A Human Geospatial Predictive Analytics Framework With Application to Finding Medically Underserved Areas

James M. Keller, *Life Fellow IEEE*
Andrew R. Buck, *Student Member IEEE*
Alina Zare, *Member* IEEE
Electrical and Computer Engineering Department
University of Missouri
Columbia, MO 65211 USA
kellerj@missouri.edu; arb9p4@mail.missouri.edu;
zarea@missouri.edu

Mihail Popescu, *Senior Member IEEE*
Health Management and Informatics Department
University of Missouri
Columbia, MO 65211 USA
popscum@missouri.edu

***Abstract*— Human geography is a concept used to indicate the augmentation of standard geographic layers of information about an area with behavioral variations of the people in the area. In particular, the actions of people can be attributed to both local and regional variations in physical (i.e., terrain) and human (e.g., income, political, cultural) variables. In this paper, we study the utility of a human geographic data cube coupled with computational intelligence as a means to predict conditions across a geographic area. This becomes a Big data problem. In this sense, we are using genotype information to predict phenotype states. We demonstrate the approach on the prediction of medically underserved areas in Missouri.**

Keywords—Human Geography, Predictive Analytics, Medically Underserved, Computational Intelligence, Human Geographic Data Cube, Big Data, Feature Selection

I. Introduction

Geography is concerned with the systematic spatial or spatio-temporal analysis of the patterns, distributions, variations, and relationships of natural and human phenomena; human geography, in particular is focused on how behavioral variations can vary spatially as a resulting interplay between local and regional variations in physical (i.e., terrain) and human (i.e., income, political, cultural) variables.

Using Agent-Based Modeling (ABM), we studied disaster evacuation scenarios [1-3] within a human geographic framework. This included decision making via bounded rationality and several communication schemes implemented as rumor spreading models. These simulations were dynamic representations of individuals or groups of people trying to move to an evacuation shelter during a natural disaster. Emotion patterns were added to agents and linguistic summaries of the emotional states of agents who made it to shelters were analyzed [4].

Recently, we have assembled a human geographic data cube of the State of Missouri, containing over 300 feature layers of mixed data that encompass economic, educational, religious, cultural attributes across the state, collected by the Geography members of our research team [5]. There are 24 basic categories of attributes: Ability to speak English, Athletic association class, Citizenship, Disability, Euclidean distances to selected places (schools, libraries, etc.), Employment, Food stamps, Geo mobility, Heating fuel, Hispanic population, Household income, Industry, Language spoken, Means of transportation, Network analysis (derived features), Occupation, Place of birth, Poverty, Income taxes, Religion, Social security assistance, Transportation, Vehicles owned, and House age, each with several associated layers. For example, Household income is broken up into 11 different groups with a layer representing the distribution of each. This data cube is treated much like one would consider a hyperspectral image except that the information is not sampled from the electromagnetic spectrum, but from numeric, ordinal and categorical aspects of the population. The data layers are not inherently co-registered, but must be aligned through GIS functions.

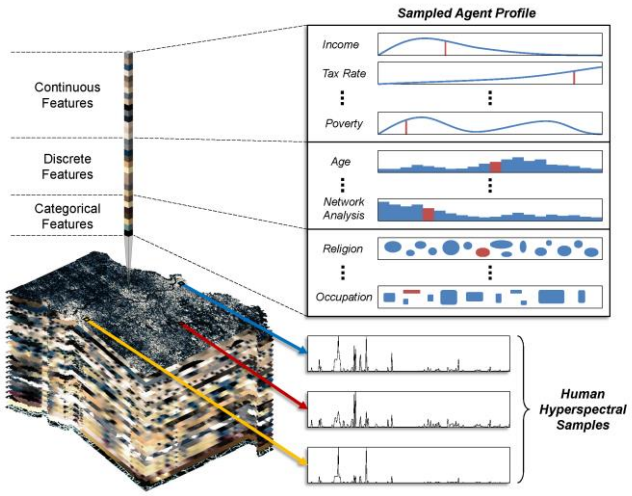

Fig.1 Human geographic data cube with human geography layers of Missouri with over 300 continuous, discrete, and categorical feature layers.

The authors wish to thank the National Geospatial-Intelligence Agency for support of this research under contract NGA HM 1582-10-C-0013.

Fig. 1 depicts the Human geographic data cube with human geography layers of Missouri with a size of 1,127,957 vectors. A companion paper [6] explores the concept of human hyperspectral unmixing and the meanings of endmembers in this new domain. Fig. 2 is an example of one of the layers in the Human geographic data cube, population density. The layer values contain considerable uncertainty and should be treated as such. With this increased layer complexity, extending the predictive analysis over larger geographic areas, and looking at more dynamic problems (disease spread, disasters), this fits the model of Big Data. It has the volume, variety, velocity, veracity, and value attributes characteristic of Big Data.

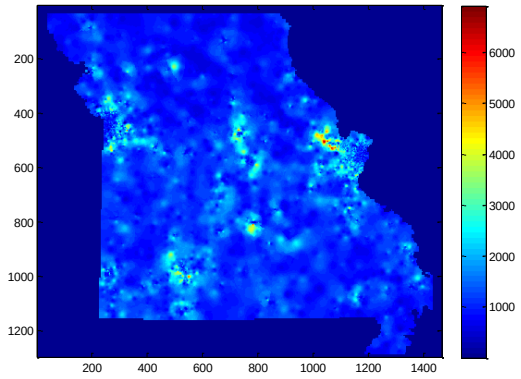

Fig.2 Population density layer from the human human geographic data cube of Missouri.

Our goal is to construct a framework to utilize the rich nature of this data to perform predictive analysis of conditions across the geographic area represented by the human geographic data cube with human geography layers. Geospatial predictive analytics (GPA) represent a large array of machine learning, statistics, pattern recognition and data mining methods employed to compute future events in a given geospatial context based on available historical information [7,8]. GPA present specific challenges related to the geospatial and sociopolitical aspects of the domain data such as size, heterogeneity, missing values and complexity of human behavior models. While domain experts are able to make reasonable predictions for particular situations or populations, their mental models are not able to deal with large amount of data over large geospatial areas during a rapid changing context. Here, we present a **pr**edictive **a**nalytic **f**ramework (PRAF) that intends to assist the domain expert in analyzing a large amount of human geographic data in a timely manner. PRAF is a general technique based on computational intelligence algorithms able to attack many problems even with fairly sparse training data. In order to understand the generalizability, we utilize PRAF to study a problem where ground truth information is available.

In this paper, we employ PRAF to predict underserved medical areas in the geographic region of the human geographic data cube. A federal Medically Underserved Area (MUA) is defined by scoring the following 4 criteria [10]:

1. percentage of population below 100% poverty
2. percentage of population age 65 and over
3. infant mortality rate
4. primary care physicians per 1,000 population

The four resulting scores are added and the sum is identified as the Index of Medical Underservice (IMU). An area with a score less than 62 is generally eligible for designation as an MUA. This is a problem that attracts much research [11-16], and provides an effective way to test the efficacy of our PRAF. We demonstrate our approach on an example that consists in predicting medically underserved areas (MUA) in Missouri based on data compiled from the United States Census Bureau (http://www.census.gov/), American community survey (http://factfinder2.census.gov), Health Resources and Services Administration (http://muafind.hrsa.gov/) and Missouri Department of Health & Senior Services (http://health.mo.gov/data/brfss/index.php). Predicting MUA in a given state can assist the local government in the strategic planning of allocation of medical resources such as clinical personnel, clinics and hospitals. Fig.3 is the generated ground truth image for the State of Missouri, obtained from the Health Resources and Services Administration web site [10]. In this figure, the scores above 62 are set to 100 indicating fully medically served areas, while those below 62 are kept at their original values, providing degrees of being medically underserved.

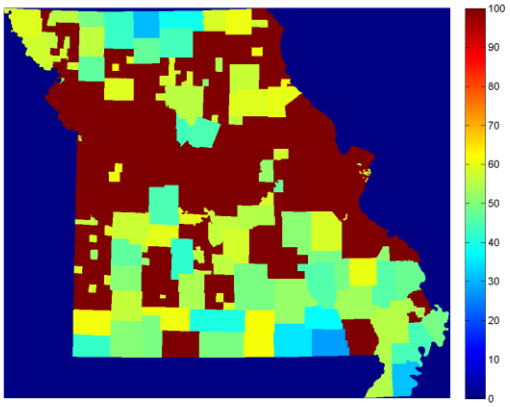

Fig.3 Medically Underserved Areas of Missouri to test PRAF. The values above 62 have been thresholded into the full membership category of Medically Served whereas the medically underserved regions are represented by their degree.

II. Predictive Analytic Framework

A. *Feature Selection*

The first step of our framework consists in feature (layer, band) selection. Feature selection (FS) is challenging in a human geographic data cube with human geography layers due to the image size (large number of data points) and to ever-increasing number of layers (features) available [18]. Since FS is NP-hard, most problems are solved using heuristic approaches. There are two main types of FS methods: filters and wrappers. Filter methods are based on some estimation of the discriminatory power of features, while wrappers use a

Algorithm 1:
RELIEF Feature Selection

Input:

X: <*N* x *P*> input data matrix, where *N* is the number of sample points (1,127,957) and *P* is the number of features (270).

T: <*N* x 1> target data matrix containing the raw MUA score in the range [0, 1]. MUAs are samples with a value <= 0.62.

Initialize:

Normalize each feature column in *X* to have 0 mean and a standard deviation of 1.

Initialize a weight vector *w* = <1 x *P*> of all zeros.

For *i* = 1 **to** *M***:** (*M* = 5000)

Pick a random sample row x_i from the input data matrix.

Compute the Euclidean distance from this point to the *N*-1 other points in feature space.

Find the nearest point from the same class (Hit) and from the other class (Miss).

Update the weight vector according to equation (1).

End For

Sort the weight vector from largest to smallest weights.

Output:

w: <1 x *P*> weight vector. The most significant features will have the highest weights.

Algorithm 2:
NN Wrapper Feature Selection

Input:

X: <*N* x *P*> input data matrix, where *N* is the number of sample points (1,127,957) and *P* is the number of features (270).

T: <*N* x 1> target data matrix containing the raw MUA score in the range [0, 1]. MUAs are samples with a value <= 0.62.

Initialize:

Initialize a score vector *s* = <1 x *P*> of all zeros.

For *i* = 1 **to** *P***:**

Randomly pick 90% of the sample points to use as training data. The remaining 10% will be used for testing.

Construct a 1x5x1 neural network, containing 5 hidden units.

Train the network using only feature *i* as input for 10 epochs.

Compute the MSE of the i^{th} feature of the testing data set and save in the score vector s_i.

End For

Sort the score vector from smallest to largest MSE.

Output:

s: <1 x *P*> MSE score vector. The most significant features will have the lowest MSE values.

classifier in the evaluation process. Filter methods tend to be faster, while wrapper ones are usually more accurate. Most popular wrapper methods are Sequential Floating/Forward Selection (SFS)[19]. Some examples of filters are RELIEF [20] and feature clustering [18]. Typically, for large datasets, several filters are applied first to reduce the number of features to a manageable number (typically 100 [18]), followed by a wrapper method.

As part of the PRAF framework, in this work we investigate several feature selection methods for large datasets. Initially, we considered RELIEF and a Neural Network wrapper method for feature selection.

1) RELIEF Feature Selection

The RELIEF algorithm assigns a weight w_j to each feature *j* based on how well it separates point $x_i \in R^P$, $i \in [1,M]$, from the closest point from opposite class (call "miss") and, in the same time, on how well it unifies (resembles) it with the closest point from the same class (call "hit"), that is:

$$w_j = \sum_{i=1}^{M} (|x_{ij} - x_{miss,j}| - |x_{ij} - x_{hit,j}|), \text{ j = 1, P.} \quad (1)$$

Although the RELIEF algorithm is many times classified as a filter, it is in fact a wrapper based method in disguise since it is based on the nearest neighbor (NN) classifier (that is, for each x_i it finds a point x_{hit} – NN from the same class, and a point x_{miss}- NN from the other class). We mention that the *M* x_i's are sampled from the entire dataset of *N* points, where *N* >> *M*. Since we have to sort the data for each sample point $i \in [1,M]$, RELIEF complexity is of the order $O(N^2MP)$. Since we are using calculations based on real numbers (like equation (1)), we pruned from the human geographic data cube several feature layers that contained terse categorical information (like the layers that showed religions). This resulted in a 270 layer image. The procedure implemented for RELIEF on this image is shown as Algorithm 1.

2) The artificial neural network (ANN) wrapper method

The ANN-wrapper method uses a classifier (in this case a small 2 layer perceptron) for quickly selecting a number of R most relevant features. The ground truth MUA data shown in Fig.3 are linearly scaled to the interval [0,1], since they represent the target values for the neural network function approximator. For each feature value, an artificial neural network ANN_i is trained using randomly chosen data and a single feature i, $i \in [1,P]$, as input. The performance s_i of network ANN_i is obtained by computing the Mean Squared Error (MSE) of a testing set of vectors. After this procedure, shown in Algorithm 2, completes the training/testing of a small ANN each feature, the scoring vector is sorted to provide an ordered list of relevant features.

B. Computational Intelligence Prediction Method

The method used to test the ability of PRAF to predict medically underserved areas was a simple multilayer perceptron. We picked this simple regression function approximator because it is simple and well-studied, and will provide information on the efficacy of the features in the human geographic data cube. The training data was sampled

randomly from the raster data cube and was varied for each run. The remaining data vectors were used to test. Mean Squared Error (MSE) was calculated and since the output of the ANN varied in the interval [0,1], Receiver Operating Characteristic (ROC) curves were calculated.

III. Results

1) Feature selection

After running both feature selection approaches, it is interesting that each method focused on different aspects of the human geographic data cube. In the top 25 weighted attributes, RELIEF emphasized Disability (7 layers chosen), Poverty (3 layers), and then features that included Home age, Travel time to work, Occupation, Place of birth, Geo-mobility, and Heating choice among a few others. The Wrapper top 25 zeroed-in on attributes like Income levels (6 layers), Taxes (4 layers), Employment, Mobility and Poverty. The only layer in common among the top 25 was one of the Income layers. Looking at the choices, these two methods only share one common attribute in their top 25, an income layer. Both of the sets of preferred layers, however, make sense intuitively for an attempt to predict medically underserved areas.

2) Prediction of Medically Underserved Areas

Now that we have ranked sets of feature layers, the final step is to determine if subsets of the features can be used effectively to predict Medically Underserved Areas. In these experiments, we considered the top 10, top 25, top 50 and top 100 feature layers from each of the two selection (ranking) approaches. Since the generalizability of PRAF to many geographic areas requires the generation of fairly detailed human geographic attributes, we are not only interested in finding small but predictive feature sets, but also in learning how small of a training set is needed to give good predictability. So, for each subset of features, we trained multilayer perceptrons with 10%, 1% and 0.1% of the available data, testing on the remaining. Again, the ground truth data shown in Fig.3 is linearly scaled to [0,1]. Each neural network configuration was trained and tested 10 times with random sampling of training sets and initialization. The hidden layer of the perceptron had $n+\sqrt{n}$ neurons and was trained with the Levenberg-Marquardt algorithm for 50 epochs. We considered two different types of estimation of MUAs.

Table I displays a summary of all of the experiments for the semi continuous ground truth, i.e., vales calculated according to [10] that were under 62 were preserved while those of 62 and above were considered medically served and set to 100 (before scaling). For each trial – choice of number of inputs and size of the training set – the table contains the average MSE on the test sets together with the standard deviations across the 10 runs. With ground truth values in the interval [0,1] the average error per pixel ranges from around 7% up to a high of 43%. Both extreme cases were for features selected by the wrapper method. From the standpoint of MSE, across all of the experiments summarized in Table I, RELIEF provided a better predictive capability than did the wrapper method. This summary information is somewhat misleading. Recall that we are trying to predict the scaled values displayed in Fig. 3. These areas represent geographical blocks (county and census track information), not individual pixels in the rasterized version of Missouri. Since the function approximator is accepting inputs that have been manipulated though a GIS system to co-register the layers, some bleeding of feature values is to be expected. This will show itself mostly at the crisp boundaries of the ground truth map. In fact, the ANN produces a soft prediction. This can be seen in Fig. 4, where the prediction output of one trial of each experiment is displayed as image with the same color scale as Fig. 3. It's easy to see the closeness between Fig.3 and the output image for 10% training data with 100 features. The ANN output images become fuzzier as the size of the training sets and the number of features decrease. Is this really bad? We might argue that in fact it is no worse than assuming that every point in the geographic block shares the same level of medical service. This is an issue that we have no ability to investigate, but it opens interesting questions. Generally speaking, this collection of output images supports the utility of PRAF even with limited training data and feature layers.

TABLE I. Mean Squared Error (MSE) of MUA Prediction

		NN Wrapper			**RELIEF**		
		Training Size			***Training Size***		
		10%	1%	0.1%	10%	1%	0.1%
# of Features	10	0.0344 ± 0.0012	0.0353 ± 0.0013	0.0483 ± 0.0052	0.0353 ± 0.0009	0.0361 ± 0.0003	0.0524 ± 0.0079
	25	0.0149 ± 0.0008	0.0175 ± 0.0008	0.1340 ± 0.0153	0.0135 ± 0.0005	0.0158 ± 0.0004	0.1190 ± 0.0127
	50	0.0075 ± 0.0003	0.0152 ± 0.0011	0.1852 ± 0.0451	0.0079 ± 0.0004	0.0150 ± 0.0005	0.1657 ± 0.0233
	100	0.0049 ± 0.0001	0.0310 ± 0.0073	0.0794 ± 0.0147	0.0055 ± 0.0001	0.0274 ± 0.0065	0.1030 ± 0.0195

An alternate way to describe the results of these prediction experiments is to examine the resulting ROC curves. For this, the values in Fig. 3 are completely thresholded into a binary image so that each geographic block is classified as either Medically Served or Medically Underserved. Then each output image is thresholded at varying levels and compared to this binary ground truth. The Probability of Detection and the Probability of False Alarm (PFA) are calculated as functions of the confidence threshold level. The ten ROC curves for each experiment are shown in Fig. 5. The area under the ROC curve (AUC) is a standard metric to compare outputs. These statistics are displayed in Table II.

TABLE II. Area Under ROC Curve (AUC) for MUA Prediction

		NN Wrapper			**RELIEF**		
		Training Size			***Training Size***		
		10%	1%	0.1%	10%	1%	0.1%
# of Features	10	0.8786 ± 0.0087	0.8761 ± 0.0104	0.8326 ± 0.0117	0.8735 ± 0.0062	0.8692 ± 0.0033	0.8145 ± 0.0085
	25	0.9774 ± 0.0025	0.9719 ± 0.0026	0.8205 ± 0.0096	0.9817 ± 0.0015	0.9760 ± 0.0012	0.8208 ± 0.0110
	50	0.9936 ± 0.0004	0.9844 ± 0.0015	0.7994 ± 0.0142	0.9931 ± 0.005	0.9836 ± 0.0011	0.8024 ± 0.0138
	100	0.9961 ± 0.0001	0.9703 ± 0.0055	0.8661 ± 0.0189	0.9956 ± 0.0001	0.9744 ± 0.0048	0.8499 ± 0.0189

As can be seen from both Table II and Fig. 5, PRAF does a good job of estimating Medical Service for binary

classification of this condition. The main message is that human geographicl layers can be used to predict geographic conditions not directly related to the collected layers and can do so with limited training exemplars. Certainly, more sophisticated function approximators/classifiers and better training methodologies will increase the accuracy of the predictions.

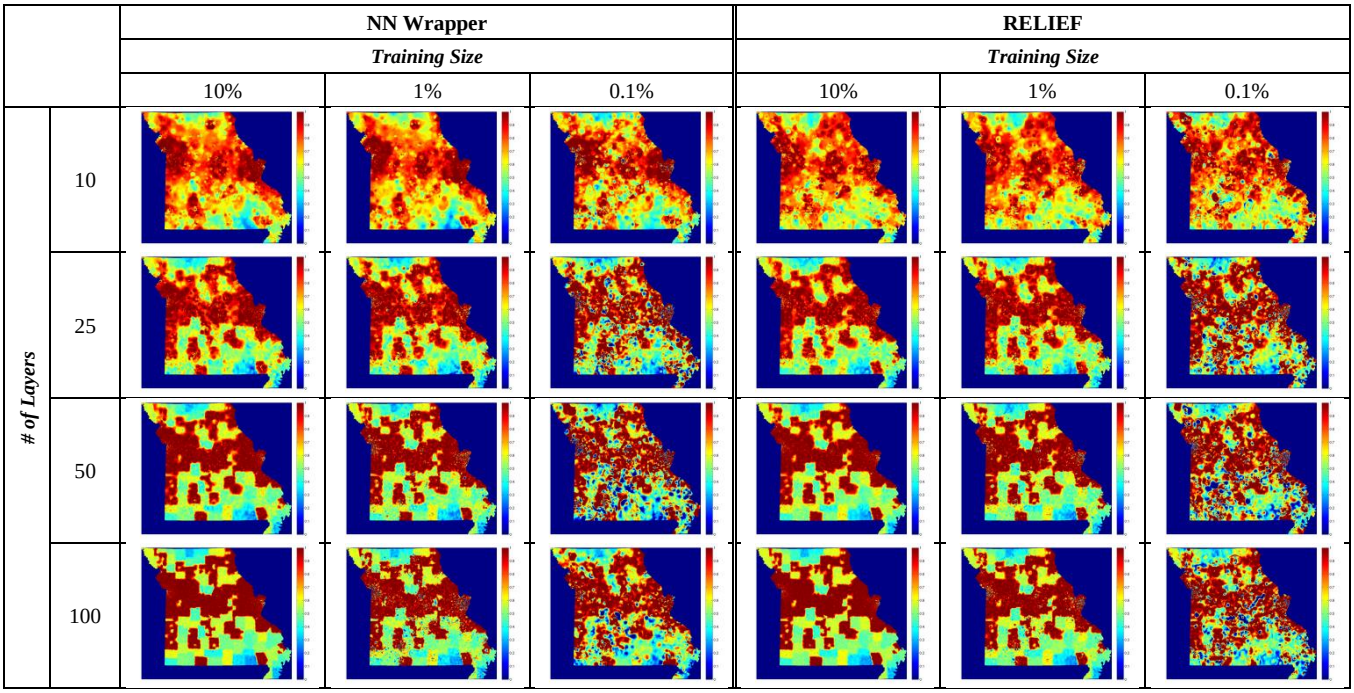

Fig. 4. MUA Prediction of NN Wrapper and RELIEF Feature Selection.

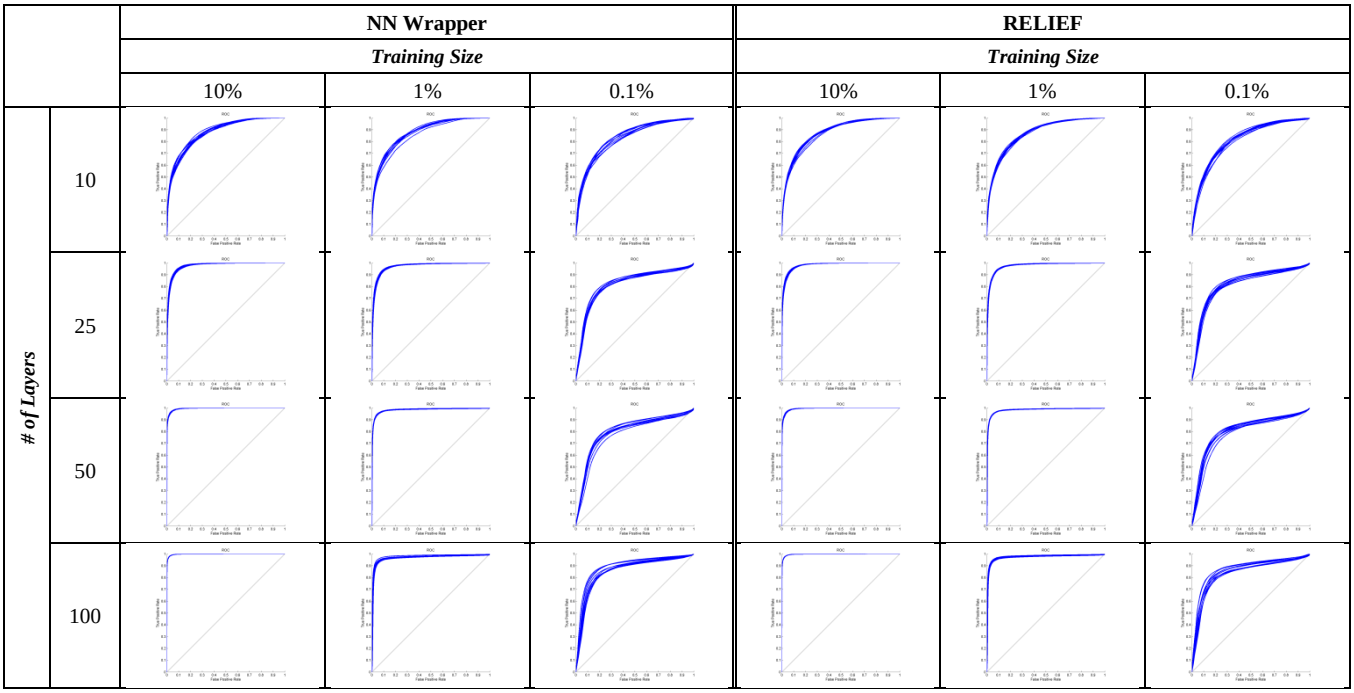

Fig. 5. ROC Curve Comparison of NN Wrapper and RELIEF Feature Selection

IV. Conclusions

In this paper, we discussed an automatic predictive analytics framework (PRAF) for geospatial human geographic data that consists in a feature selection procedure and a predictor based on a neural network. The framework handles problems that fit into a Big Data environment. To test our framework we predicted medically underserved areas in Missouri based on a 270 layer data cube. A series of cross fold validations experiment showed that excellent predictive capabilities were found as the size of the training set was reduced to 1% of the available data. More work needs to be done to refine the feature selection procedures and test more classifiers. We are investigating extensions of RELIEF to take into account soft memberships of the training data to deal with overlap of training sets. Future work will also include investigation into methods used in the hyperspectral band selection literature as an alternative to feature selection. We intend to study initialization approaches (including "pre-training" methods) for the function approximators/classifiers.

REFERENCES

[1] J. Keller, M. Popescu, D. Gibeson, "An extension of a confined space evacuation model to human geography", Proceedings, IEEE Geoscience and Remote Sensing Symposium, IGARSS 2012, Munich, Germany, July, 2012, pp. 531-534.

[2] M. Popescu, J. Keller, "Implementing bounded rationality in disaster agent behavior using OGA operators", Proceedings, IEEE Geoscience and Remote Sensing Symposium, IGARSS 2012, Munich, Germany, July, 2012, pp.5379-5381.

[3] A. Zare, Z. Fields, J. Keller, J. Horton, "Agent-based rumor spreading models for human geography", Proceedings, IEEE Geoscience and Remote Sensing Symposium, IGARSS 2012, Munich, Germany, July, 2012, pp. 5394-5397.

[4] M. Popescu, J. Keller, A. Zare, "A Framework for Computing Crowd Emotions Using Agent Based Modeling", Proceedings of the Symposium on Computational Intelligence for Creativity and Affective Computing (CICAC), as part of the Symposium Series on Computational Intelligence, Singapore, April 2013, pp. 25-31.

[5] T. Haithcoat, T. Vought, E. Mueller, "Creation of a human geographic data cube with human geography layers", *Cartography and Geographic Information Science*, to be submitted to Special Issue on: "Integrating Big Social Data, Computing, and Modeling for a Synthesized Spatial Social Science", July 2014.

[6] A. Buck, A. Zare, J. Keller, M. Popescu "Endmember Representation of Human Geography Layers", under review, Proceedings of Symposium Series on Computational Intelligence, Orlando, FL, December 2014.

[7] R. Maciejewski, R. Hafen, S. Rudolph, S. Larew, M. Mitchell, W. Cleveland, D. Ebert, D. "Forecasting Hotspots—A Predictive Analytics Approach," Visualization and Computer Graphics, IEEE Transactions on , vol.17, no.4, pp.440,453, April 2011.

[8] D. Brown, J. Dalton, H. Hoyle, "Spatial forecast methods for terrorist events in urban environments", Proceedings of the Second NSF/NIJ Symposium on Intelligence and Security Informatics, Lecture Notes in Computer Science, Tucson, Arizona, Springer-Verlag Heidelberg, June 2004, pages 426–435.

[9] G. Beauvais, D. Keinath, P. Hernandez, L. Master, R. Thurston, Element Distribution Modeling: A Primer (Version 2), Natureserve, Arlington, Virginia, June, 2006.

[10] http://muafind.hrsa.gov/, accessed June 12, 2014.

[11] P. Delamater, J. Messina, A. Shortridge S. Grady, "Measuring geographic access to health care: raster and network-based methods", International Journal of Health Geographics 2012, pp. 11-15.

[12] M. Wieland, T. Beckman, S. Cha, T. Beebe, F. McDonald, "Resident Physicians' Knowledge of Underserved Patients: A Multi-Institutional Survey for the Underserved Care Curriculum Collaborative", Mayo Clin Proc., Vol. 85, No. 8, 2010, pp. 728–733.

[13] M. Dulin, T. Ludden, H. Tapp, J. Blackwell, B. Urquieta de Hernandez, H. Smith, O. Furuseth, "Using Geographic Information Systems (GIS) to Understand a Community's Primary Care Needs", J Am Board Fam Med, Vol. 23, No. 1, 2010, pp. 13-21.

[14] J J. Runkle, H. Zhang, W. Karmaus, A. Brock-Martin, E. Svendsen, "Prediction of Unmet Primary Care Needs for the Medically Vulnerable Post-Disaster: An Interrupted Time-Series Analysis of Health System Responses", Int J Environ Res Public Health, Vol 9, No. 10, 2012, pp. 3384–3397.

[15] E. Montague, J. Perchonok, "Health and Wellness Technology Use by Historically Underserved Health Consumers: Systematic Review", J Med Internet Res., Vol. 14, No. 3, 2012, doi: 10.2196/jmir.2095.

[16] T. Horan, N. Botts, R. Burkhard, R., "A Multidimensional View of Personal Health Systems for Underserved Populations", J Med Internet Res., Vol 12, No. 3, 2010, doi: 10.2196/jmir.1355.

[17] United States Health Resources and Services Administration, http://bhpr.hrsa.gov/shortage/muaps/ .

[18] J. Bins, B. Draper, "Feature selection from huge feature sets", Proc. of ICCV 2001, vol. 2, 7-14 July, Vancouver, BC, pp 159-165.

[19] P. Pudil, J. Novovicova and J. Kittler, "Floating Search Methods in Feature Selection", Pattern Recognition Letters, Vol. 15, 1994, pp. 1119-1125.

[20] I. Kononenko, "Estimation attributes: Analysis and Extensions of RELIEF", Proc. European Conference on Machine Learning, Catana, Italy, 1994, pp. 171-182.