many different types of information, proper survey methods need to be implemented to yield valid, reliable results. Because citizen survey results are a central component of this study, understanding these methods and their constraints is essential. The accuracy of appropriately conducted surveys has been previously established (Pew Research Center, 2012). Adjustments can be made for known survey method limitations, such as non-response bias (Miller & Kobayashi, 2000). The impact of using new communication methods, such as the internet, to collect survey data has also been examined (Greenlaw & Brown-Welty, 2009).

Community surveys, particularly regarding the quality of life, have been explored in various ways. Research into community survey data, methods of collection, and types of respondents indicate that surveys, when conducted properly, are relatively accurate (Pew Research Center, 2012). Properly administered community surveys accurately represent the percentage of different political affiliations, financial status, home values, and many demographic characteristics.

When properly conducting surveys, key considerations include selecting an appropriate sample size and addressing non-response bias. The sample size is important, as there is variability in responses, and a sufficient sample is necessary to make sure the true population response is reflected. Also, non-response bias can be problematic, as there are well-established differences between respondents and nonrespondents to community surveys. For example, the typical respondent is more likely to be involved in civic activities, and surveys underrepresent people with less education (Pew Research Center 2012). However, these limitations can be managed

with responsible survey techniques, such as stratified sampling and proper weighting of responses (Pew Research Center 2012). Using the characteristics of the population of interest, results can be reweighted to accurately reflect the population (Miller & Kobayashi, 2000). For accurate reflection of the population, this reweighting method is required for most surveys (Miller & Kobayashi, 2000), and The National Research Center at Polco applied it to the NCS data as part of their data pipeline.

Another consideration is the method of survey administration. Surveys can be administered via in-person, telephone, mail, and the internet. As new methods of communication have emerged, survey techniques have evolved. According to Pew Research Center (2012), the percentage of people who respond to telephone surveys decreased from 36% in 1997 to 9% in 2012. On the other hand, the use of internet-based methods has increased, and Greenlaw and Brown-Welty (2009) demonstrated the superior balance of cost-savings and response rates with web-based methods. No method can encompass the entire target population, but the proper weighting of the responses is one strategy for addressing underrepresented populations when using a particular method.

Although collecting and analyzing survey results is important, utilizing these results to improve services is the ultimate goal. Miller and Kobayashi (2000) recommend four ways to ensure survey data is used. First, create a survey task force with community members and elected officials to provide recommendations for action in response to the results. Next, make sure the responses are discussed with the individuals who are providing the services that were assessed. Also, survey results can be used with a department's performance measures. Finally, municipal leadership can construct focus groups containing survey participants to provide additional context to the response trends.

2.2 Satisfaction and Quality of Life Measures

Measures of resident satisfaction and quality of life can be approached from various theoretical perspectives. Often, a more global measure of satisfaction is comprised of satisfaction within a variety of related subdomains. Sirgy et al. (2000) utilized the "bottom-up spillover theory," which suggests that global quality of life comprises the quality of life in different domains or components. This is similar to the NCS survey data. There is an overall satisfaction measure in each domain, such as safety or economy, and there are numerous sub-measures within each of those categories.

Another approach was employed by Barrington-Leigh and Wollenberg (2019) in their community well-being survey across Connecticut. They examined the relationship between life satisfaction and various characteristics using linear Regression. Next, using compensating differentials, they quantified the effects of changes in factors such as food security, walkability, and responsiveness of local government in terms of the equivalent increase in income that would be required to have the same impact on life satisfaction. This led to identifying the most cost-effective ways to increase life satisfaction in the community.

Objective versus Subjective Measures. One area of research examines the relationship between objective and subjective wellness measures. Historically, public

policy has focused on objective improvement measures and desire-fulfillment rather than subjective well-being measures (Dolan & White, 2007). By identifying previously researched objective characteristics associated with life satisfaction, this study can better construct predictive resident satisfaction benchmarks.

Previous research has demonstrated an association between particular characteristics and well-being. Economic prosperity and its relationship to well-being are more widely studied characteristics. There is a strong association between income and well-being when comparing nations, but the association decreases slightly when it is examined within a nation (Diener, 2013). For example, Lawless and Lucas (2011) examined measures of well-being at the county level in the United States and found that median income had a moderate correlation with well-being. Other measures related to economic prosperity, such as unemployment rate and population percentage below the poverty line, also showed associations with well-being after controlling for income (Lawless & Lucas, 2011). In a U.S. study at the state-by-state level, Rentfrow et al. (2007) also found associations between income and well-being.

Additional characteristics that are related to economic prosperity, such as education level and type of occupation, are also associated with well-being. Individuals with higher education and "professional" occupations tend to report higher life satisfaction, even after accounting for income (Lawless & Lucas, 2011).

Health is another characteristic associated with life satisfaction. Lawless and Lucas (2011) found moderate to strong county-level correlations between life satisfaction and obesity, disability, and death rates. Counties with higher rates of obesity, higher percentages of people with disabilities between the ages of 21 and 64, and higher death rates from heart disease, homicide, and diabetes had lower life satisfaction.

2.3 Expectancy-Disconfirmation Theory

Although objectively measured characteristics and performance are associated with satisfaction, individuals' expectations can also influence them. Van Ryzin (2013) experimented to explore the role of expectations in satisfaction with government services. Using the expectancy-disconfirmation theory, which states that performance expectations influence subjective judgments, Van Ryzin (2013) manipulated expectations regarding street cleanliness and then measured satisfaction with performance. The results supported the expectancy-disconfirmation theory, although effects were stronger for older adults, females, and those less politically conservative. Higher expectations amplified responses and were associated with lower satisfaction with poor performance and higher satisfaction with good performance. Overall, however, performance (in this case, actual street cleanliness) was highly correlated with satisfaction, supporting citizen surveys to measure actual performance and not just participants' expectations or predispositions.

2.4 Local Governance and Public Policy

The best way for government officials to utilize information gathered from survey data varies from government to government. Some local governments attempt to involve the

public in data collection and analysis (Sawicki & Craig, 1996). Poister and Streib (1999) found that local-level governments were motivated to collect data from their residents because of a desire to improve policy rather than fulfill any requirements set at the state or national level. The idea is that when the public is involved in analyzing data, more informed decisions for policies and regulations can be made. The presentation of data and information, along with public satisfaction, can be influential for all levels of government decision-makers.

Other local governments are performing the analysis and concluding themselves. These findings are then presented to the public for feedback on proposed changes. This type of public government relationship can allow policymakers to hear public opinion before making final decisions that could be costly or result in other adverse effects.

Of course, in democratic societies, government entities rarely propose changes without first hearing from their constituents. Survey data is an important platform that allows citizens to freely critique policies and services. Surveys can give both positive and negative feedback, letting the government know which policies are working and which are not.

2.5 Feature Importance

Deciding which features are important helped reduce the model size and increase performance. Two main types of selection were tasked with ranking variable importance.

Ridge Regression provides a penalty to the magnitude of each coefficient after the features are transformed into the same scale. The coefficients can be utilized to analyze feature importance with each variable's magnitude and directional impact to the result. Ridge Regression and Least Absolute Shrinkage and Selection Operator (LASSO) Regression performs much in the same way with a few notable exceptions. Most distinctly, LASSO can completely remove a feature's predictive power by reducing the coefficient's weight down to zero. By contrast, the ridge cannot completely remove a feature from the model. This can be beneficial in situations where it is known that there are features that are going to be more helpful but trade-off with an increase of variance in the model.

A complete test for feature importance could be conducting a permutation where each variable is randomly shuffled multiple times during evaluation and its impact on the prediction error is measured (Pedregosa et al., 2011). This approach is useful for small data sets as it allows for a complete comparison of every feature in the data. However, this can be computationally expensive, especially for data sets with a large dimensionality. This test can be performed using any model and can provide a baseline for results upon its completion.

2.6 Cluster Analysis

Clustering analysis using the Ward method and Euclidean distance was utilized by Sidorchuk et al. (2020) to classify ten different administrative districts in Moscow into five clusters based on survey results of drivers' satisfaction. Statistical analysis in

combination with cluster analysis provided valuable insights for the city of Moscow to implement parking policies, including tariffs, planning, and construction of new parking lots. This would be difficult to accomplish with other methods such as discriminant analysis or decision trees for classification because there is no known response or label ahead of time. The Ward method is part of the hierarchical clustering technique. It is an iterative process of creating hierarchical groups based on a user-defined similarity measure. This algorithm begins with each instance as individual clusters and then scans through all possible pairs that can be joined. The optimal pair among all possibilities is returned as one cluster. This cluster center is then paired with all other clusters iteratively until all clusters are connected as one (Ward, 1963).

Although Ward's hierarchical clustering method was useful for the Sidorchuck et al. (2020) parking satisfaction study, the density-based spatial clustering of applications with noise (DBSCAN) was an additional clustering method to explore. According to Schubert et al. (2017), the algorithm is density-based and useful for indexing. As it is very adaptable, DBSCAN is capable of pairing with different data types, distance functions, as well as indexing techniques. With the right parameters tuned, such as the number of samples in a neighborhood for a point to be considered a core point and radius parameter, this technique has proven effective in practice (Schubert et al., 2017). In contrast to an algorithm such as KMeans, DBSCAN does not require the number of clusters to be defined in advance. It can identify arbitrarily shaped clusters due to the density-based nature of the algorithm. Without predefined labels or the number of clusters, DBSCAN can be useful to automatically identify the number of clustered cities to be ranked based on given features.

KMeans clustering can work as a complementary clustering algorithm after DBSCAN has identified the number of clusters using density. One disadvantage of DBSCAN is that some points' density below the established threshold can still be assigned to clusters for data points in between clusters (Campello et al., 2015). In other cases, points that are not density reachable from any core points can be categorized as noise points, which are not part of any clusters (Schubert et al., 2017). These uncertainties left by DBSCAN can, however be solved by KMeans clustering. Every city in the dataset will be part of a cluster based on a defined distance rule. As KMeans requires an initial number of clusters to be defined, DBSCAN will supply such information ahead of time.

One intrinsic method for evaluation is the Silhouette score to assess the clustering result. The Silhouette score evaluates how similar an instance is with its current cluster compared to other clusters. The score itself ranges from –1 to 1, where a higher score indicates the clustering is doing better. A score of 0 means overlapping clusters, which can possibly be assigned to nearby cluster, and a negative value indicates it might have been assigned into the wrong cluster (Pedregosa et al., 2011).

2.7 Curse of Dimensionality

As clustering techniques require selecting a distance metric, factors concerning how to choose this metric for the problem in this study become important. The clustering result will serve as a direct benchmark for grouping cities together and comparing different city performances in each cluster. Aggarwal et al. (2001) noted that distance metrics

with higher degrees (higher norms) perform worse under high dimensional space to reflect closeness between data points accurately. It has also been reiterated in Domingos (2012) that even for highly relevant features, those same key features in low dimensional space would behave drastically differently when projected into higher dimensions. Domingos (2012) introduced in the article the normal multivariate Gaussian distribution we see in lower dimensions is no longer near the mean in higher dimensions but more like a distant "shell" around the mean. This means many intuitions acquired from lower dimensions no longer apply in higher dimensions.

Mathematical and empirical proofs were provided by Aggarwal et al. (2001) to demonstrate the Manhattan distance metric (L_1 Norm) is preferable to the Euclidean distance metric (L_2 Norm) for high dimensional data applications such as clustering or nearest neighbor classification problems. Aggarwal et al. (2001) compared norms one and two and provided proof for higher norms and fractional norms. Higher norms consistently performed worse in high dimensions than lower norms in all five machine learning data sets provided through the University of California Irvine (UCI) machine learning repository. Additionally, a fractional distance metric such as ($L_{0.1}$ Norm) generated better clustering or classification results in higher dimensions and was also more robust to the noise in the data set (Aggarwal et al., 2001). As effective as a fractional distance metric is in high dimensional space, it does not follow the triangle inequality rule, which declares that the sum of the length of two sides of a triangle must be greater than or equal to the remaining side (Khamsi & Kirk, 2001). By considering the factors mentioned and the number of dimensions available in this study, L_1 Norm or fractional distance metrics should be the top contenders to be considered for clustering.

One issue with L_1 Norm or fractional distance is that it is not supported by some clustering methods such as K-means algorithm from scikit-learn. According to Pedregosa et al. (2011), K-means algorithms are designed for minimizing the within-cluster sum-of-squares score when choosing ultimate centroids. To address the issue with high dimensionality spaces using Euclidean distance, it was proposed to explore Principal Component Analysis (PCA) before applying K-means. PCA is a dimensionality transformation technique, with the most variance presented in the first component and the subsequent components having increasingly less variance. (Jolliffe, 2002).

2.8 Supervised Learning

While clustering techniques are capable of grouping cities together based on given characteristics without labels, supervised learning techniques can identify or establish a relationship between those characteristics and a target. Various supervised learning algorithms can establish such relationships between features and goals. Once such a relationship is established, new cities, given trained features, can have a projected score compared to the actual score the city received. The difference between the two can be subsequently evaluated to ascertain whether the city is exceeding expectations or not based on its given features.

Ridge regression typically uses ordinary least squares as an objective function. However, Ridge regression imposes an additional penalty term, and a complexity parameter controls the penalty strength. The penalty term helps with overfitting, or

when the prediction model predicts training data well but does not generalize well into a test or future data.

K-Nearest Neighbors regression (KNN regression) is similar to KMeans clustering mentioned under cluster analysis. Consequently, many of the high dimensional problems under the curse of the dimensionality section still apply. However, instead of separating cities into groups ahead of time, KNN regression calculates at runtime to locate the defined number of cities nearby based on a given distance metric. It then weighs the option on those distances for evaluating new cities (Pedregosa et al., 2011).

Random Forests Regressor (RFR) uses the concept of an ensemble. RFR combines many weak learners (individual decision trees), and each learner is built from a sample drawn with replacement from the training set. Additionally, each split of nodes on a weak learner can be found from a subset of all the training features to decorrelate learners in the forest. RFR usually achieves better performance than many independent models (Pedregosa et al., 2011). Similarly, Gradient Boosting Regressor (GBR) uses many weak learners. However, instead of using weak learners in parallel, GBR learns in sequence using differentiable loss functions. Each subsequent learner tries to improve based on the previous weak learner's mistake.

The idea of ensemble used by Random Forests and Gradient Boosting of individual weak learners can be expanded to combine multiple machine learning algorithms. Pedregosa et al. (2011) implemented the Voting Regressor method to average the predicted values of multiple given machine learning methods. This method should even out individual models' weaknesses. Different independent models would learn a different part of the given data, and the ensemble has been empirically proven to be successful in many competitions (Zarate, n.d.).

2.9 Hypothesis

Using National Community Survey data with U.S. Census and other publicly available data to implement supervised and unsupervised techniques, a confidence in local government baseline can be developed and used to compare community performance on equal terms. Important features can be identified, and similar communities can be grouped together. Previous research and theory indicate higher confidence in government could be associated with features such as economic prosperity, higher educational attainment levels, and better health.

3 Methods

3.1 Data Overview

Polco supplied the survey data utilized in this study. The National Research Center at Polco has a wealth of statistically sampled survey information from over 500 communities around the country, in some cases going back 20 years. The surveys address community livability, governance, public trust, equity and inclusion, and public

safety. In addition to The National Research Center at Polco's survey responses, data from multiple publicly available sources, including demographics data from the American Community Survey, Gini Index of income inequality from the National Historical Geographic Information System (NHGIS), Municipal Finance Data from census data, and Internet Access Index (IAI) by Argonne National Laboratory were utilized. The additional data sources provided supplementary data about local government, such as demographics, income disparity estimates, local government finance, average income, quality of life, education, and broadband infrastructure. These attributes were utilized to calculate the expected performance of local government. These calculations were utilized as a benchmark to evaluate whether a local government was performing above or below expectations from the actual survey data.

3.2 Data Cleaning & Merging

The National Community Survey data includes information regarding the survey year, geolocation in the form of Federal Information Processing System (FIPS) codes, survey subject area, and the adjusted survey score (weighted for correct demographics). Per Polco's request, analysis focused on the five most recent years of survey results available, specifically 2017 – 2021. In addition, one area that assessed government sentiment called "Overall confidence in ABC government" was extracted. FIPS codes were used to merge the NCS data with the additional publicly available data sets, ensuring information was added to the correct city and that all data pertained to the same precise geographic locations.

The first step to combine external data sources with NCS data involved merging with American Community Survey (ACS) 2018 data. ACS data has many useful community characteristics available including total population, demographics, population aged 25 and above, number of people with high school degrees, bachelor's degrees, master's degrees, household income below poverty, and median household income. These characteristics were all potential candidates to use in the benchmark adjustment calculation. A few cities completed the survey multiple times within the chosen five-year window. By comparing scores of overall confidence in local government among the cities that conducted the survey multiple times, the results were surprisingly consistent. One of the cities, however, had a substantial increase in this score between 2019 and 2021. Apart from this city, scores did not change significantly when readministered across the five-year period. Given this consistency, all cities that included the overall confidence in local government question within the five-year window were selected for analysis, with the most recent score retained if a city administered multiple years within the window.

To directly compare cities, demographic variables reported as counts were normalized into percentage of the city population. This included raw demographic numbers, education numbers, and labor force numbers. Additionally, the number of households with incomes below poverty and number of owner-occupied housing units were normalized by dividing total number of housing units in the city. A few additional useful characteristics of a city were calculated, including total percentage of armed forces in each city, total civilian unemployment rate, and average household size of the city. Multicollinearity is a potential challenge when trying to build an interpretable

model and understand important features. Highly correlated variables that were closely represented by other variables were deleted from the feature space so the model could better illustrate the relationships between the variables and the confidence in government score. Deleted features included variables such as total civilian labor force, total occupied housing unit, and total renter occupied housing units. This process identified 207 cities across a five-year period that asked the question regarding overall confidence in local government.

Next, Gini Index data from 2019, which measures the income inequality, was merged with the data. The Gini coefficient ranges from 0 to 1, with 0 representing perfect equality and 1 representing absolute inequality (U.S. Census Bureau, 2022). For the cities contained within the data set, Gini Index ranged from 0.2972 to 0.5694 with a mean of 0.433.

From the available finance municipal data, the features total revenue and total taxes were selected as the most relevant information. These were merged with the data using FIPS codes. There were three cities from the data set missing from the finance data source, and a manual look up of their information was performed. The data of the three cities was imputed through official data sources. Like the demographic variables, financial data in raw number form is difficult to directly compare between cities, so total revenue and total taxes were converted to per capita revenue and taxes accordingly.

Finally, the internet access index was integrated into the feature space. The internet access index (IAI) was calculated by Argonne National Laboratory, a U.S. Department of Energy research center. The index combines measures of quality and availability of high-speed internet with a measure of local ability to subscribe to the service. The IAI ranges from 0 to 1, with 0 representing less access (Alexander, et al., 2021). The internet access index data available was reported by census blocks within a county. To estimate the city internet access index, the census blocks within the appropriate FIPS codes were averaged. This average was used as the estimate of a city's broadband performance. For the cities within the data set, IAI scores ranged from 0 to 0.5, with a mean of 0.33.

3.3 Clustering Analysis

Both Hierarchical Clustering using the agglomerative method and KMeans clustering using Euclidean distance were attempted on scaled features. Once cities were assigned to a cluster, their benchmark was estimated to be the mean of all the cities within the cluster. A mean squared error between the estimated benchmark and actual score was then calculated to evaluate clustering performance.

3.4 Anomaly Detection

Isolation forest was utilized to perform anomaly detection and detect unique cities, given its characteristics to provide additional insights for benchmark analysis. The algorithm itself is similar to a random forest algorithm, however it focuses on calculating the number of partitions to reach cities on average in the forest. A city which can be consistently reached by a small number of partitions in a random forest would

mean it is an anomaly. Such a city would be assigned a negative score with large magnitude indicating a stronger anomaly (Pedregosa et al., 2011).

3.5 Supervised Learning

The objective of supervised learning is to establish a relationship between the actual survey score with the given features of a city. In this case, the features were ACS survey responses, Gini Index, city finance data, and internet access index. Once a supervised learning model is trained, it can produce an expected confidence in overall government score for a city given its features. The expected score can then be compared to the actual score to evaluate whether the city performed better or worse than expected. All learning methods were performed on scaled features to ensure comparability between attributes. Supervised learning methods included ridge regression, random forest regression, decision tree, support vector regression, gradient boosting Regression, K-nearest neighbor regression, and voting regression.

3.6 Feature Importance

Several methods of feature importance were explored, including a comparison of ridge regression coefficients on scaled variables, permutation importance using Ridge Regression, permutation importance using Random Forest, and permutation importance using Voting Regressor.

Ridge regression coefficients were fit on scaled variables, so all features were scaled between zero and one. By examining the coefficients from the cross validated model with the best penalty value, the magnitude and negative or positive value of the coefficient indicated how different variables were associated with confidence in local government scores.

While ridge regression coefficients are highly interpretable and indicate directional impacts of features on the response, permutation importance uses a different strategy to evaluate each feature's importance. Each feature is shuffled during evaluation, and the impact of the random shuffle on the prediction error is measured. If a variable shuffle generates a large error, that is a direct indication of high importance of that variable. Each variable is shuffled 30 times and the average error impact is measured.

4 Results

This research intended to assess how a community's confidence in local government, as surveyed by its residents, compares to the hypothesized confidence in local government and provide a benchmark adjustment to assess performance more fruitfully. Communities targeted in this study were grouped into clusters to create scores based on the members of each cluster. Individually, each community was compared to its respective cluster to analyze how well it was performing versus similar communities. Clustering communities allows for an "apples-to-apples" comparison of communities.

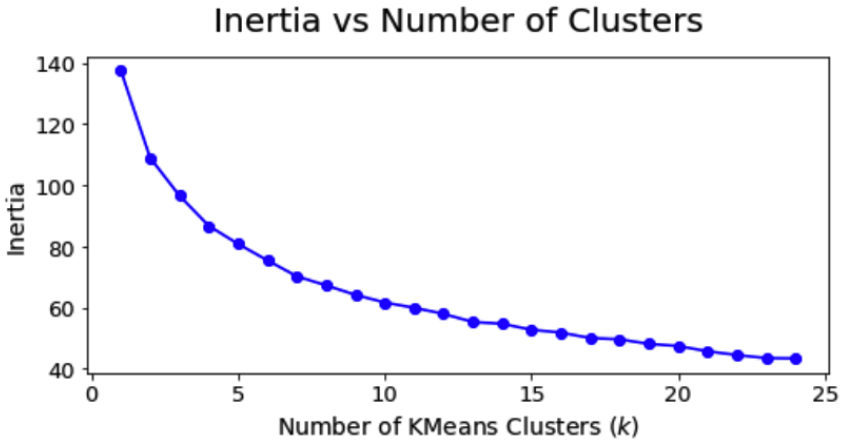
Similarly, regression models were used to predict confidence in local government scores, with actual performance above or below the prediction being an indicator of over or underperformance when compared to what would be expected in a city. Key predictive features were extrapolated from the data based on two models. This provides communities insight into which areas most influence confidence in local government.

4.1 Cluster Analysis

Initial comparison of Hierarchical and KMeans clustering was conducted by comparing the mean squared error of the two clustering techniques among cluster numbers of 1 to 16. KMeans had superior overall performance under each number of clusters, and it was deemed a superior method as a result.

To evaluate the best number of clusters to use for KMeans, the Elbow method, Silhouette coefficient, and mean squared error were used for selection. The Elbow method had no clear cutoff point where inertia suddenly changed (see Figure 1).

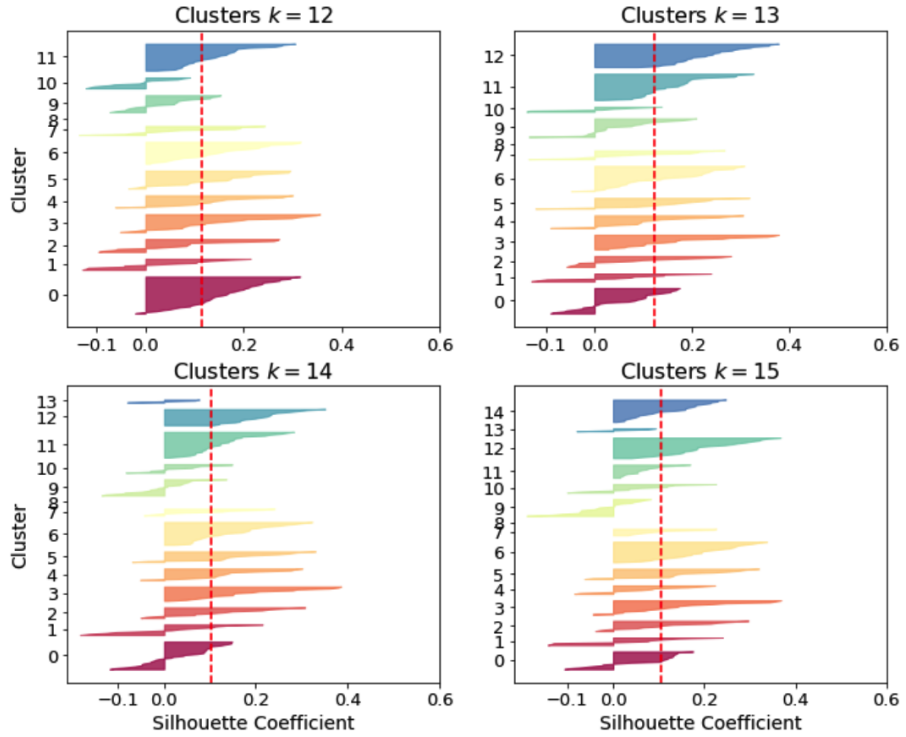
Figure 1



Silhouette coefficient performance was difficult to distinguish among the different number of clusters (see Figure 2).

Figure 2

Silhouette Coefficients for Various Cluster Values



Mean squared error identified 13 clusters to be useful. Of course, if every city were in its own cluster, the mean squared error would be the smallest. However, this is not helpful for benchmark adjustment. A cluster number at 13 was found to be a local minimum and believed to be a balance of having a reasonable number of clusters and mean squared error.

Cluster sizes range from 28 cities down to a single city. The goal for creating optimal clusters is to maximize the distance between different clusters while minimizing the distance between points inside of each cluster. This can provide a challenge when later trying to interpret how each cluster was selected. It is possible to look at the characteristics of features within each cluster and gain some clarity on what each city in a cluster had in common. Calculating the top 10 features for each cluster shows evidence that poverty level, racial demographics, and different education levels were some of the major decision-makers in cluster creation. Of course, it would be difficult to see exactly how each cluster was determined because adding more than two or three features to the model would require the ability to see into more than two or three dimensions.

Kmeans clusters were used to create baseline confidence in local government scores for each cluster. The government confidence score was averaged for each cluster, creating a baseline for comparison to a city's actual government confidence score. The results for Kmeans clustering and the Voting Regressor model performed similarly in

terms of determining if a city was overperforming or underperforming compared to the model prediction. For each of the cities where the models disagreed (such as Kmeans clustering indicating a city outperformed expectations but Voting Regressor indicating a city underperformed expectations), the predicted scores were very close to the actual scores. One model predicted slightly above the actual score whereas the other model predicted slightly below. All the top over and underperforming cities were classified as such in both models. The only notable exception was the city assigned to its own cluster. As the cluster mean was used to calculate the KMeans baseline, and there was only one city in the cluster, this city performed exactly as the KMeans model expected. This is a challenge for implementing the KMeans baseline in practice. The one city assigned an individual cluster was a poor performer, with the 7th worst confidence in government score. Effectively, any city assigned its own cluster cannot be given a meaningful adjusted KMeans baseline.

4.2 Anomaly Detection

Anomaly detection was performed using all features both with overall confidence in government score included and without. Results were consistent, with 10 of the 13 cities appearing as anomalies in both.

Unsurprisingly, the one city that was assigned its own individual cluster was identified as an anomaly. The 13 cities identified as anomalies contained many of the most extreme feature values. This included the highest household size, population, Gini score, per capita income, as well as percentage of population unemployed, in the labor force, with high school education or higher, bachelor's degree or higher, and highest percentage of households with income below the poverty line. In addition, these 13 cities contained the lowest IAI score, percentage of owner-occupied housing units, and percentage of population in the armed forces.

4.3 Supervised Learning

Ridge regression was the first supervised learning method attempted. A grid of alpha values, which control the regularization strength, was explored and tuned by cross validation to identify the best regularization strength for training. The best model achieved a mean squared error test score of 58.19. The mean squared error test score measures the average squared deviation from actual to predicted score for cross-validated test results. A smaller value indicates the model is doing a better job predicting the actual survey score and provides a more reliable benchmark for adjustment.

Random Forest Regression was the second model attempted. Tuned hyperparameters included number of estimators, minimum samples per leaf, maximum depth of each tree in the forest, and minimum samples required to split. A grid search cross validation process was used to find the best combination of parameters. The best model achieved a mean squared error test score of 49.27. Although the random forest regression model had a lower, and therefore better, test score than the ridge regression

model, it is worth mentioning that the difference between the training and test scores was much greater in the random forest regression model. This indicates that the random forest regression model is more overfit than the ridge regression model.

The third model explored was a Decision Tree model, which is a simpler model than Random Forest. Tuned hyperparameters, compared via grid search cross validation, included splitter method used, minimum samples required to perform a split, maximum features allowed, and maximum depth of the tree. Because decision trees are typically weaker models, it was expected that the best performance of the model, with a mean squared error of 68.86, was worse than any other supervised model explored.

Support Vector Regression (SVR) model was the fourth model attempted. Because the dataset size was manageable, a support vector regressor was a viable algorithm to explore. Hyperparameters explored using grid search with ten-fold cross validation included regularization parameter (C), kernel coefficient gamma, and kernel options. The radial basis function kernel performed much better than other kernels. After fine tuning other parameters, a regularization parameter of 15 and scaled gamma, which is calculated based on the size of the data frame, generated the best performance. SVR produced better performance in terms of mean squared error compared to Random Forest regression at 47.61. It was also much less overfit compared to Random Forest, as the training score was closer to the test score in comparison.

Gradient Boosting Regressor (GBR) was the fifth model attempted, with a grid of number of estimators as well as each tree's maximum depth being tuned by tenfold cross validation. It is worth noting that despite tuning the model, GBR overfit the data, but it did produce strong prediction results with a best test mean squared error of 49.56.

K-Nearest Neighbor Regression was the final solo model attempted. Hyperparameters tuned via grid search with ten-fold cross validation included number of neighbors, weight function used in prediction, and power parameter for the Minkowski metric using Manhattan distance or Euclidean distance. KNN overfit a bit more than GBR, and it produced a best test mean squared error result of 51.88.

Ultimately, Voting Regression (V.R.) was utilized to expand the idea of ensemble into multiple machine learning algorithms. V.R. combines multiple fine-tuned models, each with its best parameter combination, and uses weighted averages of the prediction results to achieve an ultimate prediction. All supervised models explored, except decision tree, were included in the V.R. Decision tree was excluded because it generated much worse results than Random Forest and Gradient Boosting. Additionally, the decision tree algorithm itself is incorporated into Random Forest and Gradient Boosting. The weight of each model in V.R. is determined by individual test score performance. For instance, in this V.R. model the Support Vector Regressor had the best performance, so it received a higher weight, while the Ridge Regression had the worst performance, and it contributed lower weight to the final prediction. Voting Regression generated the best ten-fold cross validated test mean squared error result at 47.18. V.R. was expected to produce the best results, as it combines the advantages of different prediction algorithms together to generate ultimate results with more perspectives into the data than any individual model. Table 1 provides a side-by-side comparison of model performance.

Table 1

Test Mean Squared Error Results for Supervised Models

Model	Test Mean Squared Error
Ridge	58.19
Random Forest	49.27
Decision Tree	68.86
Support Vector	47.61
Gradient Boosting	49.56
KNN Regression	51.88
Voting Regressor	47.18

The Voting Regressor model, with its superior performance among supervised methods, was used to calculate benchmark adjustments for each city. Previously, the score a city received on the ACS for overall confidence in government (or any other metric) could only be compared to the overall mean score for that metric among all cities. As previously discussed, this might not be a useful comparison. To apply the voting regressor benchmark, each city's out of fold prediction in the V.R. model was compared to the city's actual result. The difference between the prediction and the result measured the city's performance. With this method, a city that performed higher than the prediction would be overperforming, while a city that performed lower than the prediction would be underperforming.

To provide useful feedback, a margin of error can be incorporated into the benchmark. Figure 3 shows overperforming and underperforming cities more than one standard deviation from the mean, and Figure 4 shows overperforming and underperforming cities more than two standard deviations from the mean.

Figure 3

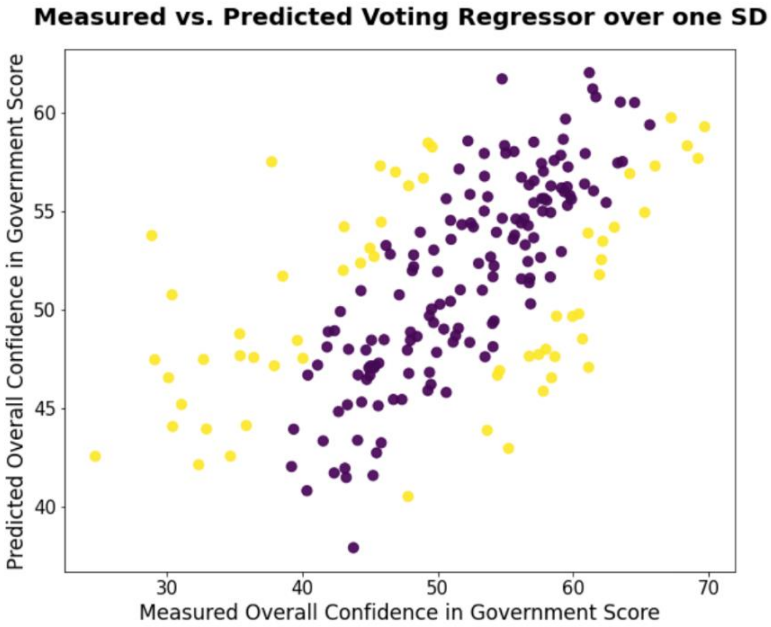
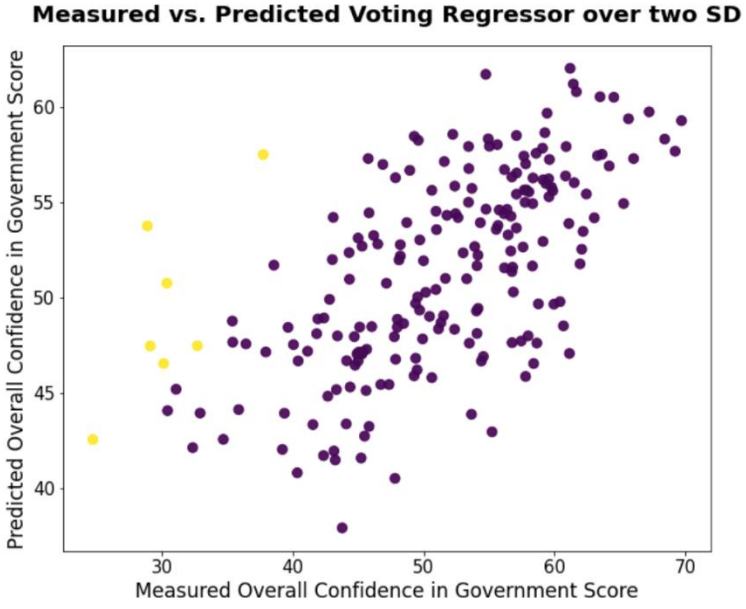

Measured vs. Predicted Voting Regressor over one SD
Predicted Overall Confidence in Government Score
Measured Overall Confidence in Government Score
40
45
50
55
60
30
40
50
60
70

Figure 4


Measured vs. Predicted Voting Regressor over two SD
Predicted Overall Confidence in Government Score
Measured Overall Confidence in Government Score
40
45
50
55
60
30
40
50
60
70

To explore how the benchmark impacts perceptions of city performance on the confidence in local government measure, it is useful to compare the highest raw scores with the cities that had the greatest overperformance when compared to the Voting Regressor benchmark. Table 2 shows rank by actual score on overall confidence in local government under "Raw Rank." With this method, the highest score is ranked number one, the second highest number two, and so on. At present, this is the way a city could receive feedback, comparing their score to the overall mean. An alternate ranking can be calculated using the predicted score from the Voting Regressor model. In this "Adjusted Rank," the city with an actual score highest above their predicted score would be ranked number one. Effectively, this benchmark adjustment gives the expected result given all the factors considered in the model. Scoring significantly above or below this prediction indicates the city could have additional contributing factors influencing the score. Perhaps local policies, methods of communication, or the quality of services provided could contribute to local sentiment.

Examining the cities with the top 10 adjusted ranks, clearly those with the top few raw scores are still outperforming their benchmark prediction. The range of predicted values for the Voting Regressor model was narrower than the actual range of scores, so this result is expected. The cities that significantly improved from raw to adjusted rank are more relevant to the overall benchmark adjustment goal. Cities A and B moved up from ranks 22 and 77, respectively. The mean overall score for confidence in local government was 51.12. Both city A and city B were predicted to score below the overall mean, but they both scored well above it. Using the current method of comparing the raw score to the overall mean to give feedback to city B, they would be under the impression that they are performing averagely, or about the same as most cities. Using the adjusted rank and new benchmark, they could learn that given the realities of all the features included in the benchmark model, they are doing much better than would be expected. City B could examine current policies and try to identify what has successfully inspired greater confidence in their local government.

Table 2

Rank by True Score (Raw Rank) vs. Rank by Prediction Error (Adjusted Rank)

City	Raw Rank	Adjusted Rank	True Score	Predicted Score
A	22	1	61.14	47.06
B	77	2	55.22	42.95
C	26	3	60.68	48.50
D	51	4	57.79	45.85
E	43	5	58.40	46.52
F	2	6	69.22	57.66
G	41	7	58.64	47.60
H	27	8	60.42	49.77
I	1	9	69.70	59.26
J	7	10	65.28	54.92

The lowest-scoring cities would also be adjusted using the benchmark. Table 3 shows the actual and adjusted rank for the lowest-scoring cities in the data set. There is

a less drastic movement among ranks for the lower scores. While there is minor rearrangement, the largest adjustment was to city X. It moved down from 15th worst score to the 3rd worst score. The benchmark adjustment predicts that city X would perform better than the overall mean, but it actually performed quite poorly. This could be useful information, indicating city X should further explore their citizens' low confidence in local government, as the variables in the model cannot account for it. For the most part, variables in the model are difficult for a city to influence, but poor performance against a benchmark score indicates there may be some factors the city can influence that are contributing to the low score. Opening a dialogue with city X citizens could identify some possibilities that could be addressed by local leadership.

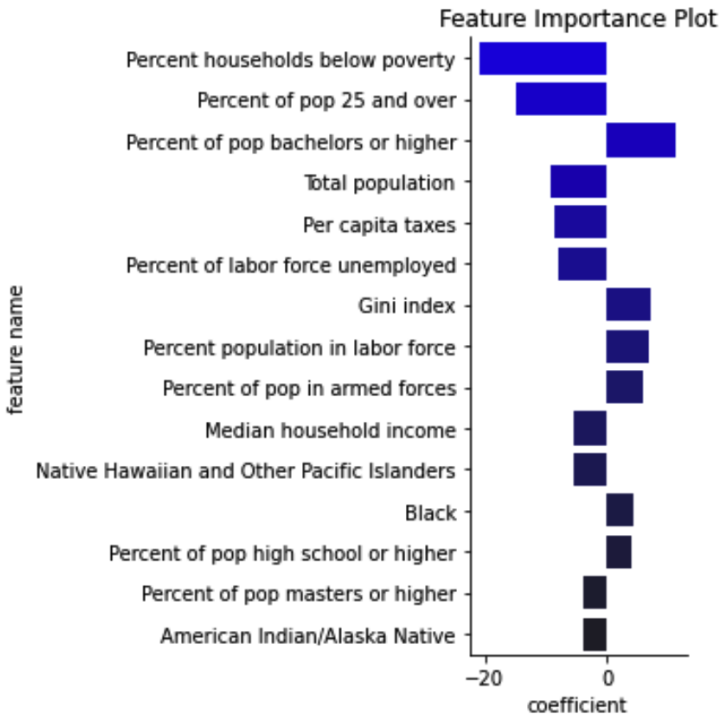
Table 3
Rank by True Score (Raw Rank) vs. Rank by Prediction Error (Adjusted Rank)

City	Raw Rank	Adjusted Rank	True Score	Predicted Score
Q	196	198	35.37	48.76
R	202	199	30.42	44.06
S	201	200	31.07	45.18
T	199	201	32.70	47.46
U	204	202	30.12	46.53
V	207	203	24.70	42.55
W	205	204	29.70	47.45
X	192	205	37.73	57.48
Y	203	206	30.37	57.48
Z	206	207	28.87	53.74

4.4 Feature Importance

Among the feature importance methods explored, ridge regression coefficients and permutation importance for Voting Regression stood out as most relevant and useful. Ridge regression coefficients were identified from the tuned ridge regression model. Figure 5 displays the 15 coefficients with the largest magnitudes from the ridge regression model. The top two features (percent of households below poverty and percent of population 25 and over) are the largest negative values. This means that an increase in either of these variables would be associated with a decrease in the confidence in government score. Conversely, the variable "percent of population with a bachelor's degree or higher" is positive, so an increase in this variable would be associated with an increase in the confidence in local government measure.

Figure 5
Ridge Regression Feature Importance Plot



A permutation feature importance method was employed to determine the important features in the Voting Regressor model. Table 4 lists the top 15 most important features, with a measurement of how the response variable changed when the feature was permuted.

Table 4
Voting Regressor Permutation Feature Importance

Feature	Importance (Mean)	2 S.D.
Percent of pop bachelors or higher	0.157	+/- 0.014
Median household income	0.144	+/- 0.014
Percent of households below poverty	0.137	+/- 0.011
Percent of labor force unemployed	0.111	+/- 0.008
Black	0.099	+/- 0.009
Percent of pop high school or higher	0.097	+/- 0.007
Gini index	0.094	+/- 0.007
Percent of pop 25 and over	0.091	+/- 0.007
Percent households owner occupied	0.085	+/- 0.004
Percent population in labor force	0.083	+/- 0.007
Average household size	0.081	+/- 0.005
White	0.076	+/- 0.004
Hispanic or Latino	0.076	+/- 0.005
Percent of pop masters or higher	0.073	+/- 0.004
Percent population over 16	0.070	+/- 0.004

An examination of ridge regression coefficients and Voting Regression permutation importance both contributed to conceptual understanding of the regression models and the associated benchmark adjustment. Although the tuned ridge regression was not the highest performing of the supervised models, the nature of ridge regression is highly interpretable. The imposed penalty term makes ridge Regression useful when features are correlated, and the ability to determine the positive or negative direction of association with the confidence in local government response provides valuable insight.

Some of the features deemed important by the Ridge model (Figure 5) can be influenced by government policies and initiatives. One focus of governments across the country is to help those below the poverty line. Another is to increase access to education, especially higher education. The Ridge model findings would suggest that these programs could be associated with an increase in confidence in local government. The percentage of a city's population in the armed forces is positively associated with confidence in government; however, this would be difficult for a local government to influence. This shows that, while there are many variables local leaders can improve to increase confidence in government, there will still be factors that are beyond their control.

Direct interpretation of coefficients in the Voting Regressor model was not possible, as it is an ensemble model. To explore feature importance in this model, the permutation method was employed. Permutation feature importance does not indicate the direction of change associated with a change in the feature, but it does measure the impact of each feature on prediction error. Table 4 lists the top 15 most important features in the Voting Regressor model. As expected, there was a great deal of overlap between the two feature importance methods.

While there is overlap, there are some interesting omissions when comparing the two methods as well. The Ridge model findings indicate per capita city taxes are a top five important feature, but the Voting Regressor findings do not indicate its importance. Interestingly, both models have racial factors contributing to confidence in government;

however, there is disagreement on the rank of importance when comparing the two methods.

The Gini index is a measurement of income inequality in a city. For instance, a city that had many high and many low-income values with only a few median values would have more disparity and thus a higher Gini index. Both the Ridge and Permutation feature importance models found the Gini index to be significant in determining overall confidence in government. However, the Ridge model goes against intuition by showing that a higher index is associated with a higher confidence in government score.

5 Discussion

5.1 Implications

Previously, local governments surveying their residents could only compare their raw scores to national average benchmarks, regardless of their population composition or resources. But as previously discussed, comparing cities using a new benchmark adjustment, rather than raw scores, could be more useful and meaningful, particularly if a city is under resourced and more diverse. To apply the new benchmark adjustment using the Voting Regression model, a city's out of fold prediction result would be compared to its actual score. A city that performed higher than predicted is overperforming while a city that performed lower than predicted would be underperforming. A margin of error can be incorporated so that only larger deviations from the prediction mean would be considered meaningful.

Ultimately, the features included in the cluster analysis and Voting Regressor models were chosen because they are variables that can be difficult to influence and were likely to have an association with confidence in government. By including these features in the model and associated baseline calculation, the new benchmark can indicate that *given all the features in the model*, the predicted score is how a city would be expected to perform. This takes into account factors known to be associated with resident satisfaction and well-being, such as high poverty levels or low education levels. If a city scores above their predicted score, it could indicate they are achieving high confidence in local government despite many factors generally associated with a lack of confidence. Their local policies and services might be effective. On the other hand, if a city scores below their predicted score, it could indicate their citizens lack confidence in local government despite potentially favorable circumstances.

Of course, cities aspiring to improve confidence in the local government could examine some of the more important features in the models. For instance, if a city contains a significant minority population associated with less government confidence, opening a dialogue with local leaders or running focus groups to solicit community grievances may provide critical information to improve trust. Scoring significantly above or below a city's predicted benchmark value might indicate something is influencing the score that is not a component in the model, and a deeper examination of current policies could be fruitful.

Other important features identified in the model can also be factors to consider for the local-level policy and budget decision-making process. For example, higher education level was associated with higher confidence in government, so new policies or budget focus on maintaining competitiveness in higher education for a city can result in higher confidence in the local government. Additionally, IAI score, which is a measure of quality and availability of local high-speed internet, was also associated with higher confidence in the government. This finding endorses the newly passed bipartisan Infrastructure Investment and Jobs Act (Infrastructure Investment and Jobs Act, 2021), which might help improve confidence in local government in the near future.

Context should be considered when interpreting important features from the benchmark adjustment model. While a higher Gini Index value was associated with higher confidence in government, this is not an indication that increasing income inequality would be a method of increasing confidence in government. This relationship may be due to the characteristics of the cities within the data set. There are very few cities with uniformly low income or massive income discrepancies that use The National Research Center at Polco's services. There are also very few large cities in the data set. If such cities were included, the expected association between higher Gini Index scores and lower confidence in government might emerge. However, until data from a wider range of cities is collected, this is speculation.

5.2 Limitations

The selected cities used for analysis were restricted to those cities that chose The National Research Center at Polco and the NCS to collect data. These cities were typically mid-sized, with populations ranging from 1,489 to 1,026,658. Huge cities collect data themselves, and tiny towns may not have the resources to devote to third-party data collection. Therefore, applying the benchmark adjustments to towns with populations outside of this range is extrapolation and should be done with caution.

While this analysis used survey data within a five-year window, the city and demographic information obtained from the U.S. Census ACS was collected in 2018. Any major changes to cities after 2018 would not be reflected in the city or demographic information used to build the models.

The Internet Access Index (IAI) was reported in individual census blocks within a county. To estimate the city's IAI score, the average was calculated for all blocks within the county, and this score was supplied as an estimate of the city's IAI score. However, this average did not consider the population within each block. Ideally with more time and resources, a weighted average by the population of each block could be calculated instead as a more accurate estimate for each city's IAI score.

The Voting Regressor (V.R.) and KMeans benchmark adjustments narrow the prediction range. Specifically, V.R. restricted the prediction range from 37.9 to 62.0, while the true score ranged from 24.7 to 69.7. KMeans restricted the prediction range from 31.1 to 59.1. Therefore, for the cities surveyed with the highest or lowest actual scores, their predicted score will automatically be lower or higher, respectively, and thus be over or underperforming. KMeans restricted the lower end less than V.R., but it restricted the higher end more than V.R. Both methods would adjust benchmarks

better for cities that are not extreme on the original scale. For cities with scores at the lowest ends, KMeans will adjust them less severely than V.R., thus could be slightly preferred for adjusting lower end extreme values. V.R. provides better adjustment for cities clustered together with fewer cities or cities with scores at the highest end of the original scoring. Both methods have limitations but can still be helpful.

It is difficult to interpret how Voting Regressor determines its final prediction score because Random Forest, Support Vector Regressor, and Gradient boosting are included as part of the model. If a city or client requires interpretation of their prediction results, ridge regression, which has reasonable prediction error, can be utilized instead. With ridge regression, parameter estimation can be extracted, and the calculation of the prediction can be easily interpreted.

5.3 Ethics

The inclusion of race indicators and other demographic information can cause ethical concerns. All information was obtained voluntarily from surveys conducted without coercion. Findings in this study show that a higher percentage of a certain race in the local population could be associated with lower confidence in government scores. According to the Ridge feature importance model, the result for this race was present as feature 11 out of the top 15 features but was not found to be as significant in the permutation test. Due to the nature of demographic data, it is difficult to separate all confounding variables from one another. This means that other features could be contributing to this result in the model. While this study aims to identify possible factors that local government officials can use to increase public confidence, it must be emphasized that correlation does not mean that causation was found.

Voluntary surveys can provide good insights into public opinion, but they also may have difficulty showing the whole picture. The surveys in this study contained data that individuals may feel unfavorably portrays them. This can lead to individuals refusing to complete a survey or not truthfully answering questions. Careful consideration was given to these constraints, and the data is felt to still be representative of the subject populations.

The manner in which the results are presented can contribute to positive or negative perceptions of a city. These results could have unintended ramifications for the subject cities without careful consideration. With this in mind, The National Research Center at Polco requested that individual city names be removed from the study. This removes researcher bias from the data and provides an extra layer of protection for the cities.

While the information gained from this study is intended to aid government officials in making fair and good policies, there are risks associated with the misuse of these findings. Policies that are made to target individuals that are part of any demographic negatively would be in stark contrast to the objectives of this study.

Concerns regarding Census data and its usage have been addressed in other studies. The previous use of racial markers obtained by the U.S. Census Bureau to discriminate against people of Japanese descent shows that misuse of data can happen. Other information contained in the Census should be handled with care as well. Items that may seem insignificant and harmless to one could be used for nefarious goals by others.

With this in mind, proper security measures were put in place for the data while at rest and while in transit. User access for this study's data has been restricted to only those who are stakeholders in this research and need access to the data source.

Furthermore, there are no notions that this study can take the place of government officials using financial advisors in terms of government spending. All financial and other decisions should be made in consultation with subject matter experts and other stakeholders.

5.4 Future Research

Future research can expand the current model to predict the other overall satisfaction measures in the NCS dataset, rather than just confidence in local government. Additionally, investigation of the key predictive features identified in the models and the ability of local governments to affect them could be explored.

Updated U.S. Census information could be applied to the methods in this study. As more current census data becomes available, it can be incorporated. Also, a better estimate for the Internet Access Index (IAI) could be incorporated by weighting each census block by the proportion of city population it contains.

6 Conclusion

A voting regression model was the most promising method of benchmark adjustment for NCS data. Using the predictions from this model, over and underperforming cities were identified. The most drastic changes in each city's performance assessment occurred for overperformers. According to the benchmark adjustment, most of the cities that performed in the top ten would have been considered average performers by raw scores. Using the benchmark model, identification of cities that have high confidence in government despite some factors that are difficult to influence is possible. While the primary goal of this study was to create more equitable and accurate benchmarks for confidence in government scores, the research also points to factors that local governments may focus on to increase confidence. While many of these factors are more difficult to change or require more non-traditional governance service approaches, they are nonetheless important considerations in local government trust building.

Acknowledgments. We would like to thank Polco for their generosity with their time and data and their willingness to collaborate with us on this project. We would also like to thank Dr. Robert Slater for his support as our advisor and Dr. Jacquelyn Cheun for her assistance and guidance with the writing process. Finally, we want to thank our loved ones for their patience with us through this process, their time spent listening to us brainstorm and discuss our project, and for being a very forgiving practice audience.

References

1. Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). *On the surprising behavior of distance metrics in high dimensional space*. Springer Berlin Heidelberg. 10.1007/3-540-44503-X_27

2. Alexander, M. R., Hyde, I., Burdi, C., Smith, B., Bergerson, J., Riddle, M., … Hutchison, J. (2021). *Internet Access Index* [White paper]. Argonne National Laboratory. https://www.anl.gov/dis/reference/internet-access-index

3. Barrington-Leigh, C., & Wollenberg, J. T. (2019). Informing Policy Priorities using Inference from Life Satisfaction Responses in a Large Community Survey. *Applied Research in Quality of Life, 14*(4), 911-924. 10.1007/s11482-018-9629-9

4. Campello, Ricardo J. G. B., Moulavi, D., Zimek, A., & Sander, J. ö. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data, 10*(1), 1-51. 10.1145/2733381

5. Diener, E. (2013). The Remarkable Changes in the Science of Subjective Well-Being. *Perspectives on Psychological Science; Perspect Psychol Sci, 8*(6), 663-666. 10.1177/1745691613507583

6. Dolan, P., & White, M. P. (2007). How Can Measures of Subjective Well-Being Be Used to Inform Public Policy? *Perspectives on Psychological Science; Perspect Psychol Sci, 2*(1), 71-85. 10.1111/j.1745-6916.2007.00030.x

7. Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM, 55*(10), 78-87. 10.1145/2347736.2347755

8. Greenlaw, C., & Brown-Welty, S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods: Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review; Eval Rev, 33*(5), 464-480. 10.1177/0193841X09340214

9. Infrastructure Investment and Jobs Act, 23 U.S.C. § 101 (2021). https://www.congress.gov/bill/117th-congress/house-bill/3684/text

10. Jolliffe, I. T. (2002). *Principal component analysis.* Springer, New York, NY. doi:10.1007/b98835

11. Khamsi, M. A., & Kirk, W. A. (2001). *An introduction to metric spaces and fixed point theory*. John Wiley & Sons, Inc. doi:10.1002/9781118033074

12. Lawless, N. M., & Lucas, R. E. (2011). Predictors of Regional Well-Being: A County Level Analysis. *Social Indicators Research, 101*(3), 341-357. 10.1007/s11205-010-9667-7

13. Miller, T. I., & Kobayashi, M.M. (2000). *Citizen surveys: How to do them, how to use them, what they mean.* Washington, D.C.: International City/Community Management Association.

14. National Community Survey (2022, February 17). https://info.polco.us/the-national-community-survey

15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B.,...Duchesney, E. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research, 12,* 2825-2830.

16. Pew Research Center. (2012, May15). *Assessing the Representativeness of Public Opinion Surveys* [Report]. https://www.pewresearch.org/politics/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/

17. Poister, T. H., & Streib, G. (1999). Performance Measurement in Municipal Government: Assessing the State of the Practice. *Public Administration Review, 59*(4), 325-335. 10.2307/3110115

18. Rentfrow, P. J., Mellander, C., & Florida, R. (2009). Happy States of America: A state-level analysis of psychological, economic, and social well-being. *Journal of Research in Personality, 43*(6), 1073-1082. 10.1016/j.jrp.2009.08.005

19. Sawicki, D. S., & Craig, W. J. (1996). The Democratization of Data: Bridging the Gap for Community Groups. *Journal of the American Planning Association, 62*(4), 512-523. 10.1080/01944369608975715

20. Schubert, E., Sander, J. ö, Ester, M., Kriegel, H., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems, 42*(3), 1-21. 10.1145/3068335

21. Sidorchuk, R., Skorobogatykh, I., Mkhitaryan, S., Voronova, T., & Ivashkova, N. (2020). Clustering Megacity Districts upon Customer Satisfaction on Parking Services. *Montenegrin Journal of Economics, 16*(1), 69-86. 10.14254/1800-5845/2020.16-1.5

22. Sirgy, M. J., Rahtz, D. R., Cicic, M., & Underwood, R. (2000). A Method for Assessing Residents' Satisfaction with Community-Based Services: A Quality-of-Life Perspective. *Social Indicators Research, 49*(3), 279-316. 10.1023/A:1006990718673
The goal of this study was to

23. United States Census Bureau Gini Index (2022, February 22). https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/gini-index.html

24. Van Ryzin, G.,G. (2013). An Experimental Test of the Expectancy-Disconfirmation Theory of Citizen Satisfaction. *Journal of Policy Analysis and Management; J.Pol.Anal.Manage, 32*(3), 597-614. 10.1002/pam.21702

25. Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*(301), 236-244.

26. Zarate, Juan Manuel Ortiz. (n.d.). Ensemble Methods: The Kaggle Machine Learning Champion. https://www.toptal.com/machine-learning/ensemble-methods-kaggle-machine-learn