International Journal of Geo-Information

MDPI

Article

A Cluster-Based Machine Learning Ensemble Approach for Geospatial Data: Estimation of Health Insurance Status in Missouri

Erik Mueller [1,*], J. S. Onésimo Sandoval [2], Srikanth Mudigonda [3] and Michael Elliott [4]

1 Integrated & Applied Sciences: Bioinformatics & Geospatial Biology, College of Arts and Sciences, Saint Louis University, St. Louis, MO 63103, USA
2 Department of Sociology and Anthropology, College of Arts and Sciences, Saint Louis University, St. Louis, MO 63103, USA; ness.sandoval@slu.edu
3 School for Professional Studies, Saint Louis University, St. Louis, MO 63103, USA; srikanth.mudigonda@slu.edu
4 Department of Epidemiology and Biostatistics, College for Public Health and Social Justice, Saint Louis University, St. Louis, MO 63103, USA; michael.elliott@slu.edu
* Correspondence: Erik.Mueller@slu.edu; Tel.: +1-314-703-5219

Received: 12 October 2018; Accepted: 20 December 2018; Published: 28 December 2018



Abstract: Mainstream machine learning approaches to predictive analytics consistently prove their ability to perform well using a variety of datasets, although the task of identifying an optimally-performing machine learning approach for any given dataset becomes much less intuitive. Methods such as ensemble and transformation modeling have been developed to improve upon individual base learners and datasets with large degrees of variance. Despite the increased generalizability and flexibility of ensemble approaches, the cost often involves sacrificing inference for predictive ability. This paper introduces an alternative approach to ensemble modeling, combining the predictive ability of an ensemble framework with localized model construction through the incorporation of cluster analysis as a pre-processing technique. The workflow not only outperforms independent base learners and comparative ensemble methods, but also preserves local inferential capability by manipulating cluster parameters and maintaining interpretable relative importance values and non-transformed coefficients for the overall consideration of variable importance. This paper demonstrates the ensemble technique on a dataset to estimate rates of health insurance coverage across the state of Missouri, where the cluster pre-processing assists in understanding both local and global variable importance and interactions when predicting high concentration areas of low health insurance coverage based on demographic, socioeconomic, and geospatial variables.

Keywords: ensemble modeling; machine learning; population health; health insurance status; spatial statistics; variable clustering

1. Introduction

The ability to simultaneously model and analyze both local and global relationships in geospatially-referenced statistical models is a challenge commonly faced by social science researchers, particularly geographers and spatial statisticians. Methods common to social scientists, such as linear and logistic regression, are limited in that they only model global relationships—entire datasets—and produce results that can only be attributed to the dataset as a whole. Other methods, such as clustering techniques, are often used to identify concentrations or "hotspots" within or between variables in a dataset, but fail to provide information about intervariable relationships or dependencies, which can

be accomplished through regression analysis [1]. As a result, researchers, in many cases, are forced to make a choice between modeling data locally or globally.

The field of educational assessment has made some important advances in techniques that simultaneously measure global relationships, while also accounting for localized trends and patterns, through the incorporation of cluster analysis as a pre-processing technique. Trivedi, Pardos, Sarkozy, and Heffernan [2–4] have explored the use of clustering as a pre-processing technique to divide data into homogenous regions, producing multiple localized models that are combined to form a global structure with improved performance over models without the cluster-based pre-processing. More specifically, Trivedi et al. [3] outlined a cluster-based method for determining whether receiving assistance on assessments resulted in increased student performance, using student information to divide students into clusters for improved modeling performance. The authors found that using the additional student information to cluster into more homogenous student groups significantly outperformed the other global models that were explored in the study. Moreover, a later publication by Trivedi et al. [2] discussed in greater detail the utility of clustering to reduce error in prediction tasks, citing that the use of clustering as a pre-processing measure improves the bias–variance tradeoff, as the cluster-based methods in conjunction with regression enable greater access to more variance in the data. Building on the work of Trivedi et al. [2–4] this study makes a unique contribution to the literature by using the clustering method as a pre-processing technique to geospatial data and regression techniques to examine global dataset trends, while also inferring local intervariable relationships.

One of the principal limitations in direct translation of the Trivedi et al. [3] research into a geospatial framework is the regression method that was chosen, which followed cluster-based pre-processing: linear regression. While linear regression may have been suitable for educational assessment, geospatial datasets often violate the statistical assumptions that are required to perform linear regression, such as multicollinearity and spatial autocorrelation [1]. To account for nonlinear relationships and datasets that potentially violate the statistical assumptions associated with linear regression, this study will employ a machine learning ensemble framework, drawing upon well-established individual machine learning methods, or "base learners" which will be evaluated and aggregated to form a dynamic learning algorithm that can provide reliable and stable performance regardless of nonlinearity, intervariable relationships, or the distribution of the dataset under study.

Justification for adapting Trivedi et al. [3] to a machine learning framework stems from the "no free lunch" theorem, which suggests that for any machine learning algorithm, elevated performance over one class of problems is offset by performance over another class [5]. In practical terms, this suggests that within a given domain or for any particular dataset, there is no single machine learning algorithm or regression approach that will always result in the most accurate learner, provided varying hyperparameters [6]. At the most basic level, each algorithm comes with a unique set of statistical assumptions that must be satisfied in order for an algorithm to reach optimal performance. The violation of learner assumptions leads to inductive bias, reducing performance and potentially allowing for another algorithm, with a different set of assumptions, to outperform the learner where assumptions have not held [6]. Intuitively, this argument leads to the conclusion that by combining multiple base learners, modeling accuracy and external validity will be improved by leveraging multiple models against a single domain or dataset [7]. By employing a clustering method in combination with an ensemble approach, we hypothesize that the cluster-based method will outperform other comparative global ensemble models, while preserving external validity and inference by the extraction of information from the base learners.

The central research questions that this study addresses are: (1) whether the utility of clustering methods as a pre-processing technique is appropriate and effective within a geospatial context; and (2) whether the cluster-based machine learning ensemble algorithm outperforms similar machine learning approaches that do not use clustering methods. Additionally, this study will also examine the extent to which inference can be drawn from the localized dataset clusters that are determined as part of the proposed approach. To answer these research questions, we will use a direct experimentation on a

geospatial dataset, which relates demographic and socioeconomic indicators to an estimation of the health insurance status for the state of Missouri. We identified the variables that most significantly impact the number of individuals without health insurance across the state of Missouri, and measured disparities in health insurance status among targeted sub-populations. The analytic outcome using the method we are proposing will identify the variables that most significantly account for the variance in health insurance coverage across the state. When used for clustering, these variables most effectively homogenize the dataset, highlighting differences in the relative variable contribution to health insurance status throughout Missouri.

2. Materials and Methods

2.1. Data Overview and Base Learner Selection

The dataset selected for the analysis presented in this paper originated from the American Community Survey (ACS) 2012–2016 five-year estimates [8], gathered by the United States (U.S.) Census Bureau. The ACS is an ongoing survey that is collected yearly throughout the United States, and is designed to be an extension of the decennial census. The decennial census only collects information on general population characteristics and housing [8]. In contrast to the decennial census, the ACS collects a wealth of demographic, socioeconomic, housing, and business information to assist in resource allocation by the federal government. ACS data are collected at hierarchical units of geography, and estimates are generated in one, three, and five-year estimates, ideally reflecting the population characteristics under study [8].

For this study, the selected dependent variable is the estimated count of individuals without health insurance aggregated by census block groups. Based on the nature of the dependent variable, all of the models that were developed for this study were run as regression models. The extent of the study area includes the entire state of Missouri, which is divided into 4506 block groups. The independent variables that were selected for the study were chosen based on a taxonomy developed by Juarez et al. [9] known as the "public health exposome". According to Juarez et al., public health issues are principally impacted by elements within the natural, built, social, and policy environments. Using those four domains to guide variable selection and adapting them based on the available geospatial datasets, the variables that were selected for this study were divided into four categories: (1) demography; (2) socioeconomic indicators; (3) geospatial characteristics; and (4) assessments for healthcare access. Variables in the demographic and socioeconomic categories were selected from the ACS, while geospatial characteristics and assessments for healthcare access were primarily derived from a network analysis using ArcGIS 10.5, which calculated distance measurements from the block group centroid locations to various points of interest. These calculated variables included distances to medical facilities, specifically designated shortage areas and government identified medically underserved areas (MUAs), educational facilities including institutes of higher education, and the overall length of roads and highways within each block group [8,10–12]. To account for variables of different types, including counts, distances, and averages, all of the variables were standardized so that they could collectively be examined for relative variable contribution in the modeling results.

Table 1 contains descriptions of all of the independent variables that were included in the analysis, which are grouped by variable category and variable type, and include both descriptive statistics and sourcing information. The variable selection for this paper was based on the idea that despite efforts to make health insurance more accessible by government and nonprofit entities, it remains a significant socioeconomic indicator, as evidenced by research suggesting continued disparities between health insurance status and socioeconomic indicators [13–16]. Viewing health insurance status as a socioeconomic indicator, it's appropriate from a theoretical standpoint to include demographics and other socioeconomic indicators as predictors, as well as other variables that lend insight to the level of access available across geographic space [17,18]. The intuition behind the variable categories is that only

those individuals who can afford insurance will be able to obtain it, and if individuals can't physically access healthcare, they have no incentive to retain it [19]. Further, working under the assumption that demography tends to cluster across space as evidenced by Tobler's Law, health insurance status is also likely to cluster across space, justifying the inclusion of geospatial and demographic characteristics [20].

Table 1. Variable Categories and Sources of Data.

Variable Name	Type	Min	Max	Mean
Dependent Variable [8]				
Variable Insurance Status—No Health Insurance [1]	Count	0	1402	149.43
Demographic Variables [8]				
Population Total	Count	0	9698	1344.80
Total Housing Units	Count	0	298	55.41
Total Male Population	Count	0	6668	659.86
Total Female Population	Count	0	4681	684.94
Race: White	Count	0	9061	1109.83
Race: African American	Count	0	3341	155.77
Race: Other	Count	0	1591	79.20
Median Age	Average	10.50	85.30	40.07
Language: English Only	Count	0	2933	490.72
Language: Other	Count	0	495	35.77
Vacant Housing Units	Count	0	2933	81.32
Socioeconomic Variables [8]				
Food Stamp/SNAP Benefits[2]	Count	0	519	68.44
No Vehicle Available	Count	0	591	38.39
Income: Less than 10,000	Count	0	575	39.63
Income: 10,001–30,000	Count	0	694	118.08
Income: 30,001–60,000	Count	0	763	151.59
Income: 60,001–100,000	Count	0	910	117.39
Income: Greater than 100,000	Count	0	1296	99.80
Education: No School Completed	Count	0	156	7.38
Education: High School Diploma or GED[3]	Count	0	1757	281.36
Education: Associates Degree	Count	0	552	68.26
Education: Bachelor's Degree	Count	0	1775	155.76
Education: Graduate Degree	Count	0	1172	94.05
Under the Federal Poverty Line	Count	0	866	77.04
Employment: Unemployed	Count	0	373	44.42
Employment: Not in Labor Force	Count	0	3607	392.41
Geospatial Variables—Distance and Mobility [8,10–12]				
Distance to Educational Facilities	Distance (km)	0	23.68	2.20
Distance to Higher Education Facilities	Distance (km)	0	62.57	11.48
Distance to Public Education Facilities	Distance (km)	0.00	23.68	2.33
Distance to Private Education Facilities	Distance (km)	0.01	47.98	5.49
Distance to Hospitals	Distance (km)	0.02	49.35	7.41
Geomobility: Same House One Year	Count	0	8452	833.96
Geomobility: Different House One Year	Count	0	2576	151.38
Length of Roads in Census Block Group	Length (m)	989.20	731,683.97	63,814.56
Length of Highways in Census Block Group	Length (m)	0	4.24	0.18
Medical Access Variables [12]				
Distance to Medical Shortage Area	Distance (km)	0.06	60.32	11.41
Medically Underserved Area (MUA)	Binary	0	1	0.46
Average MUA Score	Average	0	62	24.27

[1] Dependent Variable. [2] SNAP—Supplemental Nutrition Assistance Program. [3] GED—General Educational Development Certificate.

The next step in our study was to decide which machine learning algorithms to include as base learners in an ensemble; the largest consideration for learner-selection might be the concept of diversity. In the machine learning literature, diversity of base learners is the idea that while each learner does not necessarily need to be globally accurate individually, it must be highly accurate in the subregions of the hyperspace spanned by the predictors, outperforming each other on various subregions of the hyperspace of observations being analyzed [6]. The base learners that we used in our ensemble were selected based on their overall simplicity, their ability to infer intervariable relationships, their ability to compute varying statistical and mathematical properties, and the validity of assumptions that need to be met in the ensemble of methods. Our ensemble included seven base learners grouped into four categories based on the class of learner. Table 2 contains descriptions of each base learning technique, the category it represents within the ensemble framework, and references to the packages in the R statistical environment where the specific algorithms have been implemented.

Table 2. Categories of Machine Learning Base Learners.

Category	Base Learner	Algorithm Reference
Support Vector Regression	Support Vector Machine	e1071 [21]
Decision Tree	Random Forest	randomForest [22]
Based Regression	Boosted Tree	gbm [23]
Penalty-Based Linear	Lasso Regression	glmnet [24]
Regression	Ridge Regression	glmnet [24]
Dimension Reduction	Principal Component	pls [25]
Regression	Partial Least Squares	pls [25]

Only one of the base learners representing each learning category will advance to inclusion in the ensemble model. The ensemble algorithm begins by using each base learner to fit a global model, and then the better-performing global base learner, by category, advances to the learning category level for eventual aggregation. This method produces a diverse ensemble and avoids adding redundancy to the ensemble by including multiple base learners that essentially represent the same class of machine learning models. For example, given that lasso and ridge regression are both penalty-based linear modeling techniques, the better performing of the two models will advance to represent the penalty-based learning category for inclusion in the ensemble [6].

2.2. Statistical Methods

To illustrate the workflow associated with the cluster-based ensemble algorithm presented in this study, Figure 1 offers a generalized procedural diagram outlining the steps involved in executing the ensemble from the input to output datasets. The procedure is divided into five major steps, beginning with cluster analysis as a pre-processing technique, working through base learner aggregation, learning category aggregation, and cluster assembly or aggregation, noting that step four involves repeating the ensemble procedure on each of the four clusters derived from the original dataset.

The cluster-based ensemble algorithm begins by running a cluster analysis on the original input dataset, dividing the dataset into four clusters based on a chosen independent variable that is equal to the number of learning categories. Justification for clustering lies in the idea that for a regression problem, by running the ensemble on N clusters representing homogenous subregions within a global dataset, there will be less variation within each cluster, which will therefore result in each learner fitting a more tailored, accurate model that is local to each specific subregion in the dataset. Additionally, we have applied 10-fold cross-validation (90%/10% approach to training and testing) for each base learner in combination with aggregating cluster results to the ensemble level in order to reduce the possibility of overfitting at the cluster level [6,26]. Rather than summarizing variable importance and explanatory power across the entire dataset at the global (entire dataset) level, in our approach, interpretation and inference will instead happen at the more localized, cluster level.

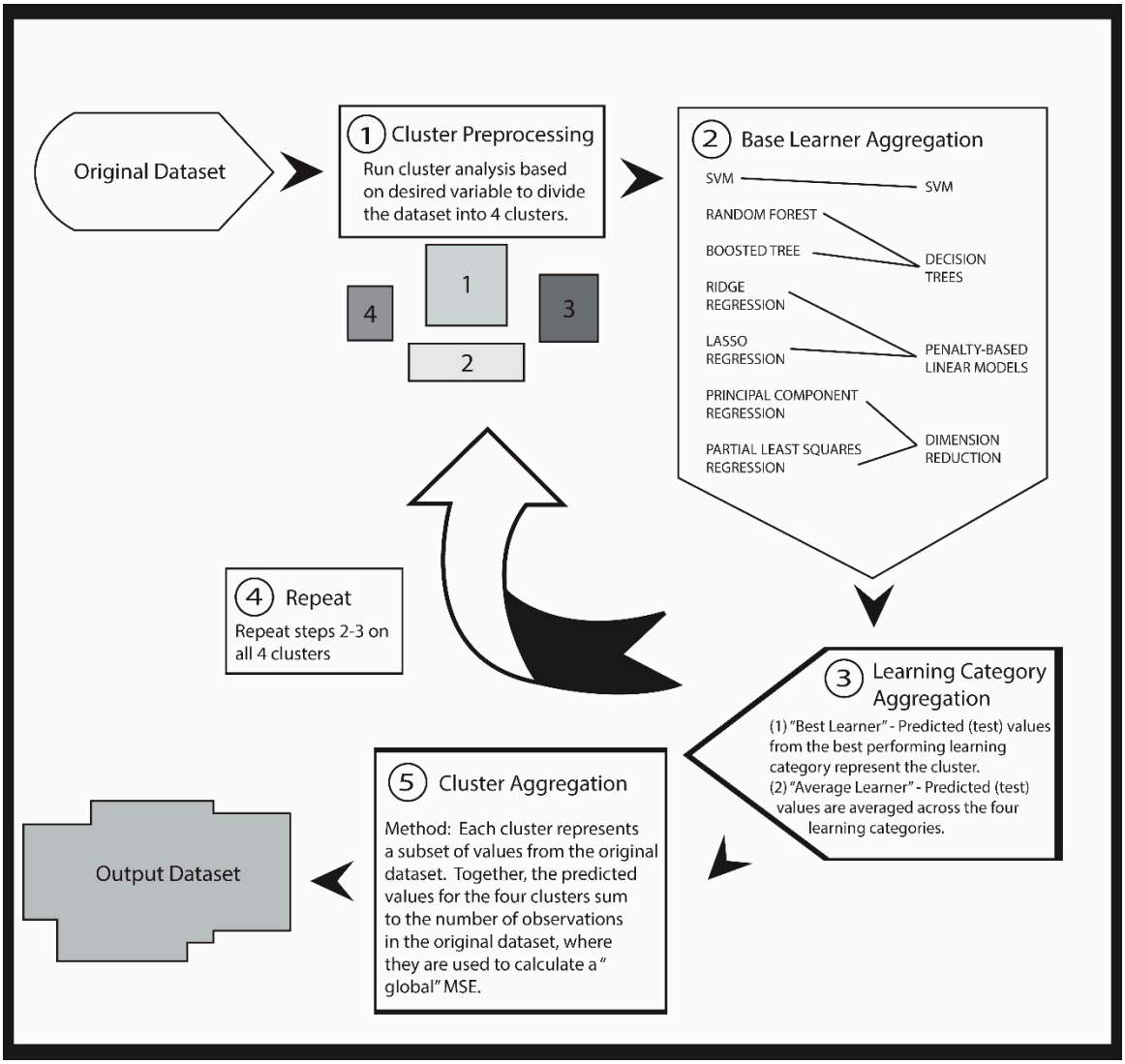


Figure 1. Cluster-Based Ensemble Procedural Outline.

In this study, clustering will be performed using the k-means approach, partitioning 'n' observations into 'k' clusters where each observation belongs to the nearest mean, which is a descriptor of the cluster [26,27]. This process essentially splits the data space into Voronoi cells, which is a method that is commonly used in geospatial statistics to partition a plane into regions based on point distances to plane subsets [28]. The k-means approach was chosen for this study due to its simple clustering method to reduce the space into disjointed subregions, as well as the relatively few hyperparameters that are required to process results [26]. In the hyperparameter assignment step, the algorithm will take in all of the dataset observations and divide the dataset into four clusters, which are equal to the number of machine learning class categories represented in the ensemble framework. Choosing an optimal number of clusters for a given dataset is a complex problem in unsupervised learning [1]. Therefore, four clusters were chosen solely for the sake of simplicity. Alternative methods for optimal cluster assignment were explored via the use of the "NBClust" library in R. However, due to wide disagreement of the optimal number of clusters, among the approximately 30 indices computed by the NBClust function, it was decided to make the number of clusters a pre-determined value that was equal to the number of categories (four) of the learners included in the ensemble [29].

We did consider other values for determining an optimal number of clusters in order to assess the impact of cluster values on the overall model performance. We found that models with fewer than four clusters had a progressive increase in mean square error (MSE), which was likely due to the

increased heterogeneity in the larger clusters. In contrast, increasing the number of clusters led to a greater variance in MSE. In many cases, the results obtained were questionable due to the progressively smaller number of observations being used per cluster as the number of clusters increased beyond six. These results indicate that while choosing a static cluster size equal to the number of base learners is feasible for this study, for datasets whose sizes are in other orders of magnitude, the number of clusters may need to be reassessed in order to ensure optimal model performance.

In the second stage of our approach, which involves the aggregation of results from base learners (step two in Figure 1), there are two themes for composing an ensemble: multi-expert combination, and multi-stage combination [6]. In multi-expert combination, the base learners work in parallel; in multi-stage combination, simple base learners attempt to fit the model individually, and if they fail to reach a desired level of accuracy, more complex learners are employed to attempt to reach the desired accuracy. The main argument for the multi-stage approach is that it minimizes the required computational power; if a model can be fitted using a simple approach, then there isn't a need to fit the models using more complex learners. In contrast, when computational power is not a constraint, such as when working with a relatively small dataset, the multi-expert method may be preferred, because it will likely result in a more accurate model compared to the multi-stage technique, since all of the possible base learners will be evaluated compared to a more bracketed or selective approach.

For this study, the cluster-based ensemble uses a multi-expert approach due to the size of the dataset and desire to reach the highest possible accuracy. Within the multi-expert combination schema, there are two approaches for learner aggregation that should be considered when developing an ensemble model. The first, or a global approach, uses all of the learners that produce an output in aggregation via techniques such as voting or stacking. In contrast, a local approach analyzes the output from each learner, and chooses a subset of the learners to be responsible for aggregation in the ensemble [6]. Provided that only one learner from each class of machine learning approaches will advance to represent each learning category, the cluster-based ensemble uses the local approach to ensemble aggregation. The multi-expert schema followed by a local approach to aggregation is evidenced in Figure 1, step two, where all seven of the base learning algorithms are assessed, and the best-performing learning algorithm in each category advances to represent the learning category in step three, learning category aggregation.

For mathematical aggregation, the ensemble models that will be used to compare performance against the cluster-based ensemble use three well-known approaches to ensemble aggregation: the average approach, the globally weighted average approach, and the minimum residual approach. The three approaches will serve as the standards for model assessment for whether the cluster-based method improves on similar global models. The average approach aggregates by taking the predicted observation of each winning base learner by class, averaging the four observations, and producing an average predicted value by observation, which is then used to calculate the MSE compared to the actual ground truth data. The globally weighted average approach is a direct extension of the average approach, where the observation for each of the four base learners is taken by class, weighted by the minimum–maximum (0–1) scaled global MSE for each learner, and then summed to obtain the weighted average predicted value. Finally, the minimum residual approach selects the estimated value across the four learners with the smallest residual to represent the observation in the ensemble [5].

The cluster-based ensemble method will use two relatively straightforward approaches to learner aggregation once four base learners have advanced to represent their respective learning categories. Figure 1, step three, summarizes the two approaches: the "best learner" approach, and the "average learner" approach. Relating to the multi-expert combination schema, the only difference between the "best learner" and "average learner" approaches is the method for learner aggregation: the "best learner" approach uses only the best of the four learning categories (in terms of performance) to represent the given cluster, where the "average learner" approach averages the four learning categories.

Following Figure 1, step four, which involves running base learner and learning category aggregation on each of the four clusters, step five involves combining the results from each of

the four clusters into a single dataset. Since the dataset was divided at the beginning into four clusters, and records within clusters persisted throughout steps two through four, retaining information referencing the original observation, step five involves appending each of the clusters into a new output dataset. MSE information by the base learning algorithm, learning category algorithm, and cluster are retained throughout the execution of the algorithm, where in step five, the estimated observations are compared against the original dataset, computing an overall MSE value representing the global performance of the cluster-based ensemble.

3. Results

As discussed previously, the cluster-based ensemble approach uses a multi-expert localized learner combination approach to determine which base algorithms to include in the final ensemble. For purposes of model evaluation, mean square error (MSE) is used to assess the relative model performance, where lower MSE values indicate improved modeling performance. Table 3 identifies each base learner and learning category, along with the MSE for each individual base learner, validated with 10-fold cross-validation followed by fitting on the entire dataset. Table 3 indicates that the tuned Support Vector Machine (SVM) model significantly outperforms all of the other base learners by an approximately 0.16 change (decrease) in scaled MSE. Additionally, following the support vector regression category, the next best-performing category consisted of dimension reduction methods, with principal components regression (PCR) and partial least squares (PLS) producing the next two smallest MSE values, respectively (0.487, 0.488). Considering the 39 variables in the dataset, the PLS and PCR methods performed well, presumably because reducing the dimensions to variance-concentrated components resulted in increased model performance. The last two machine learning categories performed roughly equally: the lasso and ridge regression models slightly outperformed the random forest and boosted tree-based models.

Table 3. Base Learner Performance. PCR: principal components regression, PLS: partial least squares, RF: Random Forest, SVM: Support Vector Machine.

Category	Model	MSE (Rank)
Support Vector Reg.	SVM $_1$	0.325 (1)
Decision Tree	RF $_1$	0.495 (5)
	Boosted	0.499 (7)
Pseudo-Linear	Ridge $_1$	0.497 (6)
	Lasso	0.488 (4)
Dimension Reduction	PCR $_1$	0.487 (2)
	PLS	0.488 (3)

$_1$ = Represents machine-learning class/category in the cluster-based global ensemble.

For the comparative ensemble models, base learners were aggregated according to the average predicted observation, the weighted-average observation, or the minimum residual method for comparison, as previously outlined in the Methods section. Table 4 is a summary of the model performance for each global ensemble aggregation approach to be compared against the cluster-based ensemble technique. From the results shown in Table 4, we find that the weighted-average approach to learner aggregation yielded a lower MSE (0.439) compared to the non-weighted-average approach (0.449), excluding the minimum residual approach, which should only be used as a reference to assess the optimal model performance in a selective aggregation scenario. A realistic expectation for ensemble performance should fall somewhere between the average aggregation (0.449) and minimum residual approach (0.267), though as close to the latter as possible. While many other methods have proven effective for global learner aggregation, potentially pushing the global MSE closer to the gold standard minimum residual MSE, the focus of this study is to compare the global and cluster-based ensemble approaches; therefore, the methods listed in Table 4 will be sufficient for comparison.

Table 4. Comparative Global Ensemble Model Performance. MSE: mean square error.

Approach	MSE (Rank)
Average MSE	0.449 (2)
Weighted-average MSE	0.439 (1)
Minimum-residual	0.267 (B)

B = Baseline.

The results of the cluster-based ensemble algorithm displayed significant performance improvements over both comparative ensemble techniques and individual base learners. Table 5 contains the results of the top performing independent variables that were chosen for cluster-based pre-processing, along with the aggregation approach that was used to compute ensemble performance. The table provides information for the top independent variables that, when used for the basis of cluster analysis, optimized the cluster-based ensemble model performance. Referencing back to the comparative ensemble MSE values from Table 4, regardless of the independent variable used for cluster-based pre-processing, the cluster-based ensemble approach outperformed the comparative global ensembles in every circumstance, except for the minimum residual aggregation method, as expected. In addition, there are significant similarities in the top five variables between the two cluster-based ensemble aggregation methods. Income, education, and distance from roads/highways were among the top performing cluster variables in both aggregation methods, leading to the conclusion that there are meaningful differences between clusters for those particular variables, resulting in varying model performance among the base learning algorithms.

Table 5. Cluster-Based Ensemble Model Performance.

Method: Average		
Variable	***MSE***	***Range***
Highway Length	0.378	0.250–0.454
Income Greater 100K	0.396	0.279–0.478
All Roads Length	0.418	0.279–0.548
Graduate Degree	0.420	0.376–0.493
Edu. Facility Distance	0.428	0.320–0.547
Method: Best Base Learner by Cluster		
Variable	***MSE***	***Range***
Hospital Distance	0.293	0.280–0.389
Bachelor's Degree	0.294	0.223–0.379
Higher Edu. Distance	0.302	0.242–0.461
Graduate Degree	0.304	0.269–0.360
Income Greater 100K	0.304	0.200–0.323

4. Discussion

4.1. Interpretation of Study Findings

The purpose of this paper was to introduce an ensemble modeling approach that localizes model fitting by splitting a dataset into homogenous subregions through the application of cluster analysis as a pre-processing step. Table 6 contains a summary of the study results and provides measures of model performance across base learners, the comparative global ensemble techniques, and aggregation measures, and the cluster-based ensemble technique and aggregation measures with the top two independent variables used for cluster-based pre-processing.

Table 6. Best-Performing Models by Category Summary.

Method	Characteristics	MSE (Rank)
Base Learner	SVM	0.325 (2)
Global Ensemble	Weighted Avg.	0.439 (4)
Global Ensemble	Min. Residual	0.267 (B)
Cluster Ensemble $_1$	Average	0.378 (3)
Cluster Ensemble $_2$	Best Learner	0.293 (1)

B = Baseline. $_1$ = Independent Variable: Highway Length. $_2$ = Independent Variable: Hospital Distance.

Overall, based on the study findings, the cluster-based ensemble algorithm aggregated by "best learner" outperformed all of other individual base learning algorithms and comparative ensemble approaches regardless of aggregation technique. This suggests that a cluster-based approach to ensemble learning may ultimately improve model performance, while at the same time maintaining increased external validity and generalizability, both of which are inherent strengths of ensemble learning. The findings from this study are consistent with those found in Trivedi et al. [2,3]. They found that cluster-based approaches to dataset division prior to model fitting may improve global performance when the cluster-based models are aggregated upon being compared to singular global models. Further, the findings of this study suggest that while cluster-based pre-processing methods have yielded superior performance on nonspatial datasets, similar procedures can also be applied to geospatial datasets with improved model performance.

The main objective for the cluster-based ensemble algorithm was to maximize the predictive ability while preserving the highest inferential capability possible. However, in order to draw any inferences, model variables required a theoretical basis for inclusion. In this study, the theoretical basis was satisfied through the inclusion of the "public health exposome" framework developed by Juarez et al. [9]. Although interpretation of cluster-based models is complex, inference is still possible, especially when using the "best learner by cluster" approach. The results from this study indicate that the cluster-based ensemble performed optimally when the *distances to hospitals* variable was used as the basis for clustering. Consequently, those modeling parameters were used for model interpretation and to demonstrate inferential capability throughout the paper.

From the results shown in Table 7, PCA was the learner used to fit the cluster with the highest mean value for the *distances to hospitals* clustering variable. The first cluster represented areas (block groups) farthest away from hospitals and medical facilities, which are primarily rural areas. The SVM method was used to fit clusters two, three, and four, implying that it outperformed the other base learners in suburban and urban areas where hospitals are relatively near. Although PCA was the best-performing learner on the first cluster, it also achieved the highest MSE, or error rate, among the other three clusters, meaning that it might be difficult to model the health insurance status for rural areas, or possibly that the first cluster contains the most variance among the four clusters in terms of the dependent variable under study.

Table 7. Independent Variable for Clustering: Distance to Hospitals Cluster Summary. PCA: principal component analysis.

Independent Variable: Hospital Distance			
Cluster	***Mean***	***Best Algorithm***	***MSE***
1	3.026	PCA/PLS	0.388
2	0.334	SVM	0.321
3	−0.592	SVM	0.280
4	1.449	SVM	0.295
Total			0.293

A comparison of relative variable importance among the four clusters found several consistencies in terms of variable significance across the four clusters. Firstly, we can infer relative variable significance in our models due to the z-scaling of the entire dataset prior to model execution (global and cluster-based ensembles). From Table 8, we can conclude that, based on the best-performing base learner for the first cluster (PCA), *race*, *language spoken at home*, and *income* seem to be the most significant variables based on model coefficients. We can further infer that higher values of the number of uninsured persons are associated with lower values of the number of white inhabitants (per block group), along with the number of individuals making over $100,000 per year, and to a lesser degree, the number of African American inhabitants.

Table 8. Cluster-Based Model Interpretation, by Cluster Using Best Base Learner Aggregation Method.

Variable	Importance Measure
Cluster: 1 (M: 3.206)	***PCA (Coefficients)***
Race White	9.22
Language English	7.51
Males Total	−6.69
Income GT 100K	−3.51
Race African Am.	−3.47
Cluster: 2 (M: 0.334)	***SVM (Variance)***
Race White	0.069
Income 10–30K	0.060
Females Total	0.057
Population Total	0.055
Race African Am.	0.053
Cluster: 3 (M: −0.592)	***SVM (Variance)***
Ed. No School	0.054
Not in Lab. Force	0.053
Males Total	0.053
Race White	0.049
Ed. Bachelors Deg	0.046
Cluster: 4 (M: 1.449)	***SVM (Variance)***
Language Other	0.080
Females Total	0.071
Race White	0.054
Race African Am.	0.051
Population Total	0.049

Results indicated that the best-performing model used the principal component analysis (PCA) method for fitting the first cluster, followed by an SVM to fit the latter three clusters, which was used for the basis of Table 8. Another finding is that the higher count of individuals speaking only English tend to be associated with a higher count of uninsured persons, specifically only in areas that have the greatest relative distance away from hospitals, where most of the inhabitants are white to begin with. Practically, this means that in the areas that are farthest away from medical facilities, race and income are major factors in the determination of health insurance status. In the latter clusters representing areas closer to medical facilities, there is significant variation in variable makeup. However, since the SVM technique measures relative variance rather than producing a standardized coefficient, the direction of the relationship cannot be determined. In the clusters closer to medical facilities, it can be inferred that income, population density, and education tend to be the most significant variables in terms of estimating health insurance status. While race was found to be a significant estimator, it can likely be explained through a lack of diversity in rural regions, where in Missouri, racial clusters tend to follow population clusters.

It should be noted that the method of interpretation described above is only valid when using the cluster-based ensemble algorithm with best base learner by cluster as the aggregation mechanism.

In contrast, using the average aggregation mechanism for the same modeling technique, interpretation would still be possible, although a method would need to be established for how to interpret the average importance across multiple models. This is a possible avenue for future research. In contrast to the best learner approach, the average aggregation mechanism would be preferable when the predictive power needs to be maximized for external validity and generalizability, given the nature of averaging multiple learners to achieve a more representative and reliable response. This is a tradeoff between predictive ability and inference; if predictive power is required and the resulting output needs to be most generalizable and externally valid, then averaging results is typically encouraged, although it will lead to decreased interpretability due to model complexity.

While the cluster-based algorithm can maintain a significant level of inference when using the best-learner aggregation approach, there are limitations to the method in terms of its potential for generalizability. However, despite the interpretation tradeoff, one thing becomes clear: the cluster-based ensemble algorithm resembles localized regression in a way that retains the original variable inputs for interpretation. Whereas a generalized additive model (GAM) model using local or spline regression approaches often leads to a complicated interpretation of transformed variables, the cluster-based approach is a suitable alternative when inferential capability is required. Maps 1 and 2 represent the clusters derived from k-means clustering using the *distance to hospitals* variable as the basis for clustering.

From map one in Figure 2, patterns can be discerned where the first cluster primarily represents rural areas, followed by the second and third clusters representing suburban areas, and the fourth cluster representing urban areas with hospitals nearby. Map 2 is characterized by the optimal performing base learner using the cluster best-learner aggregation method, where the first cluster (distances furthest from hospitals) was fitted using PCR, and the latter clusters (distances closer to hospitals) were fitted using the SVM approach. Localized regression is illustrated by the cluster-based approach, where the only caveat relates to the requirement for interpretation by cluster rather than globally, leading to more descriptive conclusions surrounding model performance and relative variable importance.

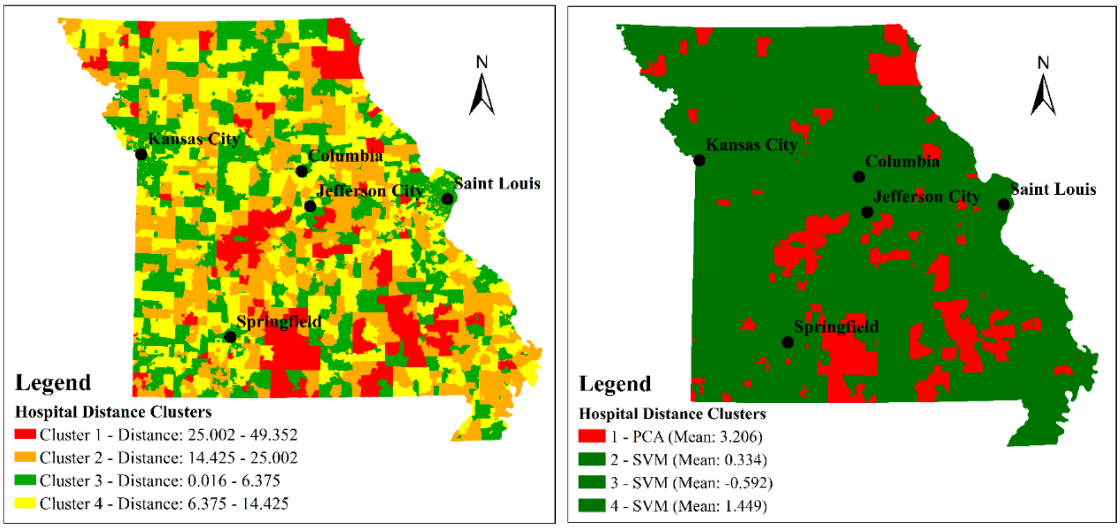


Figure 2. Map 1: Hospital Distance Clusters (**left**), Map 2: Hospital Distance Clusters by Optimal Base Learner (**right**).

4.2. Study Limitations

This study was limited by the number of machine learning approaches incorporated into the cluster-based ensemble algorithm. Despite diversity among base learners, adding more individual modeling techniques may improve modeling performance. Additionally, the number of clusters to split any independent variable also raises questions. As discussed earlier, there are multiple approaches for

determining an optimal number of clusters to split a variable, where, at best, a variety of indices return a range of optimal clusters. For this situation, it was decided early on to limit the number of clusters to equal the number of modeling classes under study. While this approach works from a practical standpoint, for true optimization, determination of the number of clusters to retain could be improved upon. More specifically, provided the variable results of MSE change as cluster values were iteratively tested, further research employing synthetic population generation could explore the relationship between the size of a given dataset and the optimal number of clusters—while maintaining variable relationships—in greater detail.

Finally, in terms of inference, some of the modeling techniques that have been described in this study lend well to statistical inference: lasso, ridge, and principal component regression specifically. However, other modeling techniques have established measures for variable importance, but lack coefficients or measures of statistical significance that could lend insight into the direction of the relationship or confidence intervals surrounding variable importance measures. This makes interpretation challenging for some of the machine learning classes such as support vector regression; future research or improvement on this technique could aim to build bootstrap-based confidence intervals around the variable importance measures, leading to better indications of the statistical significance behind each measure.

5. Conclusions

This paper outlined a machine learning ensemble algorithm using a cluster-based approach to localize modeling for added performance and inferential capability compared to individual base learners and more complex methods such as local or spline-based regression procedures. The study compared individual base learners, comparative ensemble approaches, and the cluster-based ensemble algorithm, finding overall that the cluster-based approach with best learner by cluster for aggregation was the top performing model, followed by an average aggregation method for the same cluster-based ensemble methods.

The applied research problem framed in this paper was as follows. Can using machine learning algorithms, employing cluster analysis as a pre-processing technique, serve as a methodological framework to improve upon existing methods to estimate health insurance status across Missouri using geospatial, demographic, socioeconomic, and access-to-care characteristics? The findings from this study suggest that in areas furthest from hospitals and medical facilities (cluster one), race and income seem to be the primary factors associated with estimating block groups with high concentrations of individuals without health insurance. Specifically, rural areas with high concentrations of white and only-English speaking populations are likely to experience higher rates of non-insured inhabitants, whereas rural areas with higher rates of African American populations or high median income are less likely to carry elevated rates of non-insured inhabitants. Moving to areas closest to hospitals and medical facilities (cluster three), race continues to be a factor in the determination of health insurance status, but is significantly less significant in comparison to rural areas. In contrast to rural areas, education and labor characteristics become significant in estimating health insurance coverage in more heavily populated areas, which is likely due to the types of jobs being held in urban areas and the differences in health insurance coverage being offered in urban and rural areas, such as self-purchased insurance versus insurance through employer-sponsored plans.

In practice, the cluster-based ensemble methods add inferential capability in localized regression techniques, and can assist public health officials in the decision-making processes surrounding targeted interventions among complex populations. This is especially true for large areas with diverse populations, where localizing the population into smaller clusters for tailored analysis can improve model accuracy. This paper makes a unique and significant contribution to the academic literature on clustered-based machine learning ensemble for geospatial data. First, the success of the algorithm used in this study comes from its ability to multitask in answering several research questions simultaneously, while increasing the estimative accuracy compared to base learners and comparative

ensemble approaches. Second, while there are aspects of the cluster-based ensemble technique that can be improved upon, this study lays the groundwork for future research in localized ensemble modeling. Third, collectively, this paper presents a framework that when applied to geospatial datasets can deliver both prediction and inference, as well as capture information about global performance and trends, all while retaining valuable information at the localized level to explain patterns and variations within datasets.

Author Contributions: This publication contributes to the research requirement for conferral of Erik Mueller's Ph.D. in Integrated and Applied Sciences, Bioinformatics and Geospatial Biology. E.M., with supervision from J.S.O.S., serving as primary advisor and mentor, and S.M. and M.E., serving as dissertation committee members and secondary mentors, planned, organized, carried out, and analyzed all aspects of this research study.

Funding: This research received no external funding. The article processing charge (APC) was funded by Saint Louis University through the GeoSLU Big Idea Grant, number 275100.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Waller, L.A.; Gotway, C.A. *Applied Spatial Statistics for Public Health Data*; John Wiley & Sons: New York, NY, USA, 2004; Volume 368.
2. Trivedi, S.; Pardos, Z.A.; Heffernan, N.T. The utility of clustering in prediction tasks. *arXiv* **2015**, arXiv:1509.06163.
3. Trivedi, S.; Pardos, Z.A.; Heffernan, N.T. Clustering students to generate an ensemble to improve standard test score predictions. In Proceedings of the International Conference on Artificial Intelligence in Education, Auckland, New Zealand, 28 June–2 July 2011; pp. 377–384.
4. Trivedi, S.; Pardos, Z.; Sárközy, G.; Heffernan, N. Spectral clustering in educational data mining. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, 6–8 July 2011.
5. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
6. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2014.
7. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: New York, NY, USA, 2004.
8. USCB. American Community Survey (ACS). Available online: https://www.census.gov/programs-surveys/acs/ (accessed on 17 October 2018).
9. Juarez, P.D.; Matthews-Juarez, P.; Hood, D.B.; Im, W.; Levine, R.S.; Kilbourne, B.J.; Langston, M.A.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes, M.G. The public health exposome: A population-based, exposure science approach to health disparities research. *Int. J. Environ. Res. Public Health* **2014**, *11*, 12866–12895. [CrossRef] [PubMed]
10. USCB. *The National Map: Transportation*; US Census Bureau: Washington, DC, USA, 2017.
11. USCB. *Topologically Integrated Geographic Encoding and Referencing Datasets*; US Census Bureau: Washington, DC, USA, 2018.
12. Missouri Department of Health and Senior Services. *Missouri DHSS Bureau of Health Care Analysis and Data Dissemination*; DHSS: Jefferson City, MO, USA, 2015.
13. Goldman, D.P.; Smith, J.P.; Sood, N. Legal status and health insurance among immigrants. *Health Aff.* **2005**, *24*, 1640–1653. [CrossRef] [PubMed]
14. Haas, J.S.; Lee, L.B.; Kaplan, C.P.; Sonneborn, D.; Phillips, K.A.; Liang, S.-Y. The association of race, socioeconomic status, and health insurance status with the prevalence of overweight among children and adolescents. *Am. J. Public Health* **2003**, *93*, 2105–2110. [CrossRef] [PubMed]
15. Kennedy, B.P.; Kawachi, I.; Glass, R.; Prothrow-Stith, D. Income distribution, socioeconomic status, and self rated health in the United States: Multilevel analysis. *BMJ* **1998**, *317*, 917–921. [CrossRef] [PubMed]
16. Wharam, J.F.; Zhang, F.; Landon, B.E.; Soumerai, S.B.; Ross-Degnan, D. Low-socioeconomic-status enrollees in high-deductible plans reduced high-severity emergency care. *Health Aff.* **2013**, *32*, 1398–1406. [CrossRef] [PubMed]

17. Arcury, T.A.; Gesler, W.M.; Preisser, J.S.; Sherman, J.; Spencer, J.; Perin, J. The effects of geography and spatial behavior on health care utilization among the residents of a rural region. *Health Serv. Res.* **2005**, *40*, 135–156. [CrossRef]
18. Regan, S.; Wong, S.T. Patient perspectives on primary health care in rural communities: Effects of geography on access, continuity and efficiency. *Rural Remote Health* **2009**, *9*, 1142.
19. Casey, M.M.; Call, K.T.; Klingner, J.M. Are rural residents less likely to obtain recommended preventive healthcare services? *Am. J. Prev. Med.* **2001**, *21*, 182–188. [CrossRef]
20. Tobler, W.R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970**, *46*, 234–240. [CrossRef]
21. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*; R Package Version 1.6-7; R Foundation: Vienna, Austria, 2015.
22. Breiman, L. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package Version: 2006. Available online: https://www.stat.berkeley.edu/~{}breiman/RandomForests/ (accessed on 10 October 2018).
23. Ridgeway, G. *Gbm: Generalized Boosted Regression Models*; R Package Version 1.6-3.1; R Foundation: Vienna, Austria, 2010.
24. Friedman, J.; Hastie, T.; Simon, N.; Tibshirani, R. *Lasso and Elastic-Net Regularized Generalized Linear Models*; R-Package Version 2.0-5; R Foundation: Vienna, Austria, 2016.
25. Mevik, B.-H.; Wehrens, R.; Liland, K.H. *Pls: Partial Least Squares and Principal Component Regression*; R Package Version; R Foundation: Vienna, Austria, 2011; Volume 2.
26. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 112.
27. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
28. Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892. [CrossRef]
29. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A.; Charrad, M.M. Package 'NbClust'. *J. Stat. Softw.* **2014**, *61*, 1–36.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).