Endmember Representation of Human Geography Layers

A. Buck *Student Member, IEEE*
A. Zare *Senior Member, IEEE*
J. Keller *Life Fellow, IEEE*
Department of Electrical and Computer Engineering
University of Missouri
Columbia, MO 65211
Email: {arb9p4, zarea, kellerj}@missouri.edu

M. Popescu *Senior Member, IEEE*
Department of Health Management and Informatics
University of Missouri
Columbia, MO 65211
Email: popescum@missouri.edu

***Abstract*—This paper presents an endmember estimation and representation approach for human geography data cubes. Human-related factors that can be mapped for a geographic region include factors relating to population, age, religion, education, medical access and others. Given these hundreds (or even thousands) of factors mapped over a region, it is extremely difficult for an analyst to summarize and understand the interactions between all of these factors. In this paper, a method to provide a compact representation and visualization of hundreds of human geography layers is presented. These are large data cubes containing a range of human geographic information including some represented using fuzzy values. Results on a human geography data cube compiled for the state of Missouri, USA is presented.**

I. Introduction

Human Geography is concerned with how human-related factors, e.g. cultural, economic, religious, and political, influence the spatial behavior of individuals and groups of people. The study of Human Geography is important for a number of application areas including, for example, preparing for disaster response and relief [1]–[3], identifying medically underserved areas [4], and many others [5]. In order to study the influence of these human-related factors, mathematical models along with meaningful visualization need to be developed. There are an enormous number of human-related factors that can be mapped for a geographic region. An analyst would be unable to individually summarize and understand all of the interactions between every mapped factor without automated analysis and visualization tools. Thus, this paper describes one approach to represent and visualize hundreds of human geography layers.

II. Human Geography Data Cube

In this study, in order to combine the many human geographic factors for a region, a human geography data cube was created [6]. The data cube is a three-dimensional matrix with two spatial dimensions and one human-geographic dimension as illustrated in Fig. 1. For each spatial location, the value for all of the human geographic factors of interest are collected. The data cube can be viewed in two ways: (1) by considering the human geographic *profile* for each location (i.e., one spatial location is considered over all human geography factors) as shown in Fig. 1; or (2) by examining each layer in the data cube individually (i.e., all spatial locations are considered but only one human geography factor) as shown in Fig. 2.

In the study presented in this paper, the state of Missouri is the geographic region under consideration and 270 human geographic factors were mapped. The 270 data layers included 21 basic categories of attributes: Ability to speak English, 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, Occupation, Place of birth, Poverty, Income taxes, 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. Many of these layers contain fuzzy values (e.g., distance to churches of various denominations). The data layers are not inherently co-registered, but must be aligned through GIS functions. Thus, the data for each layer needed to be either resampled or interpolated onto raster grids that were aligned across all layers. More details on the data cube generation can be found in [6].

III. Endmember Representation

Although, as discussed above, each profile and each data layer can be visualized independently, the extremely large amount of data contained in a human geography data cube is difficult for an analyst to summarize and understand the data cube without some sort of visualization and/or data reduction. The human-geographic profile for each location can be considered as a very high dimensional feature vector.

We considered the use of clustering methods to reduce the data into a small number of cluster centers and a partition matrix indicating the degree to which each data point belongs in each cluster. This would reduce the data from $\mathbf{X} \in \mathcal{R}^{N \times D}$ to $\mathbf{X}' \in \mathcal{R}^{N \times C}$ where N is the number spatial locations

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

978-1-4799-4540-5/14/$31.00 ©2014 IEEE

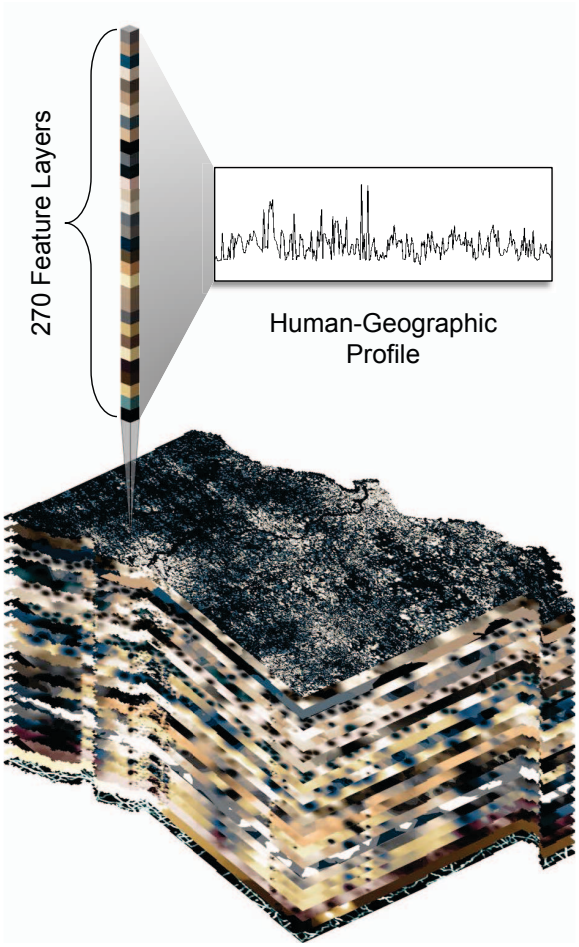
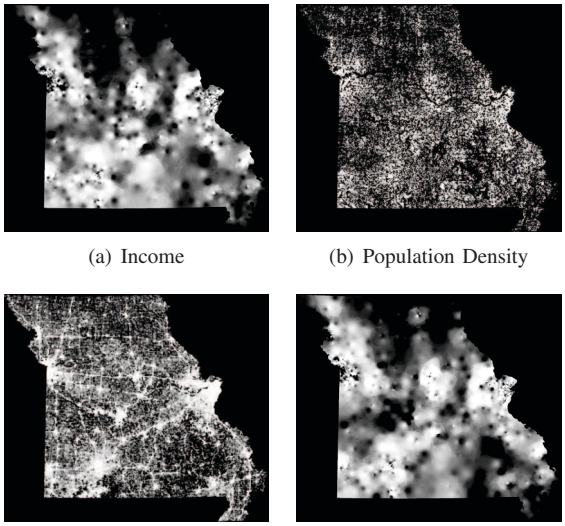


Fig. 1. Example of a Human Geography Data Cube over the state of Missouri, USA. The data cube has two spatial and one human-geographic dimension.



(a) Income (b) Population Density

(c) Road Density (d) Est. Travel to Work Time

Fig. 2. Examples of four Human Geography Layers in the Missouri Data Cube

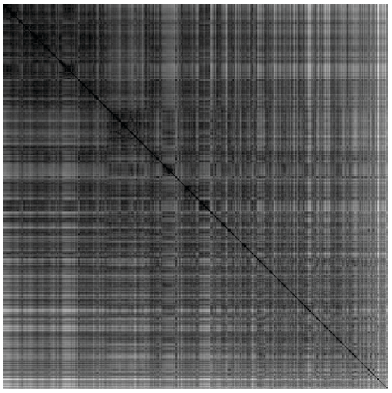


Fig. 3. VAT image of the MO data cube created using the Euclidean distance between each pair of 1000 randomly sampled human-geographic profiles from the human geography data cube.

in each rasterized data layer, D is the number of human geography layers, C is the number of clusters, and $C << D$. However, we found that the human geography data cube did not display apparent cluster structure. For example, consider the VAT (visual assessement of cluster tendency) [7] image shown in Fig. 3. A VAT image is a resorting of the pair-wise dissimilarity between each pair of data points in a data set. The rows and columns of the dissimilairty matrix are sorted such that, when a data set has a strong clustering structure or tendency, the resulting VAT image will have clear low-valued blocks along the diagonal [7]–[9]. However, the VAT image created using the pairwise Euclidean distance between 1000 randomly sampled human-geographic profiles from the MO data cube does not indicate any cluster tendency.

Furthermore, for visualization, the same 1000 randomly sampled human-geographic profiles from the MO data cube were reduced from 270 dimensions to two dimensions using multi-dimensional scaling (MDS) [10]. The Euclidean distance between each pair of human-geographic profiles was computed and, then, MDS was applied to reduce to two dimensions. A scatter plot of the resulting reduced dimensionality data is shown in Fig. 4. As can be seen by examining Fig. 4, the data does not appear to have any clear clusters.

Since the human geographic data cube did not have clear cluster structure, an *endmember* and *abundance* representation was considered instead of clustering. Endmembers and abundance representation is commonly used in the hyperspectral image analysis literature [11]. In the hyperspectral literature, the spectral signatures of the pure materials in a hyperspectral scene are often referred to as endmembers [11]. *Spectral unmixing* is the task of decomposing pixels from a hyperspectral image into their respective endmembers and abundances. Abundances are the proportions of every endmember in each pixel in a hyperspectral image. The standard model used to perform spectral unmixing is the *linear mixing model*. This model states that every pixel is a convex combination of endmembers in the scene. Thus, in this model, the endmembers are the spectra found at the corners of a convex region

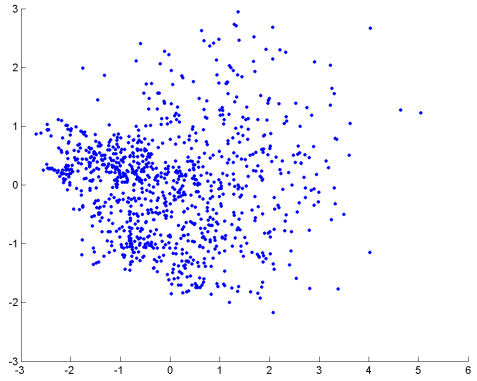


Fig. 4. MDS visualization of the MO data cube created using the Euclidean distance between each pair of 1000 randomly sampled human-geographic profiles from the human geography data cube.

enclosing all the spectra in a hyperspectral scene. This model can be written as shown in Equation (1),

$$\mathbf{x}_i = \sum_{k=1}^{M} p_{ik}\mathbf{e}_k + \epsilon_i \quad i = 1, \ldots, N \tag{1}$$

where N is the number of pixels in the image, M is the number of endmembers, ϵ_i is an error term, p_{ik} is the abundance of endmember k in pixel i, and $\mathbf{e}_k$ is the k^{th} endmember. The abundances of this model satisfy the constraints in Equation (2),

$$p_{ik} \geq 0 \quad \forall k = 1, \ldots, M; \quad \sum_{k=1}^{M} p_{ik} = 1. \tag{2}$$

By estimating endmembers, representatives located at the corners and along the edges of the data are estimated. In contrast, when estimating cluster representatives, these are located within the data at the centers of estimated clusters. By estimating endmembers and abundances, we can reduce the human geographic data cube from $\mathbf{X} \in \mathcal{R}^{N \times D}$ to $\mathbf{P} \in \mathcal{R}^{N \times M}$ where N is the number spatial locations in each rasterized data layer, D is the number of human geography layers, M is the number of endmembers, $M << D$, and each row of $\mathbf{P}$ contains the abundance values of the n^{th} data point in each of the estimated endmembers.

IV. SPICE Algorithm

In order to simultaneously estimate endmembers, abundances, and the number of needed endmembers, the Sparsity Promoting Iterated Constrained Endmembers (SPICE) algorithm was applied [12]. SPICE estimates these desired parameters by using alternating optimization to minimize the following objective function,

$$\begin{aligned} J_S(\mathbf{E}, \mathbf{P}) \quad = \quad & (1-\mu)\sum_{j=1}^{N} (\mathbf{x}_j - \mathbf{E}_i\mathbf{p}_{ij})^T (\mathbf{x}_j - \mathbf{E}_i\mathbf{p}_{ij}) \\ & +\mu \sum_{k=1}^{M-1} \sum_{j=k+1}^{M} (\mathbf{e}_{ik} - \mathbf{e}_{ij})^T (\mathbf{e}_{ik} - \mathbf{e}_{ij}) \\ & +\sum_{k=1}^{M} \gamma_k \sum_{i=1}^{N} p_{ik} \end{aligned} \tag{3}$$

such that

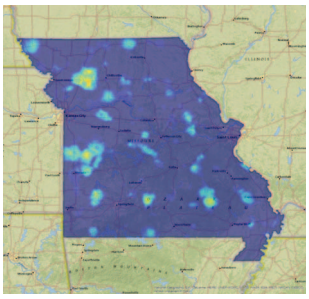
$$p_{ik} \geq 0 \quad \forall k = 1, \ldots, M; \quad \textstyle\sum_{k=1}^{M} p_{ik} = 1$$

where μ is a fixed parameter used to balance the two terms of the objective function and $\gamma_k = \frac{\Gamma}{\sum_{i=1}^{N} p_{ik}}$ computed using the abundance values from the previous iteration and fixed parameter Γ. The two terms of this objective computes the squared Euclidean distance between each input profile and their estimate using the estimated endmembers and abundance oportion values and the sum of squared differences between the estimated endmembers. The third term of this objective is a sparsity promoting term on the proportion values associated with each endmember. This term is used to determine the number of needed endmembers for an input data set by driving the abundance values for unneeded endmembers to zero.

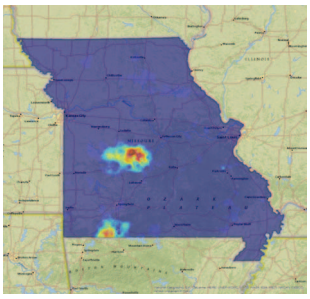
V. Results

The SPICE algorithm was applied to 100,000 randomly selected data points from the MO data cube (approximately 10% of the available data). In order to determine an appropriate set of SPICE parameters, the SPICE algorithm was run with μ varied from 10^{-4} to 10^{-1} and Γ varied from 0.1 to 100. For each set of results, two validity metrics were computed to evaluate the resulting endmembers and abundance values: (1) Residual Error: $e = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{E}\mathbf{p}_n\|_2^2$; and (2) Entropy of Abundance Maps: $H_m = -\sum_{n=1}^{N} P(p_{mn}) \log_2 P(p_{mn})$, where $P(p_{mn})$ is the probability of observing the abundance value p_{mn} when $\mathbf{P}_m$ is discretized into 256 unique values across the abundance map estimated with respect to endmember m. The resulting metrics plotted with respect to each parameter set is shown in Fig. 5.

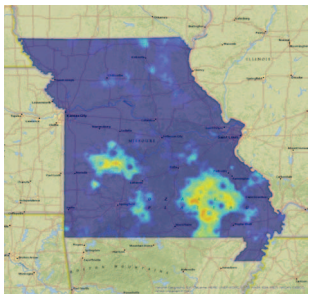
The goal is to have a minimum residual error which indicates that the estimated endmembers and abundance values can effectively fit the data. However, it is possible to *overfit* the data with endmembers and abundance values and, therefore, also represent noise in addition to the key profile information. Thus, mean entropy provides an alternative metric that can counter-balance residual error. A low entropy implies that the resulting endmembers provide for a meaningful set of abundance values (as opposed to random noise). When examining Fig. 5, the parameters were set to $\mu = 0.1$ and $\Gamma = 0.1$. These parameters were chosen after observing that a low entropy value was more important than a low residual error in producing interpretable endmembers. This resulted in



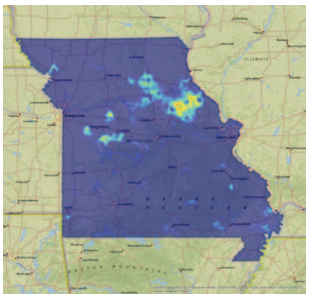
(a) Fish, Agriculture, Hunting



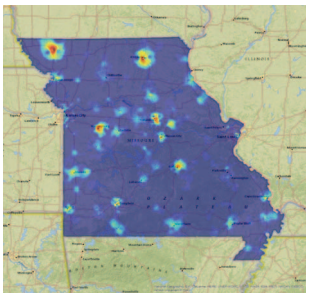
(b) Seasonal Recreational Use



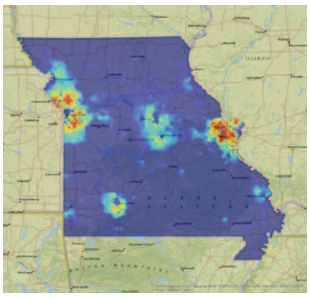
(c) Below Poverty Level



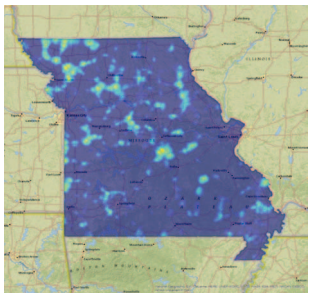
(d) Moved Within State, Not Citizen



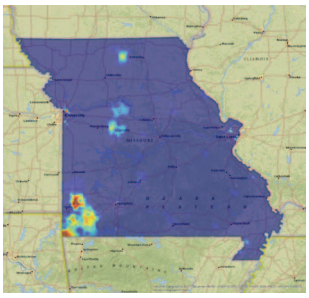
(e) Below Poverty Level, Employed



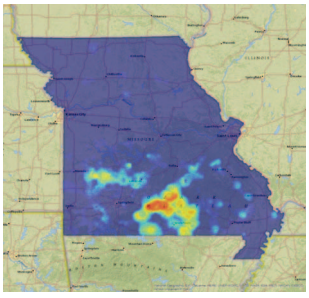
(f) Taxes: Mortgage of $3000+



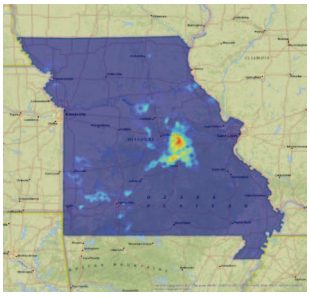
(g) Distance: Health Services



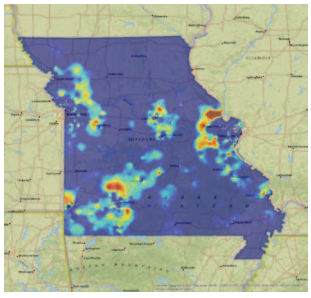
(h) Foreign, Spanish



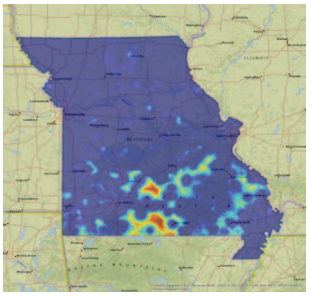
(i) Poverty, Over 65



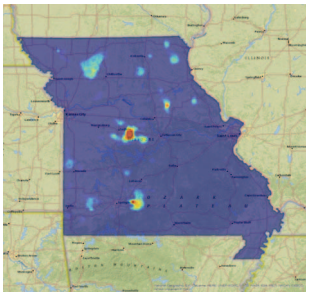
(j) Disability: Age 5-15, Self Care



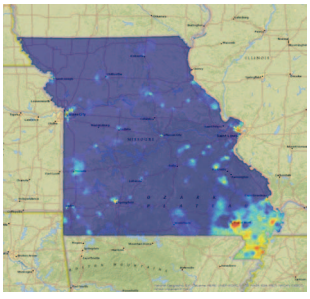
(k) Industry: Other



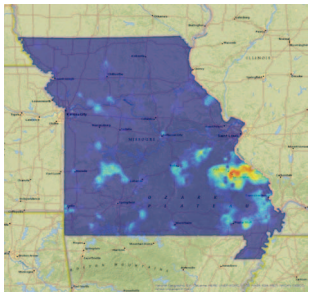
(l) Distance: Government City Hall



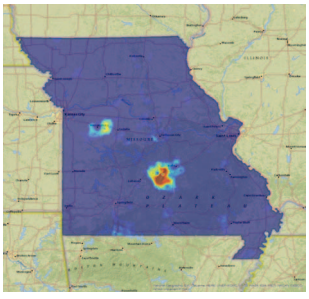
(m) Speaks Indo-European at Home



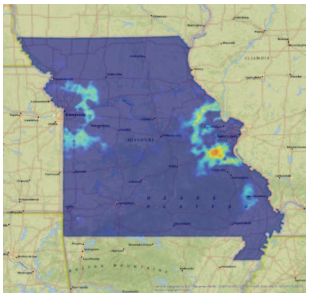
(n) Public Transport, No Vehicles



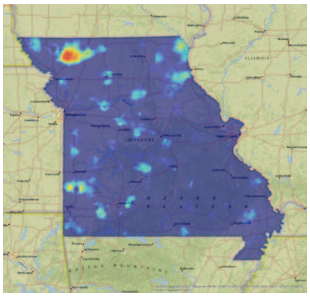
(o) Disability: 65+, Leaves Home



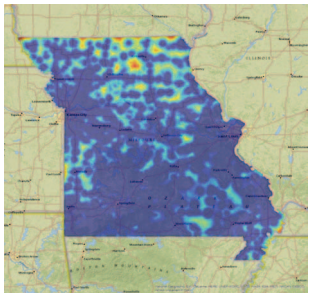
(p) Moved From Different State



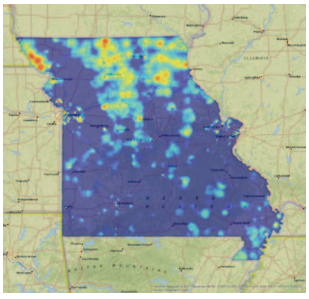
(q) Commute Time: 60-89 Minutes



(r) Moved within US, Not Citizen

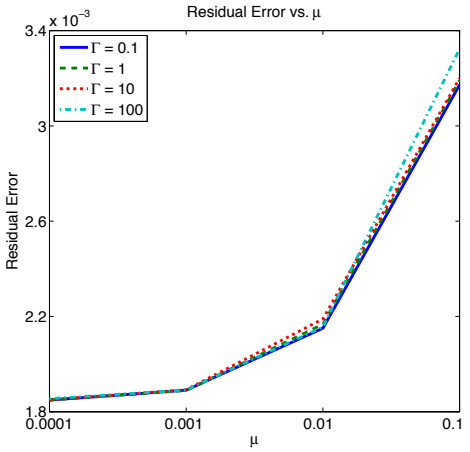


(s) Elementary Secondary Education

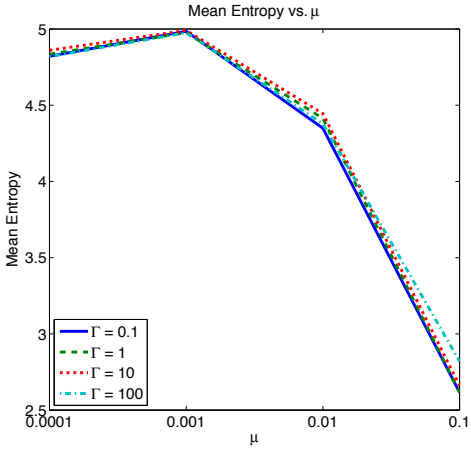


(t) Speaks only English

Fig. 6. Estimated abundance maps on the MO data cube using SPICE with $\mu = 0.1$ and $\Gamma = 0.1$.



(a) Residual Error

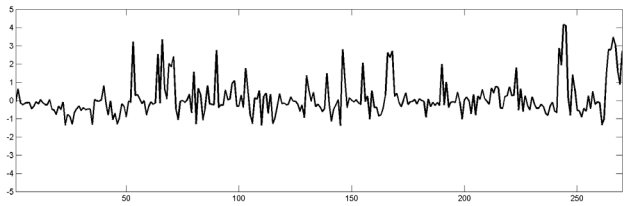


(b) Entropy

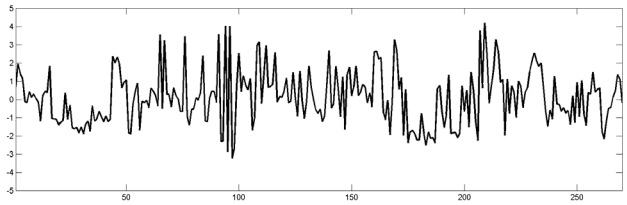
Fig. 5. Validity metrics computed for each set of tested parameters using input data, estimated endmembers and estimated abundance values.

20 endmembers. The associated abundance maps for all 20 endmembers are shown in Fig. 6.

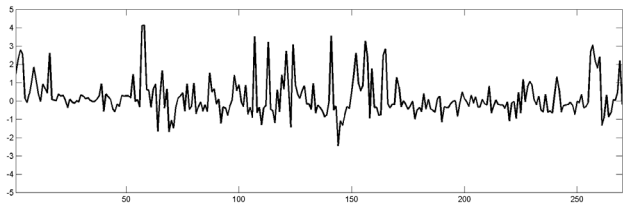
Once the abundance maps were estimated, the ability of this representation to accurately capture the information contained in the full MO data cube was needed to be examined. In order to evaluate the representation, the most informative layers associated with each endmember were identified. These were identified by computing the total squared difference between each endmember layer value to corresponding layer values in all other endmembers. Fig. 7 shows a few of the estimated endmember profiles. The horizontal axis spans the 270 data layers and the plots show the feature values, normalized across all endmembers to have zero mean and unit variance. The layer values with the largest difference from other endmembers were identified as the most representative layers for that endmember. The sub-caption of each abundance map in Fig. 6 lists the most representative layer for that endmember. In particular, consider endmember and abundance map (b). The most representative layer for this endmember is seasonal recreational use, and the abundance map clearly highlights the Lake of the Ozarks and Table Rock Lake, two of the largest recreational areas in the state. The peak at feature 244 in endmember profile (b) corresponds to this feature. Another interesting endmember is endmember (p), which highlights two military bases, Whiteman AFB and Fort Leonard Wood. The most representative layer for this endmember indicates areas with people who have moved from a different state, which seems reasonable for a military base. This is represented as the peak at feature 58 in endmember (p). Based on these examples, the estimated endmembers accurately represent these regions. Furthermore, the SPICE algorithm does not enforce any sort spatial smoothing or constraints, yet, the resulting abundance maps have spatial structures. This indicates that the described method can accurately characterize and visualize the spatial characteristics inherent in the data cube.



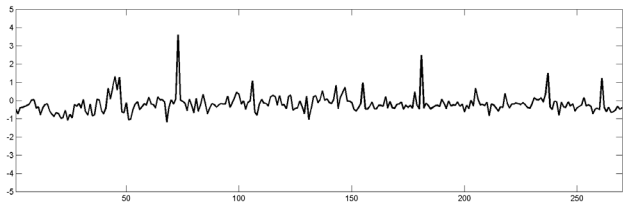
(a) Endmember (b)



(b) Endmember (f)



(c) Endmember (p)



(d) Endmember (r)

Fig. 7. Examples of estimated endmember profiles for the MO data cube.

VI. Future Work

Future work will include consider additional *mixing models* beyond the linear mixing model assumed by the SPICE algorithm. In particular, the enforcement of the sum-to-one and non-negativity constraints of the abundance values will be investigated and refined for this application and, then, related spectral unmixing algorithms will be developed. Also, further investigation into how to model individual values in each layer as a fuzzy number and use these in endmember analysis will be conducted. Further validity methods to estimate the appropriate algorithm parameter values and additional methods to identify the key layers for each endmember will also be investigated.

References

[1] J. Keller, M. Popescu, and D. Gibeson, "An extension of a confined space evacuation model to human geography," in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, July 2012, pp. 531–534.

[2] M. Popescu and J. Keller, "Implementing bounded rationality in disaster agent behavior using oga operators," in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, July 2012, pp. 5379–5381.

[3] M. Popescu, J. Keller, and A. Zare, "A framework for computing crowd emotions using agent based modeling," in *Computational Intelligence for Creativity and Affective Computing (CICAC), 2013 IEEE Symposium on*, April 2013, pp. 25–31.

[4] J. Keller, A. Buck, M. Popescu, and A. Zare, "A human geospatial predictive analytics framework with application to finding medically underserved areas," in *IEEE Symposium Series for Computational Intelligence (IEEE SSCI)*, 2014, Under Review.

[5] A. Zare, Z. Fields, J. Keller, and J. Horton, "Agent-based rumor spreading models for human geography applications," in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, July 2012, pp. 5394–5397.

[6] T. Haithcoat, T. Vought, and 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.

[7] J. Bezdek and R. Hathaway, "Vat: a tool for visual assessment of (cluster) tendency," in *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, vol. 3, 2002, pp. 2225–2230.

[8] L. Wang, X. Geng, J. Bezdek, C. Leckie, and R. Kotagiri, "Enhanced visual analysis for cluster tendency assessment and data partitioning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1401–1414, Oct 2010.

[9] T. Havens and J. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (ivat) algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 813–822, May 2012.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2001.

[11] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, April 2012.

[12] A. Zare and P. Gader, "Sparsity promoting iterated constrained endmember detection for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 3, pp. 446–450, July 2007.