Using Image Moment Invariants to Distinguish Classes of Geographical Shapes

J. F. Conley, I. J. Turton, M. N. Gahegan

[1]Pennsylvania State University
Department of Geography
302 Walker Building
University Park PA 16802
Email: [jfc173, ijt1, mng1]@psu.edu

1. Introduction

There are plenty of statistical methods and computer algorithms that can be used to find spatial disease clusters, yet simply finding the clusters is only the beginning of the knowledge discovery process (Fayyad et al. 1996). Once a cluster is found, an explanation for its existence needs to be hypothesized, and this task of identifying potential causes for the clusters is currently left to the epidemiologist. One computational approach to this problem was tried in the Geographical Explanations Machine (GEM) (Openshaw and Turton 2001), although it was largely unusable due to the very lengthy computation times involved. Instead of searching for correlations between the locations of the clusters and other variables in the dataset, as in GEM's approach, this research aims to identify potential causes based on the shape of the cluster. To do so, it adapts existing computer vision and pattern recognition techniques to the problem of geographical cluster analysis.

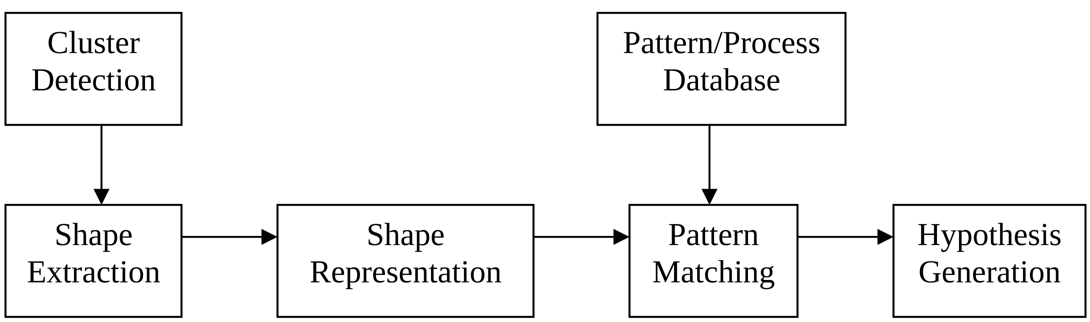

Figure 1. The software system of which this research is a part.

This works by comparing the shape of a detected cluster to those in a database containing shapes representative of various geographic processes. The entire computational process from cluster detection to hypothesis generation is shown in figure 1. This paper focuses on two sections of this system: the shape representation and pattern matching processes. Shapes are represented by a set of shape statistics that are computed for each shape and a pattern recognition system then assigns the detected cluster to one or more potential causes based on what shapes in the pattern/process database have similar values of the shape statistics. In this case, the shape statistics are the set of seven moment invariants developed by Hu (1962) and the classification system is a KNN classifier (Hastie et al. 2001, pp. 415-420). This paper reports on experiments into the ability of the Hu's invariants paired with a K-Nearest Neighbor (KNN) classifier to distinguish between different classes of geographical shapes in a pattern/process database.

2. Testing the Invariants

Two experiments were conducted with a set of shapes, classified at different levels of aggregation. The shapes are, where possible, taken from Pennsylvania. The first experiment contained twelve classes, as shown in table 1. This first experiment had a fine level of distinction between the classes, creating road, river, and soil shape classes for each of the physiographic regions of Pennsylvania because of the expected correlations between the terrain contours of an area and the shapes of roads, rivers, and soil regions in that area. Thus the different terrain patterns of the different physiographic regions could give rise to different road, river, and soil region shapes. Shapes for the roads and rivers were generated by drawing a 250-meter buffer around the centre lines from a GIS data layer. The shapes of the soil regions are taken directly from a polygon GIS data layer. The shapes for all classes were taken from publicly available GIS data layers from PASDA (Pennsylvania Spatial Data Access) and the USGS (United States Geological Survey), and there are 350 shapes in each class.

The second test condenses these twelve classes down to four classes: urban areas, soil regions, roads, and rivers, using 450 shapes from each aggregated class. The 450 shapes are evenly divided between the subgroups for roads, rivers, and soil regions.

Urban areas in the United States	Rivers in the ridge and valley region
Roads in the PA ridge and valley region	Rivers in the Allegheny plateau
Roads in the Allegheny plateau in PA	Rivers in the PA piedmont
Roads in the piedmont region of PA	Soil regions in the ridge and valley region
Roads in Pittsburgh	Soil regions in Allegheny plateau
Roads in Philadelphia	Soil regions in the PA piedmont

Table 1. Shape classes in the 12-class test.

2.1 About the Invariants

Moment invariants are derived from the theory of statistical moments (Casella and Berger 2002, pp. 59-68). Hu (1962) developed seven invariants for computer vision by applying the concepts from bivariate statistical moments to pixel values in gray-scale images.

Two-dimensional statistical moments, $m_{p,q}$ take the form of equation 1: the expected value (denoted $\mathrm{E}(\cdot)$) of the two variables, *X* and *Y*, raised to integer powers, *p* and *q*. In a geographical context, *X* and *Y* are spatial variables and $f(x,y)$ is a function that varies across space.

$$m_{p,q} = \mathrm{E}(X^p Y^q) = \sum_{x\in X}\sum_{y\in Y} x^p y^q f(x,y) \tag{1}$$

Extending these moments to image processing, $f(x, y)$ is the intensity of the pixel at (x, y). Furthermore, for black and white images, $f(x, y) = 1$ if the pixel is black and 0 if it is white. To make these moments invariant to location, so that if a shape is moved in any direction, the moment values remain constant, we shift the *x* and *y* values so that the shape's centroid is at (0, 0), giving centralized moments, $\mu_{p,q}$, in equation 2.

$$\mu_{p,q} = \sum_{x \in X} \sum_{y \in Y} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{2}$$

Next, the centralized moments are normalized with respect to area, giving normalized moments, $\eta_{p,q}$ in equation 3.

$$\eta_{p,q} = \frac{\mu_{p,q}}{(\mu_{0,0})^{\gamma}} \text{ where the normalization factor } \gamma = \tfrac{p+q}{2} + 1 \tag{3}$$

Finally, the seven invariants in equations 4 through 10, which are also made invariant to rotation, are given by Hu (1962).

$$\phi_1 = \eta_{2,0} + \eta_{0,2} \tag{4}$$

$$\phi_2 = (\eta_{2,0} - \eta_{0,2})^2 + 4\eta_{1,1}^2 \tag{5}$$

$$\phi_3 = (\eta_{3,0} - 3\eta_{1,2})^2 + (3\eta_{2,1} - \eta_{0,3})^2 \tag{6}$$

$$\phi_4 = (\eta_{3,0} + \eta_{1,2})^2 + (\eta_{2,1} + \eta_{0,3})^2 \tag{7}$$

$$\begin{aligned} \phi_5 = {} & (\eta_{3,0} - 3\eta_{1,2})(\eta_{3,0} + \eta_{1,2})[(\eta_{3,0} + \eta_{1,2})^2 - 3(\eta_{2,1} + \eta_{0,3})^2] + \\ & (3\eta_{2,1} - \eta_{0,3})(\eta_{2,1} + \eta_{0,3})[3(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2] \end{aligned} \tag{8}$$

$$\phi_6 = (\eta_{2,0} - \eta_{0,2})[(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2] + 4\eta_{1,1}(\eta_{3,0} + \eta_{1,2})(\eta_{2,1} + \eta_{0,3}) \tag{9}$$

$$\begin{aligned} \phi_7 = {} & (3\eta_{2,1} - \eta_{0,3})(\eta_{3,0} + \eta_{1,2})[(\eta_{3,0} + \eta_{1,2})^2 - 3(\eta_{2,1} + \eta_{0,3})^2] - \\ & (\eta_{3,0} - 3\eta_{1,2})(\eta_{2,1} + \eta_{0,3})[3(\eta_{3,0} + \eta_{1,2})^2 - (\eta_{2,1} + \eta_{0,3})^2] \end{aligned} \tag{10}$$

To improve classification rates, the natural logarithm of the absolute value of the seven invariants are used instead of the invariants themselves, because the invariants often have very low absolute values (less than 0.001), so taking the logarithm of the invariants reduces the density of the invariants near the origin. Also, the values of the natural logarithms are then converted into standardized z-scores before classification.

2.2 Pattern Recognition with a KNN Classifier

Once these seven invariants are computed for each shape in the database, a classification algorithm can be applied to the seven-dimensional shape vectors in the data space created by the invariants to determine how well these invariants distinguish between the twelve shape classes listed in table 1 or between the four aggregated shape classes.

This research used a K Nearest Neighbor classifier (Hastie et al. 2001, pp. 415-420) because a KNN classifier does not make any assumptions about the form of the distribution of the vectors in the data space. A preliminary examination of the distribution suggests a skewed distribution rather than a standard normal distribution, so a classifier that does not rely on a particular distribution is preferred.

3. Results

Classification accuracy rates for all values of *k* between 1 and 100 on both the 12-class and the aggregated 4-class tests were computed using a leave-one-out validation method in which each shape's class was estimated as the class that is most frequent in the *k* nearest neighbors in the rest of the dataset. Figures 2 and 3 show the accuracy rate for each class and value of *k* in the 12-class and 4-class tests respectively.

A confusion matrix for the 12-class test, which is not included because of space considerations, shows that for many of the misclassified shapes, they are still classified in the same general category, i.e., urban areas, roads, rivers, and soils.

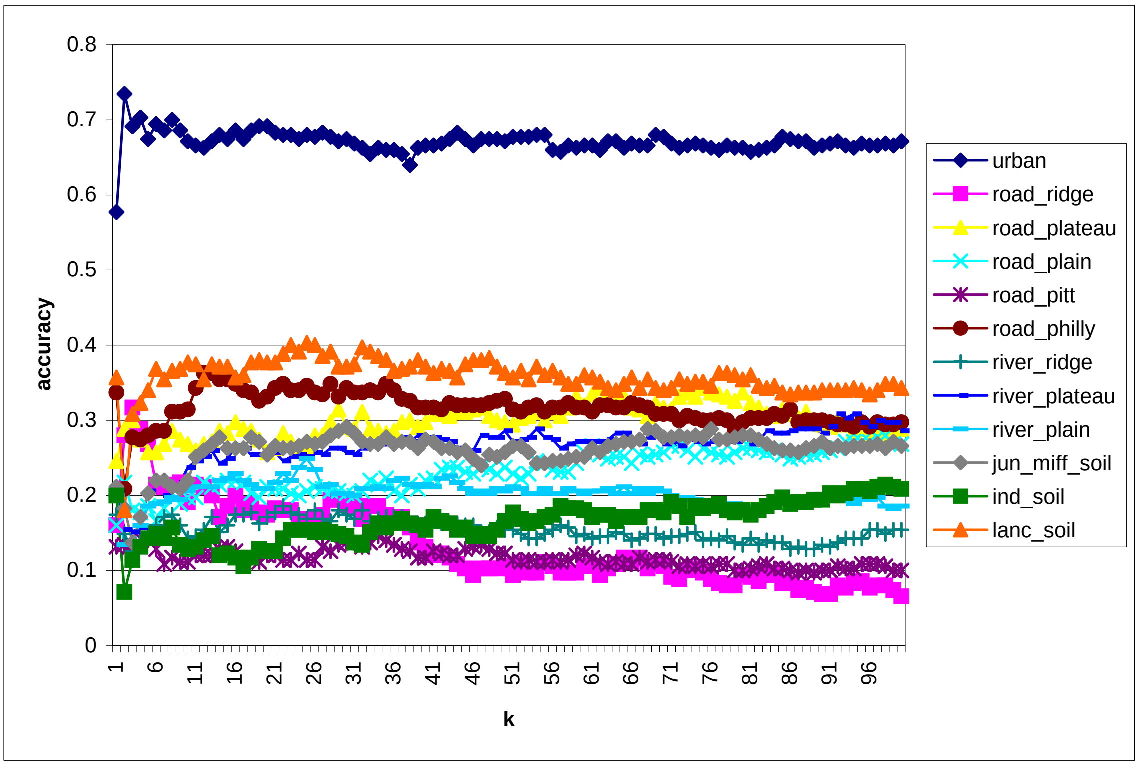

Figure 2. Accuracy rates for each class on the 12-class test.

4. Discussion

In both the 12-class and 4-class tests the KNN classification of the standardized log invariants was better than random (which would give accuracy rates of 0.083 and 0.25 respectively) for most classes and most values of *k*. The worst classification was that of soils in the 4-class test. Further investigation shows that the classification for soil region shapes in the piedmont region are often misclassified as urban area shapes, giving this sub-group an accuracy rate well below random in the 4-class test (e.g., an accuracy of 0.14 when $k = 35$), which decreases the accuracy of the combined soil class.

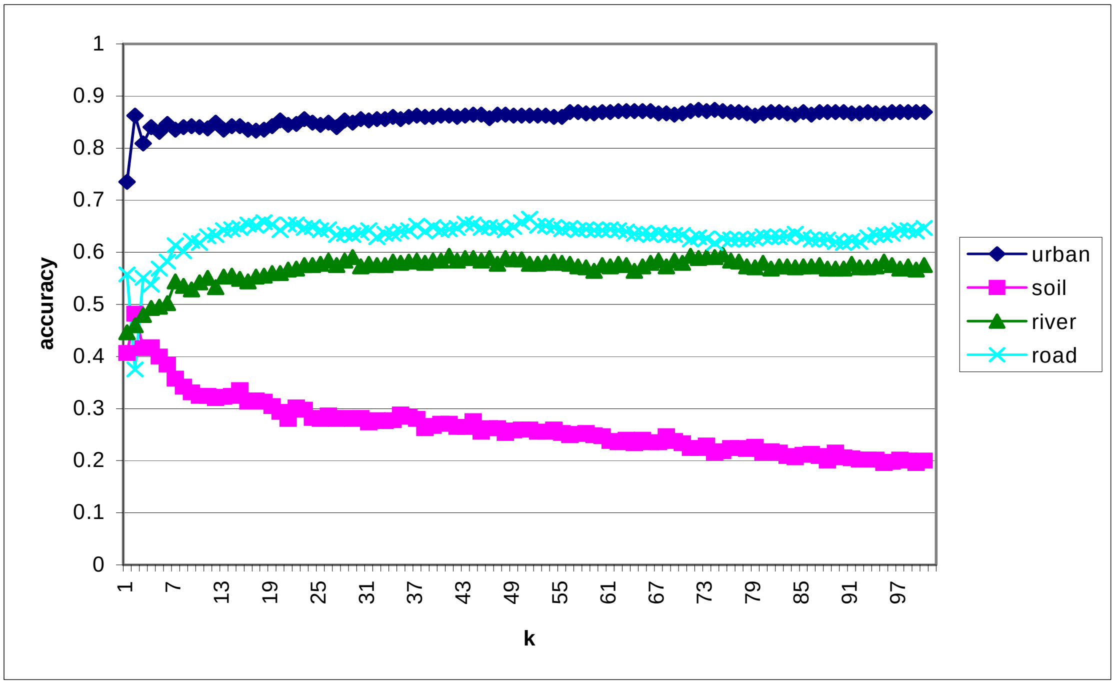

Figure 3. Accuracy rates for each class on the 4-class test.

Further testing will determine the ability of this system to identify the shape classes in the context of cluster detection and analysis. This will be done by generating a synthetic dataset containing disease clusters among a background population where the clusters are based on shapes from this pattern/process shape database, running a cluster detection algorithm on the synthetic dataset, and classifying the detected clusters based on their shapes. If these seven invariants and a KNN classifier fail, other image processing statistics, such as Legendre or Zernike moments (Teh and Chin 1988), and other pattern recognition classifiers, such as a maximum likelihood classifier (Hastie et al. 2001, pp. 229-231), can be used.

5. Acknowledgements

The authors would like to acknowledge the Academic Computing Fellowship program from the Pennsylvania State University Graduate School for financial support of the first author's doctoral research, of which this is a part.

6. References

Casella G, and Berger RL, 2002, *Statistical Inference, 2nd Edition*. Duxbury, Pacific Grove, CA, USA.

Fayyad U, Piatetsky-Shapiro G, and Smyth P, 1996, From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37-54.

Hastie T, Tibshirani R, and Friedman J, 2001, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, USA.

Hu M-K, 1962, Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8:179-187.

Openshaw S, and Turton I, 2001, Using a Geographical Explanations Machine to Explore Spatial Factors Relating to Primary School Performance. *Geographical & Environmental Modelling*, 5(1):85-101.

Teh C-H, and Chin RT, 1988, On Image Analysis by the Methods of Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496-513.