Eigenvector spatial filtering for image analysis: An efficient algorithm

Article · January 2010

CITATION
1

READS
2,893

1 author:

Melissa J. Rura-Porterfield

18 PUBLICATIONS **511** CITATIONS

SEE PROFILE

EIGENVECTOR SPATIAL FILTERING FOR IMAGE ANALYSIS:

AN EFFICIENT ALGORITHM

by

Melissa J. Rura

APPROVED BY SUPERVISORY COMMITTEE:

_____

Dr. Denis Dean, Chair

_____

Dr. Brian J. L. Berry

_____

Dr. Michael Tiefelsdorf

_____

Dr. Fang Qiu

In loving memory of Jerome Paul Tolene

(1952-2009)

EIGENVECTOR SPATIAL FILTERING FOR IMAGE ANALYSIS: AN EFFICIENT ALGORITHM

by

MELISSA J. RURA B.S., M.S.

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

GEOSPATIAL INFORMATION SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

December, 2010

ACKNOWLEDGEMENTS

I would like to thank my family for their support, my professors for their insight and patience, and my fellow students for their friendship. In particular, I would like to thank Larry Keen, Renne Fister, Rob Donnely and Michael Profit for a firm foundation in mathematics; Tom Kind, and Neil Weber for their love and understanding of physical geography; Edward Mikhail, James Bethel, Boundwin van Gelder, and Jie Shan for their many projects and exciting insight in the world of geomatics; Gilbert Strang for sharing his classes on the internet and inspiring me to hold tight to my love of mathematics; and finally to Daniel Griffith, Ronald Briggs, Brain Berry, Michael Tiefelsdorf and Fang Qiu for introducing me to the wide world of geography in social science and laying the foundation of an interesting intermingling of physical, social, and quantitative geography that became my dissertation. Especially I would like to acknowledge my husband, Eugene, and my children, Nikolai and Mikhail, for living out the reality that a wife and mother can follow her dreams and still remain "*Babes*" and "*Mama*".

August, 2010

EIGENVECTOR SPATIAL FILTERING FOR IMAGE ANALYSIS:

AN EFFICIENT ALGORITHM

Publication No. _____

Melissa J. Rura, Ph.D.
The University of Texas at Dallas, 2010

Supervising Professor: Dr. Denis Dean

Eigenvector Spatial Filtering (ESF) is an established method in social science literature for incorporating spatial information in model specifications. ESF computes spatial eigenvectors, which are defined by the spatial structure associated with a variable. One important limitation of this technique is that it becomes computationally intensive in image analysis because of the massive number of image pixels. This research develops an algorithm, which makes ESF more efficient, by using the analytical solution for the eigenvalues and spatial eigenvectors, which are essentially a series of orthogonal, uncorrelated map patterns that describe positively spatial autocorrelated patterns through negatively spatially autocorrelated patterns, and global, regional, and local patterns of spatial dependencies in a surface. A reformulation of the analytical solution reduces the required computations and allows the eigenvectors to be computed sequentially. Finally, a series of sampling methods are explored. This algorithm is applied to three example multispectral images of different sizes: small (i.e., ~200,000 pixels), medium (i.e., ~1,000,000

pixels) and large (i.e., ~110,000,000 pixels) and is evaluated in terms of output for each sampling technique and the complete spectral information. The output spatial filters of these sampling techniques compare to the filter generated with the complete spectral information. In terms of efficiency evaluation, the time is required to construct filters through sampling versus through analysis of the complete image surface is evaluated and the complexity of set-up and execution of the sampled and distributed algorithms are assessed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Spatial information is increasingly incorporated into a variety of techniques used in image analysis. Specifically, spatial information is often added to raw imagery data in the form of moving window statistics, textures, distance measures, geostatistical interpolations, and Bayesian predictions (NRC 1991). Many of these techniques are also prevalent in social science literature. Eigenvector spatial filtering (ESF) is an established method discussed in social science literature for incorporating spatial information in regression models specifications (Griffith 1996, Dray 2006, Griffith and Peres-Neto 2006, Tiefelsdorf and Griffith 2007), and here this is extended for use in image processing specifications. ESF computes spatial eigenvectors, which are defined by the spatial structure associated with a specific variable. Those spatial eigenvectors are essentially a series of "special" (i.e., orthogonal, uncorrelated) map patterns, which describe positive (i.e., clustered) through negative (i.e., dispersed), spatial dependencies in a surface.

There are several possible applications of ESF in image analysis that should be investigated. The most promising of these are in description and diagnostics of image data, thematic image and map comparisons, and image classification and change detection. Since the map patterns that are produced in ESF are orthogonal and uncorrelated, they can provide useful information about what patterns are present in an image and which of those patterns are shared between images or between sets of images. These map patterns also give insight to whether the shared or unique patterns are global, regional or local in scope. This may be particularly useful in a

change detection setting, where a series of images of the same region are compared over time, or in accuracy assessment where the spatial change between classifications is of interest.

One important limitation of ESF is that it becomes computationally intensive when applied in an image analysis setting. Often imagery consists of a massive number of pixels and since the number of eigenvectors associated with the image is equal to the number of pixels in the image as the number of pixels increase so does the computational intensity of ESF. If ESF were to be applied to imagery in the same fashion as it is applied in social science settings, the required computation time could be prohibitive because of the sheer number of pixels in the image. There are, however, several attributes of an image variable that might be used to circumvent the prohibitive nature of the algorithm. These attributes include being able to assume an underlying image surface that is a complete four sided region (i.e., no holes) and that this surface is composed of a series of regularly spaced four sided regions (i.e., pixels).

This research develops an algorithm that makes ESF more efficient to enable further research of its application in image analysis by addressing the issues of creating a massive connectivity matrix, computing a large number of eigenvectors that are not used in an analysis, and formulating a more efficient computation of the eigenvectors themselves.

The first issue addressed is overcoming the need for a massive connectivity matrix. A typical LANDSAT TM image with dimensions of 3240-by-2340 pixels (p-by-q) would have a connectivity matrix with dimensions 7,581,600-by-7,581,600 (pq-by-pq); this massive size could be a serious computational liability. Here, this challenge is addressed by using a theorem presented by Griffith (2000), which gives the analytical solution, a function which computes the exact spatial eigenvectors for the binary first-order rook or queen connectivity matrix of the

square tessellation (see Chapter (2) for more details). This analytical solution not only circumvents the issues of constructing, storing and using a massive connectivity matrix, but also allows each eigenvector to be calculated sequentially, which makes practical divide-and-conquer distributed computing techniques.

The second inefficient aspect of conventional spatial filtering is that *n* eigenvectors are computed for a spatial surface but only a candidate subset of those eigenvectors is necessary for the ESF algorithm. A large number of eigenvectors are immediately disregarded because they do not contain the desired type or amount of spatial autocorrelation. For a spatial surface with only a few hundred regions, these few unnecessary calculations are not problematic; but as the number of regions increases to thousands or millions, the number of eigenvectors being extracted that are immediately discarded because they represent too little or the wrong type of spatial dependence rapidly increases to unmanageable levels. A more efficient approach is to identify which eigenvectors contain the desired amount and type of spatial autocorrelation and construct only them. Identifying these eigenvectors is possible using the eigenvalue associated with an eigenvector as a measure of spatial autocorrelation (Tiefelsdorf et. al. 1995), which allows individual eigenvectors containing a specific type and / or amount of spatial dependence to be constructed.

The third and most time consuming aspect of the ESF algorithm is choosing which eigenvectors from a candidate set describe meaningful spatial information in an image so that they can be used to construct a spatial filter. This step is extremely time consuming, and when formulated inefficiently, computationally prohibitive. This is first addressed via a reformulation of the conventional analytical eigenvector notation that significantly reduces the number and

complexity of the required computations. Second a linear regression model with a normal error term is assumed, which allows a simple correlation between the standardized image band and the centered spatial eigenvectors to be conducted. Eigenvectors with a threshold amount of correlation with the image band are retained for construction of the spatial filter. Finally, two different strategies are explored for further reduction in computational intensity. First, is a brute force method that utilizes a simple linear regression between *all* candidate eigenvectors and *all* image bands (this process is made more efficient through the use of a divide and conquer distributed computing algorithm). Second, three sampling methods (Berry and Baker 1968; Stehman and Overton 1996) ---random areal sample, geographically stratified random sample, and systematic sample --- that reduce computation and memory requirements are tested.

This algorithm is applied to three case study images of different sizes. The first image dataset has 12 bands and 208,560 pixels, and is published in Landgrebe (2003). The second dataset is a subset of a Landsat ETM+ image with 7 bands--- the thermal band is removed--- and 1,000,000 pixels, downloaded from the Earth Science Data Interface at the Global Land Cover Facility (GLCF 2010). The third dataset is a four-band Quickbird image with 110,508,120 pixels published in Congalton and Green (2009). Each image is multi-spectral with z-score standardized digital number values. Data diagnostics are done for all bands for each multispectral image, including a Ryan-Joiner normality statistic; Moran's I and Geary Ratio spatial autocorrelation statistics, and a band multicollinearity matrix. A series of candidate threshold values (i.e., 0.25, 0.5, and 0.75) are used to identify the candidate set of spatial eigenvectors to be tested for inclusion in a spatial filter. Finally, a series of variance thresholds

(i.e., 0.001 and 0.0001) are used to determine inclusion of a candidate eigenvector in a final spatial filter are also empirically evaluated.

The following chapter gives an introduction to the concepts most pertinent to understanding eigenvector image spatial filtering. It begins with defining the spatial structure of an image, gives a discussion measures of spatial autocorrelation and their use within image analysis. The chapter further discusses the importance of orthogonal variables and the construction of spatial eigenvalues and eigenvectors. Finally Chapter 2 discusses conventional eigenvector spatial filtering and gives information on its prior applications.

CHAPTER 2

BACKGROUND

To understand the foundation of eigenvector image spatial filtering this section briefly summarizes Griffith and Fellows (1999) and Fellows (1998), who use a 250,000 pixel subsection of a Landsat TM image to attempt to construct a spatial multivariate image analysis (MIA) classification of blow-down sites in a small forested area using eigenvector spatial filters for each image band (i.e., in his case, 7 Landsat TM bands and 7 spatial filter bands). A strategy to use the analytical solution for the eigenvalues and eigenvectors to implement a stepwise regression procedure to select spatial eigenvectors that account for variation in each image band was proposed. Fellows (1998) describes problems that were encountered with this proposed methodology and how it had to be revised in order to be implemented.

One of the original goals of Fellows' study was to “use prominent eigenvectors as locational information that represents significant map patterns found in the seven spectral bands" [p.78, Fellows (1998)]. This goal was not attained because of the lack of an “efficient way of sifting through the massive number of distinct map patterns"[p.81, Fellows (1998)]. Fellows (1998) found that computing a spatial filter for an image was very computationally demanding. Although the analytical solution for the eigenvectors and eigenvalues made the large connectivity matrix unnecessary, the eigenvector calculation and selection process were both order n^2 calculations. Although the speed of processors has increased significantly since Fellows' (1998) work, this is still a daunting task in terms of number of operations.

Fellows was unable to follow through on exactly what he proposed. Instead, an exploratory analysis of a series of the most predominate eigenvectors was conducted. Initially he identified a candidate set of 62,500 eigenvectors out of a possible 250,000. He proposed constructing seven stepwise regression models, one for each of his seven Landsat image bands, using the 62,500 candidate eigenvectors as the dependent variables in every model. He found the hardware platform used "took several days to complete the stepwise regression for each band" (Fellows 1998, pg. 78). He also found "indentifying significant eigenvectors … proved nearly impossible as many accounted for only a fraction of a percentage of the total variance" (Fellows 1998, pg. 78). Fellows gives a table listing ten eigenvectors accounting for the most variation in each band, but the amount of variation accounted for by those spatial eigenvectors is not reported. It is likely that with only ten spatial eigenvectors very little of the spatial information in the image is captured by this spatial filter most likely due to a variance threshold that is too high, although this is not explained by Fellows.

The massive number of candidate eigenvectors made it impossible for Fellows to compute an image spatial filter. In chapter 5 and 6 of his thesis, Fellows explains that although he could not compute an image spatial filter he could still include spatial information in the MIA classification by using a spatial lag variable for each image band. A spatial lag variable computes an average of the surrounding pixels for every pixel in the image and might be determined through the definition of connectivity and adjacency rules. The use of spatially lagged variables allowed Fellows to bypass choosing which spatial eigenvectors are significant in a regression context and alleviated the problem of sifting through the massive number of eigenvectors. This

spatial information although useful in capturing the spatial relationships in the data is no longer orthogonal.

The study by Griffith and Fellows (1999) is an integral part of the foundation upon which this dissertation is based. This dissertation addresses the call by Griffith and Fellows (1999) for a “more efficient strategy" to identify significant spatial eigenvectors. This dissertation pursues a strategy to make both the calculation of the spatial eigenvectors themselves and the selection of the significant spatial eigenvectors more efficient. This gain in efficiency allows eigenvector image spatial filtering to be tested in a variety of image analysis applications. This chapter discusses previous literature from both social science and remote sensing, which provide insight and background into eigenvector spatial filtering. A discussion of definitions, notation conventions and terms used throughout the remainder of this document are given.

2.1 IMAGE STRUCTURE: DEFINITIONS AND NOTATION

For clarity, this discussion of image spatial filtering begins with some definitions. As is also the case for conventional georeferenced data, a multispectral image might be thought of logically as two pieces of information, the image spatial structure and the image digital number information. This section discusses the distinction between these two pieces of information and the parameters used to describe them. These concepts are also illustrated in Figure (2.1).

A multispectral image is by definition composed of multiple spectral bands, each band has the same spatial structure, but every band contains different spectral information. The *n* recorded digital number values associated with each pixel in a single image band are stored in an image band vector ($\mathbf{b}_m$), where *m* refers to the band index. For analysis purposes, all spectral

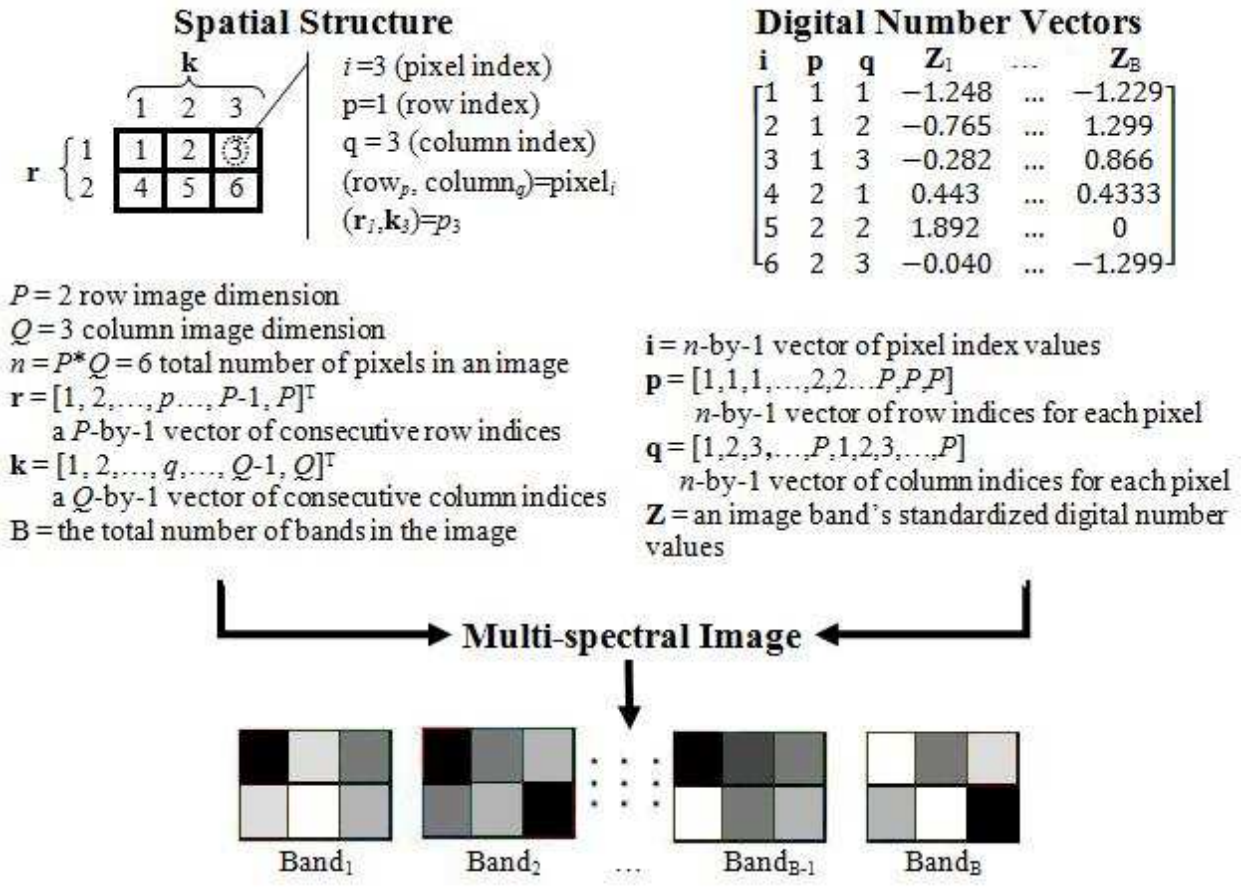


Figure 2.1. 2 pixel - by - 3 pixel multispectral image example used to illustrate the definitions and notation used through throughout the text.

information in each band of a multi-spectral image is standardized, making the band's mean zero and standard deviation one; these standardized bands are denoted $\mathbf{Z}_1, \ldots \mathbf{Z}_m, \ldots \mathbf{Z}_B$, where *B* is the total number of image bands. Next, the spatial structure of an image is defined by the conceptualization of how these pixels are connected together to create a continuous spatial surface. Unless otherwise stated, throughout this text the surface connectivity of the pixels in the image relates to the rook adjacency rule, see Section 2.2.1 for details on connectivity rules.

If the spatial tessellation for each image band is assumed to be a regular four sided region with no holes constructed of smaller regular four sided regions (i.e., pixels) then the spatial structure is defined by several parameters and a connectivity rule. First, an image's size is defined by the number of rows (P) and number of columns (Q) it contains. The total number of pixels in an image band (n) is determined by these dimensions ($n = P*Q$). Each pixel in an image band has an index (i) that is uniquely determined by its row index (p) and column index (q) through the equation $i = (p - 1)*P + q$. All row indices can be collected in a row index vector[1] called $\mathbf{r} = [1, 2, \ldots, p, \ldots, P_{n-1}, P]^{T}$. Similarly, all column indices can be collected in a column index vector called $\mathbf{k} = [1, 2, \ldots, q, \ldots, Q_{n-1,} Q]^{T}$. To link the spatial structure and the spectral information of a multi-spectral image, each entry in a standardized image (i.e., each entry in $\mathbf{Z}_m$) must be related to a pixel index in the spatial structure (i), as show in Figure (2.1). The vector **i** collects all n pixel indices of a spatial surface into an n-by-1 vector. The row and column indices that uniquely define each pixel index are collected in the n-by-1 vectors **p** and **q**, respectively (i.e., $\mathbf{i}^{T} = [1, 2, 3 \ldots, n-1, n]$, $\mathbf{p}^{T} = [\underbrace{1,1, \ldots,1}_{\text{Q times}}, \underbrace{2,2, \ldots,2}_{\text{Q times}}, \ldots, \underbrace{P, P, \ldots, P}_{\text{Q times}}]$ and $\mathbf{q}^{T} = [\underbrace{1,2, \ldots Q, 1,2, \ldots, Q, \ldots,1,2, \ldots, Q}_{\text{P times}}]$).

It is important to note the difference between vectors **r** and **k**, and vectors **p** and **q**. The values of **r** and **k** are the row and column indices numbered consecutively from 1 to P and 1 to Q, respectively, without duplication. The values of **p** and **q** are the row or column indices that relate to a particular pixel index and therefore there is repetition of row and column indices within each vector. The sizes of **r** and **k** are P-by-1 and Q-by-1, respectively, while the sizes of

[1] Throughout the text column vectors are transposed using the superscript symbol ᵀ to conserve space

p and **q** are both *n*-by-1. The distinction between these vectors becomes important in order to gain efficiency computing the eigenvectors.

Figure (2.1) shows a simple 2-row-by-3-column multispectral image using the notation discussed above, which is used throughout the remainder of this document. In the Appendix B, the example image given in Figure (2.1) is used to illustrate the methodology explained in chapter 3 and thereby make some abstract ideas more concrete. However, keep in mind that the techniques described here are designed for massive datasets with over 10,000 observations, and that in cases of moderate-to-large datasets of under 10,000 the creation of a conventional connectivity matrix is advisable.

2.2 SPATIAL SURFACE STRUCTURE

There are many conceptualizations of a spatial surface that allow for the abstract of reality into a model. For instance, an object which is discrete in space, such as a house, might be conceptualized as a point, a road as a line, or a country border as a polygon. A field conceptualization might be more appropriate for a continuous variable, which has an infinite number of point locations that must be made finite through contour lines, pixel representations, or sample point locations, such as sea level or temperature (Hainning 2003). Here a pixel is conceptualized not as a field but more like a boarder of a polygon, with the digital number representing the information stored in that polygon. It is assumed that the this digital number information is constant across the pixel (i.e., polygon). This assumption introduces error into the model since it is known that area on the ground covered by the pixel will not always be homogenous.

Based on one of these conceptualizations of a spatial surface, it is possible to conceptualize the spatial structure associated with that surface. This structure might be based on interactions between point samples, the sharing of common lines between nodes, the existence of a common border between regions, the percentage of common boundary a region shares with its neighbor, or many other possible forms of interaction. The following section gives an overview of defining spatial structure, emphasizing adjacency connectivity and the creation of a binary connectivity matrix. For more discussion on spatial surface conceptualizations see Hainning (2003), and for further discussions of spatial structure Tiefelsdorf (2000).

2.2.1 Connectivity

As a part of understanding the spatial surface of a variable, it is important to consider how objects or fields are connected. For definitions based on adjacency, this can be done by specifying the “case” and “order” of the connectivity. Generally speaking, regions can be connected with other neighboring regions either along an edge or at a shared point. Regions connected along edges and not at single points are referred to as having the rook's case adjacency; regions connected along edges as well as at single points are referred to as having the queen's case adjacency. Figure (2.2) illustrates these relationships for both a regular square tessellation and an irregular tessellation. For the square tessellation, the distinction between a point and an edge is easily visualized, while for an irregular tessellation the distinction is less important since it happens less often that polygons meet only at one point. As can be seen in Figure (2.2) there is only one region difference, for this example, between the rook and queen specification for the irregular surface, while for the regular square tessellation there is always a four neighbor difference.

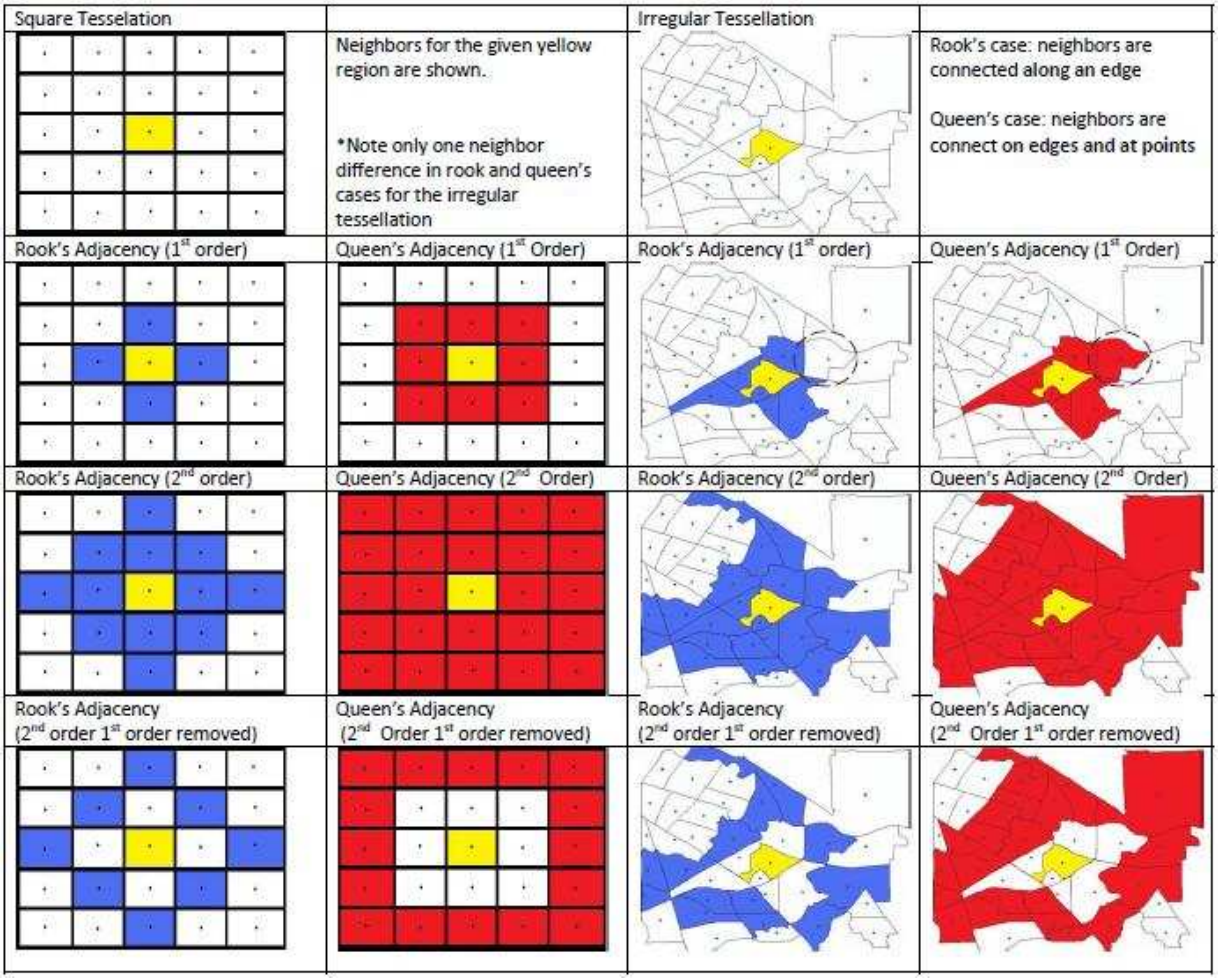


Figure 2.2. Rook's case (blue) and Queen's case (red) adjacency and order of adjacency for a particular region

Also illustrated in Figure (2.2) is the order of adjacency (i.e., the number of lags), which refers to how many cells away from the original region the neighbors may be counted. Adjacency of order 1 allows only cells directly adjacent to the original cell as neighbors, while second-order adjacency allows a distance of two cells. All orders of adjacency must not be included as neighbors; if in second-order adjacency the first order neighbors are not included this creates a sort of buffer-zone of non-neighbors around the original cell. For more information on

connectivity specified with distance measures for the general specification see Cliff and Ord (1981) and for its use in spatial filtering Patuelli et al. (2006).

2.2.2 Spatial Weight Matrices

Once the case and order of adjacency are conceptualized, a binary connectivity matrix can be constructed. If two regions are connected, then a one is placed in the matrix where the column and row of those regions intersect.

The example given in Figure (2.1), a 2-by-3 spatial tessellation from which the 6-by-6 **C** and **W** connectivity matrices shown in Equation (1) are constructed assume the rook's contiguity rule and an order of one.

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \rightarrow \mathbf{W} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}. \quad (1)$$

Region 2 is connected to region 5, so matrix entries (2,5) and (5,2) in Equation (1) are coded 1. This coding also insures that the matrix is symmetric around its main diagonal. Also, by definition, a region is not connected to itself; therefore, the matrix's main diagonal contains only zeros.

A binary connectivity matrix is often denoted by **C**, whereas a row standardized spatial weights matrix, **W,** has each entry in a row is divided by the total number of neighbors for that region - the row sum. This adjustment insures that every row sums to one, which makes sense spatially: if a region has many neighbors, each individual neighbor would have relatively less

influence. A non-symmetric square matrix, such as **W** can be made symmetric when pre- and post- multiplied by a projection matrix such as $\mathbf{D}^{-0.5}$ where **D** is the diagonal linkage degree matrix.

Tiefelsdorf et al. (1999) give these two connectivity specifications as the extremes of a whole family of spatial weights specifications that take the form:

$$A_a = \frac{n}{d^{a+1}} D^a C \tag{2}$$

where **A** is a general adjacency matrix and **d** - the vector of linkage degrees - is defined to be the number of neighbors (i.e., links) a region has (i.e., the row sums of **C**) and **D** - the linkage degree diagonal matrix – is a diagonal matrix with **d** as entries in its main diagonal. The element-wise power, parameter *a*, varies between 0 and -1, creating a series of connectivity matrices. For example, by specifying the *a* parameter to be -1, the **W** matrix is created (i.e., $\mathbf{A}_{-1} = \mathbf{W}$); similarly $a = 0$ specifies a normalized **C**, which is the binary connectivity matrix standardized by the total number of regions divided by the total number of links degrees (i.e., $\mathbf{A}_0 = (n/\mathbf{d}_{(0+1)})\mathbf{C}$). Tiefelsdorf et al. (1999) give Equation (2) and argue that specifying $a = 0.5$ creates a variance stabilized weight matrix. Using the example surface from Figure (2.1), the linkage degree for each region is $\mathrm{d} = [\,3\ 3\ 2\ 3\ 3\ 2\,]^{\mathrm{T}}$, and letting the parameter $a = 0.5$, an example calculation for a general connectivity matrix is as follows:

$$\mathrm{A}_{-0.5} = \frac{6}{d^{(-0.5+1)}} \mathrm{D}^a \mathrm{C}$$

$$\mathbf{A}_{-0.5} = \begin{bmatrix} 3.46 \\ 3.46 \\ 4.24 \\ 3.46 \\ 3.46 \\ 4.24 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}^{-0.5} \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{A}_{-0.5} = \begin{bmatrix} 3.46 \\ 3.46 \\ 4.24 \\ 3.46 \\ 3.46 \\ 4.24 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0.707 & 0 & 0.707 & 0 & 0 \\ 0.577 & 0 & 0.577 & 0 & 0.577 & 0 \\ 0 & 0.577 & 0 & 0.577 & 0 & 0.577 \\ 0.577 & 0 & 0.577 & 0 & 0.577 & 0 \\ 0 & 0.577 & 0 & 0.577 & 0 & 0.577 \\ 0 & 0 & 0.707 & 0 & 0.707 & 0 \end{bmatrix}$$

$$\mathbf{A}_{-0.5} = \begin{bmatrix} 0 & 3 & 0 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 2 & 0 & 2 \\ 0 & 0 & 3 & 0 & 3 & 0 \end{bmatrix}$$

The conceptualization of the connectivity matrix is important for eigenvector spatial filtering because the eigenvectors are extracted from the connectivity matrix. For the purposes of this dissertation the binary **C** adjacency with first order rook connectivity is employed.

2.2.3 Edge Effects

Any spatial surface has regions along its boundaries. These regions can be considered edge polygons or pixels. Generally these regions are connected to other outside regions not included in the study area. The outside regions can influence the edge regions causing errors in the conceptualization of the surface and the estimation using the data associated with the edges

regions. This means that although an outside pixel on the edge of an image has no variable data recorded it still has an influence on the data value in the recorded edge pixel. In image analysis the edge of an image is a result of the sensor system making the boundary arbitrary and independent of the variable information on the ground. This arbitrary boundary can include unnecessary information or exclude necessary information simple as a result of the sensor system design.

The unaccounted for influence of the outside edge pixels can cause edge effects that are systematic errors along edges when geographic information for a variable is considered [see Griffith (1983); Ripley (1981); Griffith (1988)]. Several ways have been proposed to account for the distortions caused by edge effects, which are discussed in Griffith and Amrhein (1983) and Griffith (1985). Although the edge effects get smaller as the number of regions increases, they do not go to zero and are instead asymptotic to zero [Griffith (1982)]. But the conclusion of these papers states emphatically that all of the methods fail to “cope with the biases" caused by edge effects, and that a researcher must acknowledge that edge effects exist in the data and cause inferential error in any analysis conducted on a bounded surface. In this context, where the number of regions is massively large the edge effect should be small.

2.3 SPATIAL AUTOCORRELATION IN IMAGE ANALYSIS

Spatial autocorrelation might be interpreted in many different ways: Cliff and Ord (1973, pg. 1) explain spatial autocorrelation as “[i]f the presence of some quality in a county of a country makes its presences in neighboring counties more or less likely,… the phenomenon exhibits spatial autocorrelation.” Anselin (1988, pg. 11) defines spatial dependence as “the existence of a functional relationship between what happens at one point in space and what

happens elsewhere." Burt and Barber (1996, pg. 411) state that "spatial autocorrelation refers to the correlation for a variable with itself through space." They also describe autocorrelations as "a systematic pattern in the spatial distribution of a variable." Griffith (2003) gives a list of interpretations of spatial autocorrelation as:

1. Spatial process mechanism
2. Diagnostic tool
3. Nuisance parameter
4. Spatial spillover effect
5. Outcome of areal unit demarcation (Modifiable Areal Unit Problem MAUP)
6. Redundant information
7. Map pattern
8. Missing variable indicator / surrogate
9. Self- correlation

Here spatial autocorrelation is discussed using the interpretation of spatial autocorrelation as self-correlation and systematic patterns. Spatial autocorrelation is quantified by a variety of measures and these measures are dependent upon the conceptualization of the spatial surface structure. Qualitatively a spatial surface might be described as "clustered" if like values tend to located close together in a surface, whereas a spatial surface might be described as "dispersed" if unlike values tend to locate close together on a map. A quantification of these qualitative observations about a spatial surface is if nearby values tend to be similar, the pattern is said to be positively spatially autocorrelated, while if neighboring areas tend to be dissimilar the pattern is said to negatively spatially autocorrelated. If a spatial surface tends to have neither positive or negative spatially autocorrelated patterns then a surface pattern is said to be random.

Many measures have been developed to determine where on the spectrum from highly positively spatial autocorrelated to highly negatively spatially autocorrelated a particular map

pattern might be located. There are measures of global spatial autocorrelation, which summarize the amount of spatial autocorrelation into a single number, and local spatial autocorrelation, which give a measure for every observation in the map. Choosing the appropriate measure to quantify the spatial autocorrelation in a particular surface requires knowledge about the measurement level of the numerical information. Table (2.1) gives a summary of the appropriate measurement scale for some measures of spatial autocorrelation most often used in remote sensing applications. The following section describes these measures theoretically, and their applications in remote sensing.

Table 2.1. Measures of spatial autocorrelation classified by measurement scale.

Measurement level	Nominal	Interval/Ratio/ Continuous
Local	LICD	*LISA*
Global	Join Count	*MC,GR* ,semi-variogram

2.3.1 Global measures of spatial autocorrelation

Global measures of spatial autocorrelation consider the neighbor relationship of all regions in a surface and summarize those relationships into a single measure. The join count statistic is a measure of global spatial autocorrelation useful for nominal variables (Cliff and Ord 1973). Nominal variables are categorical variables without natural ordering, such as favorite types of music (e.g., country, classical, jazz, rock) or political affiliation (e.g., Republican, Democrat, Independent) (Agresti 2002). Nominal variables are often binary (i.e., all observations are 0 either or 1), determined by whether an event has or has not occurred. Counting the number of times two “joined” regions (i.e., neighbors in a spatial surface) have the same or different nominal labels in a surface is the basis of the join count statistic. For the binary surface (i.e., black and white surface) example, “joined” regions could both have black labels, both have

white labels or the two regions might be split between black and white labels. Using that information Iyer (1948) and Dacey (1968) give the formulation for the join count statistic for both a binary and k-color categorical map. Iyer (1948) gives the binary specification, already derived by Moran (1946), and derives the k-color join count statistic for the rectangular lattice. Dacey (1968) extends this for the case of an irregular lattice through the use of a connectivity matrix and Cliff and Ord (1973) give a weighted specification.

The specification of the standard join count statistic is given in Tables (2.2) and (2.3), where a two-dimensional surface of regular or irregular tessellation has *N* observations. For the regular square tessellation $N = m*n$, if the regular tessellation dimensions *m*-by-*n*. Both the rook's and queen's case adjacency is given, where *A* is the total number of neighbors and *D* is the number of regions with no regions in common.

Table 2.2. Number of neighbors and number of regions with no neighbors in common for a regular square and irregular surface

Regular square tessellation
Rook's Case $A = 2mn - (m + n),$ $D = 4[3mn - 3(m + n) + 2],$
Queen's Case $A = 4mn - 3(m + n) + 2$ $D = 8[7mn - 9(m + n) + 11]$
Irregular tessellation
L_k = number of links for region k $A = \frac{1}{2}\sum_k L_k$ $D = \sum_k L_k\,(L_k - 1)$

Table 2.3. Join count equations for sampling with and without replacement

Binary Case, k=2
Sampling with replacement $\mu_{BB} = Ap_1^2$ $\mu_{WW} = Ap_2^2$ $\mu_{BW} = Ap_1p_2$ $\sigma_{BB}^2 = Ap_1^2 + 2Dp_1^3 - (A+2D)p_1^4$ $\sigma_{WW}^2 = Ap_2^2 + 2Dp_2^3 - (A+2D)p_2^4$ $\sigma_{BW}^2 = 2(A+D)p_1p_2 - 4(A+2D)p_1^2p_2^2$
Sampling without replacement [2] $\mu_{BB} = \frac{An_1^{(2)}}{N^{(2)}}$, $\mu_{WW} = \frac{An_2^{(2)}}{N^{(2)}}$, $\mu_{BW} = \frac{An_1n_2}{N^{(2)}}$, $\sigma_{BB}^2 = \frac{An_1^{(2)}}{N^{(2)}} + \frac{2Dn_1^{(3)}}{N^{(3)}} + \frac{[A(A-1)-2D]n_1^{(4)}}{N^{(4)}} - \left[\frac{An_1^{(2)}}{N^{(2)}}\right]^2$, $\sigma_{WW}^2 = \frac{An_2^{(2)}}{N^{(2)}} + \frac{2Dn_2^{(3)}}{N^{(3)}} + \frac{[A(A-1)-2D]n_2^{(4)}}{N^{(4)}} - \left[\frac{An_2^{(2)}}{N^{(2)}}\right]^2$, $\sigma_{BW}^2 = \frac{An_1n_2}{N^{(2)}} + \frac{2Dn_1n_2(n_1+n_2-2)}{N^{(3)}} + \frac{4[A(A-1)-2D]n_1^{(2)}n_2^{(2)}}{N^{(4)}} - 4\left[\frac{An_1n_2}{N^{(2)}}\right]^2$

The join count statistic for the binary case (i.e., k = 2) is defined for sampling with/without replacement in Table (2.3), where μ_{BB} is the join count statistic for black on black joins and σ_{BB}^2its standard deviation.

Two limitations of the join count statistic are given by Cliff and Ord (1973). The first limitation is topological invariance, where the relative strength between connections is ignored, which makes the statistic invariant under certain topological transformations (see also the illustration in Cliff and Ord (1973)). The second limitation given by Cliff and Ord (1973) relates to only defining first order neighbors, which they suggest doing a correlogram analysis and including information on higher order lag neighbors, which was adopted in both Congalton

[2] $n^{(b)} = n(n-1)\ldots n(n-b+1)$

(1988) and Pugh and Congalton (2002) for their analyses. Cliff and Ord also suggest including a non-binary weight matrix to overcome both limitations.

On an interval or ratio scale global measures of spatial autocorrelation often employed in remote sensing applications include Moran's *I* (i.e., *MC*) or Geary's *C* (i.e., *GR*) (Moran 1950 and Geary 1954). The *GR* statistic is applied by Waren and Shank (1997) for use in feature selection, by ranking image bands according to their combined spatial and spectral information content. From each band a ratio of the original band with all other bands in the image is constructed and then the *GR* computed. This is used as a metric of the spectral information present in the bands. This was expanded upon by Derksen et.al (1998) and LeDrew et al. (2004).

GR, given in Equation (3), is defined as the squared difference between observations of the random variable, similar to the *black-white* measure of the join count statistic.

$$GR = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad (3)$$

where y_i is an observation of an interval or ratio random variable and c_{ij} is the connectivity of observations *i* and *j*.

Read and Lam (2001) use Moran's *I* to characterize the spatial complexity of a TM image and compare this with the image fractal dimension. This was further studied by Emerson et al. (2005). *MC* given in Equation (4) is defined as the cross products of the deviations of the observations from the mean of the random variable, analogous to the BB join count statistic.

$$MC = \frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \frac{\sum_{i=1}^{n}(y_i - \bar{y})\sum_{j=1}^{n} c_{ij}(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (4)$$

Interesting to note is that the numerators of both *MC* and *GR* measure the covariance among the observations and the denominators measure the variance (Cilff and Ord 1973).

Cliff and Ord (1973) give two assumptions: normality or randomization, under which *MC* and *GR* may be evaluated. These measures share the important limitation with the join count statistic of topological invariance. Once a connectivity matrix is defined the size and shape of the regions in the system, and the relative strength between connection is ignored, making the measures invariant under certain topological transformations.

For a global measure of spatial autocorrelation for a variable measured on a continuous scale often a model for autocovariance or a model for a semi-variogram is employed (Matheron 1963, and Cressie 1993). Haining (2003) beginning on page 74 defines and compares these models and discusses the difference in the conceptualization of the continuous field as points or as pixels.

2.3.2 Local measures of spatial autocorrelation

Local measures of spatial autocorrelation consider the neighborhood structure of each region separately and give a summary measure of spatial autocorrelation for each region in a surface. Boots (2003) proposes local measures of spatial autocorrelation for a variable measured on a nominal scale called Local Indicators for Categorical Data (LICD). LICD statistics are a local measure similar to the join count statistic. Boots (2003) proposes that under the assumption of no global spatial autocorrelation the binomial distribution gives the probability of a cell being black. Under this assumption a test for significant presence or absence of black cells can be

conducted at some given significance level. This is called local composition and parallels the approach given by the Getis statistic for interval and ratio data. Next he asks the question given the number of black cells in a subregion, does the number of joins of a specified type differ from what would be expected. This is evaluated by assuming a normal approximation of the sampling distribution of the join counts which Boots (2003) states is reasonable for $n > 30$, where n is the number of observations, and p_b or $(1\text{-}p_b) < 0.2$, where p_b is the proportion of black joins.

For the interval or ratio scale often Local Indicators of Spatial Autocorrelation (*LISA* statistics; e.g., Local Moran's *I* and Local Geary's c_i) are employed (Anselin 1995) as are Getis and Ord's *G* (Getis and Ord 1992). LISA statistics such local Moran's I_i for a given surface, gives each region in that surface a measure of local spatial autocorrelation I_i that can be interpreted as follows. Positive local spatial autocorrelation indicates a cluster of data values around *i*, that are similar to *i*, that deviate strongly from the mean (i.e., $\bar{y}_j$) either positively or negatively. Negative local spatial autocorrelation describes the same situation only the sign of region *i* is opposite of that of its neighbors. Goovaerts et al. (2005) use local Moran's I_i for image anomaly detection by comparing geostatistical techniques, the RX detector and Local Moran's I_i. Equation (5) gives one formulation of local Moran's I_i, where $z_j = (y_j - \bar{y})$ and $w_{ij}$are the spatial weights.

$$I_i = z_i \sum_{j=1}^{n} w_{ij} z_j \tag{5}$$

In contrast, another LISA statistic, local Geary's c_i given in Equation (6), indicates positive local spatial autocorrelation when the values in the neighbor around *i* are similar to the value at *i* and negative local spatial autocorrelation when the neighborhood values are dissimilar to the

value at *i*, independent of the data values mean value. Myint et al (2007) compare a moving window version of the global Geary's c and local Getis and Ord's G_i statistic in the context of LC / LU applications. They explore how well these statistics distinguish hot and cold stops in remotely sensed imagery. Equation (6) gives the formulation for local Geary's c_i, where $z_i = (y_i - \bar{y})$ and all the other variable are defined as above.

$$c_i = \sum_{j=1}^{n} w_{ij}(z_i - z_j)^2 \tag{6}$$

Boot (2002) notes that these measures are "strongly influenced" by the number of regions in the neighborhood of *i* and are also dependent of the specification of the spatial weights matrix. Getis and Ord give the *G* and the *G** statistics as measures of local spatial autocorrelation (Getis and Ord 1992; Ord and Getis 1995*)*. The difference in these two measures is that G_i sums the data values in a neighborhood around region *i* relative to the sum of all data values excluding y_i. G_i^* in contrast sums the data values in the neighborhood around region *i* relative to the sum of all data values. Boots (2002, pg. 171) explains the differences as "G_i and G_i^* can be considered to measure clustering around *i* and at *i*, respectively." Wulder and Boots (1998) use the Getis statistic to measure a pixels relative spatial dependence and give a summary table of some early contributions of spatial autocorrelation in the context of remotely sensed imagery. Equation (7) gives G_i^* where all variables are defined as above.

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} y_j} \tag{7}$$

Both Anselin (1995) and Boots (2001) give more detail on the expected value and variance of these measures along with significant testing and distributional assumptions. There

is also further discussion of these measures in Boots and Tiefelsdorf (2000) and Tiefelsdorf and Boots (1997).

One of the early references to spatial autocorrelation in the context of satellite imagery is the assessment by Craig (1979) of the presence of spatial autocorrelation along scan lines and columns of MSS imagery. This was followed by further studies by Craig and Labovitz (1980), Craig (1981), and Craig (1984) extending this to TM imagery and exploration of the sources of the autocorrelation such as surface land cover, terrain, instrument specifications, and sun angle.

The presence of spatial autocorrelation in image data has often has been acknowledged in remote sensing literature. Table (2.4) gives a representative sample of theoretical and application papers from three perspectives, spatial statistics, geostatistics, and textures, on how spatial autocorrelation can be assessed and included in image analysis. The distinction between these perspectives is not necessarily straightforward and much overlap exists between them, for this reason a paper from each perspective that compares across perspectives is also included, see Carr and Miranda (1998), Myint (2003), and Maillard (2003). The techniques used in this paper relate most directly to the spatial statistical perspective.

Congalton (1988) introduces the join count statistic in remote sensing as a measure of spatial dependence within an image difference map, to quantify the spatial error in the two classified maps. Pugh and Congalton (2002) further explore the use of the join count for assessing map accuracy and include a correlogram analysis of higher order spatial lag neighbors. Kabos and Csillag (2001) also give a modified join count statistic for use in remote sensing, based on probability distribution of colors at each location. The join count statistic could have several interesting applications within remote sensing and might be used in conjunction with a 2-

way confusion table to give a measure of spatial dependence the measures of confusion. Also an application in change detection is possible used and a measure of spatial autocorrelation within and between the change pixels. Future research concerning the connect between EISF and the join count statistic is also interesting.

Table 2.4. Selected references for theory (T), applications (A), comparisons (C), and literature reviews (R) in spatial statistical, geostatistical and texture methods in remote sensing

Spatial Statistics in Remote Sensing		**Geostatistics in Remote Sensing**		**Texture measures in Remote Sensing**	
Description	*Reference*	*Description*	*Reference*	*Description*	*Reference*
Assessment of spatial autocorrelation (T)(A)	Craig 1979, Craig and Labovitz 1980, Cambell 1981	Semi variance as an image processing technique (T)(A)	Carr and Meyer 1984	Texture and autocorrelations function (T)	Haralick et. al 1973, Haralick 1979
Join Count Statistic for classification errors (T)(A)	Congalton 1988, Pugh and Congalton 2001	Semi variance theory and application in remote sensing (T)(A)	Curran 1988 Jupp et.al 1988,1989	Intraclass local variance for Image scale (T)(A)	Woodock and Strahler 1987
Getis statistic in TM imagery (T)(A)	Wulder and Boots 1998	Integrating spatial statistics and remote sensing (T)(R)	Cressie 1993, Stein et. al. 1999	Texture in land cover classification (A)	Gong, Marceau and Howarth 1992
Geary's *c* in image analysis (T)(A)	Warner and Shank 1997, Derksen et.el.1998, Ledrew et. al. 2004	Directional semi-variogram (T)(A)	St-Onge and Cavyas 1997	Texture for multipolarization crop discrimination (A)	Anys and He 1995
Moran's I for characterizing spatial complexity in TM imagery (A)	Read and Lam 2002, Emersom et. al. 2005	Semi variogram and co-occurrence matrix comparison (C)	Carr and Miranda 1998	Image Texture review (R)	Tuceryan and Jain 1998

Table 2.4 continued-

Spatial Statistics in Remote Sensing		**Geostatistics in Remote Sensing**		**Texture measures in Remote Sensing**	
Description	*Reference*	*Description*	*Reference*	*Description*	*Reference*
Moran's *I*, Geary's *c*, fractal dimension, simple standard deviation and mean texture comparison (C)	Myint 2003	Geostatistics for DEM conflation and accuracy assessment (T)(A)	Kyriakidis et. al 1999	Filtering for texture classification: A comparative study (C)	Randen and Husoy 1999
Local Moran's *I* for anomaly detection (T)(A)	Goovaerts et.al. 2005	Introduction to geostatistical classification in image analysis (R)	Atikinson and Lewis 2000	Fractals / fractal Brownian motion (T)(A)	De Cola 1989, Lam 1990, Xia and Clarke 1997
Getis and Geary Indices for LC/LU mapping (C)	Myint et. al. 2007	Error detection accuracy assessment (T)(A)	Cressie and Kornak 2003	Semi-variogram, Fourier transform and co-occurrence matrix comparison (C)	Maillard 2003
Eigenvector spatial filtering on ISODATA classification of image (T)(A)	Jacob et. al. 2008	Indicator geostatistics (T)(A)	Boucher and Kyriakidis 2006	Wavelet texture for feature extraction (A)	Ouma 2006

2.4 ORTHOGONAL VARIABLES

A spatial filter is constructed by adding a series of "special" weighted map patterns together. These map patterns are "special" because they are orthogonal. Orthogonal map patterns ensure that each map pattern used in a spatial filter adds new spatial pattern information to the filter. The follow section explains orthogonality mathematically and discusses how spatial eigenvectors are constructed so that they are orthogonal.

To better understand conceptually orthogonal variables Rogers et.al. (1984) explain them in terms of their relationship to linearly independent and uncorrelated variables --- two ideas with which they are often confused. The algebraic definition for all three terms can be stated as follows:

Let the variables $\mathbf{H}_1$ and $\mathbf{H}_2$ be observed vectors.

Then

1. $\mathbf{H}_1$ and $\mathbf{H}_2$ are linearly independent if and only if there exists no constant *a* such that $a\mathbf{H}_1 - \mathbf{H}_2 = \mathbf{0}$ (when $\mathbf{H}_1$ and $\mathbf{H}_2$ non-null vectors).
2. $\mathbf{H}_1$ and $\mathbf{H}_2$ are orthogonal if and only if $\mathbf{H}_1^{\mathrm{T}}\,\mathbf{H}_2 = \mathbf{0}$
3. $\mathbf{H}_1$ and $\mathbf{H}_2$ are uncorrelated if and only if $(\mathbf{H}_1 - \bar{\mathrm{H}}_1\mathbf{1})(\mathbf{H}_2 - \bar{\mathrm{H}}_2\mathbf{1})^{\mathrm{T}} = \mathbf{0}$; $\bar{\mathrm{H}}_1$ and $\bar{\mathrm{H}}_2$ are the means of $\mathbf{H}_1$ and $\mathbf{H}_2$ respectively, and $\mathbf{1}$ is a vector of ones.(Rodgers 1984, pg. 133)

The relationships of these three types of variables are portrayed clearly using a Venn diagram (Rodgers 1984). Both uncorrelated and orthogonal variables are always linearly independent, but if a variable is uncorrelated it is not necessarily orthogonal and vice versa. The important

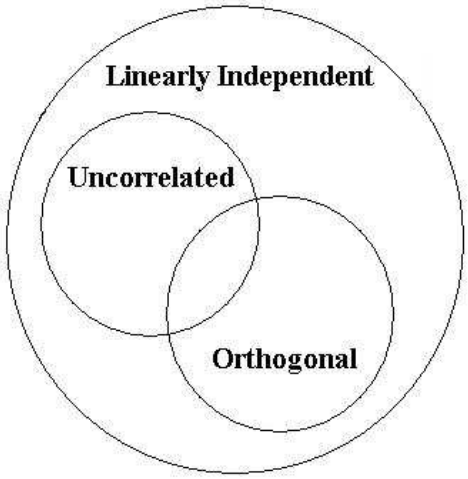


Figure 2.3. Comparison of linearly independent, orthogonal, and uncorrelated variables

difference between orthogonal and uncorrelated variables is that an uncorrelated variable must be mean centered. Rodgers et al. (1984) give a clear and concise explanation of this distinction. In the context of spatial filtering it is important to note that the orthogonality and uncorrelatedness of the eigenvectors is insured by the geographic connectivity matrix being symmetric and pre- and post- multiplied by the projection matrix, $\mathbf{M} = \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^{\boldsymbol{T}}}{\boldsymbol{n}}\right)$. This removes the mean from the geographical connectivity matrix while not losing its symmetry. If a geographic connectivity matrix is symmetric, then the transpose of the eigenvectors is equal to the inverse of the eigenvectors (i.e., $\mathbf{E}^{\mathrm{T}} = \mathbf{E}^{-1}$) and its eigenvectors are real and orthogonal (i.e., $\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{0}$). If the mean is not removed, the eigenvectors, although orthogonal, could still be correlated. Pre and post multiplying by the projection matrix does not make a non-symmetric spatial weights matrix symmetric but does insure that the mean can be removed without losing the symmetric property of the connectivity matrix. Section 2.2.2 explains how to make a non-symmetric connectivity matrix such as the row summed normalized connectivity matrix **W,** symmetric.

Conventional spatial filtering first centers the connectivity matrix by pre and post multiplying by the projection matrix, and then extracting the spatial eigenvectors from the centered MCM matrix. The following section discusses how the orthogonal spatial patterns are constructed and how they have been used in previous spatial filtering applications.

2.5 EIGENFUNCTIONS OF GEOGRAPHICAL CONNECTIVITY MATRICES

Eigenfunctions arise naturally in a variety of fields of study, including electrical systems, genetics, chemical reactions, quantum mechanics, mechanicals stress, economics, and geometry (Anton and Rorres, 1994). The eigenfunction problem is presented in many texts from both a geometric and linear algebra prospective. The following discussion is limited to the computation of eigenvalues and eigenvectors from symmetric connectivity matrices and their use in geographical contexts. Anton and Rorres (1994) describe the eigenfunction problem as follows: If **A** is an *n*-by-*n* matrix and **x** is an *n*-by-1 vector, generally there is no geometric relationship between the vector **x** and the vector **Ax**. There are, however, certain nonzero vectors **x** such that **x** and **Ax** are scalar multiples of each other. Thus Anton and Rorres (1994) define the eigenfunction problem as follows:

> *If **A** is a n-by-n matrix, then a nonzero vector **x** in [the reals] is called an eigenvector of **A** if **Ax** is a scalar multiple of **x**; that is,*
>
> $$\boldsymbol{Ax} = \lambda \boldsymbol{x} \qquad (8)$$
>
> *for some scalar λ. The scalar λ is called an eigenvalue of **A**, and **x** is said to be an eigenvector of **A** corresponding to λ.*

This is always true and yet uninteresting when **x** = **0**. The more interesting and useful question is what are λ and $\boldsymbol{x}$ when $\boldsymbol{x} \neq \mathbf{0}$. Solving Equation (8) for zero gives$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = \mathbf{0}$. Since $\mathbf{x} \neq \mathbf{0}$ this equation can only be true if the determinate of $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is zero. Finding nonzero values for λ is known as the eigenvalue problem. The eigenvector problem uses the eigenvalues to solve for the *n* vectors each of size (*n*-by-1) that satisfy the equation $(\boldsymbol{A} - \lambda_j \boldsymbol{I})\boldsymbol{E}_j = \mathbf{0}$ each of these vectors is determined only up to an arbitrary constant.

The computation of the eigenfunctions (i.e., eigenvalues and associated eigenvectors) does not change because the matrix from which they come is a geographic connectivity matrix, and yet because all the entries of the geographic connectivity matrix are real and the matrix is or can be made symmetric[3], the eigenfunctions posses several useful properties. These properties are important for the formulation of both conventional eigenvector spatial filtering and the specification for use in image analysis. Only two such properties, which relate directly to the methodology are discussed here. A more comprehensive list of eigenvector properties can be found in Griffith and Amrhein (1997) and Griffith (2003).

2.5.1 Eigenvalues of a Spatial Surface

The eigenvalue problem is to find the *n* roots of the equation det(**A**-λ**I**) = **0**, which is often rewritten as an n^{th} order polynomial equation:

$$b_n(-\lambda)^n + b_{n-1}(-\lambda)^{n-1} + \cdots + b_{n-r}(-\lambda)^{n-r} + \cdots + b_0 = 0 \qquad (9)$$

where **A** is a matrix, here a general symmetric geographic connectivity matrix, λ are eigenvalues, $b_n = 1$, b_{n-1} = trace(**A**), b_{n-r} = the sum of all principle minors of order *r*, and b_0 = det(**A**).

[3] See Section 2.2.2

Generally, the roots of this equation for a geographic connectivity matrix are real, not necessarily distinct, and some may be zero (Griffith, 2003). Here the set of all eigenvalues of a surface is denoted by $\boldsymbol{\lambda}$ and the set can be arranged:

$$\lambda_n \leq \lambda_{n-1} \ldots \leq \lambda_2 \leq \lambda_1,$$

in descending rank order (i.e., λ_1 = the maximum eigenvalue and λ_n = the minimum eigenvalue). Two useful properties of eigenvalues are (1) the sum of all eigenvalues of a surface equals zero (i.e., $\sum_{i=1}^{n} \lambda_i = 0$) and (2) the sum of the square of all eigenvalues equals the sum of all neighbors in the link matrix (i.e., $\mathbf{1}^T\mathbf{C1}$, where **C** is a connectivity matrix and **1** is a vector of ones) (Griffith 2003).

Griffith (2003) gives several theorems that relate to the eigenvalues of a spatial surface. The first concerns the principle eigenvalue (i.e., λ_1) of **C** and the projection matrix ($\boldsymbol{I} - \frac{\mathbf{11}^T}{n}$).

> *For the binary geographic connectivity matrix **C**, multiplication by the projection matrix* ($\boldsymbol{I} - \frac{\mathbf{11}^T}{n}$) *replaces the principle eigenvector associated with* λ_1 *from **C** with 0 in the set of eigenvalues for matrix* $(\boldsymbol{I} - \frac{\mathbf{11}^T}{n})\boldsymbol{C}(\boldsymbol{I} - \frac{\mathbf{11}^T}{n})$.

Replacing the first eigenvalue with zero inserts the intercept vector into the set of eigenvectors. If **C** is not multiplied by the projection matrix this is apparent in the pattern of the eigenvectors; the first eigenvector resembles a bullseye or a hill, which is essentially the mean. This is illustrated in Figure (2.4).

The rank order of the eigenvalues of **C** and $(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})\mathbf{C}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})$ tend to converge after the principle eigenvector of **C** is replaced with $\frac{1}{\sqrt{N}}$ and the eigenvectors are mean centered; this is discussed further in Griffith (2000). This follows from a lemma, given in Durbin and Watson (1950) that states that the eigenvalues of **C** and $(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})\mathbf{C}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})$ interlace. From this Griffith (2003) gives the following theorem:

> *For the binary geographic connectivity matrix* ***C*** *and its transformed counterpart* $(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})\boldsymbol{C}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})$, λ_2 *from* $\boldsymbol{C} \leq \lambda_1$ *from* $(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})\boldsymbol{C}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})$.

This is important in that it describes the systematic relationship between the uncorrelated eigenvalues of $(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})\boldsymbol{C}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})$ and the eigenvalues of **C**.

2.5.2 Eigenvalue and Moran's I

A very useful property of eigenvalues of a geographic connectivity matrix is that an eigenvalue of a given eigenvector corresponds exactly to the Moran Coefficient value of that eigenvector. Tiefelsdorf and Boots (1995) show, using a lemma from Durbin and Watson (1950), that $\boldsymbol{\lambda}(\frac{n}{\mathbf{1}\boldsymbol{A}\mathbf{1}^T})$, can express Moran's *I* (MC) for the eigenvector associated with λ_i, where $\boldsymbol{\lambda}$ are the eigenvalues, *n* is the number of observations, **A** is a general connectivity matrix, from which $\boldsymbol{\lambda}$ is computed, **1** is a vector of ones, and T is the transpose of a vector. Tiefelsdorf (2000) discusses in detail the mathematics behind this correspondence.

2.5.3 Analytical Eigenvalues of the Square Tessellation

The analytical solution for the eigenvalues of **C** is a function given in Equation 10 and 11 that produces the exact eigenvalues of **C** given the row and column indices for any square tessellation. This analytical solution for the eigenvalues of the binary connectivity matrix of the square tessellation **C** was first given in Ord (1975). Ord (1975) and Griffith (2000) formulate the computation of the rook's case eigenvalues of the square tessellation as:

$$\boldsymbol{\lambda} = 2\left[\cos\left(\frac{\pi \boldsymbol{p}}{P+1}\right) + \cos\left(\frac{\pi \boldsymbol{q}}{Q+1}\right)\right]. \tag{10}$$

Similarly Griffith (2003) gives the analytical solution for eigenvalues of the square tessellation queen's case adjacency as:

$$\boldsymbol{\lambda} = 2\left[\cos\left(\frac{\pi \boldsymbol{p}}{P+1}\right) + \cos\left(\frac{\pi \boldsymbol{q}}{Q+1}\right) + 2\cos\left(\frac{\pi \boldsymbol{p}}{P+1}\right) * \cos\left(\frac{\pi \boldsymbol{q}}{Q+1}\right)\right]. \tag{11}$$

Interestingly the eigenvectors for both the queen's and rook's case adjacency are the same (Griffith, 2003). Chapter 3 discusses modifications to these formulations for more efficient implementation.

2.5.4 Eigenvectors of a Spatial Surface

At least initially, the most useful way of understanding the eigenvectors of a spatial surface is to map them Figure (2.4), Figure (2.5), Figure (2.6), and Figure (2.7) show the 100 eigenvectors extracted from a 10-by-10 regular square spatial surface using a rook's adjacency rule. The patterns are ordered from highest positive spatial autocorrelation to highest negative spatial autocorrelation. The eigenvectors in Figure (2.4) show highly clustered patterns (i.e., show

positive spatial autocorrelation), while the patterns in Figure (2.5) and Figure (2.6) show less systematic spatial autocorrelation. They are a mixture of clustered pattern and dispersed patterns tending to show more local or regional correlation. The patterns in Figure (2.7) are systematically dispersed patterns (i.e., show negative spatial autocorrelation).

Of note is that these eigenvectors are constructed from **C.** This is apparent because the first eigenvector resembles a bullseye or a hill, when the connectivity matrix is first centered by.

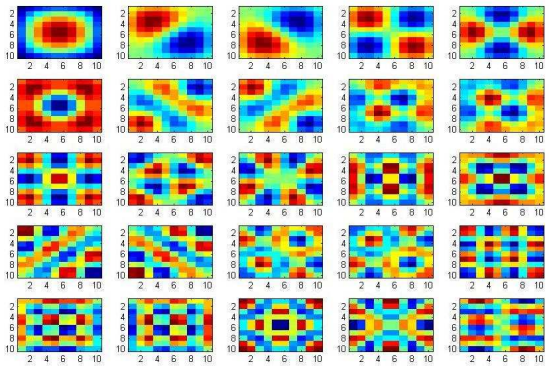


Figure 2.4. First 25 spatial eigenvectors of a 10-by-10 square tessellation demonstrate global and regional positive SA

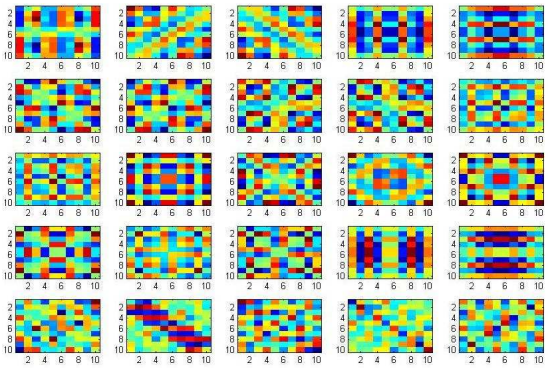


Figure 2.5. Spatial eigenvectors 26-50 of a 10-by10 square tessellation demonstrate regional and local SA

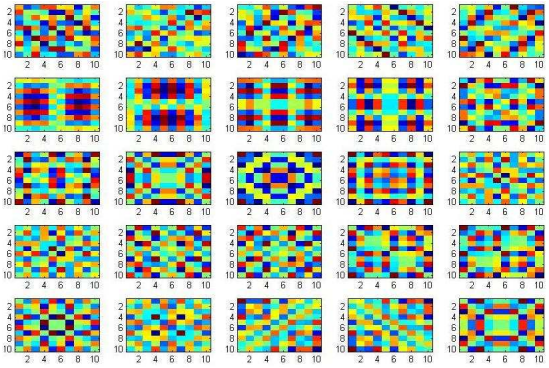


Figure 2.6. Spatial eigenvectors 51-75 of a 10-by-10 square tessellation demonstrate regional and local SA

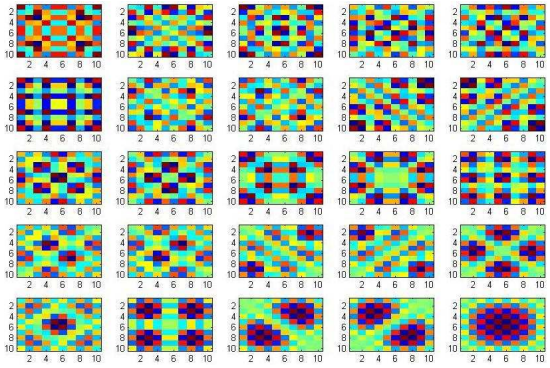


Figure 2.7. Spatial eigenvectors 76-100 of a 10-by-10 square tessellation demonstrate regional and global negative SA

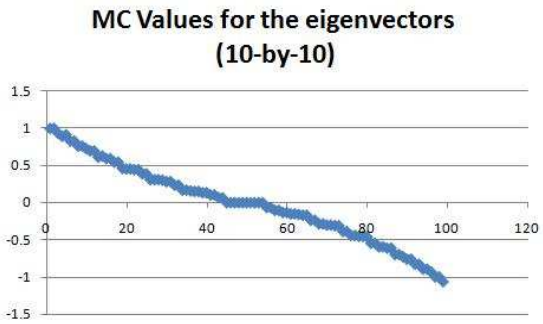


Figure 2.8. Moran Coefficient of the 10-by-10 square tessellation eigenvectors

the projection matrix this global trend is removed and replaced with an eigenvector that corresponds to a surface with zero spatial autocorrelation The spatial autocorrelation in the patterns in Figure (2.4), Figure (2.5), Figure (2.6), and Figure (2.7), can be quantified by computing the Moran Coefficient or similarly by examining the normalized eigenvalue for each eigenvector surface. Figure (2.8) shows a scatterplot of the normalized eigenvalues for each surface against the MC values in descending rank order of eigenvalues. The normalized eigenvalue correspond exactly to the MC value of the eigenvector surface and plot shows that the initial inspection that the eigenvector surfaces was correct they range from highly positively spatially autocorrelated to highly negatively spatially autocorrelated.

Griffith (2004) uses Tiefelsdorf and Boots (1995) connection between the eigenvalue and *MC* to interpret the eigenvectors as a series of orthogonal map patterns in the following way:

> *The first eigenvector, $\boldsymbol{E}_1$, of expression $(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})\boldsymbol{C}(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{n})$is the set of numerical values that has the largest MC achievable by any possible set of numerical values, for the arrangement of locations given the geographic connectivity matrix **C**. The second eigenvector is the set of numerical values that has the largest achievable MC by any set of numerical values that is uncorrelated with $\boldsymbol{E}_1$. This sequential construction of eigenvectors continues through $\boldsymbol{E}_n$, which is the set of numerical values that has the largest negative MC achievable by any set of numerical values that is uncorrelated with the preceding (n-1) eigenvectors.*

The eigenvectors of **MCM** can be interpreted as orthogonal and uncorrelated synthetic variables that relate to the defined spatial surface of a variable. It is from this interpretation that conventional eigenvector spatial filtering techniques have been applied in a regression setting.

2.5.5 Analytical Eigenvectors of C for the Square Tessellated Surface

The analytical solution for the eigenvectors of **C** is a function given in Equation (12) that exactly produces the eigenvectors of **C** given the row and column indices of any square tessellation. These are given in Gasim (1989) and Griffith (2000). Similar to the analytical solution for the eigenvalues, the eigenvectors relate to a trigonometric function of the row and column indexes of the square tessellation. Griffith's (2000) notation, written in matrix notation, is for $p = 1$ to P and $q = 1$ to Q:

$$\mathbf{E}_i = \mathbf{E}_{p,q} = \frac{2}{\sqrt{(P+1)(Q+1)}}\left[\mathbf{sin}\left(\frac{p*\mathbf{p}*\pi}{P+1}\right)*\mathbf{sin}\left(\frac{q*\mathbf{q}*\pi}{Q+1}\right)\right], \tag{12}$$

where **p**, **q**, p, q, P and Q are defined as in Figure (2.1) and the multiplication between p and **p** and q and **q** are scalar elementwise multiplications.

Since **C** is not centered before the eigenvectors are constructed it is important to verify the properties of orthogonality and uncorrelatedness of the eigenvectors. To do this their relationship to the orthogonal and uncorrelated eigenvectors of **MCM** are considered. The analytical solution for the eigenvectors given in Griffith (2000) are for **C** and not **MCM**, for that reason it is important to explore the mean and correlation of the eigenvectors of **C** and describe their relationship to the eigenvectors of **MCM**. Theorem 3.6 of Griffith (2000) describes this relationship as:

> *Suppose that symmetric binary (0-1) matrix* $\boldsymbol{C}$ *is an n-by-n incidence matrix representing a planar partitioning of some two-dimensional geographic surface into n polygonal units. Let* $\boldsymbol{E}$ *be the matrix of eigenvectors of* $\boldsymbol{C}$*,* $\boldsymbol{E}_1$ *denote the principal eigenvector in matrix* $\boldsymbol{E}$*,* $\boldsymbol{I}$ *be the identity matrix,* $\boldsymbol{1}$ *be an n-by-1 vector of ones, and k be a positive constant. If the non-principal eigenvectors of matrix* $\boldsymbol{C}$ *are centered and renormalized (i.e., replaced with* $k(\boldsymbol{I}-\frac{\boldsymbol{11}^T}{n})\boldsymbol{E}$*), then as n increases they converge upon the eigenvectors of matrix* $(\boldsymbol{I}-\frac{\boldsymbol{11}^T}{n})\boldsymbol{C}(\boldsymbol{I}-\frac{\boldsymbol{11}^T}{n})$*, but with the centered principal eigenvector* $(\boldsymbol{I}-\frac{\boldsymbol{11}^T}{n})\boldsymbol{E}_1$ *being replaced with vector* $(\frac{1}{\sqrt{n}})\mathbf{1}$.

Conventionally, when extracting the eigenvectors of matrix **C**, $\mathbf{E}_1$ is replaced with **1** to insert the intercept into the set of eigenvectors. When using the analytical solution for the eigenvectors of **C, 1** is multiplied by $(\frac{1}{\sqrt{n}})$ to ensure they converge on the eigenvectors of $(\boldsymbol{I}-\frac{\boldsymbol{11}^T}{n})\mathbf{C}\left(\boldsymbol{I}-\frac{\boldsymbol{11}^T}{n}\right)$. The proof of this theorem is given in Griffith (2000).

A property that relates to the mean of the eigenvectors of **C** is given in Griffith (2000) as Theorem 2.4:

> *Suppose a regular square tessellation planar surface partitioning is finite* $(P < \infty;\ Q < \infty)$*. Then a number of eigenvectors,* $\boldsymbol{E}$*, of matrix* $\boldsymbol{C}$ *have a zero mean.*

Griffith (2000) gives the proof of this and also calculates the number of eigenvectors with means that are exactly zero Table (2.5).

Table 2.5: Number of zero mean eigenvectors for analytical eigenvectors of C

	P_{even}	P_{odd}
Q_{even}	$\frac{3}{4}PQ$	$\frac{Q(3P-1)}{4}$
Q_{odd}	$\frac{P(3Q-1)}{4}$	$\frac{3PQ-P-Q-1}{4}$

It is also important to consider the correlation between the eigenvectors of **C**. Griffith (2000) gives the following theorem on the correlations:

> *Let n = PQ, the number of units into which a surface is partitioned. Then as both P and Q go to infinity, all correlations (ρ) amongst the non-principle eigenvectors of matrix* $\mathbf{C}_{pq}$ *converge on zero.*

This shows that if the eigenvector has an even row or column index its correlation is zero. Here it is shown that if the row and column values of both eigenvectors are odd than the correlation, given by Equation(25) (in the appendix because of its size), is asymptotically zero as *n* gets large, assuming the analytical eigenvectors have been mean centered and $\mathbf{E}_1$ is excluded.

The following pilot study evaluates the asymptotic nature of the correlation of the analytical eigenvectors of **C** with the mean removed and the first eigenvector replaced with $(\frac{1}{\sqrt{n}})$ **1** and shows that the maximum correlation is essentially zero for *n* greater than 200,000. The analytical solution for the eigenvectors of the binary connectivity matrix C are orthogonal but could be correlated since **C** is not pre- and post- multiplied by the projection matrix (i.e., not **MCM**). Griffith (2000) provides a proof showing that asymptotically these eigenvectors are not correlated for large n if the mean is removed and the first eigenvector replaced with $1/\sqrt{n}$. This is evaluated here empirically by finding the pairwise correlation of the complete set of analytical eigenvectors. This is done for a series of square tessellations with dimensions 2-by-2 through 500-by500, for the square case, and 12-by-2 through 510-by-500, for the rectangular case. The correlation of two eigenvectors $\mathbf{E}_j$ and $\mathbf{E}_k$ extracted from the same square spatial tessellation can be calculated by substituting each eigenvector into the Pearson Correlation Coefficient equation as variables and simplifying, as in Equation (13).

$$\frac{\mathbf{E}_j^T(\mathbf{I}-\frac{\mathbf{1}\mathbf{1}^T}{n})\mathbf{E}_k}{\sqrt{\mathbf{E}_j^T(I-\frac{\mathbf{1}\mathbf{1}^T}{n})\mathbf{E}_j}\sqrt{\mathbf{E}_k^T(I-\frac{\mathbf{1}\mathbf{1}^T}{n})\mathbf{E}_k}} = -\frac{\mathbf{E}_j^T\mathbf{1}(\frac{\mathbf{1}^T\mathbf{E}_k}{n})}{\sqrt{1-\mathbf{E}_j^T\mathbf{1}(\frac{\mathbf{1}^T\mathbf{E}_j}{n})}\sqrt{1-\mathbf{E}_k^T\mathbf{1}(\frac{\mathbf{1}^T\mathbf{E}_k}{n})}} \qquad (13)$$

The analytical solution for these eigenvectors are substituted into Equation (13), giving Equation (25) in Appendix (C). Examining the numerator of Equation (25) if either $\frac{kp_j}{P+1}$ or $\frac{lq_j}{Q+1}$ is a multiple of 2, the numerator and the correlation will be zero, since the sine of any multiple of 2π is zero. Table 2.5), gives equations for determining exactly how many of the correlations will be exactly zero.

For those combinations of row and column indexes that do not produce an angle multiple of 2π, Equation (25) must be further explored. Consider the first term in the numerator of Equation (25),

$$\sum_{j=1}^{P} \sin\left(\frac{kp_j\pi}{P+1}\right).$$

This may be simplified using the trigonometric identity in Equation (14) [see Spiegel (1968) for more details].

$$\sum_{n=1}^{N} sin(N\theta) = \frac{sin\left(\frac{1}{2}Nx\right) sin\left[\frac{1}{2}(N+1)x\right]}{sin\left(\frac{1}{2}x\right)} \tag{14}$$

Letting $\mathbf{N}_1 = P$ and $\theta_1 = \frac{k\pi}{P+1}$ then the first term of the numerator of Equation (25) may be rewritten as

$$J_{P\theta_1} = \frac{\sin\left(\frac{1}{2}P\theta_1\right) \sin\left[\frac{1}{2}(N+1)\theta_1\right]}{\sin\left(\frac{1}{2}P\right)}$$

The next three terms in the numerator

$$\sum_{i=1}^{Q} \sin\left(\frac{lq_i\pi}{Q+1}\right) * \sum_{j=1}^{P} \sin\left(\frac{kp_j\pi}{P+1}\right) * \sum_{\mathrm{i}=0}^{\mathrm{Q}} \sin\left(\frac{lq_i\pi}{Q+1}\right)$$

follow similarly, giving all the terms necessary for both the numerator and the denominator, $J_{Q\theta_2} = \frac{\sin(\frac{1}{2}Q\theta_2)\sin[\frac{1}{2}(Q+1)\theta_2]}{\sin(\frac{1}{2}Q)}$, $K_{P\theta_3} = \frac{\sin(\frac{1}{2}P\theta_3)\sin[\frac{1}{2}(P+1)\theta_3]}{\sin(\frac{1}{2}P)}$ and $K_{Q\theta_4} = \frac{\sin(\frac{1}{2}Q\theta_4)\sin[\frac{1}{2}(Q+1)\theta_4]}{\sin(\frac{1}{2}Q)}$.

Substituting all of these reductions back into Equation (25) gives:

$$\frac{J_{P\theta_1} * J_{Q\theta_2} * K_{P\theta_3} * K_{Q\theta_4}}{\sqrt{n - J_{P\theta_1} * J_{Q\theta_2} * J_{P\theta_1} * J_{Q\theta_2}}\sqrt{n - K_{P\theta_3} * K_{Q\theta_4} * K_{P\theta_3} * K_{Q\theta_4}}} \tag{15}$$

Using Equation (15), the maximum correlation between analytical eigenvectors is found from the series of complete square tessellations for both the square and rectangular cases. Figure 2.9) plots the maximum correlation of these eigenvectors and shows that the maximum correlation goes to zero asymptotically as *n* gets larger and is essentially zero after 200,000 observations (i.e., surface of 400-by-400). For an image with dimensions smaller than 400- by-400 (i.e., less than

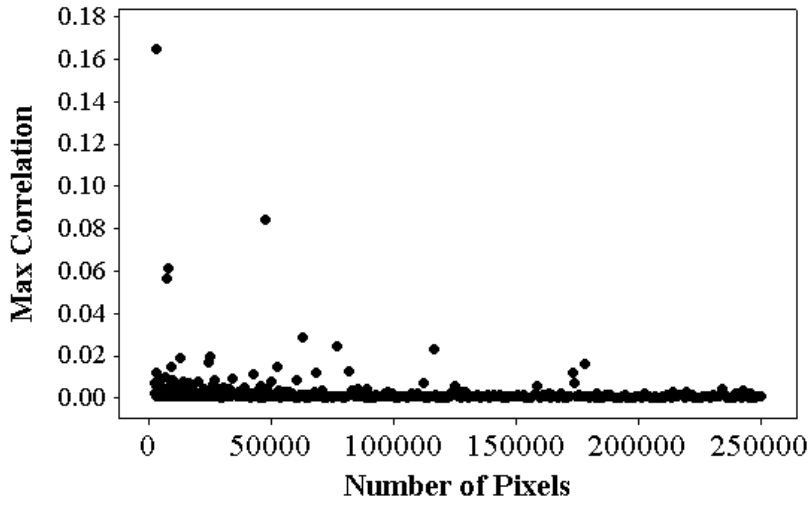


Figure 2.9. Maximum correlation for eigenvectors of the square tessellation 50-by-50 to 500-by-500

200,000 pixels) consider using conventional eigenvector spatial filter since the uncorrelated property of the eigenvectors might not be satisfied. But for images above 200,000 pixels the correlation of the analytical eigenvectors is essentially zero and should not cause problems in the analysis.

This shows pairwise independence but there are several methods to examine the multicolinearity of variables. Another is the condition number, which is the square root of the ratio between the maximum eigenvalue and the minimum eigenvalue. The larger the condition number the more linear dependence.

$$\text{condition number} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

This condition number is evaluated for a series of square tessellations of both square or rectangular dimensions with row dimensions from 3 to 50 rows and column dimensions from 3 to 50 and graphed in Figure (2.10). By approximately 1,500 observations the condition number

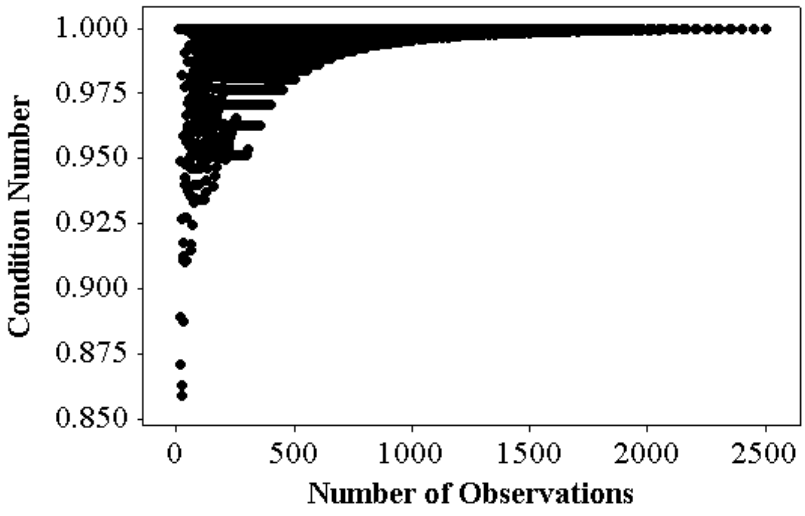


Figure 2.10. Condition number for the analytical Eigenvectors of surfaces up to 2500.

is essentially one. For the small case study image with image dimensions 948-by-220 the condition number is 1.0001 and for the medium case study image with dimensions 1000-by-1000 the condition number is 1.0000. Another measure of multicolinarity is the determinant of a matrix (e.g., **X**) of the variables. If **X** is orthogonal than the transpose of **X** (i.e., $\mathbf{X}^T$) is the inverse of **X**. Therefore $\mathbf{X}^T\mathbf{X}=\mathbf{I}$ taking the det of both sides gives $\det(\mathbf{X}^T\mathbf{X})=\det(\mathbf{I})$ and since it is well known that det(AB)=det(A)det(B) and the $\det(\mathbf{I}) = 1$, then $\det(\mathbf{X}^T) = 1$ and $\det(\mathbf{X}) = 1$, thus proving that if **X** is orthogonal than the determinant of **X** should be one. The determinant of **X** might be computed by finding the product of the eigenvalues of the matrix (see Anton and Rorres 1994). For the same surfaces of 3 to 50 rows and 3 to 50 columns, the determinant is essentially one with a ranging between 0.99 and 0.94. This empirical study illustrates using pairwise correlation, the matrix condition number and determinate, that the eigenvectors of C are asymptotically orthogonal.

2.6 CONVENTIONAL EIGENVECTOR SPATIAL FILTERING

Eigenvector spatial filtering has its roots in the thematic maps of social science. Fields like demography, economics and human geography often work with variables associated with locations. Generally the spatial distribution of a variable is not random; similar values tend to be close together. This tendency makes a variable spatially dependent on itself, or spatially correlated with itself (i.e., spatial autocorrelated) Spatial filtering was developed to account for this autocorrelation in social data models (Griffith 1996).

The spatial filtering technique combines information about data connectivity and data values to choose a linear combination of distinct map patterns that represents the spatial information of a variable. Eigenvector spatial filtering has been applied across a variety of

disciplines including geography (Tiefelsdorf and Griffith, 2007), image analysis (Griffith and Fellows, 1999), demography (Tiefelsdorf and Griffith, 2007), migration (Chun, 2008), economics (Kosfeld and Dreger, 2006; Getis and Griffith, 2002) applied mathematics and statistics (Griffith, 2002, 2000) environmental sciences (Diniz-Filho and Bini, 2005), epidemiology (Griffith, 2005), and ecology (Dray et al., 2006; Griffith and Peres-Neto, 2006). Data in many fields are often georeferenced and eigenvector spatial filters allow spatial statistical analysts to use conventional linear regression models in which parameters are estimated by ordinary least squares (Getis and Griffith, 2002). The mechanics of spatial filtering, its underlying assumptions and its implementations are discussed in the following sections.

2.6.1 Data structure - Definitions and Notation

Spatial data might be logically thought of as two pieces of information: the observed variable, and its spatial arrangement. This conceptualization of georeferenced variables, also discussed in Haining (2003) and Goodchild (1989), is the same as the one employed in Geographical Information Systems (GISs). The variable information is numerical or categorical information related to locations in a spatial surface that inventories some phenomenon. The spatial arrangement is a single underlying surface configuration that covers an area of interest; it can be thought of as an empty outline base map, which the variable information fills in. A key part of the conceptualization of a spatial surface is how the individual regions or points in a surface are connected or arranged in relation to each other.

Any given model could have many different variables (i.e., attributes) filling in a single common underlying spatial surface, like several bands of one multi-spectral image. As a part of initial data diagnostics of regression variables, an evaluation of the nature and degree of spatial

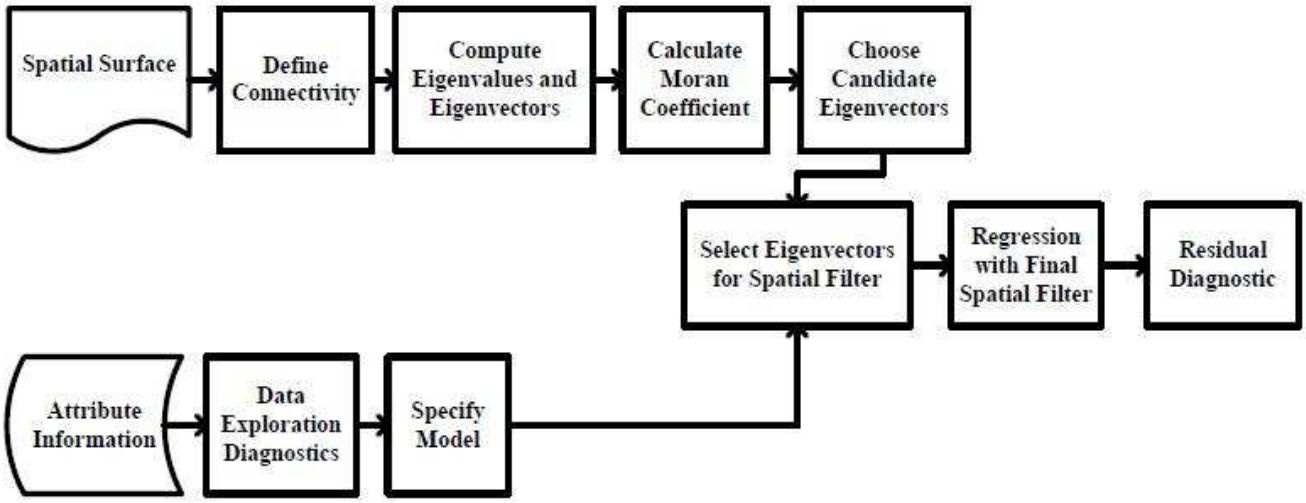


Figure 2.11. Conventional Spatial Filter Methodological Flow Chart

autocorrelation present in each variable should be included in the evaluation of whether model data conforms to model assumptions. In the following section, the steps for specifying a spatial regression model for these types of georeferenced random variables using eigenvector spatial filtering is discussed in more detail. Figure 2.11) shows a flow chart of the steps involved in this process.

2.6.2 Spatial Filtering Model

The model used in a regression analysis involving an eigenvector spatial filter is similar to a conventional regression model. Let a response variable (i.e., the variable being modeled) be denoted by **Y**, a vector of size *n*-by-1, where *n* is the number of observations, and let the explanatory variables (i.e., the variables modeling **Y**) and the intercept be denoted by **X**, a matrix of size *n*-by-*l*, where *l* is the number of covariates plus 1 for the intercept. As should be standard procedure in an aspatial regression specification, exploratory data diagnostics should be conducted to assess the data model assumptions. The initial data diagnostics might point a

researcher immediately to a spatial specification, for instance if the data shows strong spatial autocorrelation. If not, a conventional aspatial regression equation might be specified as $\mathbf{Y} = \mathbf{X}\beta_x + \xi$, where $\boldsymbol{\beta}_x$ is the *l*-by-1 coefficient vector of **X**, and ξ is an *n*-by-1 error vector. For the normal linear model, values in this error vector are assumed be independent and normally distributed. This assumption must be tested. If the residuals show signs of autocorrelation and the researcher has reason to believe spatial processes might be at work, a spatial regression model should be specified. Tiefelsdorf (2000) gives more details about Gaussian processes and testing for spatial autocorrelation in residuals.

Eigenvector spatial filtering is one method of including spatial information in a regression model. To add this spatial information, a set of candidate eigenvectors, **E**, is added as another term to the conventional aspatial regression equation in order to model the spatial information, yielding

$$\mathbf{Y} = \mathbf{X}\beta_x + \boldsymbol{E\beta_E} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is an error term free of spatial autocorrelation, and $\boldsymbol{\beta}_E$ is the eigenvector coefficient vector of size *k*-by-1, where *k* is the number of candidate eigenvectors, discussed in Section 2.6.3. Using this specification, the subset of eigenvectors in **E** that account for spatial autocorrelation in the data is chosen. Two methodologies have been implemented for defining this set, which are both discussed further in Section 2.6.4. Regardless of how this set is chosen, a final regression equation is written to include the subset of eigenvectors **E**, here denoted as **F**, such that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_\mathbf{x} + \mathbf{F}\boldsymbol{\beta_F} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_F$ is the final spatial filter coefficient vector of size *f*-by-1, where *f* is the number of chosen eigenvectors.

The residuals of this model should also be tested for independence and normality. Although this discussion refers to a normal linear model specification for numerical data, eigenvector spatial filtering is applied in a variety of alternative situations, depending on the type of data under investigation. Alternatives include - Poisson / negative binomial models for counts and rare events, and binomial for percentage and presence/absence data (see Chun, 2008; Griffith, 2006, 2004). The output spatial filter and the chosen eigenvectors might be examined for possible patterns that could be explained by missing variables in the model. The following sections discuss choosing a candidate set of eigenvectors, **E**, and briefly describe two methods for choosing which candidate eigenvectors reflect spatial information. Finally, a discussion is given on the linear combination of the eigenvectors which create a spatial filter.

2.6.3 Candidate Eigenvector

The candidate set of eigenvectors is a subset of all possible orthogonal eigenvectors of a spatial surface. This subset should contain the eigenvectors that model the amount and type of spatial autocorrelation present in data. To date there is no single universally accepted way of determining the candidate set, but a few rules of thumb have been suggested (Griffith, 2003). Eigenvectors that have virtually no autocorrelation (i.e., their patterns are essentially random) are not likely to be useful in modeling spatial pattern in the data and therefore are not included in the candidate set. Which eigenvectors fall into this category can be determined by examining the *MC* values of each of the eigenvectors. If the null hypothesis of $MC = 0$ is not rejected, then the eigenvector in question does not portray a significant amount of negative or positive spatial

autocorrelation and thus may excluded from the candidate set. It might also be the case that only a specific type of autocorrelation is of interest, for instance data diagnostics might show the data has high positive spatial autocorrelation. The candidate set of eigenvector may only include eigenvectors portraying positive spatial autocorrelation.

2.6.4 Choosing Eigenvectors / Linear Combination

After establishing a set of candidate eigenvectors, this set must be further refined to identify only those vectors that contain spatial information inherent in the response variable. Two approaches for choosing these final eigenvectors have been implemented, both of which are briefly discussed here. The first approach chooses eigenvectors by selecting those which account for the most spatial autocorrelation in the regression residuals; this is termed the Minimizing Residual Autocorrelation (MRA) method. MRA's objective function seeks to minimize the spatial autocorrelation left in the residuals. This is accomplished through an iterative process. The eigenvector that accounts for the most spatial autocorrelation in the residuals is chosen and the residuals re-evaluated. This is repeated until only a small threshold amount of spatial autocorrelation is left in the residuals.

Another approach chooses eigenvectors that account for the most variation in the response variable; this is termed Minimizing Response Variation or MRV. A standard stepwise regression is used to identify the relevant covariate variables and the eigenvectors from a candidate set, which account for a threshold amount of variation in the response variable.

Both these approaches are discussed in more detail in Tiefelsdorf and Griffith (2007), where detailed descriptions of the algorithms are given and case study examples of both

techniques are compared. The algorithm described in this dissertation relates to the MRV approach.

Once the model variables and parameters are estimated, the spatial filter is simply the eigenvectors multiplied by their estimated regression coefficients, summed together. This linear combination should show the residual spatial pattern in the data not accounted for by the covariate information. It is advisable to critically examine a spatial filter and selected eigenvectors including the sign and magnitude of their coefficient value. This information may point to variables missing from the model, and provide insight into the spatial processes causing the spatial patterns observed in the filter.

2.6.5 Image Spatial Filter

This section briefly summarizes Griffith and Fellows (1999) and Fellows (1998), who use a 250,000 pixel subsection of a Landsat TM image to classify forest blow-down in Adirondack State Park in New York. In his master's thesis, Fellows (1998) proposes to include spatial information in a multivariate image analysis (MIA) classification of blow-down sites in a small forested area by constructing eigenvector spatial filters for each image band. He proposes a strategy to use the analytical solution for the eigenvalues and eigenvectors, which is known for the binary connectivity matrix for the square tessellation, and to implement a stepwise regression procedure to select the eigenvectors that account for variation in each image band. He further proposes using both the spectral and spatial information (e.g., in his case, 7 Landsat TM bands and 7 spatial filter bands) in the MIA classification. Fellows (1998) describes problems that were encountered with this proposed methodology and how it had to be revised in order to be implemented.

One of the original goals of Fellows' study was to “use prominent eigenvectors as locational information that represents significant map patterns found in the seven spectral bands" [p.78, Fellows (1998)]. This goal was not attained because of the lack of an, “efficient way of sifting through the massive number of distinct map patterns"[p.81, Fellows (1998)]. Fellows (1998) found that computing a spatial filter for an image was very computationally demanding. Although the analytical solution for the eigenvectors and eigenvalues made the large connectivity matrix unnecessary, the eigenvector calculation and selection process were both order n^2 calculations. Although the speed of processors has increased significantly since Fellows' (1998) work, this is still a daunting task in terms of number of operations.

Fellows was unable to follow through on exactly what he proposed, instead, an exploratory analysis of a series of the most predominate eigenvectors was conducted. Initially he identified a candidate set of 62,500 eigenvectors out of a possible 250,000. He proposed constructing seven stepwise regression models, one for each of his seven Landsat image bands, using the 62,500 candidate eigenvectors as the dependent variables in every model. He found the hardware platform used “took several days to complete the stepwise regression for each band” (Fellows 1998, pg. 78). He also found “identifying significant eigenvectors … proved nearly impossible as many accounted for only a fraction of a percentage of the total variance” (Fellows 1998, pg. 78). A smaller set of the most prominent eigenvectors with the largest eigenvalues were indentified for each band in the image. Fellows gives a table listing of ten eigenvectors accounting for the most variation in each band, but the amount of variation accounted for by those eigenvectors is not reported. It is interesting to note that there is some overlap in the eigenvectors chosen for each band, showing multicollinearity between bands.

The massive number of candidate eigenvectors made it impossible for Fellows to compute an image spatial filter. In chapter 5 and 6 of his thesis, Fellows explains that although he could not compute a spatial filter he could still include spatial information in the MIA classification by using a spatial lag variable for each image band. A spatial lag variable computes an average of the surrounding pixels for every pixel in the image and might be determined through the definition of connectivity and adjacency rules. The use of spatially lagged variables allowed Fellows to bypass choosing which eigenvectors were significant in a regression context and alleviated the problem of sifting through the massive number of eigenvectors.

The study by Griffith and Fellows (1999) is an integral part of the foundation upon which this research is based. Here the call by Griffith and Fellows (1999) for a “more efficient strategy" to identify significant eigenvectors. A strategy is pursued to make both the calculation of the eigenvectors themselves and the selection of the significant eigenvectors more efficient. This gain in efficiency allows eigenvector spatial filtering to be tested in a variety of image analysis applications.

CHAPTER 3

METHODOLOGY: CONSTRUCTING AN IMAGE SPATIAL FILTER

The discussion of spatial structure, connectivity, the eigenfunction problem and the specification of eigenvector spatial filtering for less than 10,000 regions in Chapter 2 lays the foundation for the following section, which outlines the steps necessary to efficiently construct a spatial filter for a massively large number of square regions, like the pixels in a remotely sensed image. The construction of spatial filters for image analysis is necessarily more laborious than conventional spatial filtering due to the massive number of regions in an image. Cressie (1993 [Chpt.7, pg. 502]) says, “Because remote-sensing data sets are frequently very large, computationally efficient statistical techniques are required for pixel [analysis]." Three aspects of conventional spatial filtering are addressed in this research to make the technique applicable in an image analysis setting.

3.1 ALGORITHM FOR MASSIVE NUMBER OF SQUARE REGIONS

The algorithm for constructing an image spatial filter is given in the flow chart in Figure Figure (3.1). The flow chart begins with a multispectral image. This image is then conceptually divided into the spectral information (i.e., digital number information) for each band and spatial structural information for the entire image. Although the processes are shown as divided here because they may be run independently on a computer, these processes are not conceptually independent.

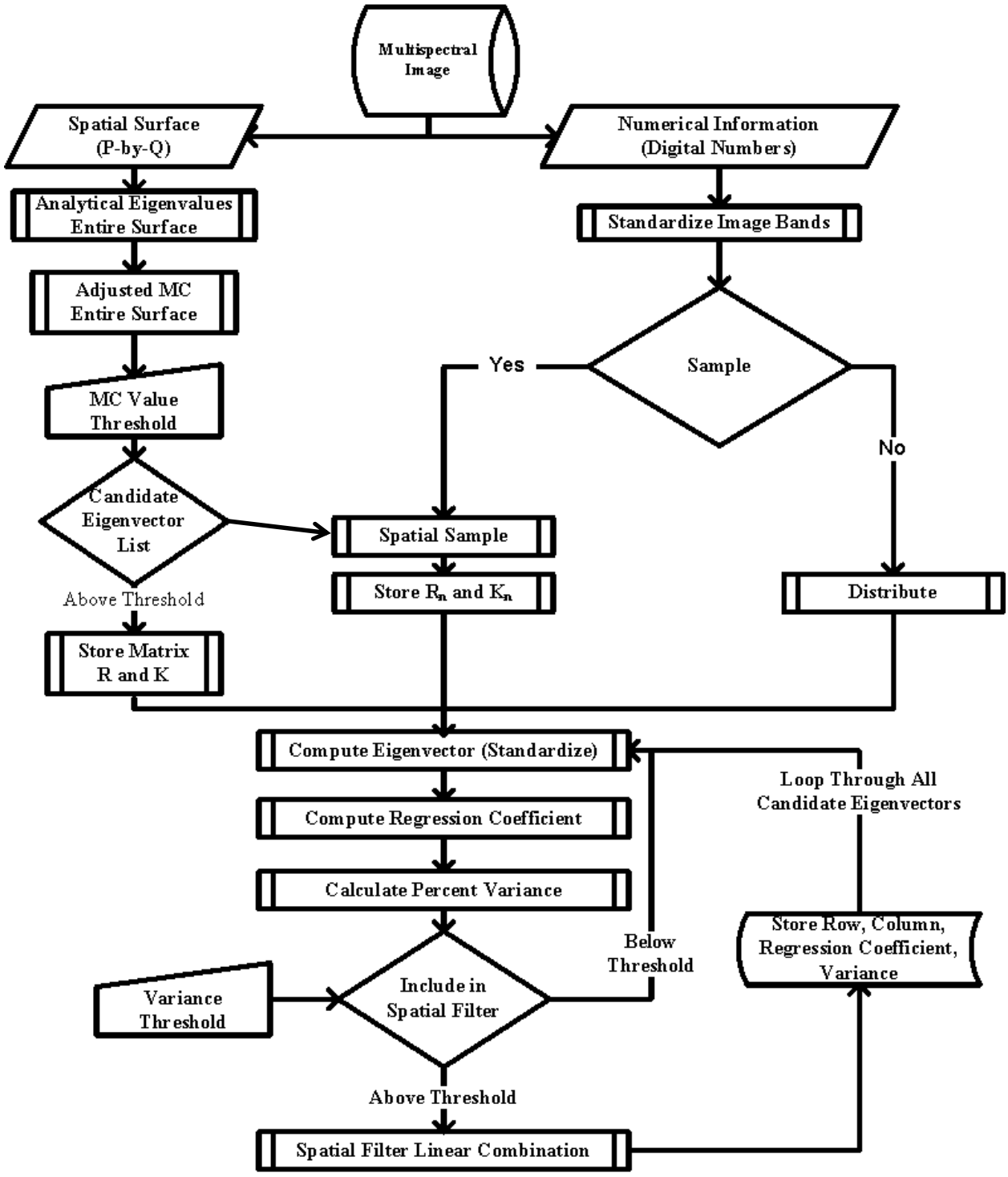


Figure 3.1. Image Spatial Filter Methodological Flow Chart

This is because the definition of spatial structure is dependent upon the spatial relationships displayed by the numerical information and should be theoretically justifiable through understanding information on the ground in an image. The following considers first the digital number work flow, then the spatial surface structure work flow and finally the loop to create an image spatial filter.

The digital number information is stored in standardized *n*-by-1 vectors, as shown in Figure (2.1), making a vector's mean zero and standard deviation 1 for each band. Then a series of data diagnostics is done on the image bands, to explore their distributional properties, the amount and type of spatial autocorrelation present, and outliers that might exist. Finally, a choice between sampling the spectral information or retaining the complete spectral information is made. If no sample is taken, then the complete spectral information is used in a distributed computing environment. If sampling is chosen the spatial sample must be determined and the matrices $\mathbf{R}_s$ and $\mathbf{K}_s$, discussed further in Section 3.2.2, constructed and stored.

Concurrently in the flow chart, the spatial surface structure of an image is assumed to be complete (i.e., no missing pixels) regular (i.e., all pixels the same size and shape) square tessellation with the image size defined to be *P*-by-*Q*. The rook's or queen's connectivity rule is then defined as the neighbor structure for this spatial surface. This knowledge allows the analytical eigenvalues to be computed for the entire surface. These assumptions which allow the analytical solution to be employed address the first inefficient aspect of the algorithm by bypassing the need to store or make computations with a large connectivity matrix. These analytical eigenvalues are then normalized to compute the *MC* value for every eigenvector associated with this assumed spatial surface structure. The *MC* value is a measure of the amount

and type of spatial autocorrelation present in each analytical spatial eigenvector. This allows a candidate set of analytical spatial eigenvectors to be determined before a single analytical spatial eigenvector is constructed, insuring that a large number of unnecessary eigenvectors are never computed. This addresses a second inefficient aspect of the conventional algorithm. Finally the matrices **R** and **K**, which eliminate repetitive calculations when constructing the analytical eigenvectors for the complete image surface are constructed so that the reformulation of the computation of analytical spatial eigenvectors may be employed.

The loop that constructs an image spatial filter begins first by constructing an analytical spatial eigenvector from the list of candidates determined using the *MC* value and choice of sampling. This analytical spatial eigenvector is constructed by a Kronecker product of one column in **R** and one column in **K** (one column in $\mathbf{R}_s$ and one column in $\mathbf{K}_s$ if the spectral information is sampled) and then standardizing the vector. Next, a simple linear regression of the standardized image band on the standardized analytical spatial eigenvector is preformed. The analytical spatial eigenvector must account for a minimum amount of variance in a standardized image to be included in an image spatial filter. If the analytical spatial eigenvector is to be included, it is multiplied by its regression coefficient and added to the previously chosen analytical spatial eigenvectors, and the spatial filter is constructed sequentially; if not, the next candidate analytical spatial eigenvector is tested. This continues until all candidate analytical spatial eigenvectors have been tested. The following section provides more detail on individual steps within this workflow and where within the workflow gains in efficiency are achieved, including the analytical solution of the spatial eigenvalues and eigenvectors of the complete

regular square tessellated surface, the processes of selecting eigenvectors for inclusion in the image spatial filter, and constructing the final image spatial filter.

3.2 GAINS IN EFFICIENCY THROUGH REFORMULATION

The computation and selection of the analytical spatial eigenvectors used in an image spatial filter is the most computationally intensive part of the image spatial filter algorithm. Methodologies for making this procedure more efficient outlined in the following section show large gains in efficiency by the reformulation and storage techniques. A discussion of several sampling techniques to reduce the number of operations required is required. Finally, a conceptual discussion of implementing this algorithm in a distributed environment without sampling is also given. Distribution requires no assumption about the homogeneity or type of autocorrelation of the spectral information.

3.2.1 Analytical Eigenvalues of the Square Tessellation

The formulation of eigenvalue problem remains the same for any definition of a symmetric connectivity matrix **A**. But for the special case of the binary connectivity matrix for a complete regular square tessellation, Ord (1975) gives an analytical solution as a sum of two cosine functions. This analytical solution is important for the application of Eigenvector Image Spatial Filtering (EISF), because under the assumption that an image has a complete regular square tessellation, the massively large connectivity matrix need not be constructed to find its eigenvalues. As in conventional eigenvector spatial filtering, in EISF the amount of spatial autocorrelation in each analytical spatial eigenvector is measured by a normalized eigenvalue associated with an analytical spatial eigenvector.

The analytical solution for spatial eigenvalues and eigenvectors is based on the row and column indexes of the square tessellation; a spatial eigenvalue and its associated spatial eigenvector have the same row and column indexes.

Equation (16) gives the spatial eigenvalues for the rook's adjacency connectivity definition, which uses the summation of two cosine functions of the row (**r**) and column (**k**) indexes vectors:

$$\boldsymbol{\lambda} = 2\left[\cos\left(\frac{\pi \mathbf{r}}{P+1}\right) \oplus \cos\left(\frac{\pi \mathbf{k}}{Q+1}\right)\right] \tag{16}$$

where $\oplus$ is the Kronecker addition of two vectors and is defined as $\begin{bmatrix} a \\ b \end{bmatrix} \oplus [a \quad b \quad c] = \begin{bmatrix} a+a & a+b & a+c \\ b+a & b+b & b+c \end{bmatrix}$ (Laub, 1995), and *P* and *Q* are the image dimensions. Although the change in notation here between the formulation given in Ord (1975) and Griffith (2000) may seem minor, from a computational standpoint it makes a substantial difference. If the cosine function is considered a single operation, the number of operations required to compute λ using the Ord (1975) notation requires 8 **p***q*+2, while the this notation requires (3 * *p* + 1) operations for the first term, (3 * *q* + 1) operations for the second term, and *n* additions and *n* multiplications, for a total of 3*p* + 3*q* + 2*n* + 2 operations.

Equation (17) gives the spatial eigenvalues for the queen's case adjacency connectivity definition (Griffith, 2003):

$$\boldsymbol{\lambda} = 2\left[\cos\left(\frac{\pi \mathbf{r}}{P+1}\right) \oplus \cos\left(\frac{\pi \mathbf{k}}{Q+1}\right) \oplus \cos\left(\frac{\pi \mathbf{r}}{P+1}\right) \circ \cos\left(\frac{\pi \mathbf{k}}{Q+1}\right)\right] \tag{17}$$

where ∘ is the Hadamard or elementwise multiplication of matrix elements. Either of the analytical spatial eigenvalue specifications, given in Equations (16) and (17), can be used to compute a measure of spatial autocorrelation for each eigenvector, since surprisingly the

eigenvectors for the regular square tessellation are the same for both the rook and queens adjacency specifications, see Griffith (2003) for details.

A single *n*-by-1 vector of each eigenvector's *MC* value is then computed using $\boldsymbol{\lambda}$ as given in Equation (18):

$$\mathbf{MC}=\frac{PQ}{2[P(Q-1)+P(Q-1)]}\boldsymbol{\lambda}. \tag{18}$$

Of note is that each value in the **MC** vector is a constant multiple of the eigenvalues for the entire surface [see Section 2.5.2 and Tiefelsdorf and Boots (1995) for more details].

Since images generally have highly positive *MC* value, a threshold may be established that identifies eigenvectors that contain more than a specified amount of spatial autocorrelation. This substantially reduces the number of analytical spatial eigenvectors that are constructed and the required number of linear regressions to construct an EISF. The determination of a value for this threshold requires external input, and should be guided by the amount and type of spatial autocorrelation in an image band.

3.2.2 Storing R and K

Many of the calculations required to construct an analytical spatial eigenvector for a square tessellation are repetitive, this comes from the nature of the surface. By considering critically these computations a reformulation of Griffith's (2000) notation can be given that more efficiently stores and computes the eigenvectors, by first storing two matrices **R** and **K**, and in a second step taking the Kronecker product of one column from each to compute one eigenvector at a time. This formulation does not require four "for loops" of length *n* for each eigenvector as

the formulation in Griffith (2000), which significantly lowers the number of operations needed to compute the eigenvectors. The matrices **R** and **K** are constructed as in Equation (19)

$$\mathrm{R} = \frac{\sin(\mathbf{r} * \mathbf{r}^T * \pi)}{P+1}, \quad \mathrm{K} = \frac{\sin(\mathbf{k} * \mathbf{k}^T * \pi)}{Q+1} \tag{19}$$

where **r**, **k**, *P*, and *Q* are defined as in Figure (2.1). The number of operations required to calculate these matrices would be $4P^2 + 1 + 4Q^2 + 1$. An eigenvector (e.g., $\mathbf{E}_i = \mathbf{E}_{pq}$ where *i* is the pixel index and *p* and *q* are the row and column indices that relate to that pixel index), is computed by the Kronecker product (i.e., $\otimes$) of the column *p* from **R** and column *q* from **K** such that $\mathbf{E}_{p,q} = \mathbf{r}_p \otimes \mathbf{k}_q$. This requires $p * q = n$ operations for a total of $4p^2+1+4q^2+1+n$ operations required to compute one eigenvector. This reformulation drastically reduces the number of repetitive calculations and required number of sine function calls.

3.2.3 Analytical Eigenvectors of the Square Tessellation

The analytical solution for the spatial eigenvectors of **C** for the square tessellation is the catalyst that enables eigenvector spatial filtering to be applied to image analysis. The solution given in Griffith (2000) as an extension of the work done in Gasim (1988) makes it unnecessary to store or do calculations with a massively large connectivity matrix, and allows each eigenvector to be computed individually. The section on storing **R** and **K** considers how the reformulation of the analytical spatial eigenvectors reduces the number of operations required to construct an analytical spatial eigenvector substantially. In this section, the analytical solution and a reformulation of its notation are presented, and some properties of the analytical eigenvectors are discussed.

The analytical solution for the eigenvectors of **C** given in Cliff and Ord (1975) can be reformulated as given in Equation (20) to construct all spatial eigenvectors of C simultaneously. The full set of n eigenvectors is constructed as

$$\mathrm{E} = \underbrace{\frac{2}{\sqrt{(P+1)(Q+1)}}}_{z}\left[\underbrace{\sin\left(\frac{\pi}{P+1}\mathrm{r}\mathrm{r}^{T}\right)}_{\mathbf{R}} \otimes \underbrace{\sin\left(\frac{\pi}{Q+1}\mathrm{k}\mathrm{k}^{T}\right)}_{\mathbf{K}}\right] \tag{20}$$

where **r**, **k**, *P*, and *Q* are defined as in Equation (12) and $\otimes$ is the Kronecker or direct product - defined as $\begin{bmatrix} a \\ b \end{bmatrix} \otimes [a \quad b \quad c] = \begin{bmatrix} a*a & a*b & a*c \\ b*a & b*b & b*c \end{bmatrix}$ - of two matrices **R** and **K** which are square matrices of size *p*-by-*p* and *q*-by-*q* respectively. In general the memory requirements for constructing the complete set of all candidate eigenvectors simultaneously is very large but one of the advantages of the analytical solution is that it is unnecessary to construct all analytical eigenvectors simultaneously each can be constructed sequentially. This is done efficiently in two steps: first compute and store **R** and **K** as in Equation (19), then compute the Kronecker product between $\mathbf{r}_p$ (i.e., column *p* in **R**) and $\mathbf{k}_q$ (i.e., column *q* in **K**) in each matrix, as in Equation (21). A single candidate eigenvector is constructed using the Kronecker product of column *p* in **R** with column *q* in **K** such that:

$$\mathrm{E}_i = \left(\mathrm{r}_p \otimes \mathrm{k}_q\right), \tag{21}$$

where *i* is the pixel index of the candidate eigenvector, and $\mathbf{r}_\mathrm{p}$ and $\mathbf{k}_\mathrm{q}$ are column vectors in **R** and **K**, respectively, and *p* is the row and *q* the column index associated with pixel index *i*.

Counting the number of operations required to construct one analytical spatial eigenvector in Equation (12) yields 11 operations for each entry, *n* entries in each vector, and approximately 1/4*n* eigenvectors computed (assuming the candidate set, which constitutes ¼ of the possible spatial eigenvectors, is computed). This requires $2.75n^2$ operations, in total, where *n*

is $p*q$. Consider the number of operations required to compute an eigenvector using Equations (20) and (21), **R** requires $4p^2 + 1$ and **K** requires $4q^2 + 1$ operations next the Kronecker product adds n operations for a total of $4p^2 + 4q^2 + 2 + n$ operations.

It is also important to note that the number of sine functions drops substantially in the reformulated calculation. The sine function, although counted above as a single operation, is generally implemented as a polynomial approximation requiring at minimum 10 multiplications and additions (Muller, 2006, pg. 76). The reformulated notation requires only $p + q$ sine function calls, while the previous notation uses $2pq$. Even for the average desktop computer, either formulation of the analytical eigenvalues, even for very large n, is computed in seconds since both are order n calculations. The formulation change here allows the eigenvalue calculation to remain consistent with the analytical spatial eigenvector construction. The reformulation in notation changes a computation from an order n^2 to an order n calculation makes a much more substantial impact on the algorithm efficiency.

To make the difference in computing time more apparent, consider the image used in Griffith and Fellows (1999) for a 500-by-500 image, constructing the spatial eigenvectors as given in Griffith (2000), the number of operations $=2.75 * 250{,}000^2 = 171{,}880{,}000{,}000$; in terms of time for a machine with 550 MFLOPS (Mega (10^6) FLoating-OPoint Operation per Second), this requires ~5.21 minutes of processing time. By storing **R** and **K** the same calculation requires only 2,250,000 operations and can be done on the same machine in ~0.00006 seconds.

More compelling is to consider a typical Landsat image (3240-by-2340). **R** would have dimensions 3240-by-3240 and **K** 2340-by-2340, while **E** has dimensions 7,581,600-by-7,581,600. The previous method requires $2.75 * 7{,}581{,}600^2 = 1.58 * 10^{14}$ operations and 79.57

hours, whereas by storing **R** and **K** the number of operations is only 4 * 10,497,600 +4 * 5,475,600 +2+75,816,000 = 714,744,002 and the time required ~0.0022 seconds. The computation of the eigenvector is the second most computationally intensive step in the construction of an EISF. The next section discusses the selection of analytical spatial eigenvectors for inclusion in the spatial filter, which is the most computationally intensive step in the algorithm.

3.3 EIGENVECTOR SELECTION

Griffith and Fellows (1999) state selecting the eigenvectors as the largest problem for implementing spatial filtering in image analysis. One strategy for identifying which spatial eigenvectors from the candidate set model spatial structure in the spectral image data is to perform a simple bivariate linear regression between the standardized spectral image band and each centered eigenvector of the candidate set. More formally, a simple linear regression of the standardized image band ($\mathbf{Z}_b$) on the standardized eigenvector $\mathbf{E}_i$ is preformed. Let $\mathbf{Z}_b = \mathbf{E}_i\beta + \varepsilon$ and solve for $b = (\mathbf{E}_i^T \mathbf{E}_i)^{*}\mathbf{1}\mathbf{E}_i^T \mathbf{Z}_b$. This can be simplified by the orthogonality property ($\mathbf{E}^T \mathbf{E} = \mathbf{I}$) of the eigenvectors (discussed in Section 2.4) to $b = \mathbf{E}_i^T \mathbf{Z}_b$. The regression coefficient, *b*, is computed for each candidate eigenvector. The square of *b* can be interpreted as the amount of variation accounted for in an image by the spatial eigenvector. Under this interpretation b^2 is used to determine which eigenvectors should be included in a spatial filter, following the MRV approach for selecting spatial eigenvectors. The calculation of the regression coefficient is the most time consuming step in this algorithm. Although the calculation itself is not complex, the sheer size of the vectors and the repetition of the calculations make the computation very slow.

A threshold amount of variance must be accounted for in the spectral information by a spatial eigenvector in order for it to be included in the set of spatial eigenvectors used in constructing a EISF. Similar to the threshold for the candidate eigenvectors, the threshold for the amount of variance accounted for in the spectral information must be evaluated empirically. If the spatial eigenvector does account for more than the threshold amount of spatial autocorrelation, it and its coefficients are retained for use in a final spatial filter.

Once a spatial eigenvector is indentified as accounting for at least a threshold amount of variance in an image band, this spatial eigenvector is multiplied by its computed regression coefficient and added to any prior selected spatial eigenvector multiplied by its regression coefficient for that band. The linear combination is essentially the sum of all selected eigenvectors weighted by their associated regression coefficients. Figure (3.2) illustrates this, where *F* is the total number of eigenvectors indentified to model a threshold amount of spatial information in the image and *b* is the regression coefficient associated with selected spatial eigenvector *f*. From a storage perspective, the linear combination requires only one vector of size *n*-by-1 to be stored. An EISF is constructed separately for each band of an image.

$$\mathbf{SF}_1 = b_1 * \quad (22)$$

$$\mathbf{SF}_F = \mathrm{SF}_{f-1} + b_f *$$

Figure 3.2. The sequential linear combination of the chosen eigenvectors to create a spatial

The question might be posed asking whether the eigenvectors would cancel each other out when summed, since the regression coefficients might be positive or negative. This is not a problem since their orthogonality property insures that this does not happen. Orthogonal

variables are a special case of linearly independent variables; they not only do not fall long the same line and also fall perfectly at right angles to each other (i.e., the cosine of the angle between the two vectors is zero). In order for the vectors to cancel each other out, they must be the same magnitude, lie on the same line and have opposite signs. The eigenvectors may or may not be the same magnitude and might have opposite signs but since they are orthogonal, they will never lie on the same line and therefore never cancel out.

3.4 GAINS IN EFFICIENCY THROUGH SAMPLING AND DISTRIBUTION

The analytical solution for the eigenvectors of **C** makes the connectivity matrix unnecessary, the computation of the spatial eigenvalues trivial, and allows the spatial eigenvectors to be constructed sequentially. The reformulation of the spatial eigenvectors computation substantially reduces the required number of calculations to construct an analytical eigenvector and using the MC threshold to determine a candidate set substantially lowers the number of analytical spatial eigenvectors to be constructed. Nevertheless as the number of pixels grow so does the computational intensity of this algorithm.

Cressie et al. (1996, p. 115) discuss how “the sheer size of a massive dataset may challenge and ultimately defeat a statistical methodology that was designed for smaller datasets,” and points out that as datasets get massive, the statistical techniques used to analyze them tend toward simpler descriptive statistics. He suggests this is due to the fact that simpler tools, such as measures of central tendency, variability, and pairwise associations continue to work efficiently on large datasets.

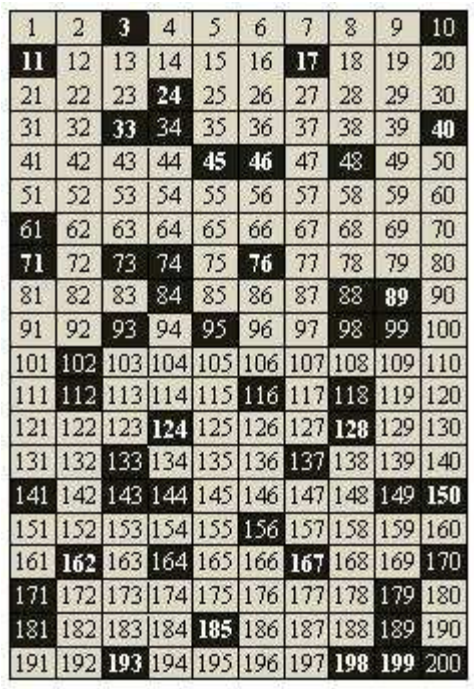
When faced with a massive dataset, the first question one might ask is whether all the data are really necessary. A massive dataset like an image often has a great deal of data redundancy. Cressie et al. (1996) suggest that accessing an entire population is unnecessary when there exists a great deal of redundancy in data. Cressie et al. (1996, pg.117) argue further that "sampling (e.g. systematic, adaptive) is particularly appropriate when there are redundancies …". He also argues (pg. 117) that "Sampling offers a way to analyze massive datasets with some statistical tools we currently have. Data that exhibit statistical dependence do not need to be looked at in their entirety for many purposes because there is much redundancy. " Under the assumption that a sample of the spectral information could represent the complete spectral surface well because of the high amount of data redundancy in an image, a procedure of sampling might be implemented to make the selecting of spatial eigenvectors more efficient. Although, sampling does not change overall the computational intensity involved in the algorithm it does lessen the number of required computations by using fewer pixels spectral values. The actual gain of efficiency in the algorithm would depend on the sampling technique chosen and its implementation.

To effectively implement a sampling algorithm, first the amount of redundancy in the form of spatial dependence should be measured by computing the MC and GR of the image data. If these measures reject the null hypothesis of no spatial dependence then the effective sample size can be computed, as given in Griffith (2005) and discussed in Griffith (2003), to approximate the amount of independent data in an image. This then could be used as a lower bound for the number of sample pixels that should be included in a particular sampling technique.

When choosing any particular sampling method it is important to consider what assumptions are being made about the population of interest (i.e., the target universe) and how that affects the conclusions that might be drawn. First consider what is being sampled, the spectral information of a *P*-by-*Q* image. Steheman and Overton (1996) define the sampling universe, **U,** as the spatial extent or region of interest in a sampling problem. Here the sampling universe would relate to the *P*-by-*Q* image pixels, which makes the target universe finite and discrete. Overton and Stehman (1995) discuss finite sampling and give the inclusion probabilities for discrete populations. The sampling frame (i.e., the units from which a sample is drawn), generally a list frame (i.e., a list of sampling units) or a map frame (i.e., a list of sampling units ordered in a geographical way) used in the study must also be defined. This might be the all of the elements of **U**, such as in simple random sample, or groups of elements of **U**, as in stratified sampling. Stehman and Overton (1996) argue that to obtain a spatially well distributed sample, a systematic spatial or a geographically stratified sample might be used. Three sampling methods are evaluated in this study including a simple random sampling, as is often employed in image analysis applications, the geographically stratified random sampling and systematic sampling to insure spatial distribution across the complete surface. A brief discussion is given of each of these sampling methods for illustration in the following section, more detailed discussion can be found in Demming (1950), Berry and Baker (1968) and Stehman and Overton (1996).

3.4.1 Simple Random Sampling

Simple random sampling is discussed in many texts including a simple overview in Gonick and Woolcott (1993), a more classic text Demming (1950), and more specifically for geography in Burt and Barber (1996). Essentially simple random sampling uses a list frame



1	2	**3**	4	5	6	7	8	9	10
11	12	13	14	15	16	**17**	18	19	20
21	22	23	**24**	25	26	27	28	29	30
31	32	**33**	34	35	36	37	38	39	**40**
41	42	43	44	**45**	**46**	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	**76**	77	78	79	80
81	82	83	84	85	86	87	88	**89**	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110
111	112	113	114	115	116	117	118	119	120
121	122	123	**124**	125	126	127	**128**	129	130
131	132	133	134	135	136	137	138	139	140
141	142	143	144	145	146	147	148	149	**150**
151	152	153	154	155	156	157	158	159	160
161	**162**	163	164	165	166	**167**	168	169	170
171	172	173	174	175	176	177	178	179	180
181	182	183	184	**185**	186	187	188	189	190
191	192	**193**	194	195	196	197	**198**	**199**	200

Figure 3.3. Simple Random Sample

where a pre-determined number of samples are randomly chosen by generating a random number and selecting the pixel associated with that random number. Here simple random sampling without replacement is implemented, meaning a pixel may only be chosen once for inclusion in the sample as shown in Figure (3.3). The geographical distribution of samples chosen in a sample random fashion may be clustered, random, or dispersed.

An advantage to a simple random sample is it requires minimum advance knowledge of the population other than the frame. Its simplicity also makes it relatively easy to interpret. Simple random sampling best suits situations where not much information is available about the population and data collection can be efficiently conducted on randomly distributed items, or where the cost of sampling is small enough to make efficiency less important than simplicity. By taking a simple random sample of the spatial eigenvector the reformulation of the spatial

eigenvectors computation can no longer be implemented because the row and column values of the sample pixels are random. This requires the simple random sample spatial eigenvectors to be constructed using the formulation given in Griffith (2003) in a series of four loops.

It is likely that the gain made by sampling the spectral information will be lost by reverting back to the nested loop approach and this sampling technique will not gain much in terms of time over the non-sampled filters.

3.4.2 Geographically Stratified Random Sampling

In geographically stratified random sampling an image is overlaid with a fine resolution square tessellation – a map frame. This map frame consists of well defined boundaries in which the spatial objects of interest (i.e., the pixels of an image) must be within (Stehman and Overton

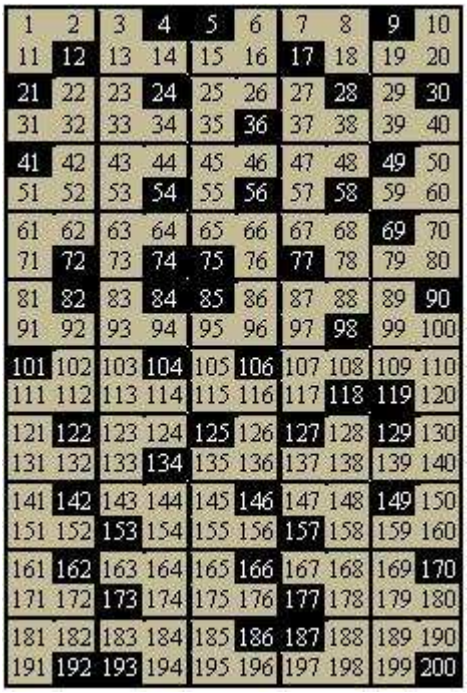


Figure 3.4. Geographically Stratified Random Sample

1996). Here, the resolution of the map frame is defined by the number of pixels a grid cell of the map frame encompasses. For example, a 10-by-20 pixel image subset could have a map frame of 2-by-2 pixels, such that the map frame would have 10-by-5 grid cells [see Figure (3.4)]. The map frame orders the pixels in groups of geographically close pixels. From these groups of pixels a sample pixel is randomly chosen. This sampling approach insures a geographically dispersed sample over the image, improving on a simple random sample, which might be clustered or disperse. Also, since the pixels are randomly selected, they are less likely to systematically miss spatial structure in an image.

One drawback to this approach is that the systematic spatial structure in an image could be missed because of its random nature. Another drawback is that the reformulation of the analytical eigenvectors cannot be implemented for this sampling method. Similar to simple random sampling, the lost of this more efficient computational technique will most likely lose any computation time that might have been gained through sampling.

3.4.3 Systematic Sample

Similar to simple random sampling, systematic sampling does not incorporate geographical position into the selection of sample pixels. Pixels are sampled from a list sampling frame, ordered by pixel index, by defining a sampling interval. Figure (3.5) shows a systematic sample with a sampling interval of two, where every other pixel is selected. Systematic samples assume underlying homogeneity of near pixels and can systematically miss spatial or spectral information in an image. This sampling technique is desirable because it is

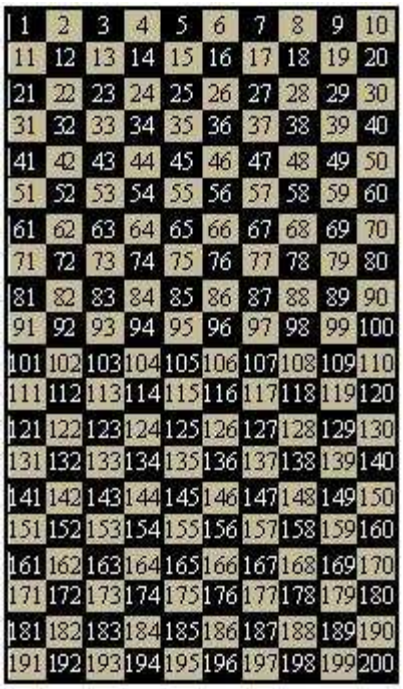


Figure 3.5. Systematic sampling with a sampling interval of two.

shown to be orthogonal and can be implemented using a modification of the reformulation of the spatial eigenvector computation. The geographical distribution of this sampling method is disperse since a standard distance is defined for the whole image.

An advantage to systematic sampling is that there is an efficient reformulation of the matrices **R** and **K**, which allows a sample eigenvector to be directly constructed by removing rows and columns from **R** and **K**. In the case of systematic sampling this formulation can further simplify the algorithm since computed rows and columns are removed from the **R** and **K** matrices, to create $\mathbf{R}_s$ and $\mathbf{K}_s$, as in Equation (23):

$$\mathrm{R_s} = \frac{\sin(\mathbf{r_s} * \mathbf{r_s}^T * \pi)}{P+1}, \qquad \mathrm{K_s} = \frac{\sin(\mathbf{k_s} * \mathbf{k_s}^T * \pi)}{Q+1} \tag{23}$$

where $\mathbf{r}_s$ and $\mathbf{k}_s$ are the row and column indexes of the systematic sample, respectively.

3.4.4 Geographically Stratified Systematic Sampling

In a geographically stratified systematic sample the target universe- in this case an image- is overlaid with a fine resolution square tessellation – a map frame. This map frame consists of well defined boundaries in which the spatial objects of interest (i.e., the pixels of an image) must be within (Stehman and Overton 1996). Here, the resolution of the map frame is defined by the number of pixels a grid cell of the map frame encompasses. For example, a 10-by-20 pixel image subset could have a map frame of 2-by-2 pixels, such that the map frame would have 10-by-5 grid cells. [see Figure (3.6)]. To choose the sample, a pixel within the first region of the map frame is randomly chosen and becomes the pointer cell. This cell then is shifted to each region of the map frame to yield a systematic sample with a randomized pointer cell. For the example in Figure (3.6)

Figure (3.6), if pixel 12 is randomly chosen to be the pointer grid cell, then the sample would consist of the yellow pixels. Equal probability is insured by equal cell size and shape, regularity of the grid, and the randomization procedure. Overton and Stehman (1995) give the inclusion probabilities (i.e., the probability of selecting a given sample), for systematic sampling as $\frac{1}{K}$, where *K* is the sampling interval. In Figure (3.6), *K*=4 since four pixels are included in a map frame grid cell, so the inclusion probability is $\pi_u = \frac{1}{4}$. A pixel's spectral information may only be included once in the sample therefore this sampling method uses sampling without replacement. This sampling design is especially convenient in an image setting because the

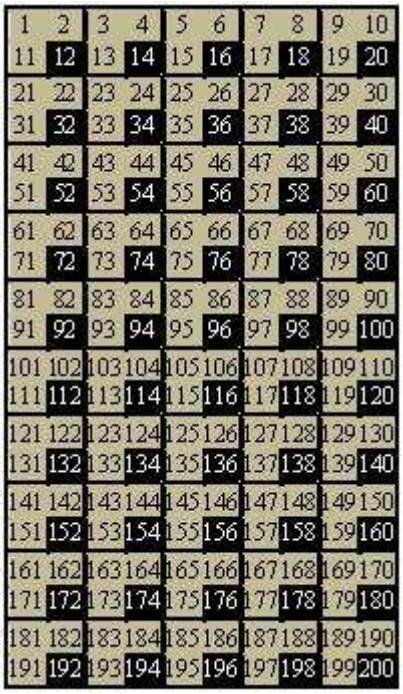


Figure 3.6: Geographically stratified systematic sampling

pixels of the original data and the map frame are “compatible,” meaning all the pixels in the image correspond exactly one grid cell in the map frame.

If an image has dimensions P-by-Q than the corresponding unsampled image vector has length n = PQ. If a map frame for that image is constructed with dimensions H-by-K, where Ha=P and Kb=Q and a and b are the sampling frame grid dimensions then the corresponding sampled image vector has length s=HK (see Figure (3.6): P = 20, Q=10, n=200, H=10, K=5, a=2, b=2, and s = 50). The spatial eigenvector is of length n –one observation for each pixel in the image. Both the spatial eigenvector and the spectral image vector must be of the same length if a simple linear regression is to be conducted.

To remedy the difference in vector length of the sampled image vector and the spatial eigenvectors, the spatial eigenvectors are constructed using only the row and column values of

the pixels chosen in the spectral image vector sample. For the geographically stratified systematic sample an efficient reformulation of the matrices **R** and **K**, allows a sample eigenvector to be directly constructed. Since the sample is systematic, only the sampled row and column indices appear in r_s and k_s. For the example surface in Figure (3.6), $\mathbf{r}_s = [2\ 4\ 6\ 8\ 10]^T$ and $\mathbf{k}_s = [2\ 4\ 6\ 8\ 10\ 12\ 14\ 16\ 18\ 20]^T$ and the sample eigenvector would be constructed as in Equation (24). Since this sampling technique preserves the ability to construct a sampled **R** and **K,** the gain in efficiency afforded by **R** and **K** is preserved**.**

$$R_s = \frac{sin\left(\begin{bmatrix}\mathbf{2}\\\mathbf{4}\\\mathbf{6}\\\mathbf{8}\\\mathbf{10}\\\mathbf{12}\\\mathbf{14}\\\mathbf{16}\\\mathbf{18}\\\mathbf{20}\end{bmatrix} * \begin{bmatrix}\mathbf{2}\\\mathbf{4}\\\mathbf{6}\\\mathbf{8}\\\mathbf{10}\\\mathbf{12}\\\mathbf{14}\\\mathbf{16}\\\mathbf{18}\\\mathbf{20}\end{bmatrix}^T * \pi\right)}{P+1} \quad K_s = \frac{sin\left(\begin{bmatrix}\mathbf{2}\\\mathbf{4}\\\mathbf{6}\\\mathbf{8}\\\mathbf{10}\end{bmatrix} * \begin{bmatrix}\mathbf{2}\\\mathbf{4}\\\mathbf{6}\\\mathbf{8}\\\mathbf{10}\end{bmatrix}^T * \pi\right)}{Q+1} \tag{24}$$

3.4.5 Drawbacks to Sampling

If an image has dimensions P-by-Q, then the corresponding unsampled image vector has length *n* = PQ. If a map frame for that image is constructed with dimensions H-by-K, where H*a*=P and K*b*=Q and *a* and *b* are the sampling frame grid dimensions, then the corresponding sampled image vector has length s=HK (see Figure (3.6): P = 20, Q=10, n=200, H=10, K=5, a=2, b=2, and s = 50). The spatial eigenvector is of length *n* –one observation for each pixel in the image. Both the spatial eigenvector and the spectral image vector must be of the same length if a simple linear regression is to be conducted.

A remedy to the difference in vector length of the sampled image vector and the spatial eigenvectors must be found. One solution for this discrepancy is to sample the spatial eigenvectors to the same length as the sampled spectral information. This requires computing the complete spatial eigenvector and then selecting those pixel values that correspond to the sampled pixel values from the spectral information. This solution is hardly ideal since again many unnecessary computations are being done to construct the complete spatial eigenvectors and then sampling from it. A better solution is to construct the sampled eigenvector directly. Direct computation of the sampled eigenvector is possible for both systematic sampling techniques since rows and columns can be systematically removed from the **R** and **K** matrices. This is not possible for the simple random sampling methods.

3.4.6 Orthogonality and Uncorrelatedness of the Sampled Eigenvector

Sampled eigenvectors are not used to construct a EISF, they are only used in the linear regression to determine which unsampled spatial eigenvectors should be included in a EISF. If the orthogonality of the spatial eigenvectors are corrupted, then the assumption of orthonormality (i.e., $\mathbf{E}\mathbf{E}^{T}=\mathbf{I}$) is not valid and the simplification of the normal regression formula to a simple correlation between the image and the spatial eigenvector is corrupted. For this reason the orthogonality of the spatial eigenvectors for each sampling method is considered, by computing the condition number and the determinate of the matrix of all the sampled eigenvectors of a particular surface, similar to the empirical evaluation the orthogonality of the unsampled eigenvectors of **C**. A series of square tessellated surfaces with square and rectangular

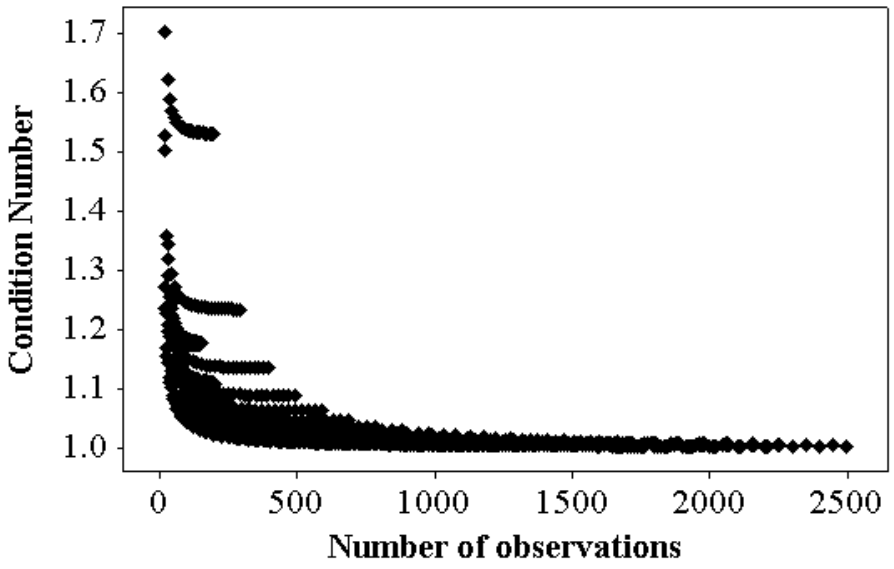


Figure 3.7. The condition number for the sampled eigenvectors

dimensions (row and column dimensions ranging from 3-50) are evaluated for each sampling method.

First consider the systematic sample. Intuitively this sampling technique is most likely to preserve the orthogonality of the spatial eigenvectors. This comes from the fact that the analytical solution for the eigenvectors of **C** is derived from the sine function. The periodic nature of the sine function curve lends itself well to systematic sampling technique. Again to evaluate this, the condition number is computed and shown in Figure (3.7). There is deviation from one for the smaller surfaces but as the number of observations increase to over 2000 the condition number goes to 1. The determinant of each matrix of sampled eigenvectors for each surface is computed with the maximum value of the determinant being 0.0156 which occurs under 200 observations. This same analysis is conducted for both the simple random sample and

the stratified random sample and it is found that neither type of sampled spatial eigenvectors were orthogonal. Both exhibited condition numbers in the thousands.

3.4.7 Distributed Computing

Although a sample may reduce the required computation time of the algorithm there is the possibility that important information is lost in the sampling process. An alternative to taking a sample of the spectral information might be to parallelize the algorithm. By using the complete spectral information and dividing the computation over many processors, the time required to construct a spatial filter might be greatly reduced. Turton (2000) discusses parallel computing from the prospective of its use in solving geographical questions. Before employing parallel computing, he asks a researcher to first consider two questions:

1. How often is this code to be run in production?
2. How long does it take to run at the moment?

He suggests if the answers to these questions are not too often and not too long, then parallel programming is unnecessary and just allowing a computer to run a week might be faster than programming in parallel. The first question requires a characterization of how often the code might be used. Part of the contribution of this dissertation is that the code be written in such a way that it is distributable in order to further investigate the spatial filtering techniques in image analysis. This could mean that the code might be run a large number of times by others investigating the algorithm. But even only in terms of the spatial filters required for the experimentation in this dissertation, a spatial filter for each candidate and variance threshold scenario is constructed for each band of the three multispectral image case studies.

Table 3.1. Estimated processor time required to construct a spatial filter for each of the case study images

Time	Case study 1	Case study 2	Case study 3
no – sample	1.30 minutes	30.30 hours	256.99 days
systematic sample	0.32 minutes	7.58 minutes	64.25 days

The second question addresses the time required to compute that spatial filter, which might be estimated by using the number of required calculations and assuming a 550 MFLOP machine, the processor time required to construct a filter for each of the case studies is given in Table (3.1). It might be argued that the medium and small image datasets do not require distributed computing, but a large image dataset would require nearly two months of processor time even to construct a spatial filter using the sampling procedure. Although the large image dataset has a large number of pixels, it is not large in terms of satellite imagery. It is actually only a subset of a full Quickbird image.

The algorithm described here could be considered an embarrassingly parallel problem as described by both Densham and Armstrong (1998) and Turton (2000). In this case this means that each loop in the algorithm is independent and might easily be distributed to many processors. Densham and Armstrong (1998) discuss parallel computing in spatial statistics and give examples of Dijkstra's shortest path algorithm (Dijkstra, 1959) and the Getis *G* statistic (Getis and Ord, 1992) as embarrassingly parallel computing problems. Turton (2000) notes this kind of embarrassingly parallel problem almost always produces speedups of nearly *f*, where *f* is the number of parallel processors.

Turton (2000) suggests the simplest model for parallelizing an embarrassingly parallel problem is to divide the problem evenly between processors. Reorganizing Griffith's (2000)

series of constants *k*, *l*, *p*, and *q* into vectors and restating the equation using a Kronecker product, a divide and conquer approach to distributed computing is simplified. The Kronecker product is a series of column vector dot products of a dense matrix calculation. Depending on the number of processors, the work load might be divided by bands and then further into sets of candidate eigenvectors. In this way each processor computes a linear combination of eigenvectors chosen from a subset of the candidate eigenvectors to model spatial information in an image, and then the final filter is the addition of those linear combinations.

CHAPTER 4

ASSESSMENT

To evaluate the methodology described in Chapter 3 for constructing eigenvector image spatial filters an empirical assessment is designed using 3 images of increasingly larger numbers of pixels, a small image with approximately 200,000 pixels, a medium image with approximately 1,000,000 pixels and a large image with approximately 110,000,000 pixels. For each of these datasets a set of scenarios is constructed by systematically changing the spatial autocorrelation threshold value (i.e., candidate eigenvector threshold, here 0.25, 0.50, 0.75) and the spatial filter eigenvector inclusion threshold (i.e., a variance threshold, here 0.001 and 0.01, for a spatial eigenvector to be included in a final spatial filter's linear combination of spatial eigenvectors). Table (4.1) summarizes the scenarios for each dataset and gives characteristics and naming convention used in each scenario. All of the scenarios given in Table (4.1) are applied to the complete spectral information (C), and to the spatially sampled spectral information (SR- simple random, GR - geographically stratified random sample, SS- systematic sample, and GS-geographically stratified systematic sample) and evaluated in terms of hardware, software, computational intensity and output EISF. Table (4.2) shows the number of candidate eigenvectors used in each case study scenario and how that number changes in the sampled cases. Although the sampling technique changes for different scenarios, the number of samples taken is kept constant for each sampling technique in order to make them more comparable.

Table 4.1. Naming conventions for the scenarios given a particular sampling frame, which adds the suffix C (complete spatial surface), SR (simple random sample), SS (systematic sample), GR (geographic random sample), and GS (geographic systematic sample).

Image	***Candidate Threshold***	***Variance Threshold***		***Filters per Method***
		0.01	*0.001*	
Small	*0.25*	k25s01	k25s001	72
12 bands	*0.50*	k50s01	k50s001	
~200,000 pixels	*0.75*	k75s01	k75s001	
Medium	*0.25*	k25s01	k25s001	36
5 bands	*0.50*	k50s01	k50s001	
~1,000,000 pixels	*0.75*	k75s01	k75s001	
Large	*0.25*	k25s01	k25s001	24
4 bands	*0.50*	k50s01	k50s001	
~110,000,000 pixels	*0.75*	k75s01	k75s001	

Three candidate eigenvector thresholds are selected for testing; these thresholds allow only spatial eigenvectors with a MC value greater than 0.25, 0.50, or 0.75 to be considered for inclusion in an EISF. The spatial eigenvectors included in 0.75 candidate set is a subset of the spatial eigenvectors included in the 0.50 candidate set and similarly the 0.50 candidate set is a subset of the 0.25 candidate set of spatial eigenvectors (i.e., $0.75 \subseteq 0.50 \subseteq 0.25$, where $\subseteq$ is subset). Griffith (2003) gives these threshold values as relating to weak, moderate, and strong positive spatial autocorrelation depicted in the random variable, **Y** (e.g., here the image). The candidate spatial eigenvectors contained in the 0.25 candidate set include global, regional and local patterns and is therefore more likely to model the spatial structure present in an image. The 0.25 candidate set will likely choose a more complete set of spatial eigenvectors to model the spatial structure of an image, but also will require more computational intensity. The 0.50 candidate set includes mostly global and regional patterns. This threshold substantially lowers the number of candidate eigenvectors to be tested for inclusion in the final spatial filter. The exclusion of the finer local patterns with this candidate set results in a loss of detail in the final

Table 4.2. Number of Candidate Eigenvectors for scenarios and sampling

Number of Candidate Eigenvectors			
Case Study	***0.25***	***0.5***	***0.75***
Small	103,399	102,674	101,992
Medium	495,839	492,198	489,012
Sampled Small	25,835	25,664	25,495
Sampled Medium	123,957	123,041	122,249

EISF. The 0.75 candidate set eliminates regional and local spatial patterns, leaving spatial eigenvectors with global patterns which fail to capture the detail in the image's spatial structure.

In this empirical study, the spatial eigenvectors inclusion criterion is based on a threshold amount of variance accounted for in an image by a spatial eigenvector. Here, two arbitrary selection criteria thresholds are tested, 0.01 and 0.001. If a simple linear regression between the a spatial eigenvector and the standardized image bands produce a squared coefficient value above this threshold the spatial eigenvector is multiplied by its coefficient value and linearly combined with all other selected spatial eigenvectors. A candidate spatial eigenvector must account for at least 1% (i.e., 0.01 selection criteria) or 0.1% (i.e., 0.001 selection criteria) of the variance in image. These thresholds are much smaller than those typically employed in conventional spatial filtering analysis because of the massive number of candidate spatial eigenvectors.

The number and type of spatial eigenvectors included in the final spatial filter and the amount of spatial structure captured by the final spatial filter are evaluated. The efficiency of this implementation and hardware and software considerations are discussed as well as the computational intensity of the algorithm.

The following questions are considered in this assessment:

1. Final EISF
    a. What are the similarities and differences in the output spatial filters constructed using the different sampling methods compared to those constructed using the complete spectral information?
    b. Is there some systematic nature to the common spatial eigenvectors included in an EISF created from the sampled image?
    c. Are there spatial eigenvectors that appear in an EISF constructed from the sampled surface and that do not appear in an EISF constructed from the complete spectral information?
        i. Is there some systematic nature to these spatial eigenvectors?
        ii. Are they all global, regional, local patterns or a mixture patterns?
    d. How much variance does an EISF constructed from the sampling methods account for in an image compared to an EISF constructed from the complete spectral information?
2. Computational intensity
    a. How much CPU time is required to construct an EISF for the complete spectral information compared to the CPU time required for the sampled spectral information?
    b. How does CPU time change according to candidate and variance threshold values?

c. Is computational efficiency gained by computing and selecting from a set of sampled candidate spatial eigenvectors and sequentially computing the complete spatial eigenvector for inclusion in the final spatial filter?

3. Hardware / Software
    a. What are the hardware specifications for the computers used to test this algorithm? Why were they chosen? Are there specific hardware concerns that should be considered before executing the algorithm?
    b. What software is used to implement the methodology and for what reasons?

4.1 DATA DIAGNOSTICS

4.1.1 Small Image Case Study

Flightline C1 (FLC1), a dataset of an agricultural landscape located in the southern part of Tippecanoe County, Indiana, is found in Landgrebe's *Signal Theory Methods in Multi-spectral Remote Sensing* (2003). This dataset has been studied many times in other published work and used in testing a variety of classification methods [see Ingram and Actkinson (1973); Starks et al. (1977); Tadjudin and Landgrebe (2000); Landgrebe (2003); Karakahya et al. (2003); Dundar and Landgrebe (2004)]. FLC1 was collected by an airborne data scanner in June of 1966. The image has 12 spectral bands and image dimensions of 948-by-220 (208,560 pixels), and is displayed in side by side bands in Figure (4.1). This image's total number of pixels is relatively small in terms of most remotely sensed images. The scanner's instantaneous field of view (IFOV) is 3 miliradians, and the scan was at an altitude of approximately 2600 ft above the terrain. The scan

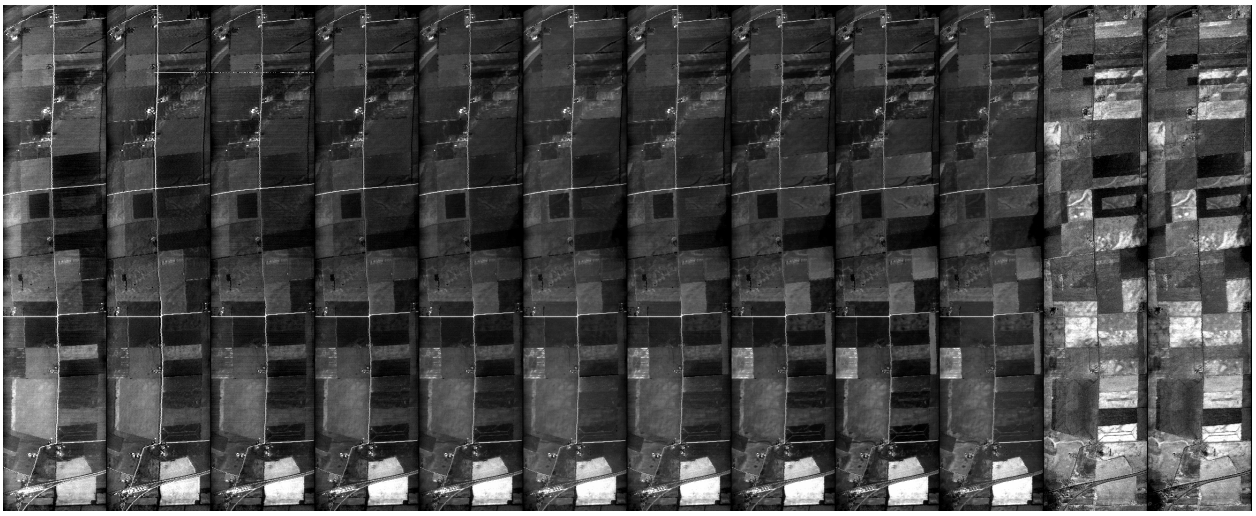


Figure 4.1. Flightline C1 all 10 visible bands and 2 inferred bands side by side

is approximately ±45º about nadir, and each pixel was digitized to 8 bit precision. No radiometric adjustments were made to the data for atmospheric or other observational effects (Landgrebe, 2003). The spectral range of each band is given in Table (4.3). This image consists almost exclusively of agricultural fields, which are relatively homogeneous in nature, and includes virtually no shadow or slope to affect spectral response.

Since this data is to be used in a statistical analysis it is important to evaluate characteristics of the data that might impact assumptions made by a statistical model. The algorithm employed on this case study image uses a simple linear regression to determine if an eigenvector should be used in a spatial filter. This statistical model assumes linearity, constant variance, normality and independence. To explore these assumptions all the bands have been standardized, meaning each has a mean of zero and standard deviation of one. Table (4.4) gives the Ryan/Joiner statistic which is similar to the Shapiro/Wilk test for data normality, the closer this value is to one the more likely the data conform to a normal distribution. The Ryan Joiner

Table 4.3. Wavelength of each band for the small case study image

Wave-length, µm	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12
Min	0.40	0.44	0.46	0.48	0.50	0.52	0.55	0.58	0.62	0.66	0.72	0.80
Max	0.44	0.46	0.48	0.50	0.52	0.55	0.58	0.62	0.66	0.72	0.80	1.00

statistic is significant at a p-value of 0.01 for all the bands but, as can be seen in Table (4.4), there is some deviation from normality for all the bands.

Table (4.3) also gives the maximum and minimum wavelength values for each band in micrometers. It is of note that only band 11 and band 12 are outside the visible light range. Also given in Table (4.4) are the Moran Coefficient and Geary ratio values for each band and their Z-score test against random. The closer the *MC* value is to one the higher the positive spatial autocorrelation is in each band. All of the bands are highly positively spatial autocorrelated. This is also verified by the Geary Ratio which approaches 0 when measuring a highly positively

Table 4.4. FLC1 data characteristics: Band Number, Ryan Join statistics (measure of conforming to a normal distribution), Moran Coefficient, Geary Ratio, and Effective sample size

	Small Dataset Standardize Image Bands					Effective
Band	**R/J**	**MC**	**Z Sig.(Rand.)**	**GR**	**Z Sig.(Rand.)**	**sample size**
b1	0.923	0.826	532.711	0.174	524.565	10,534
b2	0.862	0.761	491.083	0.239	489.968	11,616
b3	0.797	0.679	437.944	0.322	436.897	13,440
b4	0.866	0.873	562.834	0.128	561.614	10,069
b5	0.867	0.889	573.112	0.112	572.059	9,869
b6	0.876	0.899	579.824	0.101	579.226	9,777
b7	0.872	0.898	578.906	0.102	578.162	9,777
b8	0.884	0.913	588.800	0.087	587.737	9,689
b9	0.895	0.920	593.671	0.080	592.432	9,606
b10	0.878	0.928	598.259	0.072	597.736	9,527
b11	0.985	0.912	588.075	0.088	587.338	9,689
b12	0.982	0.919	592.600	0.081	591.981	9,606

Table 4.5. Image band multicollinearity---the visible bands (b1- b10) have high collinearity and the inferred bands (b11-b12) also show high collinearity while the collinearity between the visible and inferred bands is low

Band Multicollinearity

	b2	*b3*	*b4*	*b5*	*b6*	*b7*	*b8*	*b9*	*b10*	*b11*	*b12*
b1	0.85	0.78	0.82	0.85	0.77	0.72	0.68	0.62	0.558	-0.08	0.15
b2		0.81	0.9	0.87	0.85	0.82	0.79	0.75	0.7	-0.05	0.09
b3			0.81	0.84	0.78	0.77	0.77	0.73	0.7	-0.03	0.07
b4				0.96	0.96	0.96	0.94	0.91	0.87	-0.07	0.09
b5					0.95	0.93	0.93	0.89	0.87	0.002	0.04
b6						0.98	0.93	0.88	0.94	0.14	0.11
b7							0.96	0.93	0.96	0.07	0.04
b8								0.98	0.96	-0.07	0.09
b9									0.96	-0.15	0.15
b10										0.08	0.08
b11											0.94

spatially autocorrelated geographic distribution. The effective sample size gives an estimate of how many independent samples are in each band of the image and gives a lower bound for the number of samples that should be taken for the sampling algorithm.

Between band correlation generally is very high in remotely sensed images. Table (4.5) shows the correlation coefficient for all pairs of standardized bands in the image. Interestingly there is high between band correlations between the visible and visible bands and between the inferred and inferred bands but low between band correlation between the visible and inferred bands. In summary all bands are significantly close to normal, a significantly high amount of positive spatial autocorrelation is present in all the bands and a large amount of between band correlations are present.

4.1.2 Medium Image Case Study

The medium dataset consists of a subset for a Landsat ETM+ image. This dataset is available for download free from the Earth Science Data Interface (ESDI)[4]. Bands 1-5 are employed in this analysis. The image was collected over Austria, Germany, Italy, Liechtenstein, and Switzerland on September 13th, 1999. Landsat is a spaceborne sensor with an IFOV is 42.5 miliradians, and the scan is at an altitude of approximately 705 km above the terrain. Each pixel is digitized to 8 bit precision. No radiometric adjustments are made to the data for atmospheric or other observational effects. Displayed in the top left corner of Figure (4.2) is the complete Landsat ETM+ image with a square showing the 1000-by-1000 pixel subset used here. The five

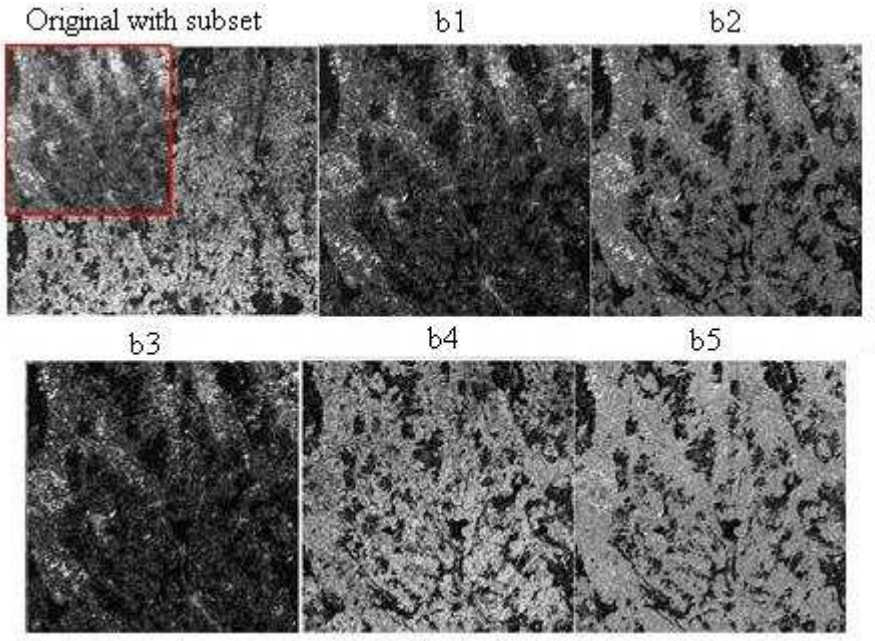


Figure 4.2. Top left image is the complete ETM image with the 1000-by-1000 pixel subset shown by inset box, top 2 right images and bottom row images show 5 standardized z-score bands for the subset used in the analysis.

---

[4] Global Land Cover Facility Earth Science Data Interface
http://glcfapp.umiacs.umd.edu:8080/esdi/index.jsp Last accessed 08/13/2009

Table 4.6. Data Diagnostics for subset of Landsat ETM image

	Medium Dataset Standardize Image Bands					Effective sample
Band	**R/J**	**MC**	**Z Sig.(Rand.)**	**GR**	**Z Sig.(Rand.)**	**size**
b1	0.900	0.871	1230.59	0.129	1229.983	48,251
b2	0.960	0.910	1286.00	0.090	1285.53	46,437
b3	0.890	0.886	1252.76	0.114	1252.366	47,298
b4	0.950	0.924	1306.37	0.076	1306.21	46,038
b5	0.990	0.906	1280.89	0.094	1280.553	46,437

subset bands are shown side by side in the rest of Figure (4.2). This image subset consists of some mountainous areas to the south with relatively flat areas with urban development in the center and north in the image and is relatively heterogeneous in nature including valleys, mountains and urban areas.

This data is also to be used in a statistical analysis and so an evaluation of some data characteristics that might impact on model assumptions is necessary. All the bands have been standardized, meaning each has a mean of zero and standard deviation of one. Table (4.6) gives the Ryan/Joiner statistic which is significant at a p-value of .01 for all the bands but, as can be seen in Table (4.6); there is some deviation from normality for all the bands; at least in part this is probably due to the very high amount of spatial autocorrelation in the bands. Also given in Table (4.6) are the *MC* and *GR* values for each band. These statistics show all of the bands are highly positively spatial autocorrelated.

Table (4.7) shows the correlation coefficient for all pairs of standardized bands in the image subset. The between band correlation is especially high for the visible bands 1, 2, and 3. Interestingly band 5 the mid-IR band also shows high multicollinearity with the visible for this image, while the near- IR little between band correlation

Table (4.7). Band multicollinearity matrix for Landsat ETM subset image

Band Multicollinearity

	b2	b3	b4	b5
b1	0.891	0.935	0.046	0.617
b2		0.887	0.392	0.841
b3			0.005	0.650
b4				0.650

4.1.3 Large Image Case Study

The large case study image is a Quickbird image of an area in the vicinity of Panama City, Florida used by Congalton and Green (2009) as a case study on accuracy assessment. The dataset belongs to NOAA and is used in their Coastal Change Analysis Program (C-CAP). The area has little elevation change but is characterized by diverse landuse and landcover types.

This Quickbird image was collected by a satellite sensor on November 9th, 2004. The scanner's IFOV is 2.44 meters for the multispectral data. The image's swath width is approximately 16.5 km. The scan is approximately ±25^{o} about nadir, and each pixel was

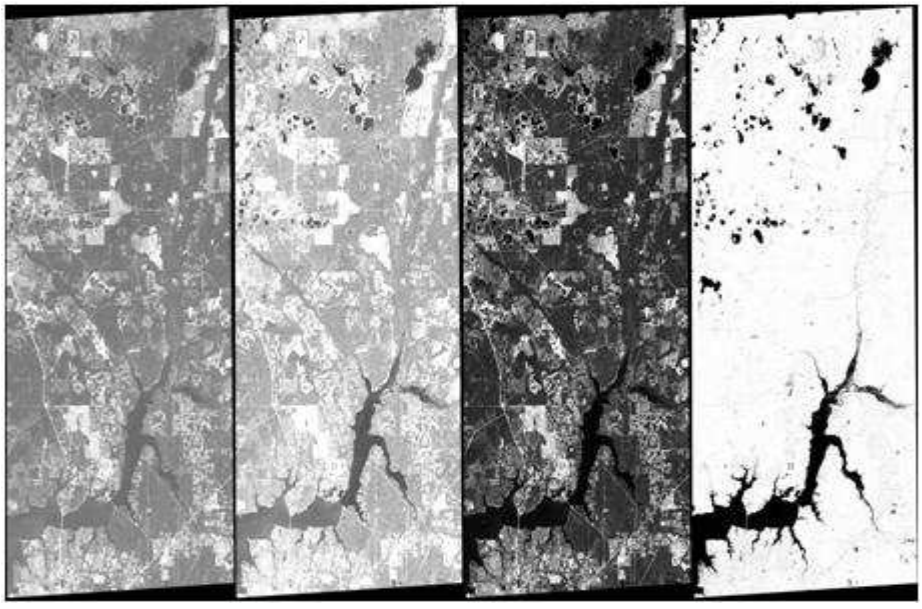


Figure 4.3. Case study 3 Quickbird image 4 bands image

Table 4.8. Data diagnostics for large dataset image

	Normality	Spatial autocorrelation		Effective	Wavelength, µm	
Band	**S/W**	**MC**	**GR**	**sample size**	**Min**	**Max**
1	0.9617	0.973	0.025	4,886,906	0.45	0.52
2	0.9649	0.959	0.040	4,934,509	0.52	0.60
3	0.9699	0.955	0.045	4,948,742	0.63	0.69
4	0.9694	0.916	0.084	5,193,739	0.76	0.90

digitized to 11 bit precision. No radiometric adjustments were made to the data for atmospheric or other observational effects. The image has 4 spectral bands and image dimensions of 7,380-by-14,974 (i.e., 110,508,120 pixels), and is displayed as in Figure (4.3). All Quickbird imagery have 3 visible bands, a near inferred band, and one panchromatic band.

Table (4.8) gives some summary statistics for this dataset, including the Shapiro/Wilk test for data normality, which is significant for all the bands with a p-value of 0.01. All four bands show high positive spatial autocorrelation as measured by *MC* and *GR* and have significant Z-scores. The effective sample size is approximately 5,000,000 pixels for all the bands. The between band correlation is very high for the visible bands 1, 2, and 3, while the near inferred band 4 has still significant but lower correlation with the visible bands, as shown in Table (4.9).

Table 4.9. Large dataset image between band collinearity

Band Multicollinearity			
	b2	b3	b4
b1	0.975	0.949	0.394
b2		0.977	0.509
b3			0.465

4.2 HARDWARE AND SOFTWARE

Using this data and empirical study design, the following sections evaluate the methods by which the EISF algorithm is made efficient enough for application research in image analysis. In terms of hardware this empirical study is limited to the computers available for use within the Geospatial Information Science Department at the University of Texas at Dallas. All scenarios are executed on Dell desktop computers with an Intel Core 2 Quad processor with 2.4 GHz CPU and 3.25 GB RAM running 32 bit windows and located in the EPPS room 3.602 GIS Lab.

In terms of software, MultiSpec, a free image analysis software from researchers at Purdue's LARS laboratory, is used to extract digital number values from the multispectral images, and for the creation of most of the image figures shown in this document. ENVI also is used for data display and for part of the initial image diagnostics performed, such as Moran's I and Geary's G computations. The EISF algorithm, as presented in this text, was initially programmed in MatLab using small datasets and then written in SAS for implementation with large datasets. SAS was chosen for its efficient handling of large datasets including a robust ability to connect to a variety of database systems. Although it might be true that a software package like R does not require licenses fees and is therefore more accessible to a larger community of researchers. The SAS program itself is open source and well commented and might be readily translated, updated or changed to another programming language as best suits the researcher.

4.3 COMPUTATIONAL INTENSITY / OUTPUT EISF

The results of the scenarios for each case study image are described for each algorithm methodology (i.e., complete spectral surface EISF (C-EISF), simple random sample spectral surface (SR-EISF), systematic sample spectral surface (SS-EISF), Geographically stratified simple random sampled spectral surface (GR-EISF), geographically stratified systematic sample (GS-EISF), and finally parallel processing (P-EISF) for the large case study image) discussed and the example scenario k25s001 is evaluated in terms reproducing spatial structure of the image and type of spatial eigenvectors included in the final EISF.

CHAPTER 5

CONCLUSIONS

5.1 SAMPLING

Cressie (1996) argues that when much redundancy is present in data often a complete dataset is unnecessary and a sample of the data can be sufficient for modeling proposes. To evaluate this, the complete spectral surface is used to construct a series of EISFs and these are compared to a series of sampling methods.

5.1.1 Complete Spatial Surface (C) / Small Case Study

Figure (5.1) shows, as number of pixels in the image remains constant and the *MC* threshold decreases and the number of candidate spatial eigenvectors increase, that the time required to construct a C-EISF increases linearly. The more candidate eigenvectors to be tested and the higher the selection criteria, the more CPU time required to construct a C-EISF. The scenario with a *MC* value threshold of at least 0.25 and a variance threshold of 0.001 is constructed in 2 hours and 40 minutes, whereas the scenario requiring at least a *MC* value of 0.75 and 0.01 variance threshold is constructed in 34 minutes

Figure (5.2) shows the amount of variance accounted for by each scenario. All scenarios with variance threshold of 0.001, hereafter called the S001 scenarios, account for substantially higher total variance in the image than scenarios with variance threshold 0.01, hereafter called

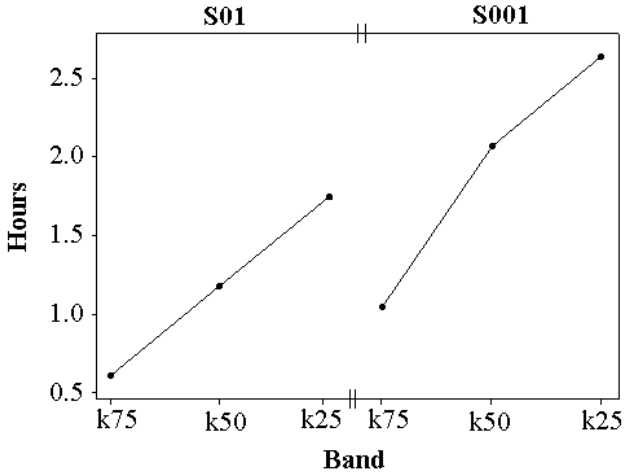


Figure 5.1. Time in hours required to construct EISF for 12 band small case study image.

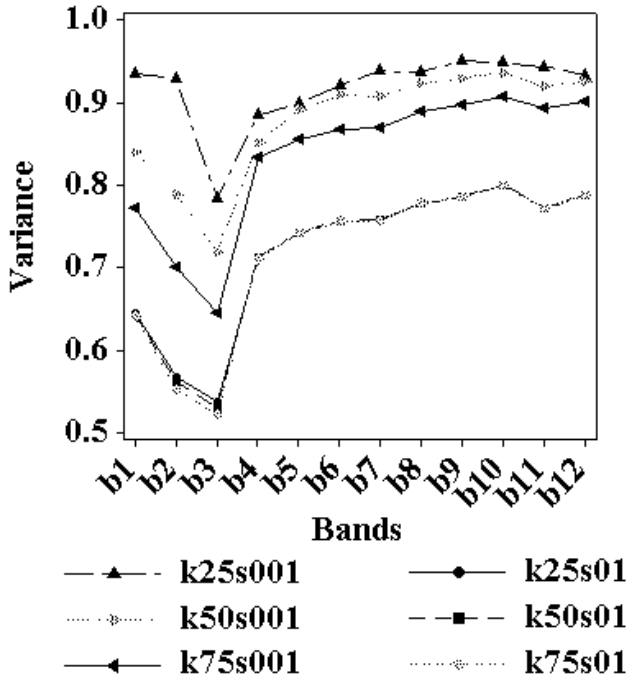


Figure 5.2. Percent variance account for in each band, for each scenario in the small case study

the S01 scenarios. The S01 scenarios also group together systematically increasing the variance accounted for in the image by the value of the *MC* threshold criteria. The k25s001 C-EISF of band 3 accounts for the least amount of variance of all the bands (78%), while band 10's C-EISF accounts for the most at 95%. Band 3 also had the lowest *MC* value; this may indicate that the candidate eigenvectors chosen should include more regional and local eigenvectors or possibly eigenvectors containing negative spatial autocorrelation. Considerably less variance is accounted for by the S01 scenarios and these scenarios are not systematically grouped by candidate eigenvector threshold as are the S001 scenarios. These results are intuitive; the scenarios with the most candidate eigenvectors (i.e., k25) and the lowest variance threshold (i.e., S001) accounted for the most variance in the image for every band.

The number of eigenvectors included in the C-EISFs for all scenarios is illustrated in Figure (5.3). This number changes systematically, dropping as the number of candidate spatial eigenvectors drops and as the selection criteria becomes more rigorous. Each EISF constructed

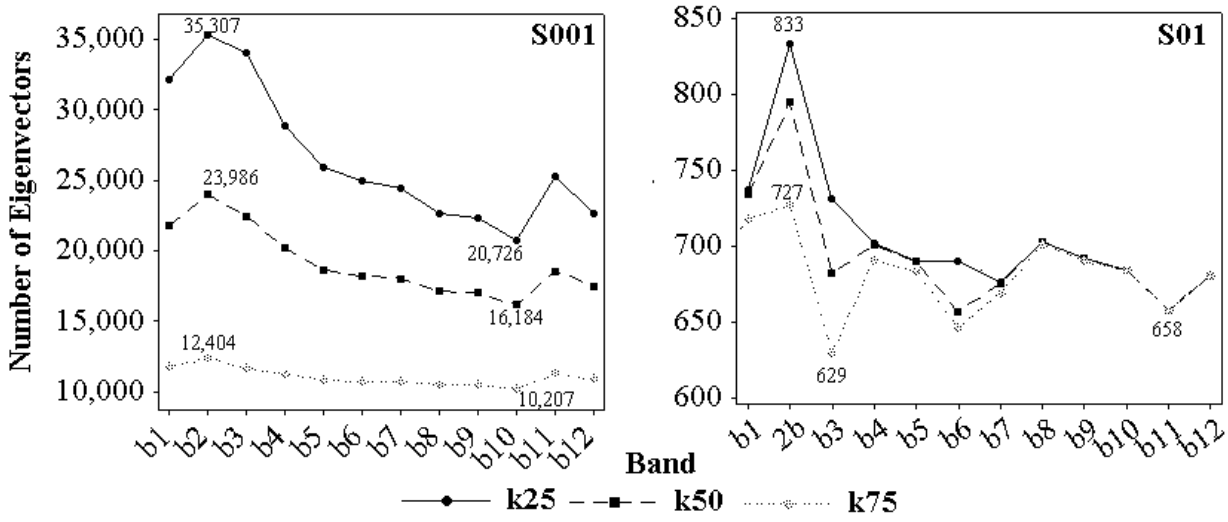


Figure 5.3. The number of spatial eigenvectors included in the final EISF for each scenario and each band in the small case study image

from the S001 scenario has more than 10,000 spatial eigenvectors, with band 2's C-EISF including 35,307 spatial eigenvectors (49% of the candidate spatial eigenvectors and 16.9% of all possible eigenvectors). The S01 scenarios include less than 1,000 spatial eigenvectors in the final C-EISFs. It is interesting to note that the number of spatial eigenvectors varies systematically among the scenarios similar to the amount of variance accounted for in the image. For example, all scenarios include more spatial eigenvectors in band 2 than any other band even though the selection and candidate criteria changed. Moreover the k25s001 selected eigenvectors not only include the most spatial eigenvectors but also included all of the eigenvectors from k50s001, k75s001, and all the S01 scenarios. As the selection criteria are being relaxed more spatial eigenvectors are being added to the final C-EISF but all previous selected eigenvectors are still included.

Figure (5.4) shows the C-EISFs for each scenario constructed for the first band. There is a clear distinction between the C-EISFs constructed using S001 scenarios on the left and the C-EISFs constructed using S01 scenarios on the right. Much of the spatial detail present in S001 scenario filters is absent in the S01 scenario filters. This could be attributed at least partially to substantially fewer spatial eigenvectors used to construct the S01 scenarios. There are also subtle differences between the scenarios as the *MC* value threshold changes. This becomes more apparent when the list of spatial eigenvectors chosen for each C-EISF is compared. Even visually it is possible to notice some differences between the C-EISFs of the same variance threshold when the images are compared at their highest resolution showing that both k50s001 structure of the image is captured in all scenarios although the level of detail changes. The S01

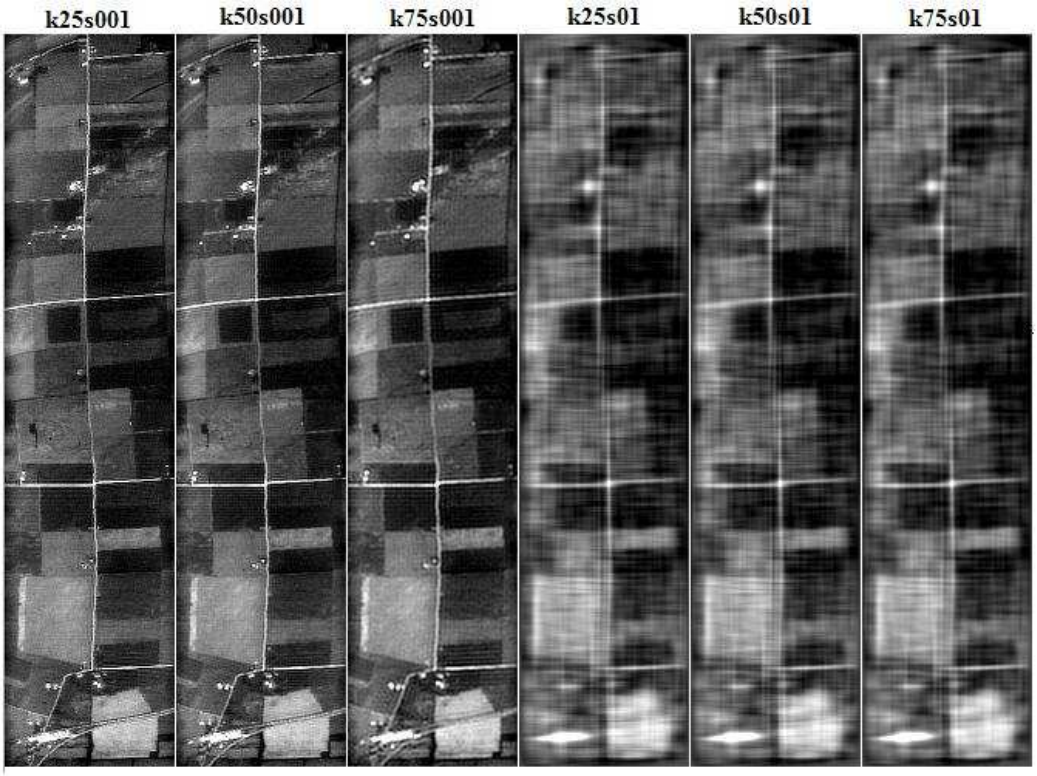


Figure 5.4. EISF of band 1 for each scenario using the complete spectral information and k75s001 are more pixilated than the k25s001 scenario. Important to note is that the spatial scenarios appear similar to the original image after a low pass filter applied, with the roads and fields distinguishable within the image. Although the number of spatial eigenvectors included in the S01 scenario filters is under a thousand, for all bands the spatial structure of the image is preserved by the final C-EISF.

The C-EISFs for scenario k25s001 account for on average across all bands over 95% of the variance in the image. They capture the spatial structure, shown in Figure (4.1), of the image essentially replicating it, as can be seen in Figure (5.5) . The percent of global, regional, and local bands included in the C-EISF for each band is given in Table (5.1). On average over all the bands the C-EISFs are split ~42% global, 30% regional and 27% local spatial eigenvectors.

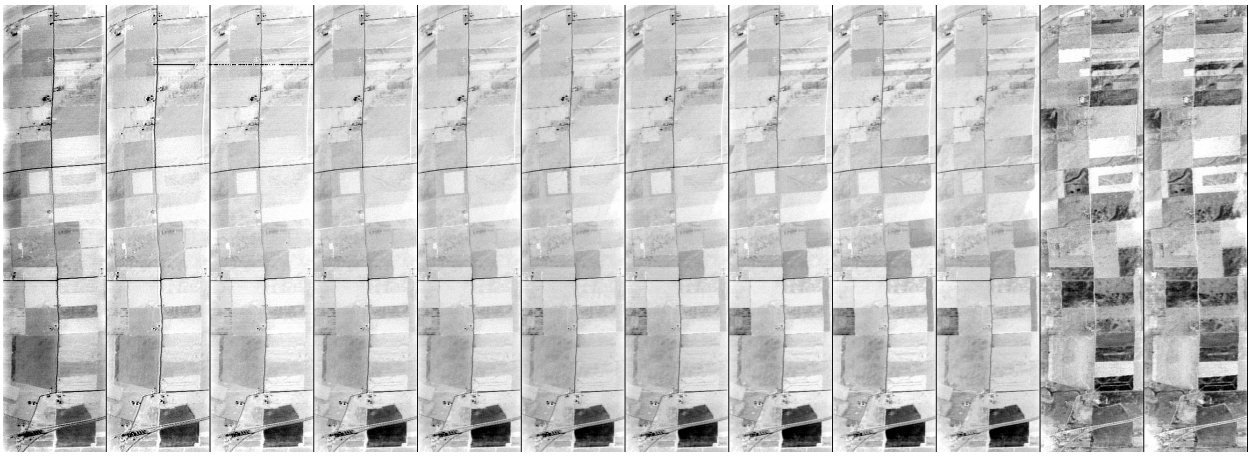


Figure 5.5. The C-EISFs for each band side-by-side for the k25s001 scenario.

Therefore although there are fewer spatial eigenvectors chosen of a particular type of spatial eigenvector their overall contribution to the EISF is higher because of their higher coefficient value. For this scenario there are 82 spatial eigenvectors with a coefficient value higher than 0.001 which account for 47.65% of the variance in the image,

Table 5.1. Global, regional and local number and percent of total spatial eigenvectors included in the C-EISF for each band.

No Sample (C)						
Bands	**Global**		**Regional**		**Local**	
b1	44.12%	11,735	29.77%	10,060	26.11%	10,400
b2	46.42%	12,404	29.25%	11,582	24.33%	11,321
b3	47.02%	11,643	29.13%	10,816	23.84%	11,582
b4	49.43%	11,261	28.51%	8,947	22.06%	8,646
b5	44.87%	10,801	28.77%	7,854	26.36%	7,256
b6	48.19%	10,711	28.84%	7,518	22.96%	6,737
b7	36.66%	10,724	31.08%	7,328	32.26%	7,328
b8	35.42%	10,472	32.60%	6,687	31.98%	5,502
b9	34.43%	10,454	31.69%	6,565	33.88%	5,307
b10	39.23%	10,207	30.87%	5,977	29.90%	4,542
b11	41.90%	11,293	30.11%	7,282	27.99%	6,673
b12	43.07%	10,877	29.94%	6,586	26.99%	5,200

while the remaining 32,113 spatial eigenvectors included in the EISF account for 41.39% of the variance. In every band more than 10,000 global and over 5,000 regional and local eigenvectors are chosen.

As the number of pixels in an image increases so does the number of candidate spatial eigenvectors and therefore the required computational intensity of the algorithm. For the 1,000,000 pixel image the number of candidate spatial eigenvectors increases to 308,248 for *MC* > 0.25, 184,660 for *MC* > 0.50 and 84,985 for *MC* > 0.75. Figure (5.6) shows the hours required to construct the C-EISFs for this medium sized image. Similar to the small case study as the number of candidate spatial eigenvectors increase the required computation time increase linearly. The shortest computation time relates to the k75s01 scenario which executes in 17.4 hours, whereas the k25s001 scenario requires 50.3 hours to complete. In contrast to the C-EISFs of the small case study, which has approximately ¼ as many pixels this medium sized image requires 35 times as much CPU time.

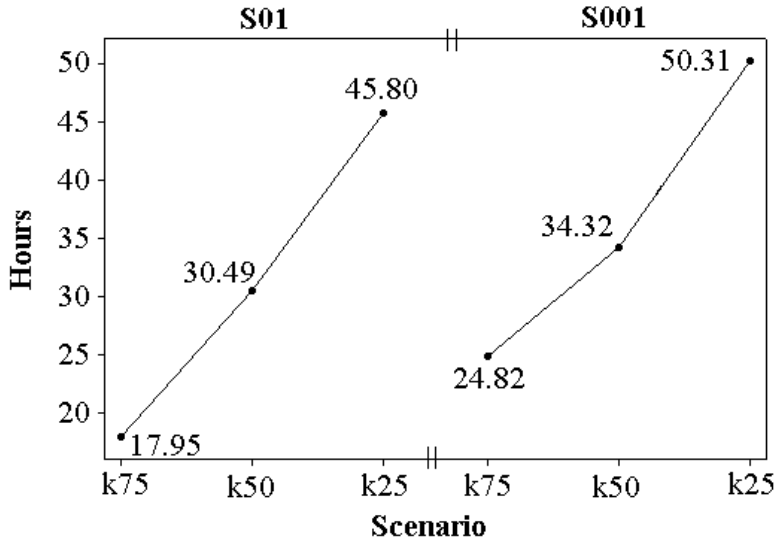


Figure 5.6. Time required in hours to construct EISF using the complete spectral information for the medium case study image.

Similar to the small case study image the amount of variance accounted for in the medium image by the C-EISFs varies systematically with the scenarios. Figure (5.7) illustrates how the amount of variance accounted for in the image by the C-EISFs is substantially higher for the S001 scenarios, and also follows a systematic ordering according the *MC* value threshold. However, for this medium sized image the scenarios k25s001 and k5s001 account for nearly the same amount of variance in the image, but the k50s001 scenario requires only 30 hours of CPU time while the k25s001 requires over 50 hours. The amount of variance accounted for in the image by the S01 scenarios is lower than the S001 scenarios but still between 40% and 70%. The MC threshold has no impact on the S01 scenarios. This could indicate that the coefficient values of the selected regional and local spatial eigenvectors are very small.

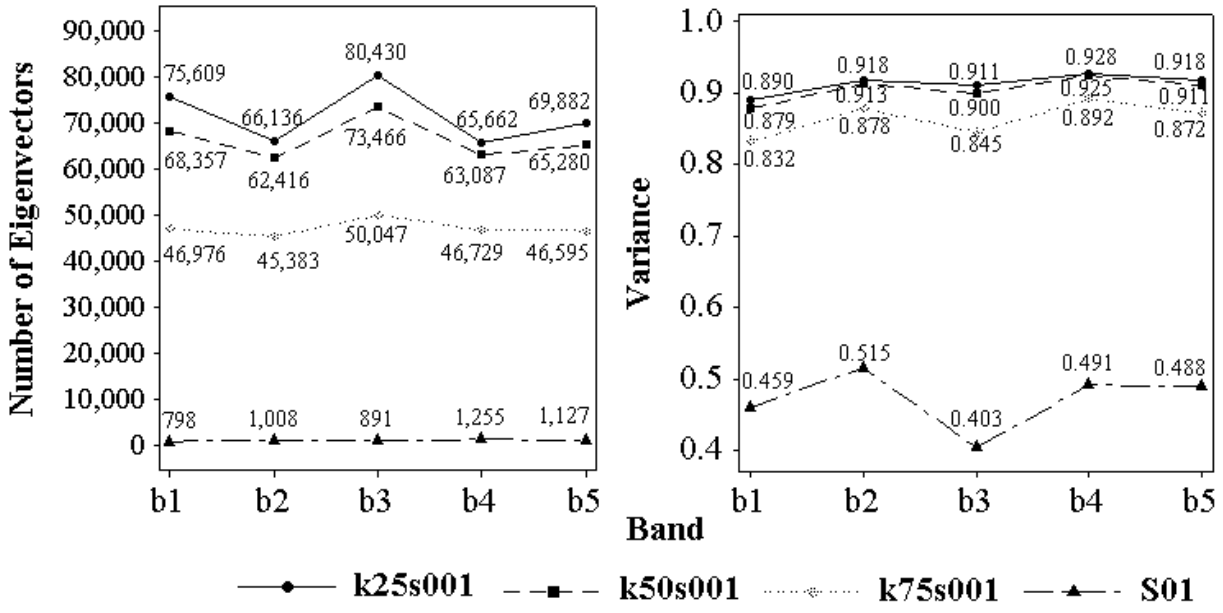


Figure 5.7. Number of spatial eigenvectors chosen for the final C-EISFs for each scenario for the medium case study image (left). Variance accounted for in the image by the C-EISFs for each scenario for the medium case study image.(right)

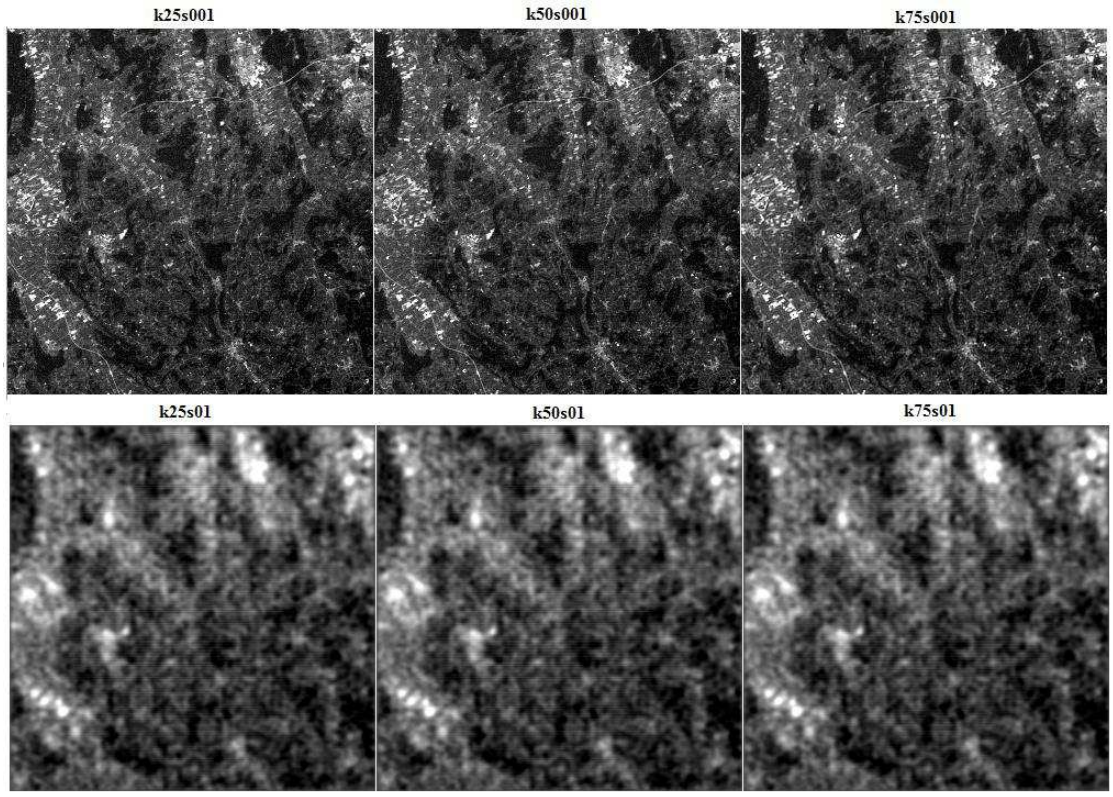


Figure 5.8. C-EISFs for the first band of the for the medium case study image for each scenario.

Similar to the small case study the S001 C-EISFs included many more spatial eigenvectors than the S01 C-EISFs, as shown in Figure (5.7) right. The S001 C-EISFs included between 40,000 and 80,000 spatial eigenvectors, whereas the S01 C-EISFs included only between 700-1300 spatial eigenvectors. Again there is a systematic grouping of the number of eigenvectors selected but the systematic nature changes between the S001 and S01 scenarios.

Figure (5.8) shows each scenario for the C-EISFs, for the first band of the medium case study image. The S001 filters show more detail and essentially reconstruct the spatial structure in the image, whereas the S01 filters show less detail while still preserving the overall spatial structure of the image which is still visually recognizable using less than 2000 spatial

eigenvectors. Again similar to the small image the S01 scenarios look visually like an image after a low pass spectral filter has been applied.

For this medium image the k25s001 scenario selects on average across all bands of 66.16% global, 26.96% regional and 6.88% local spatial eigenvectors. There is a tendency to choose more global spatial eigenvectors and comparatively fewer local patterns, as shown in Table (5.2). On average across all bands 0.1% of the spatial eigenvectors have a coefficient value above 0.001 and they represent 28% of the variance accounted for in the image. Figure (5.9) shows the C-ESIFs constructed for the k25s001 scenario for each band. These C-EISFs essentially replicate the spatial structure of the original image.

The scenarios for the large case study image could not be executed on the desktop computers available in the EPPS labs, since the memory was exhausted. Some data diagnostics were completed using frequencies of the digital number information, these were shown in the proposal of this document. This large case study will be further explored within a parallel computing environment or computers with larger Random Access Memory at some point in the future.

Table 5.2. The number and percent of spatial eigenvectors selected for inclusion in the C-EISF broken down in to global, regional and local patterns for each band in the medium image.

	Number of Spatial Eigenvectors			**Percent of Total**		
Complete	***Global***	***Regional***	***Local***	***Global***	***Regional***	***Local***
b1	46,976	21,381	7,252	62.13%	28.28%	9.59%
b2	45,383	17,033	3,720	68.62%	25.75%	5.62%
b3	50,047	23,419	6,964	62.22%	29.12%	8.66%
b4	46,729	16,358	2,575	71.17%	24.91%	3.92%
b5	46,595	18,685	4,602	66.68%	26.74%	6.59%

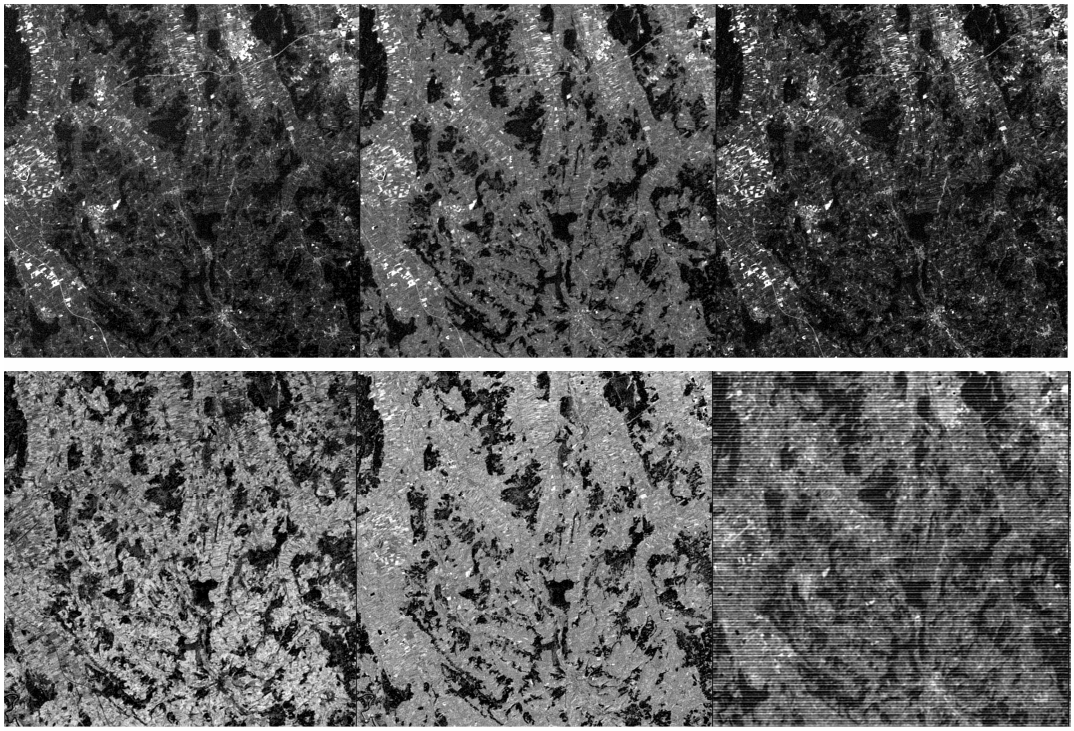


Figure 5.9: C-EISFs for scenario k25s001 bands 1 – 5.

5.1.2 Simple Random Sample (SR)

A simple random sample is often employed in statistical analysis because it is well known to insure an unbiased sample. It is employed here because of its prevalence in remote sensing literature and applications. It seems unlikely that a simple random sample of the image spectral information would be able to capture the spatial structure of that spectral information because there is no guarantee of the geographical dispersion of the sample pixels which might be clustered purely by chance leaving certain geographical areas un-sampled or under-sampled. For this empirical study, a seed of 1978 is used in SAS's random number generator to choose a sample of ¼ of the image pixels. The same randomly selected pixel index values are used to choose samples in each band. There is no systematic nature to the row and column indices of the pixels chosen by the simple random sampling procedure, meaning the sampled eigenvector

must be constructed using notation given in Griffith (2004). This unfortunately slows down the algorithm substantially because of the repetitive computations required to construct the sampled eigenvector.

Figure (5.10) contrasts the hours required to construct the SR-EISFs and the C-EISFs. The SR-EISFs all require more time to construct than the C-EISFs because the reformulation of the analytical eigenvectors cannot be implemented. The fastest SR-scenario, k75s01, requires 3 hours and 56 minutes to execute and all scenarios are completed in plus or minus 10 minutes of 3 hours and 50 minutes. The longest running C-EISF, k25s001, requires 2.6 hours CPU time while the C-EISF k75s01 requires only 26 minutes to execute. This is a result of the inability to implement the reformulation of the analytical eigenvectors when taking a simple random sample. The sample is not as efficient as simply doing less computation by factoring the analytical eigenvector equation.

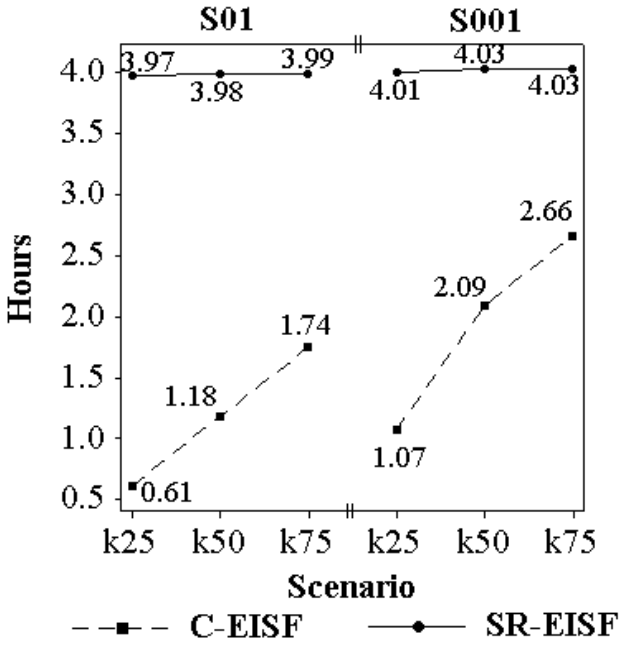


Figure 5.10. Time in hours required to construct the final EISFs for the small case study image using the simple random sampling technique

With the increase in required CPU time for the SR-EISFs comes a drastic decrease in the amount of variance accounted for in the image by the SR-EISFs, the highest being 0.7% variance. The *MC* value threshold had no impact on the amount variance accounted for in the image or the number of spatial eigenvectors included in the SR-EISF only the variance threshold changed the final SR- EISF, which is apparent from Figure (5.11) right. Figure (5.11) left shows the total number of spatial eigenvectors included in the SR-EISFs. The S001 scenarios identify a similar number of eigenvectors as do the S01 scenarios for the complete spectral information but the variance accounted for in the image is substantially lower for the SR-EISFs. The S01 scenarios are unable to identify even one spatial eigenvector for any band except bands 11 and 12, which identify 7 and 9 spatial eigenvectors respectively.

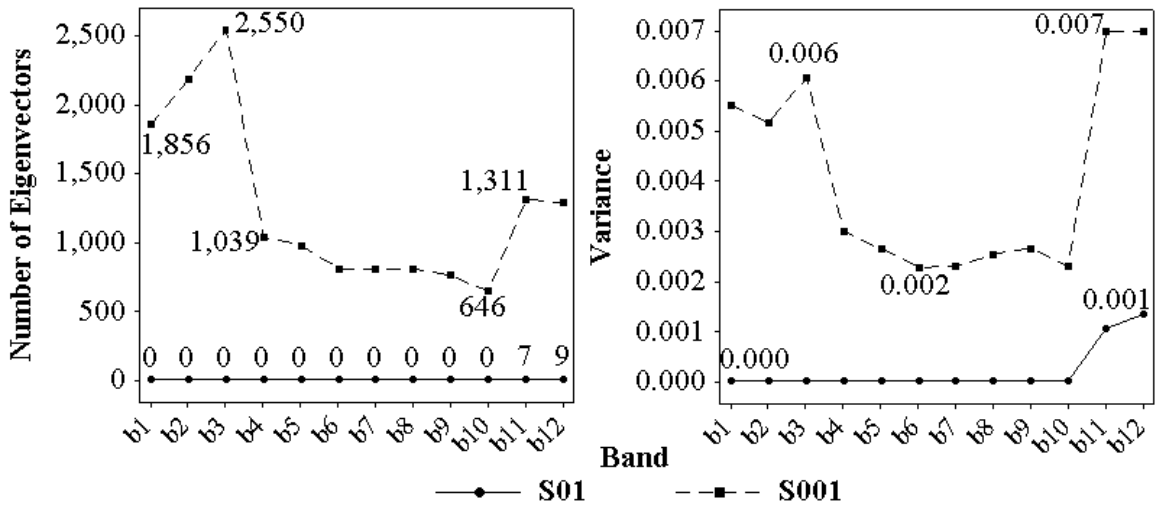


Figure 5.11. Variance accounted for in small case study image using simple random sampling (left). The number of spatial eigenvectors included in final EISF (right).

The SR-EISFs for the first band of each scenario are shown in Figure (5.12). All S001 scenarios construct the same SR-EISF and all S01 scenarios did not select any spatial eigenvectors for the first band. The spatial structure of the image is not visually discernable using this sample method. It is clear that the S001 scenarios SR-EISFs capture the random nature of the sample but the spatial structure of the image has been lost. Although the number of spatial eigenvectors selected for inclusion in the SR-EISFs are similar to the number chosen by the S01 scenario for the C-EISFs, which capture the spatial structure of the image with less local information, the SR-EISFs do not capture the spatial structure of the image at all.

For each sampling method the scenario k25s001 is compared to the same scenario for the C-EISFs in terms of number and type of spatial eigenvectors included in the constructed EISFs. Figure (5.13) shows the SR-EISFs for the small case study scenario k25s001. Visually these patterns look random across all the bands, although the MC value for the first band's SR-EISF is 0.61 and significantly different from random, none of the spatial structure of the original image is captured and no systematic nature is visually apparent. This positive spatial autocorrelation in the band may be a result of a combination of both regional and local clustering that is not immediately apparent in inspecting Figure (5.13), but becomes more apparent when examining the type spatial eigenvectors selected for inclusion in the final EISF. Table (5.3) gives the number of spatial eigenvectors included in the C-EISF and the SR-EISF (C1SR1), next those included in the SR-EISF and not the C-EISF (C0SR1) and those included in the C-EISF and not in the SR-EISF (C1SR0).

It is clear that C1SR0 has consistently more spatial eigenvectors but it also apparent that little overlap exists between the C-EISF and SR-EISF. The number of spatial eigenvectors

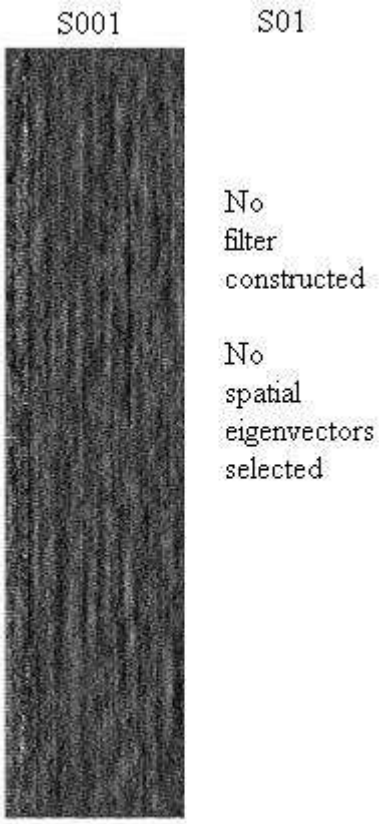


Figure 5.12. Band 1 EISFs constructed using a simple random sample of the spectral information for each threshold scenario

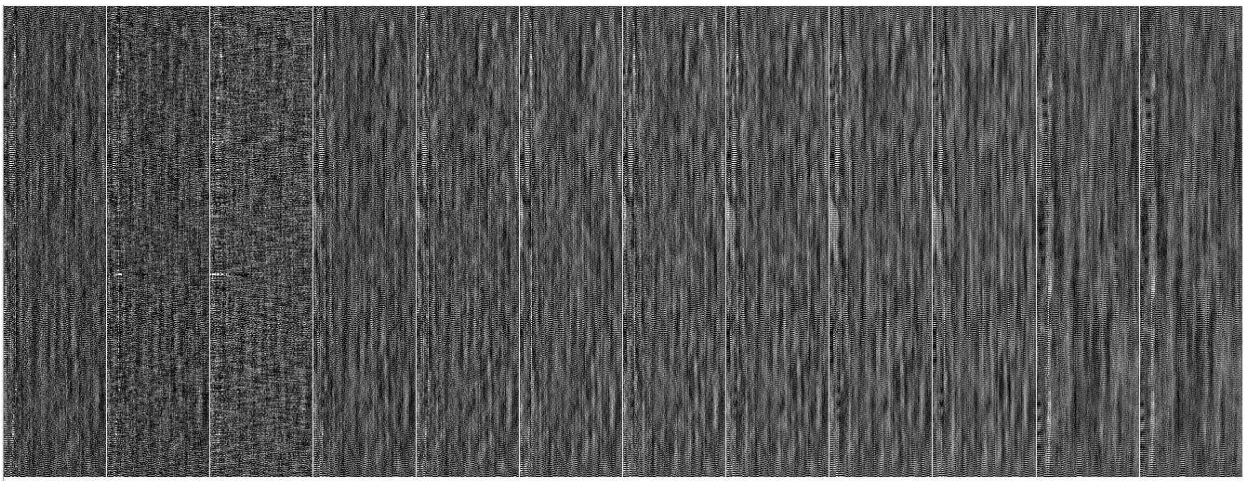


Figure 5.13. The SR_EISFs for the k25s001 scenario for small case study image

common to the SR-EISF and C-EISF is nearly equal to the spatial eigenvectors which are distinct between them. Table (5.3) shows that the percent of total global, regional and local spatial eigenvectors selected is similar to the C-EISF but the total number selected is substantially smaller. On average over all the bands, 34.77% of the spatial eigenvectors included in the SR-EISFs were global, 16.19% were regional and 48.44% were local. This seems to show a tendency for more local patterns to be chosen by the SR-EISFs unlike the C-EISFs.

The SR-EISFs for the small case study image require more CPU time to construct, account of less variance in the image, and capture no apparent spatial structure of the image. Although the image contains a large amount of redundant information a simple random sample of the spectral information does not seem to represent the complete spectral information well enough to construct an EISF for this small image and these scenarios

Table 5.3. Columns (2-4) show the number of spatial eigenvectors the SR-EISFs have in common with or distinct from the spatial eigenvectors of the C-EISFs. Columns (5-10) show the type of patterns chosen by percent total and the raw counts of spatial eigenvectors.

Common and Distinct EVs				Random Sample (SR)					
Bands	C1SR1	C0SR1	C1SR0	Global		Regional		Local	
b1	801	31394	1055	670	36.10%	285	15.36%	901	48.55%
b2	977	34330	1214	759	34.64%	342	15.61%	1090	49.75%
b3	1050	22991	1500	846	33.18%	396	15.53%	1308	51.29%
b4	453	28401	586	358	34.46%	171	16.46%	510	49.09%
b5	388	25523	589	339	34.70%	153	15.66%	485	49.64%
b6	321	24645	488	384	47.47%	130	16.07%	295	36.46%
b7	315	24102	496	274	33.79%	140	17.26%	397	48.95%
b8	296	22365	508	258	32.09%	147	18.28%	399	49.63%
b9	292	22034	469	236	31.01%	150	19.71%	375	49.28%
b10	229	22497	417	203	31.42%	120	18.58%	323	50.00%
b11	492	24756	819	455	34.71%	211	16.09%	645	49.20%
b12	434	22239	856	435	33.72%	217	16.82%	638	49.46%

.

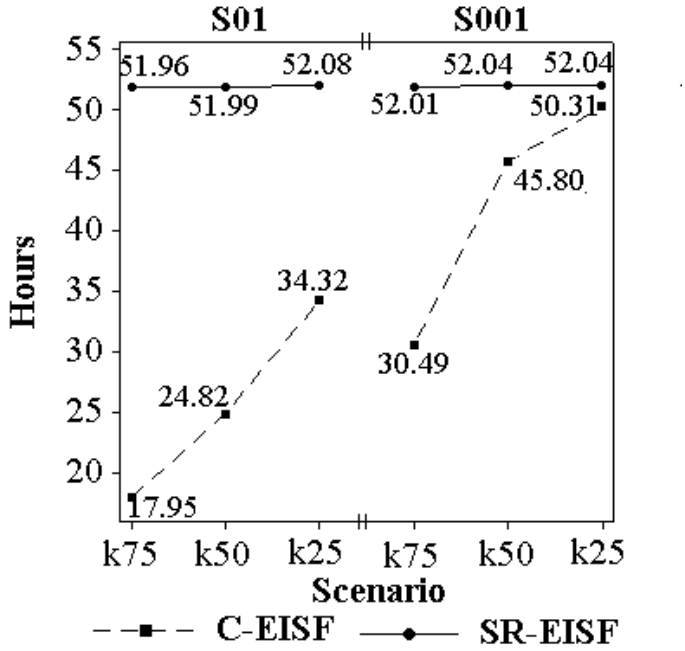


Figure 5.14. Required CPU time for the SR-EISFs for the medium case study image for all scenarios

For the LandSat data of the medium case study image, the CPU time increases from a range of ~17-50 hours per image for the C-EISFs 52 hours plus or minus 10 minutes for all scenarios. This is mostly due to the fact that reformulation of the eigenvector calculation cannot be used when simple random sampling is implemented. Although sampling lowered the number of pixels in the image on which calculation must be done the required number of calculations is higher since the repetitive computations cannot be avoided by using the reformulated algorithm. The number of hours required to construct the SR-EISFs compared to the C-EISFs for each scenario is shown in Figure (5.14).

Similar to the small case study image, with the gain in the CPU time there is a drastic decrease in the amount of variance accounted for in the image by the SR-EISFs. Band 2's SR-EISF accounted for 0.0044 of the variance in the image, which is the most for all the bands.

Again similar to the small case study image, there is no distinction made between the scenarios by the *MC* threshold value as shown in Figure (5.15). All bands except band three include less than 1000 spatial eigenvectors in the S001 SR-EISFs and the S01 scenarios selected no more than 5 spatial eigenvectors. There is a systematic ordering to the amount of variance accounted for in the image by the EISFs of each *MC* value threshold. Although the S001 scenarios select between 670-1,048 spatial eigenvectors, while the S01 EISFs choose between 1 -5 spatial eigenvectors, the difference in the amount of variance accounted for in the image is only between 0.0026- 0.0030.

Figure (5.16) shows the SR-EISFs the first band for each scenario, since the *MC* threshold did not change the amount of variance accounted for in the image up to 4 decimal places only the k25s001 and k25s01 scenarios are shown. There seems to be very little spatial structure for the image captured by the SR-EISF although

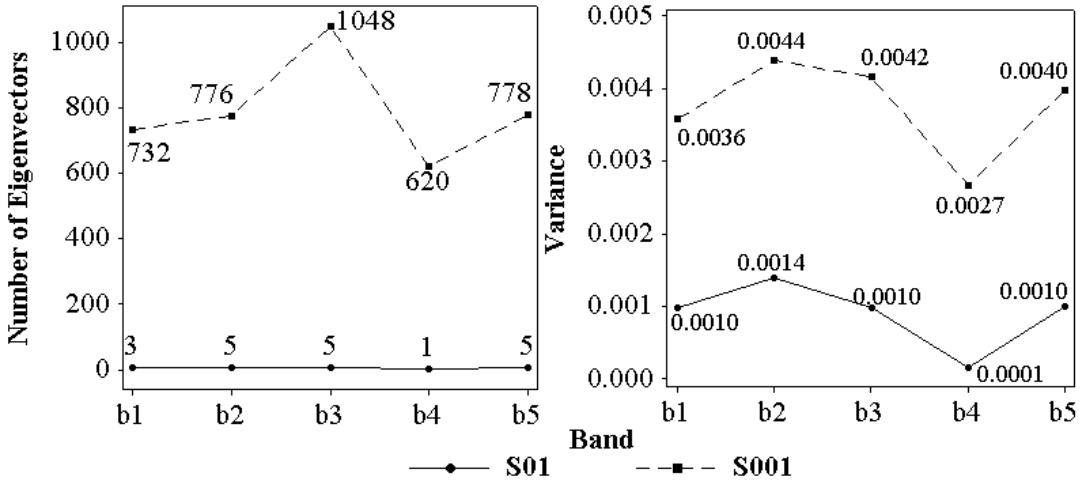


Figure 5.15. (Left) The number of spatial eigenvectors selected for inclusion in the SR-EISFs for the S001 and S01 scenarios. (Right) The percent variance accounted for in the image by the SR-EISFs of the S001 and S01 scenarios.

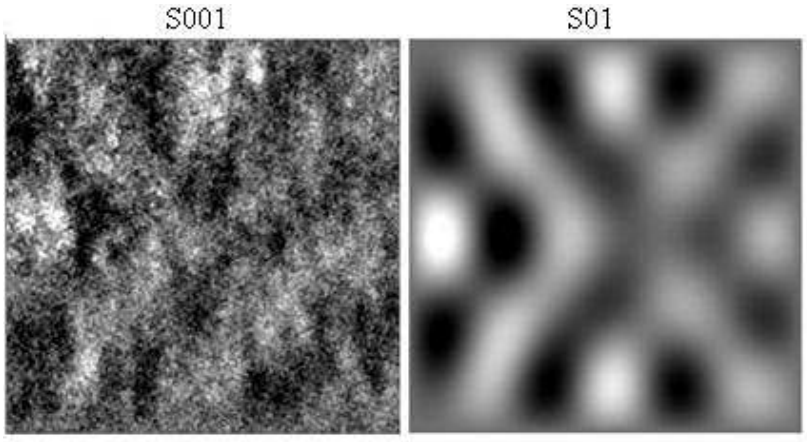


Figure 5.16. Band 1 SR-EISFs for each scenario for the medium case study image.

the top left corner is generally brighter than the bottom right, which is also apparent in the original image. The S01 filters again appear visually to be a smoother version of the S001 filters that show regional variation across the image.

Figure (5.17) shows the SR-EISFs for each band of the k25s001 scenario. Each captured very little variance and do not visually resemble the spatial structure of the original image, but do differ from each other both visually and by the spatial eigenvectors included in the EISFs. Given in Table (5.4) is the number of spatial eigenvectors common and distinct to the C-EISF and the SR-EISF for the medium case study image. Approximately 43.19% of the selected SR-EISFs are also included in the C-EISF while 0.34% of the C-EISF are included in the SR-EISF. This is consistent with what was found for the small case study image.

The global, regional and local spatial eigenvectors for scenario k25s001 are given in Table (5.4). On average over the bands, 53% of the included spatial eigenvectors are global, 23% are regional and 22% are local. This differs from the small case study image which included higher levels of local pattern

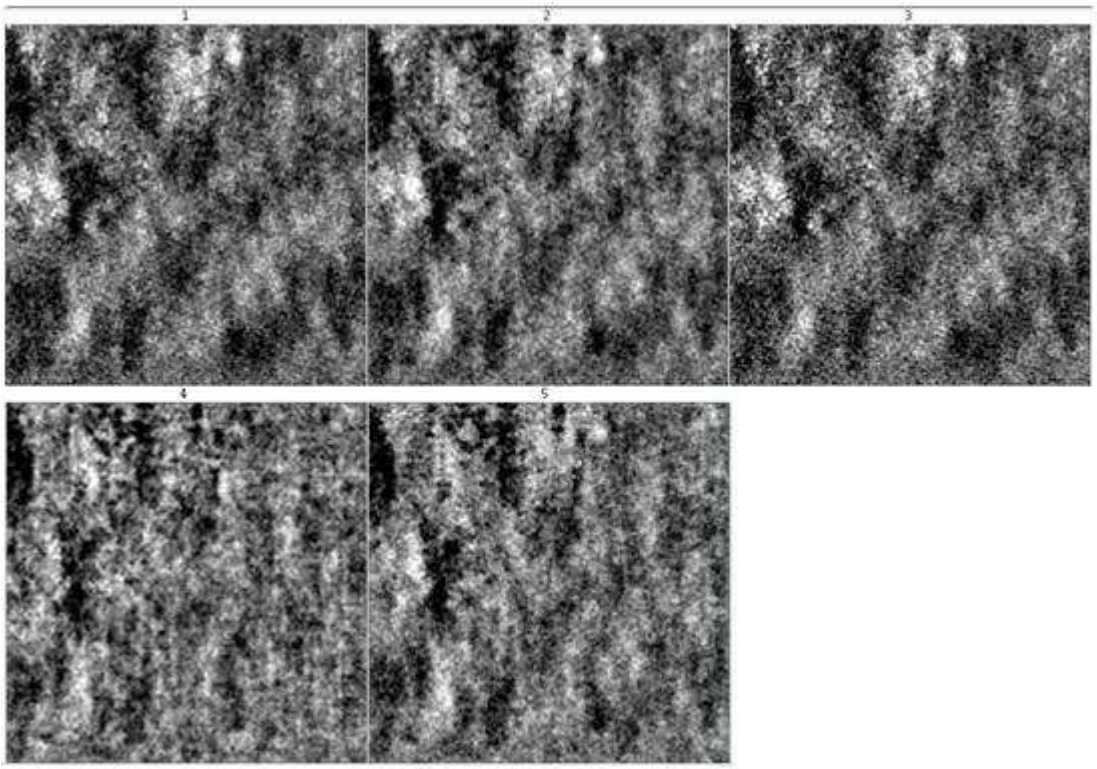


Figure 5.17. The k25s001 scenario SR-EISFs for all bands of the medium case study image.

By applying simple random sampling to both the small and medium case study images there is a loss in computational efficiency. The amount of required CPU time increases for all scenarios and the change in *MC* value makes virtually no difference in a final SR-EISF. Virtually no variance is accounted for in either image by the SR-EISFs and visually no spatial structure of the original images is apparent. There is no systematic nature to which spatial eigenvectors are included both in the C-EISF and SR-EISF. These 2 case study images show that although there is a large amount of redundancy in the spectral information of the images a simple random sample of the spectral information cannot be used to select spatial eigenvectors for spatial filter construction.

Table 5.4. The number and percent total of the type of spatial eigenvector included in the SR-EISFs (left) and the number of spatial eigenvectors common and distinct to the SR-EISF and the C-EISFs (right).

SR	Global		Regional		Local		C0SR1	C1SR0	C1SR1
b1	285	54.18%	122	23.19%	119	22.62%	502	75379	230
b2	285	51.63%	136	24.64%	131	23.73%	538	65898	238
b3	410	54.45%	169	22.44%	174	23.11%	739	80121	309
b4	246	53.71%	112	24.45%	100	21.83%	407	65449	213
b5	290	52.16%	137	24.64%	129	23.20%	547	69651	231

5.1.3 Geographically Stratified Random Sample (GR)

To evaluate whether the poor performance of the SR-EISFs might relate to the poor geographical distribution of sample points across the image, a geographically stratified random sample is implemented. This method ensures that the random samples are geographically disperse and therefore might be more representative of the complete image spatial structure. One large drawback to this method, similar to simple random sampling, is that the row and column values of the geographically stratified random sample do not conform to the reformulation of the spatial eigenvector computation. This requires that the reformulation for computing the spatial eigenvectors described in Griffith (2004) be implemented, which is found to significantly slow computation time in the simple random sampling algorithm. To geographically stratify the case study images, an image is divided into blocks of four pixels, a seed (i.e., here 1978) is used in the SAS random number generator to randomly select a single pixel sample from within that block; this selects ¼ of the image pixels to be used to construct the GR-EISFs.

Similar to the SR-EISFs, the GR-EISFs require more time to construct than the C-EISFs using the reformulated eigenvector computation. Each scenario is completed within ten minutes of 3 hours 56 minutes. In some cases this is over 3 hours longer than when the complete spectral

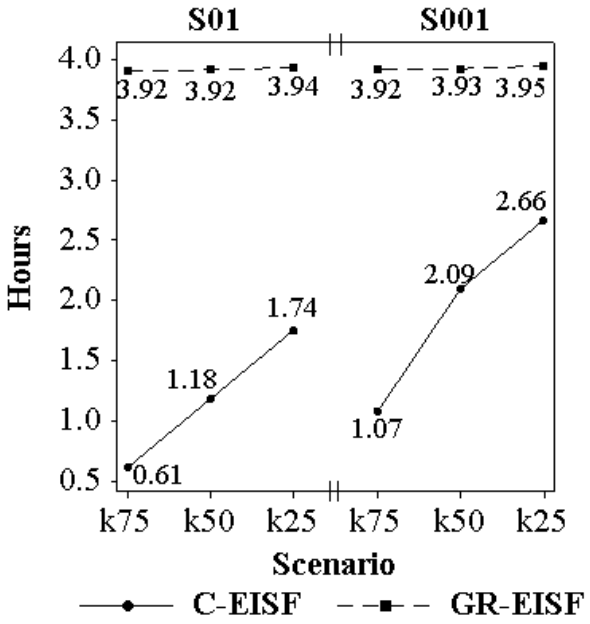


Figure 5.18. The number of hours in CPU time required to construct the GR-EISFs for the small case study image.

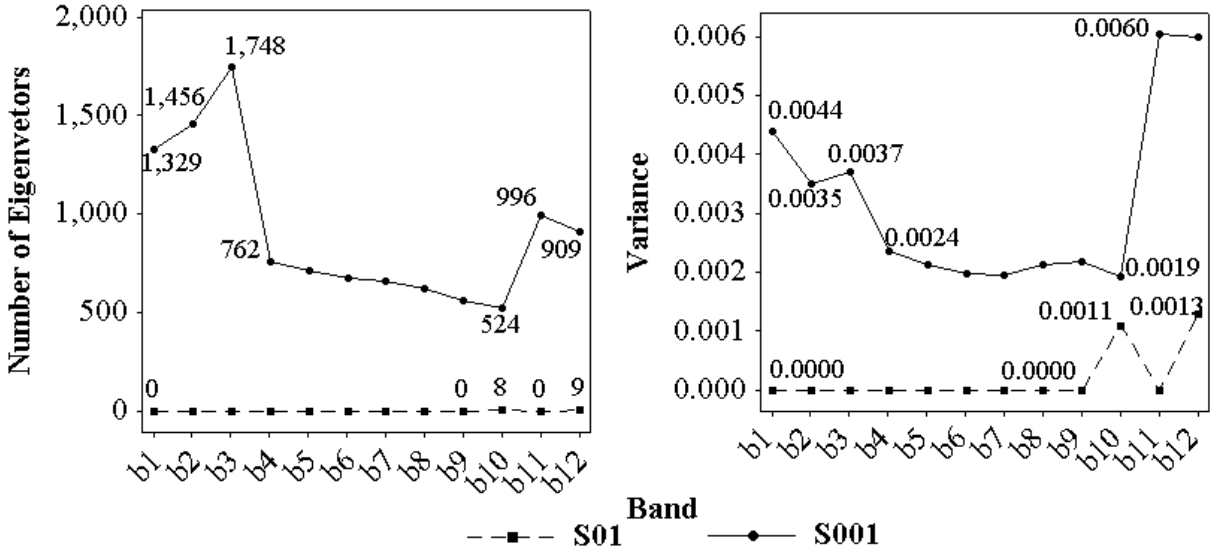


Figure 5.19. The number of spatial eigenvectors chosen for the GR-EISFs for each band in the small case study image (Left). The amount of variance accounted for in the image by the GR-EISFs for each band. (Right)

information is used. A comparison of the required CPU time for both C-EISFs and the GR-EISFs are shown in Figure (5.18). The extra time gains nothing in the amount of variance accounted for in the image. On the contrary the amount of variance accounted for in the image is substantially lower than for the C-EISFS. The highest amount of variance accounted for in the image is 0.0045, in band 12 for the S001 scenarios, as shown in Figure (5.19). The GR-EISFs and SR-EISFs account for essentially the same amount of variance in the image and follow the same pattern across the bands of the image, although fewer overall spatial eigenvectors are chosen for the GR-EISFs.

Similar to the SR-EISFs, the GR-EISFs show no distinction between the *MC* threshold values up to four decimal places. The filters change only in terms of the variance threshold. The GR-EISFs S01 scenarios do not identify a single spatial eigenvector for inclusion in the GR-EISFs with the exception of bands 10 and 12, which indentify 8 and 9 spatial eigenvectors respectively, as is illustrated in Figure (5.19). The GR-EISFs S001 scenarios select fewer spatial eigenvectors than the SR-EISFs S001 scenarios and yet both are within a similar range to the number of spatial eigenvectors chosen for C-EISFs S01 scenarios, but unlike the C-EISFs these scenarios account for virtually no variance in the image.

The GR-EISFs for the first band of each scenario are shown in Figure (5.20). The S001 scenarios show the GR-EISFs visually resemble the SR-EISFs but are not the same. The spatial structure of the original image is lost, the GR-EISF appears mostly random, although it has an *MC* value of 0.64 which is significant, similar to the SR-EISFs. If compared closely, the GR-EISFs, for this example, seem to be slightly more systematic in visual appearance than the SR-

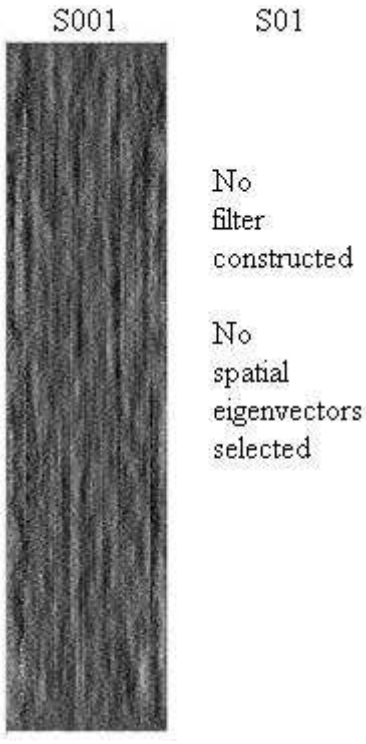


Figure 5.20. The GR-EISFS for the first band of each scenario for the small case study image.

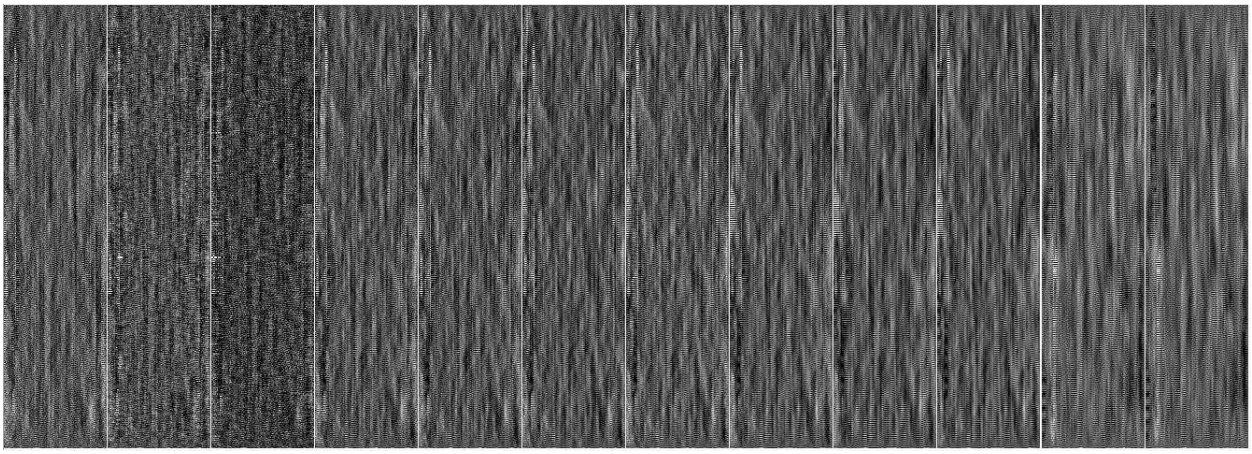


Figure 5.21. GR-EISFs of scenario k25s001 for all bands for the small case study image

Table 5.5. Coulmn (2-4) show the number of spatial eigenvectors the GR-EISFs have in common with the C-EISFs. Columns (5-10) show the type of patterns chosen by percent total and the number chosen for the SR-EISFs

Common and Distinct EVs				**Sample (GR)**					
Band	*C1GR1*	*C0GR1*	*C1GR0*	*Global*		*Regional*		*Local*	
b1	574	31621	755	466	35.06%	212	15.95%	651	48.98%
b2	675	34632	781	217	35.16%	237	16.28%	707	48.56%
b3	726	33315	1022	585	33.47%	296	16.93%	867	49.60%
b4	334	28520	428	282	37.01%	129	16.93%	351	46.06%
b5	158	25753	558	129	18.02%	65	9.08%	522	72.91%
b6	276	24690	399	239	35.41%	125	18.52%	311	46.07%
b7	269	24148	389	232	35.26%	112	17.02%	314	47.72%
b8	245	22416	380	229	36.64%	111	17.76%	285	45.60%
b9	233	22093	327	195	34.82%	109	19.46%	256	45.71%
b10	218	20508	306	190	36.26%	98	18.70%	236	45.04%
b11	373	24875	623	84	34.34%	157	15.76%	497	49.90%
b12	320	22343	589	323	35.53%	136	14.96%	450	49.50%

EISFs. The S01 scenarios do not identify a single spatial eigenvector and therefore is not constructed.

As expected the k25s001 scenario accounts for the most variance in the image. Figure (5.21) shows all of the GR-EISFs constructed for the small case study image. The lack of spatial structure from the original image is apparent in all of the SR-EISFs constructed for the k25s001 scenario. On close inspection some local systematic pattern is apparent but non-consequential. On average over all the bands 33.92% of the spatial eigenvectors are global 16.45% are regional, and 49.64% are local, as shown in Table (5.5). This is similar to the SR-EISFs except in band 5 where the percent of local spikes. Also shown in Table (5.5) is the number of spatial eigenvectors common between the C-EISFs and the GR-EISFs and those distinct. This is also similar to the numbers found for the SR-EISFs.

For the medium case study image, the CPU time required to construct the GR-EISFs drops from approximately 52 hours for the SR-EISFs to approximately 42 hours. This is an

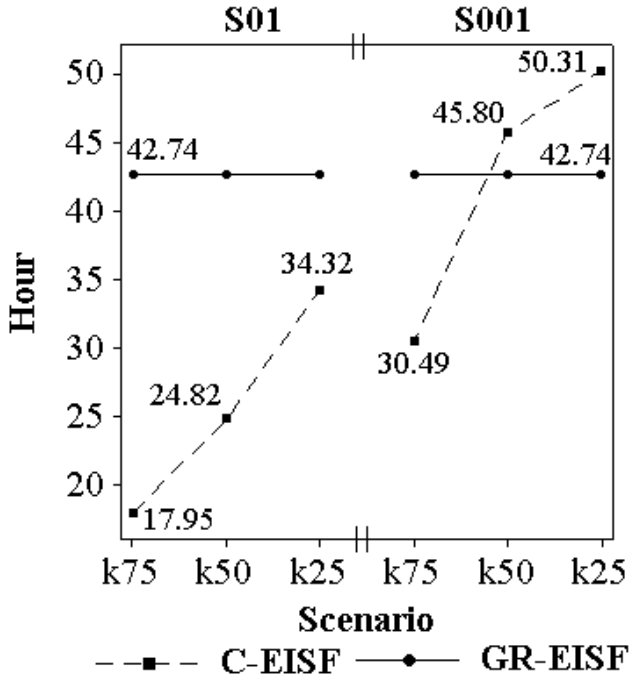


Figure 5.22. CPU time in hours to create the GR-EISFs for each scenario.

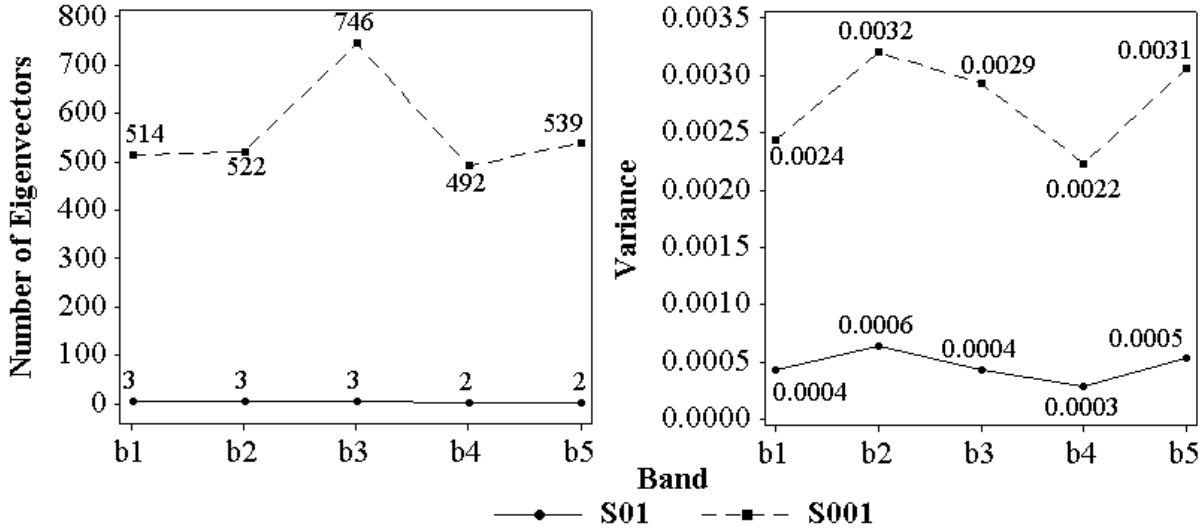


Figure 5.23. (Left) The amount of variance accounted for in the image by each scenario by the GR-EISFs. (Right) The number of spatial eigenvector included in each scenario for the GR-EISFs

improvement but is in contrast to the CPU time to construct the C-EISFs, which range from 17.4 hours to 50.3 hours, the improvement is small at best. This loss of efficiency is again due to the inability to implement the reformulation of the spatial eigenvector calculation for the random sample, therefore although the number of observations requiring computation is lowered the actual number of computations required is not reduced. Figure (5.22) shows the hours required to construct the GR-EISFs in contrast to compared the C-EISFs.

Moreover the amount of variance accounted for in the image by the GR-EISFs is virtually nothing (i.e., 0.0013 for the scenario accounting for the most variance in the image), similar to the small case study GR-EISFs and the SR-EISFs. Also shown in Figure (5.23) the number of eigenvectors chosen for the S001 scenarios is within the same range as the C-EISFs for S01 scenarios but the spatial structure of the original image is not captured by the filter. This is apparent in Figure (5.24) where the GR-EISF for the first band of each scenario is shown. Again there is no distinction between the scenarios with different MC threshold values in term of the number of eigenvectors chosen or the amount of variance accounted for in the image.

Similar to both the C-EISFs and the SR-EISFs, the GR-EISFs for S01 show more detail while the S01 scenarios seem to be constructed from more regional and local patterns. The SR-EISFs and the GR-EISFs are visually similar but the GR-EISFs seem to show more regional and local patterns. The k25s001 scenario of the geographically stratified random sample is shown for each band in Figure (5.25). None of the spatial structure of the original image is preserved in the GR-EISFs. Similar to the SR-EISFs although the images look visually random there is more regional pattern apparent in the medium case study GR-EISFs. Band 1 of the k25s001 scenario

has a *MC* value of 0.61, which is significantly different from random, similar to the bands from the SR-EISFs k25s001 scenario.

Table (5.6) gives the number of spatial eigenvectors selected and the percent of the total spatial eigenvectors selected for inclusion in a GR-EISF for each type of spatial surface. On average over all the bands, 56% of the spatial eigenvectors included in the GR-EISFs are global, 24% are regional and 19% are local. This shows an increase in the number of global spatial eigenvectors chosen and a decrease for both regional and local spatial patterns. This increase in global spatial eigenvectors is similar to what is found for the SR-EISFs for the medium case study image and different from the C-EISFs which remained fairly evenly distributed between each spatial eigenvectors type. There is also a drop in the number of selected spatial eigenvectors common to both the GR-EISFs and the C-EISFs. Only an average of 35% of the GR-EISF selected spatial eigenvectors are also included in the C-EISF, which constitutes on average only 0.2% of the C-EISF. There seems to be little gained by geographically distributing the random sample. The CPU time is generally longer and the final spatial filter does not capture the spatial structure of the original image. Although there is a large amount of redundancy in the image spectral data neither random sampling method are able to identify spatial eigenvectors that capture the spatial structure of the original image

Table 5.6. (Left) The number and percent total of the spatial eigenvectors chosen for the GR-EISFs (Right). The number of common and unique spatial eigenvectors between the GR-EISF and the C-EISF

GR	**Global**		**Regional**		**Local**		**C0GR1**	**C1GR0**	**C1GR1**
b1	214	55.44%	95	24.61%	77	19.95%	343	75430	179
b2	214	56.46%	85	22.43%	80	21.11%	338	65960	176
b3	307	57.28%	116	21.64%	113	21.08%	493	80177	253
b4	201	54.18%	102	27.49%	68	18.33%	318	65488	174
b5	230	57.07%	104	25.81%	69	17.12%	338	69681	201

.

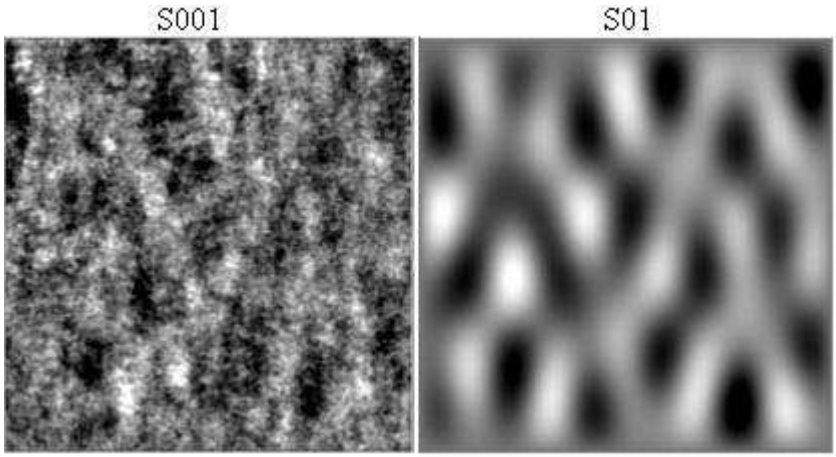


Figure 5.24. The GR-EISFs for all scenarios for the medium case study image.

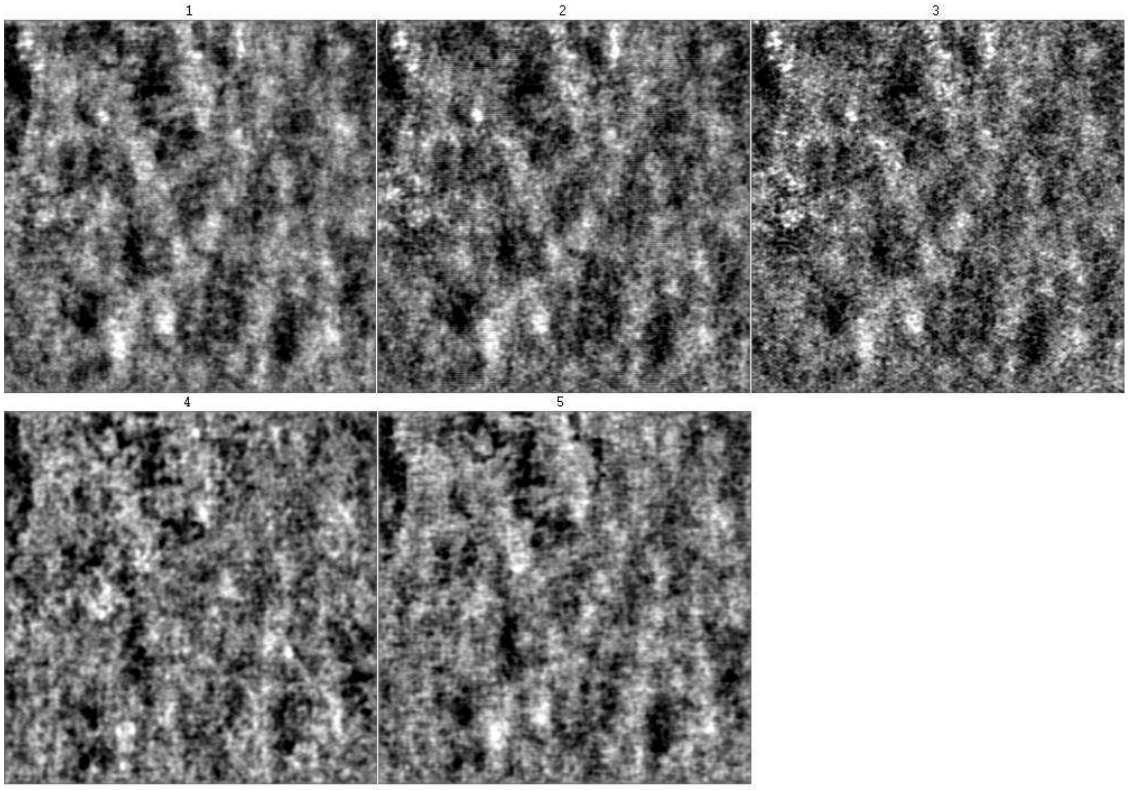


Figure 5.25. The GR-EISFs for the k25s001 scenario for all bands of the medium case study image.

5.1.4. Systematic Sample (SS)

Steheman (1995) states, “to obtain a sample that is spatially well distributed, either stratified or systematic sampling may be used.” Since stratification of the random sampling procedure did not produce EISFs that capture spatial structure, it might be useful to construct a systematic sample. A systematic sample would not only provide a spatially well distributed sample but also may better capture spatial structure in an image. Although the systematic nature of the sample might hide some spatial structure it may be more successful than a simple random sample at identifying patterns. There are several other advantages to a systematic sample as opposed to a simple or stratified random sample. The first advantage is that the reformulation of the spatial eigenvector computation may be implemented for a systematic sample. This is because the systematic sample simply removes columns from the **R** and **K** matrices. So, unlike the random sampling approaches, systematic sampling lowers the number of computations through sampling and through the removal of redundant calculations. Another advantage to systematic sampling is that the periodic nature of the sine function from which the analytical spatial eigenvector is constructed can be sampled systematically and retain the characteristics of orthogonality. This is shown empirically in Chapter 3. A systematic sample is selected as discussed in Chapter 3, here using a random number seed of 1978 and a sampling rate of 4 to select ¼ of the image pixels.

Taking a systematic sample of the small case study’s spectral information substantially changes the computation intensity of the algorithm. The SS-EISFs take substantially less CPU time to construct, with a range of 2.1 to 3.6 minutes compared to a range of 34 minutes to 2

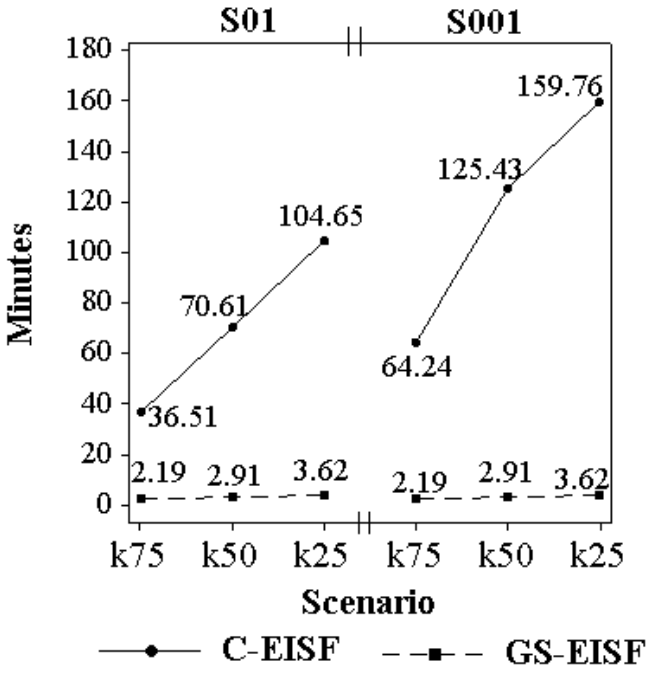


Figure 5.26. Time required to construct the SS-EISF for the small case study image.

hours and 40 minutes for the C-EISFs for this small case study image, as illustrated in Figure (5.26). This large reduction in computation time stems from the ability to combine sampling with the spatial eigenvector reformulation.

Although systematic sampling addresses the computational intensity issue, this systematic sample similar to the other sampling methods fail the produce EISFs that account for a substantial amount of variance in an image. Figure (5.27) right shows the amount of variance accounted for in the image by the SS-EISFs for each scenario. The SS-EISFs account for a similar amount of variance in the image as do the SR-EISFs or GR-EISFs and the total amount is very small. There is in general a systematic ordering to the amount of variance accounted for in the image by each scenario, similar to what is found for the C-EISFs but the patterns of the graphs for each

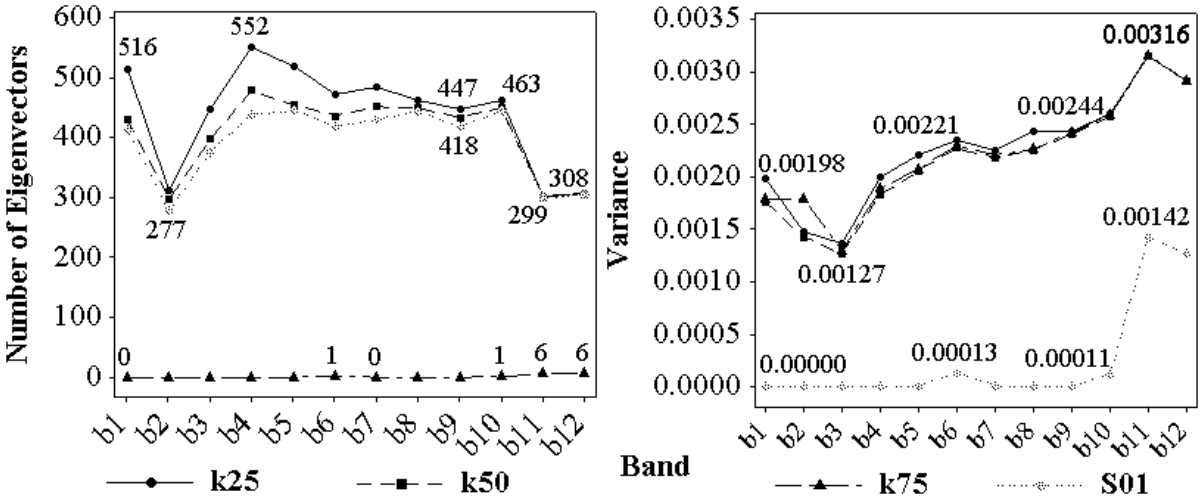


Figure 5.27. Number of spatial eigenvectors included in the final EISF for each scenario and each band in the small case study image (left). Variance accounted for in the small case study image by the SS-EISF (right).

technique are different. For example, the C-EISF of band 3 accounts for the most variance in the image, whereas the SS-EISF of band 3 accounts for the least.

The total number of spatial eigenvectors included in the SS-EISFs is under 600 for each band. This is generally less than is included in the SR-EISFs or the GR-EISFs and considerably less than even the S01 scenarios for the C-EISFs. The SS-EISFs S01 scenarios selected no spatial eigenvector for inclusion in a spatial filter, except bands 6 and 10, which selected one spatial eigenvector respectively and bands 11 and 12, which selected 6 spatial eigenvectors respectively. Figure (5.27) left shows how many spatial eigenvectors are selected for each scenario and for each band.

Although the SS-EISFs account for very little variance in the image, similar to the filter constructed using the random sampling methods, the SS-EISFs do not visually resemble the SR-EISFs or the GR-EISFs. The SS-EISFs do not appear random but instead have an apparent

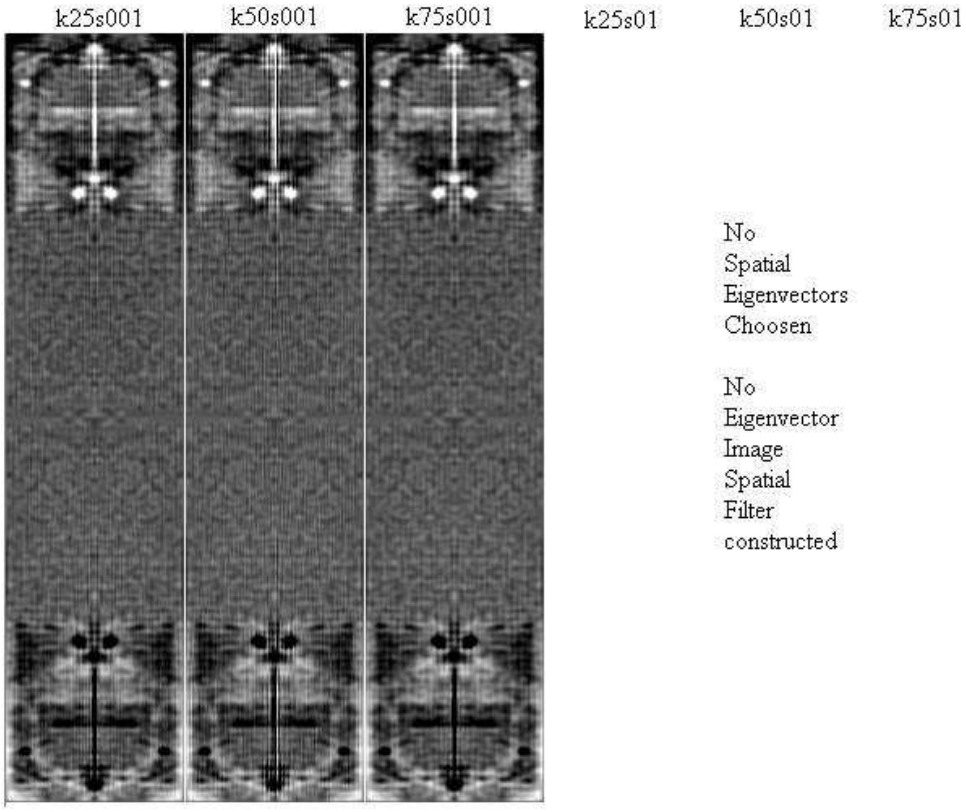


Figure 5.28. Band 1 SS-EISF for each scenario using the systematic sampling technique

systematic pattern. Figure (5.28) shows the SS-EISFs for the first band for each scenario evaluated. Unlike the random sampling methods, the *MC* threshold value has an impact on the construction of the SS-EISFs, although the filters are still very similar. There is some visual similarity of this filter and the original image. At the top and the bottom of the SS-EISFs bright pixels parallel a road present in the original image. There is also some color variations similar to what is seen in the fields in the original image. The center portion of the image also shows systematic structure, which is not very clear in Figure (5.28), although it has little resemblance to the original image. The S001 scenarios for the SS-EISFs seem visually to have a mixture of both.

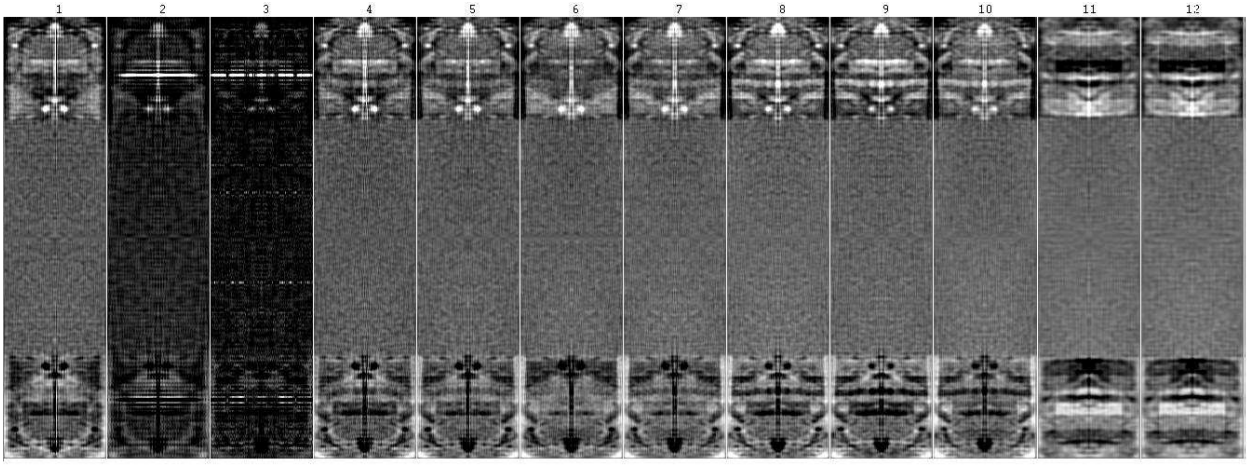


Figure 5.29. Eigenvector image spatial filter bands for the scenario 0.25 candidate eigenvector threshold, 0.001 selection criteria implementing systematic sampling technique

local and global patterns. The S01 scenario for the first band of the SS- EISFs did not identify any spatial eigenvectors for inclusion in the final filter. The k25s001 scenario for the SS –EISFs are shown in Figure (5.29). The SS-EISFs show some spatial resemblance to the image at the top and bottom. There is also systematic pattern in the middle of the image but the contrast in the pixel values make it difficult to see in print. The SS-EISFs are a mirror reflection horizontally and vertically through an axis running through the center of the image. Interestingly, a scanline artifact that is present in the original image is captured in band 2 and 3. There is also a change in pixel brightness that is consistent with the original image apparent across all the bands. Also of note is that the patterns of the SS-EISFs constructed from bands in the visual spectrum are very similar while the patterns of the SS-EISFs constructed from the inferred bands are similar to each other and different from the visual bands SS-EISFs. This is consistent with the multicollinearity diagnostics done for this dataset.

The types of spatial pattern included in the SS-EISFs tend to be global. Table (5.7) shows counts and percent of the total spatial eigenvectors for each type of spatial pattern for the

Table 5.7. Columns 1 -3 give the number of spatial eigenvectors common and not common between the C and S filters. Columns 4-9 show the number of global, regional and local spatial eigenvectors chosen for inclusion in the C and the S filters respectively.

Common and Distinct EVs				**Sample (SS)**					
Bands	*C1SS1*	*C1SS0*	*C0SS1*	*Global*		*Regional*		*Local*	
b1	446	31749	70	80.23%	414	3.29%	17	16.47%	85
b2	274	31921	39	88.50%	277	6.71%	21	4.79%	15
b3	391	33650	58	83.96%	377	4.68%	21	11.36%	51
b4	469	28385	83	79.53%	439	7.25%	40	13.22%	73
b5	450	25461	69	85.74%	445	2.12%	11	12.14%	63
b6	412	24554	62	88.40%	419	3.80%	18	7.81%	37
b7	421	23996	63	89.05%	431	4.34%	21	6.61%	32
b8	400	22261	63	95.90%	444	1.30%	6	2.81%	13
b9	388	21938	59	93.51%	418	3.58%	16	2.91%	13
b10	393	20333	70	96.33%	446	1.08%	5	2.59%	12
b11	270	24978	32	99.01%	299	0.66%	2	0.33%	1
b12	278	22385	20	99.35%	306	0.00%	0	0.65%	2

SS-EISFs. The across band average shows that SS-EISFs include 89% global 3% regional and 6% local spatial eigenvectors. This is especially apparent in the inferred bands which almost exclusively select global spatial eigenvectors. This tendency to choose global patterns is not observed in the other sampling methods or in the complete spectral information. This might explain the systematic nature of the SS-EISFs and is likely to be attributable to the homogeneity assumption of the systematic sampling.

The number of spatial eigenvectors common and distinct to the SS-EISFs and the C-EISFs for all the bands is also given in Table (5.7). The number of spatial eigenvectors only belonging to the C filters (C1SS0) is substantially higher as is the case in other sampling techniques. Interesting though is that on average over all the bands, 87 % of the spatial

eigenvectors selected for the SS-EISFs for this scenario are common to the C-EISFs. Most of the spatial eigenvectors included in the SS-EISFs are also included in the C-EISFs.

In the medium case study image systematic sampling substantially changes the computational intensity of the algorithm. For this LandSat dataset the required CPU time drops substantially from a range of 17 to 50 hours for the C-EISFs to a range of 1 to 3.4 hours as shown in Figure (5.30). Similar to the C-EISFs there is a linear increase in computational time required to construct the SS-EISFs, but the slope of the line is not as steep as for the C-EISFs and therefore is increasing more slowly. This is a consequence of both the sampling and the reformulation of the spatial eigenvector calculation.

The amount of variance accounted for in the image by the SS-EISFs for the medium case study is similar to that of the small case study SS-EISFs, very small. However, the amount of variance accounted for in the image did change enough between scenarios that separate EISFs are constructed, unlike the small case study SS-EISFs. There is a systematic ordering of variance accounted for in the image similar to the small case study image and again the S001 scenarios account for more variance in the image than the S01 scenarios. When compared to the other sampling techniques slightly more variance is accounted for in the image by the SS-EISFs although there is still virtually no spatial structure pressured in the final SS-EISF.

The number of selected spatial eigenvectors for the SS-EISFs S001, shown in Figure (5.31), scenarios is within a similar range to the medium case study image for C-EISF S01 scenarios and slightly higher than the other sampling methods. No spatial eigenvector selected for the SS-EISFs had a coefficient value above 0.001. Similar to the SS-EISFs small case study scenarios the *MC* threshold value has an impact of the medium cases study's SS-EISFs and a

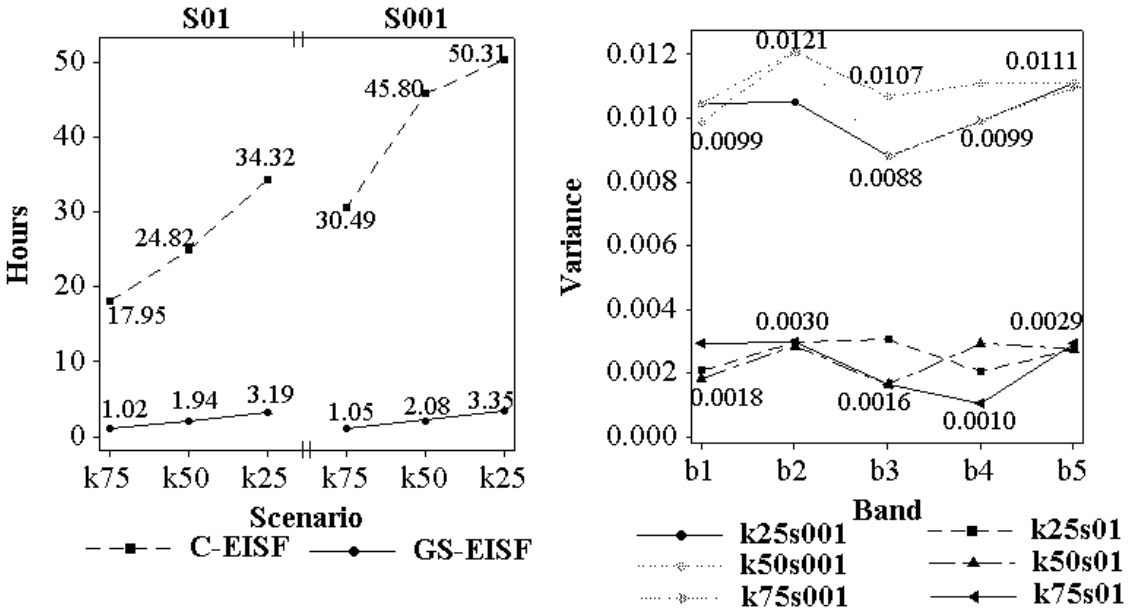


Figure 5.30. CPU time required to construct medium case study image SS-EISFs (left). The amount variance accounted for in the image by the SS-EISFs (right).

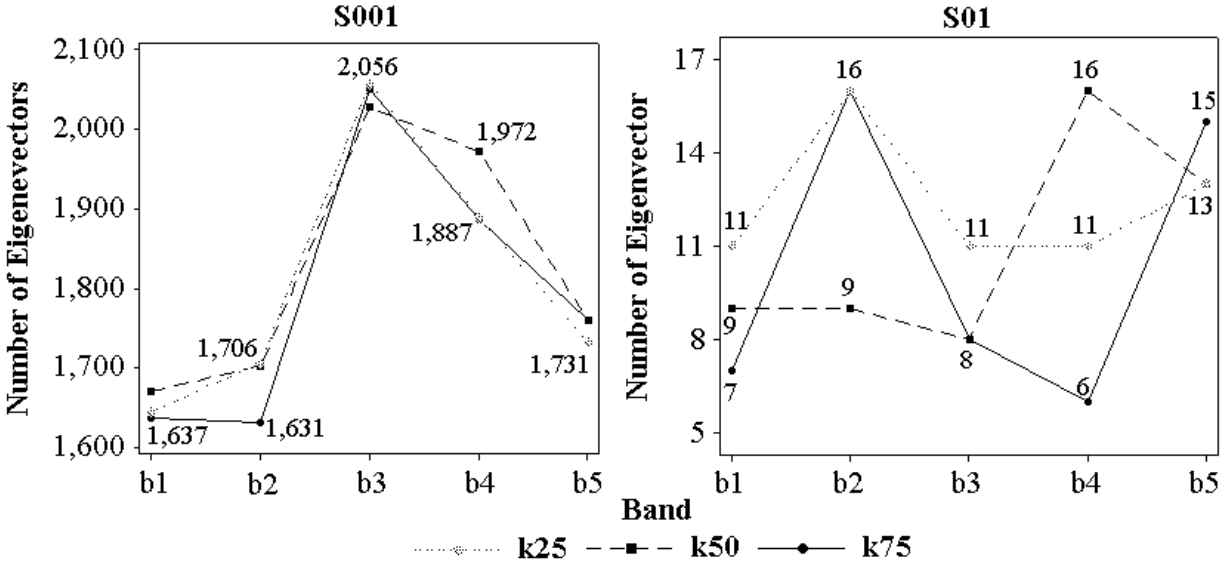


Figure 5.31. The number of spatial eigenvectors selected for inclusion for all 6 SS-EISFs scenarios

spatial filter can be constructed for every scenario, although the SS-EISFs S01 scenarios indentify very few spatial eigenvectors and account for virtually no variance in the image.

Figure (5.32) shows the first band for each SS-EISF scenario. The systematic nature of these images is apparent along with the symmetry along the horizontal and vertical axis through the center of the image. There is little resemblance to the original image spatial structure. The S001 SS-EISFs visually show more local and regional variation while the S01 SS-EISFs appear to have only regional or global variation. These global and regional patterns are not a smoothed version of the S001 SS-EISFs as seem the case for the C-EISFs. The patterns are fundamentally different.

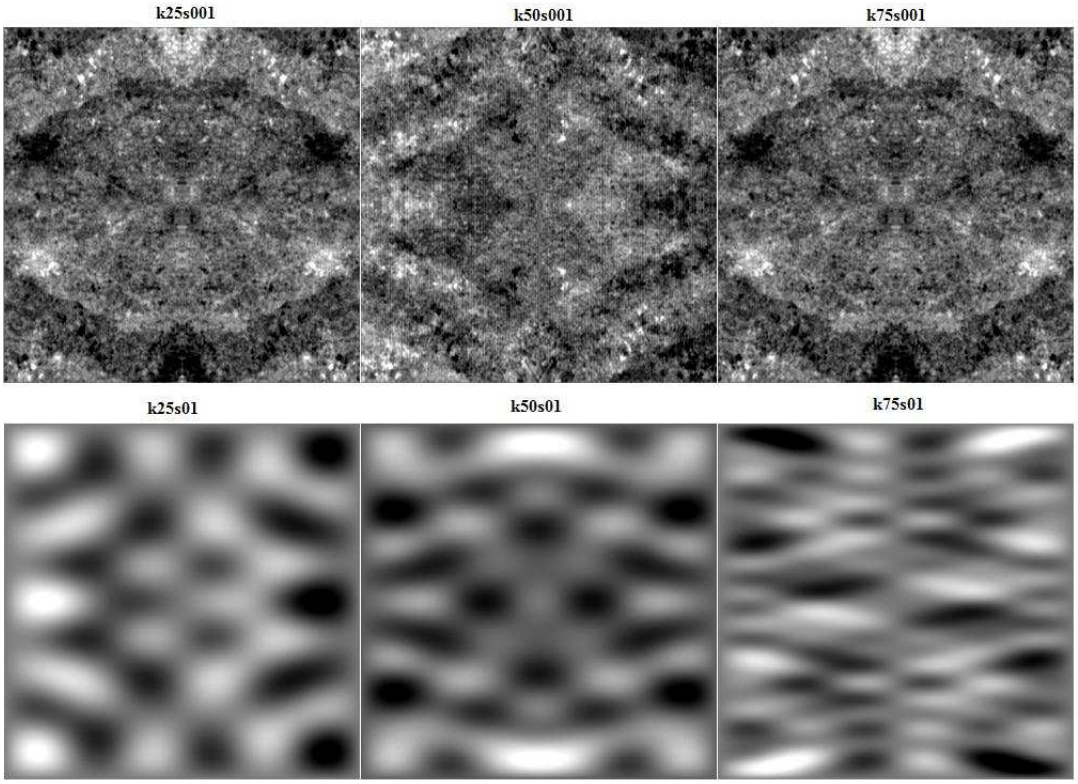


Figure 5.32. SS-EISFs for each scenario for the medium case study image

Table 5.8. Columns 2 -7 show the number of global, regional and local spatial eigenvectors chosen for inclusion in the C and the SS filters respectively. Columns 8-10 give the number of spatial eigenvectors common and distinct to the C and SS filters for the medium case study image.

	Global		**Regional**		**Local**		**C0SS1**	**C1SS0**	**C1SS1**
b1	1637	99.57%	4	0.24%	3	0.18%	2	73967	1642
b2	1701	99.71%	1	0.06%	4	0.23%	3	64433	1703
b3	2051	99.76%	3	0.15%	2	0.10%	3	78377	2053
b4	1887	99.89%	2	0.11%	0	0.00%	0	63773	1889
b5	1729	99.88%	2	0.12%	0	0.00%	0	68151	1731

When the k25s001 scenario, shown in Figure (5.33), is considered more carefully, the systematic nature of all the filters is visually apparent. Similar to what was found in each scenario, all of the SS-EISFs are symmetric about a transect that passes through the center of the filter. Again the similarities between patterns seen in the SS-EISFs reflect somewhat the collinearity found in the data diagnostics, SS-EISF for band 1 and 2 are the most similar.

Finally a breakdown of the type of spatial eigenvector included in the SS-EISFs and the number of spatial eigenvectors common and distinct to the SS-EISFs and the C-EISFs are given in Table (5.8). This shows that 99% of all the spatial eigenvectors selected for inclusion in the SS-EISFs are global. This is contrary to what is expected since in general an essentially even distribution across the 3 types of spatial pattern is expected as is found in the C-EISFs. This is due at least in part to the assumption of homogeneity of surrounding pixels that must be made to extract a systematic sample from the spectral information. More interesting though is that nearly all the selected spatial eigenvectors are common to the C-EISFs and for band 4 and band 5 all selected spatial eigenvectors are also included in the C-EISFs. Although this is interesting it is not very promising since only ~2% of the total spatial eigenvectors included in the C-EISFs are selected for SS-EISFs.

In summary, taking a systematic sample of the image spectral data and implementing the reformulation of the spatial eigenvector construction substantially reduces the computation time and changes the steepness of the linear increase of computation time as the scenarios change. The SS-EISFs for either case study however did not produce filters that account for much variation in the image. The filters constructed showed systematic pattern and selected nearly only global spatial eigenvectors. Aside from the reduced computation time the most promising aspect of this method is that of the spatial eigenvectors it selected nearly all of them were also selected by the C-EISFs.

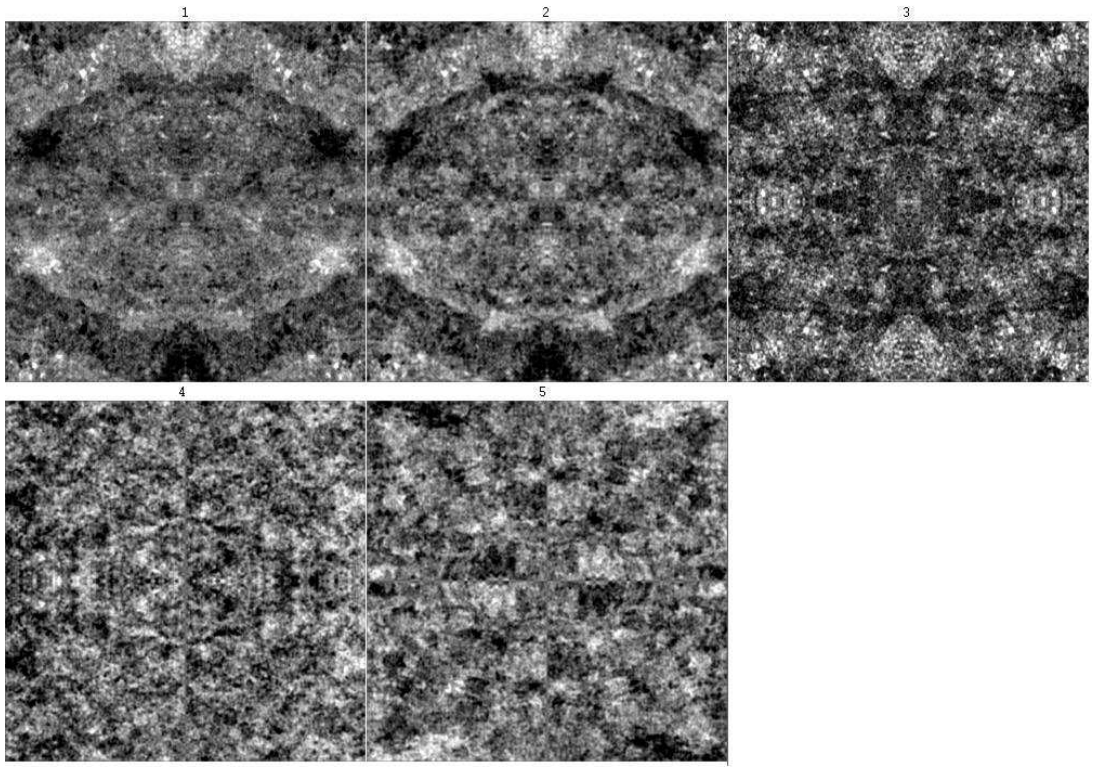


Figure 5.33. Medium case study image scenario k25s001 SS-EISFs.

5.2 DISTBUTED COMPUTING AND THE LARGE CASE STUDY DATASET

Parallelizing seems to be the most promising avenue for implementation of this algorithm, but this proves to be difficult in many ways. The GPU system acquired by UTD is installed and setup by the IT department at UTD and access is granted to the author. Upon investigation of programming the algorithm for use in a GPU environment the expense of acquiring a SAS licence necessitated exploring other programming options. Both R or MatLab, promote compatibility and/or tools which take advantage of the GPU capacity for parallel processing. Upon inspection of the available resources it was found that at present neither really access the GPU capacity of the machine and programming would need to be done in C and CUDA. Although this is an obstacle that may be overcome at present this research will be postponed for a future implementation. This requires that the investigation of the large case study image also be postponed until it might be executed on a parallel system.

5.3 FINAL CONCLUSIONS

This dissertation formulates an efficient algorithm for implementing spatial filtering in an image analysis setting to make possible further investigations of spatial filter for tasks like image classification, change detection, and accuracy assessment. The algorithm presented here incorporates several strategies for a more efficient implementation, which incorporates both efficient programming techniques as well as methodological changes. This is then assessed empirically using a small and a medium sized image and a series of threshold scenarios.

Conventional spatial filtering constructs a spatial weights matrix, pre and post multiplies that matrix by a projection matrix centering it and then extracts a complete set of spatial

eigenvectors from this matrix. This method allows for a great deal of flexibility in the definition of a spatial weights matrix and insures orthogonality and uncorrelatedness of the eigenvectors. This is the way the algorithm should be implemented for any reasonably small dataset (i.e., under 100,000 observations). This is not practical or efficient for the massive number of pixels in an image. For this reason the analytical solution for the eigenvectors of the binary rook's adjancy matrix for a complete regular square tessellation as given in Gasim (1989) is implemented. Since this is the analytical solution for a matrix that is not pre- and post-multiplied by the projection matrix the eigenvectors could be correlated. It is shown here empirically and proven in Griffith (2003) that asymptotically the analytical eigenvectors of **C** are both orthogonal and uncorrelated if the eigenvector is centered and the first eigenvector replaced with $\frac{1}{\sqrt{n}}$.

Much is gained in terms of efficiency by implementing the analytical solution for the spatial eigenvectors. First, the analytical solution allows a single spatial eigenvector to be constructed at a time. This makes sequential computation possible greatly reducing memory requirements. Second, by evaluating the spatial eigenvalues for the complete surface and indentifying a particular level or type of spatial autocorrelation of interest a candidate set of spatial eigenvectors (i.e., a small set in contrast to all possible spatial eigenvectors for that surface) might be constructed directly.

These two steps toward efficiency for constructing eigenvector image spatial filters are also implemented by Fellows (1996) in his master's thesis. He found, much as this dissertation does, that these techniques make the construction of an image spatial filter more tractable for both computational memory and implementation but the computation of even the candidate set of analytical spatial eigenvectors is daunting and determining which spatial eigenvectors should

be used to construct an EISF remained difficult because each spatial eigenvector accounted for an extremely small amount of variance in an image. At the time Fellows (1996) was not able to construct an eigenvector image spatial filter because of the extremely long run time and spatial eigenvector selection problems so he used a spatial lag variable for his spatial principle components analysis.

This dissertation begins to address these issues of the computational inefficiency by reformulating the implementation of the analytical eigenvector solution eliminating a large number of repetitive calculations. This reformulation breaks the computation of the spatial eigenvectors into two steps. First the matrices **R** and **K** are constructed. Next, the Kronecker product of a column from each of these matrices constructs a spatial eigenvector for the image surface. The row and column indices used to construct the conventional analytical eigenvectors respectively index the column in **R** and the column in **K** used in the Kronecker product. This reformulation is shown in the empirical analysis to make a substantial impact on the CPU time required to construct a EISF. Although no C-EISFs are constructed without implementing the reformulation, both random sampling methods are implemented without the reformulation since the approaches are incompatible. The required computation time for a quarter of the image pixels is greater than the time for the complete surface when the reformulation cannot be implemented as is shown by both random sampling methods. Since the reformulation produces exactly the same spatial eigenvectors just constructed in a more algebraically efficient manner the final EISFs are unaffected. This is illustrated by the complete spectral surface case studies.

Although the CPU time for the EISFs becomes more reasonable using the reformulation of the spatial eigenvectors the medium case study image, which is only a subset of a LandSat

image requires over 40 hours of computation time. For that reason Cressie (1993) suggests that sampling large datasets with much data redundancy in order to make more efficient sophisticated statistical analysis is evaluated for constructing an eigenvector image spatial filter. Three sampling procedures are evaluated for each case study, simple random sampling, stratified random sampling, and systematic sampling.

The first sampling method evaluated, simple random sampling, made implementing the spatial eigenvector reformulation impossible. The sample proved less efficient than the reformulation by at least 2 hours for all threshold scenarios and both case studies. Moreover the spatial filter constructed from the randomly sampled spectral data capture none of the image spatial structure. The number of spatial eigenvectors and the amount of variance accounted for in the image is very small. The SR-EISF resembles an image of noise most likely reflecting the random nature of the sample. The spatial eigenvectors included in the spatial filter are just as likely to be in the C-EISF as not. For the small case study the spatial eigenvectors selected were more local while the medium case study more global spatial eigenvectors are selected. In both cases the selected spatial eigenvectors include spatial eigenvectors of each type. Overall using a simple random sample to select which spatial eigenvectors should be included in an eigenvector image spatial filter failed to construct an EISF which capture image spatial structure or gain efficiency in CPU over simply using the complete spectral data and the spatial eigenvector reformulation.

To ensure a more geographically distributed sample a geographically stratified random sample is assessed as well as a systematic sample. The geographically stratified systematic sample has many of the same problems the simple random sample faced. Because of the random

nature of the sample **R** and **K** cannot be constructed and the reformulation cannot be implemented. This loses what gain sampling may have won in computation time. Similar to the simple random sampling approach, the GR-EISFs visually resemble noise and display no spatial structure of the original image. The number of spatial eigenvectors selected and the amount of variance they account for is very small. The selected spatial eigenvectors are distributed fairly evenly between global, regional and local patterns and are slightly more likely to not be included in the C-EISFs than to be included. Overall this sampling approach performs similar to that of the simple random sample and does not capture the spatial structure of the cases study images.

Systematic sampling also insures a sample is spatially well distributed throughout an image. To its advantage the systematic sample allows the **R** and **K** matrices to be constructed and therefore can combine both sampling and the reformulation of the spatial eigenvector calculation. This combination reduces computation time from 2 hours to 3 minutes in the small case study and from 50 hours to 3 hours in the medium case study. Unfortunately the number of spatial eigenvectors selected and the amount of variance accounted for in the image are very small. Interestingly though the selected spatial eigenvectors for both small and medium case study images are almost exclusively global and moreover almost exclusively also chosen for inclusion in the C-EISF. Visual inspection of the SS-EISFs shows a systematic pattern in the final filter.

Although Cressie's (1993) suggestion for sampling could be useful in many statistical settings it does not seem to be effective in constructing eigenvector image spatial filters for the original image. The sampling methods explored here seem to be unable to identify spatial structure in the original image and seem to capture the structure of the sample itself. This makes

sense because the selected eigenvectors will reflect the collective spatial structure of the surface as determined by the sample. If the surface is randomly sampled then it should follow that the eigenvectors selected to model the image information would also produce a random pattern. Effectively the orthogonality of the spatial eigenvectors ensure that the EISF reflects the spatial structure present in the sample, if the sample does not preserve the spatial structure of the original image then the EISF will not be able to reproduce the original image structure.

Large gains in efficiency are made by reformulating the computation of the analytical solution for the eigenvector for the complete spectral information. This makes future research into parallel processing feasible. Some promising aspects of systematic sampling might also be explored in the future. By adjusting the sampling rate or by sequential taking systematic samples (i.e., for a sampling rate of four, as implemented here take sequentially all four possible samples of the surface). Construct a SS-EISF for each sample and sum all the SS-EISFs together to construct a final SS-EISF. This might be interesting but it is likely to miss much of the regional and local structure of the image. The most promising future research should focus on parallel computing. Using the complete spectral information clearly constructs the best EISF.

APPENDIX A

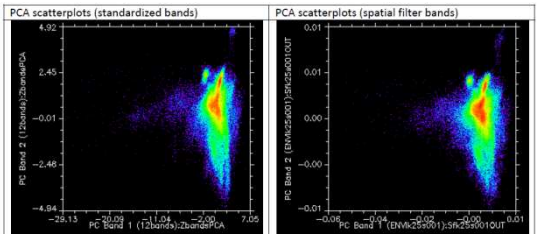
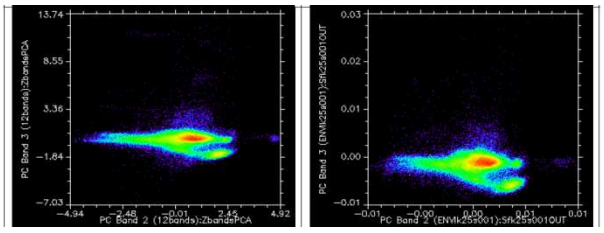
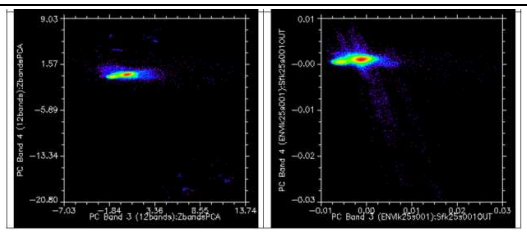
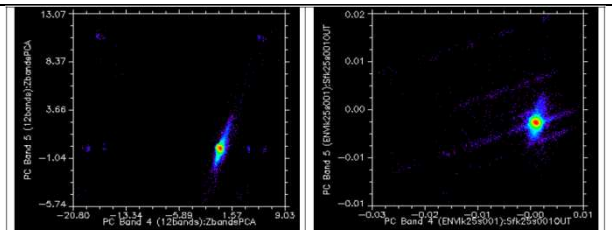
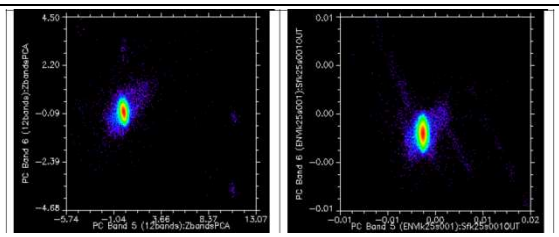
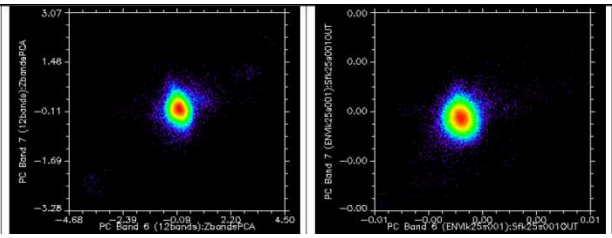
SPATIAL PCA: FELLOWS (1998) REVISITED

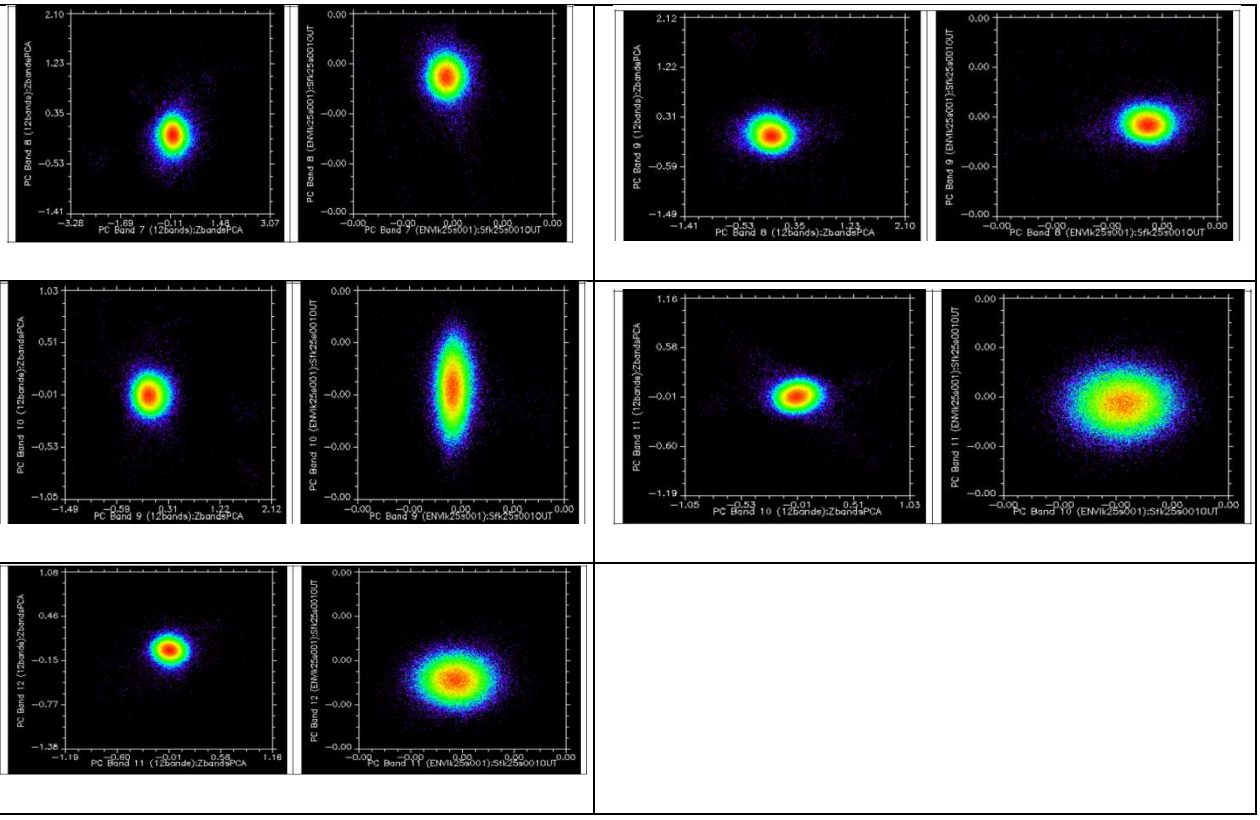
Fellows (1998) proposed the use of spatial filters in a Principle Components Analysis (PCA) as a way of accounting for spatial autocorrelation in the data but found the spatial filters too cumbersome to construct and the eigenvectors too difficult to identify. A study was conducted to evaluate that problem and spatial filters were constructed. The following pilot study uses those spatial filters as Fellows proposed in a PCA and explores the differences between PCA on only spectral information and that on the constructed spatial filters. Using the scenario k25s001 (i.e., selection criteria of 0.001 and a MC threshold of 0.25) a spatial PCA is done and this compared to a PCA of the spectral bands. In this example the spatial filter bands push more of the variance into the first two principle component bands; this is shown in Table (A.1). This is also evident when comparing the principle component scatterplots for the spectral data and principle components of the spatial filters. The scatterplots for the PCA for both the standardized spectral bands and the spatial filters can be seen in Table. These plots, shown in Table (A.2), are very similar except the spatial filter plots are more dispersed making the pixel conglomerations easier to distinguish. This could be useful when selecting endmembers.

Table A.1. Percent variance accounted for by the standard PCA for the standardized image bands and the spatial filter bands

	% Variance PCA (Z-score)	% Variance PCA (Spatial Filter)	Variance Difference
b1	72.09	73.59	1.50
b2	16.83	17.50	0.67
b3	6.16	5.81	-0.36
b4	2.05	1.09	-0.96
b5	1.15	0.75	-0.39
b6	0.56	0.45	-0.11
b7	0.46	0.25	-0.21
b8	0.28	0.21	-0.07
b9	0.20	0.17	-0.03
b10	0.09	0.07	-0.01
b11	0.07	0.06	-0.01
b12	0.07	0.05	-0.01

Table A.2. PCA scatterplots for the standardized image bands (left) and the spatial filter bands (right)

PCA Z-score band	PCA SF band	PCA Z-score band	PCA SF band
			
			
			


PC Band 8 (12bands):ZbandsPCA
PC Band 7 (12bands):ZbandsPCA
PC Band 8 (ENMk25s001):Sfk25s0010UT
PC Band 7 (ENMk25s001):Sfk25s0010UT
PC Band 9 (12bands):ZbandsPCA
PC Band 8 (12bands):ZbandsPCA
PC Band 9 (ENMk25s001):Sfk25s0010UT
PC Band 8 (ENMk25s001):Sfk25s0010UT
PC Band 10 (12bands):ZbandsPCA
PC Band 9 (12bands):ZbandsPCA
PC Band 10 (ENMk25s001):Sfk25s0010UT
PC Band 9 (ENMk25s001):Sfk25s0010UT
PC Band 11 (12bands):ZbandsPCA
PC Band 10 (12bands):ZbandsPCA
PC Band 11 (ENMk25s001):Sfk25s0010UT
PC Band 10 (ENMk25s001):Sfk25s0010UT
PC Band 12 (12bands):ZbandsPCA
PC Band 11 (12bands):ZbandsPCA
PC Band 12 (ENMk25s001):Sfk25s0010UT
PC Band 11 (ENMk25s001):Sfk25s0010UT

APPENDIX B

EXAMPLE IMAGE EIGENVALUE AND EIGENVECTOR CALCULATION

The following over simplified example illustrates the analytical solution for the eigenvalues and eigenvectors of a regular square tessellation.

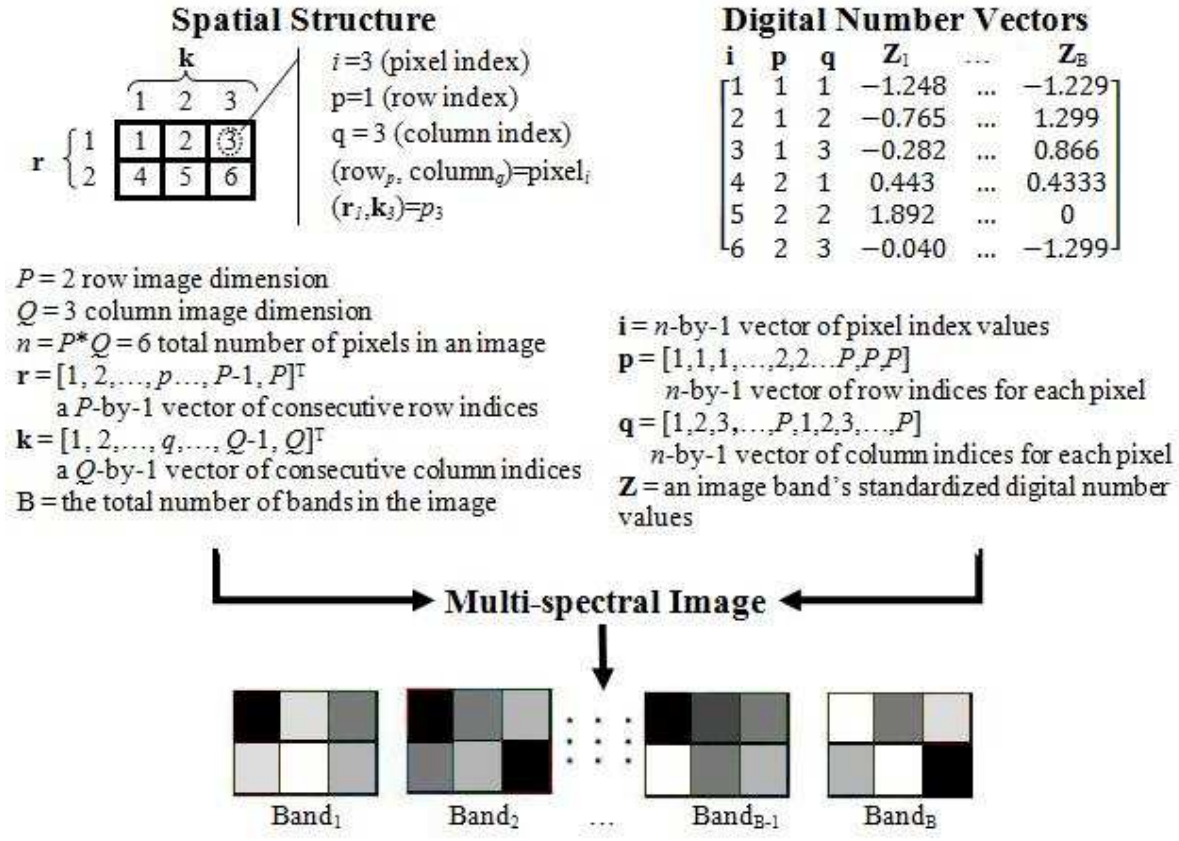


Figure 2.1: Reprint

Eigenvalue for the square tessellation example

From the example surface in Figure (2.1), the eigenvalues and MC values are computed as follows. The image dimensions are $P = 3$ and $Q = 2$, and the vectors indexing the rows and columns are $\mathbf{r} = \begin{bmatrix}1\\2\end{bmatrix}$ and $\mathbf{k} = [1 \quad 2 \quad 3]^T$

The set of all eigenvalues, $\boldsymbol{\lambda}$ may be calculated as follows:

$$\boldsymbol{\lambda} = 2\left[\cos\frac{\pi\begin{bmatrix}1\\2\end{bmatrix}}{2+1} \oplus \cos\frac{\pi\begin{bmatrix}1\\2\\3\end{bmatrix}}{3+1}\right] = 2\begin{bmatrix}\begin{pmatrix}0.4990\\0.4990\\0.4990\end{pmatrix} + \begin{pmatrix}0.7065\\-0.0017\\-0.7089\end{pmatrix}\\ \begin{pmatrix}-0.5020\\-0.5020\\-0.5020\end{pmatrix} + \begin{pmatrix}0.7065\\-0.0017\\0.7080\end{pmatrix}\end{bmatrix} = \begin{bmatrix}2.41\\1.00\\0.41\\-0.41\\-1.00\\-2.41\end{bmatrix}$$

There is the same number of eigenvalues as regions in the surface, six, and each eigenvalue is associated with a particular eigenvector extracted from the geographic weights matrix. The set of eigenvalues can be used to approximate the amount of spatial autocorrelation in the associated eigenvectors as in the following equation:

$$\mathbf{MC} = \frac{3*2}{2[3(2-1)+2(3-1)]}\boldsymbol{\lambda} = \begin{bmatrix}1.03\\0.42\\-0.18\\0.18\\-0.43\\-1.03\end{bmatrix}$$

MC can be adjusted to having a maximum of 1 through standardizing by the maximum *MC* value. Here $MC_{max} = 1.03$ and the adjusted *MC* values (MC_{adj}) are computed by the element-wise division of *MC* by

$$\mathbf{MC}_{adj} = \frac{1}{\mathrm{MC}_{max}}\mathbf{MC} = \begin{bmatrix}1\\0.41\\-0.17\\0.17\\-0.41\\-1\end{bmatrix}$$

This adjusted *MC* value can now be used to establish a threshold value for determining a set of candidate eigenvectors.

Eigenvector for the square tessellation

For the 2-by-3 example of the square tessellation given in Figure (2.1), where $P = 2, Q = 3$, $\mathbf{r} = [1 \quad 2]^T$ and $\mathbf{k} = [1 \quad 2 \quad 3]^T$ the notation given in Griffith (2000) computes an example eigenvector:

$$\mathbf{E}_4 = \mathbf{E}_{2,1} = \frac{2}{\sqrt{(2+1)(3+1)}} * \begin{pmatrix} \sin\left(\frac{2*1*\pi}{2+1}\right) * \sin\left(\frac{1*1*\pi}{3+1}\right) \\ \sin\left(\frac{2*1*\pi}{2+1}\right) * \sin\left(\frac{1*2*\pi}{3+1}\right) \\ \sin\left(\frac{2*1*\pi}{2+1}\right) * \sin\left(\frac{1*3*\pi}{3+1}\right) \\ \sin\left(\frac{2*2*\pi}{2+1}\right) * \sin\left(\frac{1*1*\pi}{3+1}\right) \\ \sin\left(\frac{2*2*\pi}{2+1}\right) * \sin\left(\frac{1*2*\pi}{3+1}\right) \\ \sin\left(\frac{2*2*\pi}{2+1}\right) * \sin\left(\frac{1*3*\pi}{3+1}\right) \end{pmatrix} = \begin{bmatrix} -0.8080 \\ 0.7411 \\ 0.1283 \\ -0.6725 \\ 0.6168 \\ 0.1068 \end{bmatrix}.$$

The reformulated method first computes matrix **R** and **K**:

$$\mathbf{E} = \frac{2}{\sqrt{(2+1)(3+1)}} \left[\sin\left(\frac{\pi}{2+1}\begin{bmatrix}1\\2\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix}\right) \otimes \sin\left(\frac{\pi}{3+1}\begin{bmatrix}1\\2\\3\end{bmatrix}\begin{bmatrix}1 & 2 & 3\end{bmatrix}\right) \right]$$

$$= \frac{2}{\sqrt{12}} \sin\left(\frac{\pi}{3}\begin{bmatrix}1 & 2\\2 & 4\end{bmatrix}\right) \otimes \frac{2}{\sqrt{12}} \sin\left(\frac{\pi}{4}\begin{bmatrix}1 & 2 & 3\\2 & 4 & 6\\3 & 6 & 9\end{bmatrix}\right)$$

$$= \underbrace{\begin{bmatrix} 0.5403 & -0.9093 \\ -0.9093 & 0.5403 \end{bmatrix}}_{\mathbf{R}_{P-by-P}} \otimes \underbrace{\begin{bmatrix} 0.8887 & -0.815 & -0.1411 \\ -0.815 & 0.9445 & -0.2794 \\ -0.1441 & -0.2794 & -0.4121 \end{bmatrix}}_{\mathbf{K}_{Q-by-Q}}.$$

The fourth eigenvector may then be computed $\mathbf{E}_4 = [\mathbf{r}_2 \otimes \mathbf{k}_1]$:

$$\mathbf{E}_4 = \begin{bmatrix} -0.9093 \\ 0.5403 \end{bmatrix} \otimes \begin{bmatrix} 0.8887 \\ -0.8150 \\ -0.1411 \end{bmatrix}$$

$$= \begin{bmatrix} -0.9093 \begin{pmatrix} 0.8887 \\ -0.8150 \\ -0.1411 \end{pmatrix} \\ 0.5403 \begin{pmatrix} 0.8887 \\ -0.8150 \\ -0.1411 \end{pmatrix} \end{bmatrix} = \begin{bmatrix} -0.8080 \\ 0.7411 \\ 0.1283 \\ -0.6725 \\ 0.6168 \\ 0.1068 \end{bmatrix}$$

Since *p* is even, it is unnecessary to remove the mean from this eigenvector because the mean is zero. If the mean were not zero then it would next be centered using the projection matrix (**I**-**11**^{T}/*n*). The 6 values in this example eigenvector can be linked to the original square tessellation to produce a distinct map pattern, which is portrayed below. $\mathbf{E}_4$ is the first map pattern on the second row; the values from the eigenvectors are mapped top left to bottom right along the rows.

$$\mathbf{E}_4 = \begin{array}{|c|c|c|} \hline -0.8080 & 0.1283 & 0.6168 \\ \hline 0.7411 & -0.6725 & 0.1068 \\ \hline \end{array}$$

In a small example, the patterns can be harder to identify visually; this is due to the large amount of edge effects in the small number of regions, but the top left hand corner of Figure (C.1) has the highest positive spatial autocorrelation and the bottom right the highest negative spatial autocorrelation, [see Figures (2.4-7) for the 100 eigenvectors from a 10-by-10 image the range of *MC* values in the eigenvectors is more visibly apparent].

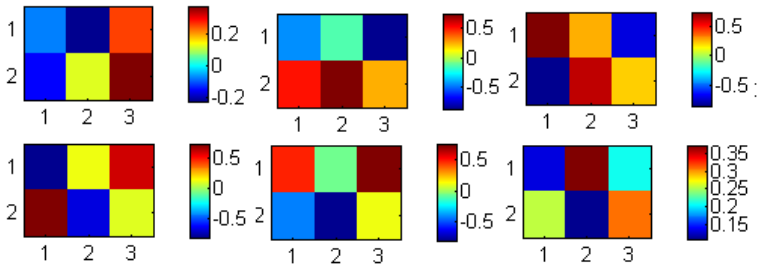


Figure B.1. Six map patterns corresponding to the six eigenvectors for the 6-by-6 sample image in Figure (2.1)

APPENDIX C

Substituting the analytical solution for the eigenvectors in Equation (13) gives the following formula:

$$\frac{\sum_{p=1}^{P}\sin\left(\frac{kp_j\pi}{P+1}\right)*\sum_{q=1}^{Q}\sin\left(\frac{lq_j\pi}{Q+1}\right)*\frac{\sum_{p=1}^{P}\sin\left(\frac{kp_k\pi}{P+1}\right)*\sum_{q=1}^{Q}\sin\left(\frac{lq_k\pi}{Q+1}\right)}{n}}{\sqrt{1-\sum_{p=1}^{P}sin\left(\frac{kp_j\pi}{P+1}\right)*\sum_{q=1}^{Q}sin\left(\frac{lq_j\pi}{Q+1}\right)*\frac{\sum_{p=1}^{P}sin\left(\frac{kp_j\pi}{P+1}\right)*\sum_{q=1}^{Q}sin\left(\frac{lq_j\pi}{Q+1}\right)}{n}}\sqrt{1-\sum_{p=1}^{P}sin\left(\frac{kp_k\pi}{P+1}\right)*\sum_{q=1}^{Q}sin\left(\frac{lq_k\pi}{Q+1}\right)*\frac{\sum_{p=1}^{P}sin\left(\frac{kp_k\pi}{P+1}\right)*\sum_{q=1}^{Q}sin\left(\frac{lq_k\pi}{Q+1}\right)}{n}}} \quad (25)$$

BIBLIOGRAPHY

Agresti, Alan. 2002. *Categorical data analysis*. Probability and statistics. Hoboken, New Jersey: Wiley

Anselin, L. 1995. Local indicators of spatial association - LISA. *Geographical Analysis* 27, (2): 93-115

Anselin, Luc. 1988. *Spatial econometrics: Methods and models*. Dordrecht; Boston: Kluwer Academic Publishers.

Anton, Howard, and Chris Rorres. 1994. *Elementary algebra applications version.* John Wiley and Sons.

Anys, Hassan, and Dong-Chen He. 1995. Evaluation of textural and multipolarization radar features for crop classification. *IEEE Transactions on Geoscience and Remote Sensing* 33, (5): 1170-81.

Atkinson, Peter M., and P. Lewis. 2000. Geostatistical classification for remote sensing: An introduction. *Computers & Geosciences* 26, (4) (5): 361-71.

Berry, Brian, and Alan M. Baker. 1968. Geographic sampling. In *Spatial analysis: A reader in statistical geography.* eds. Brian Berry, Duane Francis Marble, 91-100 Prentice-Hall.

Boots, Barry N. 2003. Developing local measures of spatial association for categorical data. *Journal of Geographical Systems* 5, (2): 139-160.

———. 2002. Local measures of spatial association. *Ecoscience* 9, (2): 168-76.

———. 2001. Using local statistics for boundary characterization. *GeoJournal* 53, (4): 339-45.

Boots, Barry N., and M. Tiefelsdorf. 2000. Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems* 2, (4): 319-48.

Boucher, Alexandre, and Phaedon C. Kyriakidis. 2006. Super-resolution land cover mapping with indicator geostatistics. *Remote Sensing of Environment* 104, (3): 264-82.

Burt, James E., and Gerald M. Barber. 1996. *Elementary statistics for geographers.* New York: Guilford Press.

Campbell, James B. 1981. Spatial correlation effects upon accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing* 47, (3): 355-63.

Carr, James R., and F. P. De Miranda. 1998. The semivariogram in comparison to the co-occurrence matrix for classification of image texture. *IEEE Transactions on Geoscience and Remote Sensing* 36, (6): 1945-52.

Carr, James R., and Donald E. Myers. 1984. Application of the theory of regionalized variables to the spatial analysis of Landsat data. Paper presented at Proceedings - PECORA 9: Spatial Information Technologies for Remote Sensing Today and Tomorrow. Sioux Falls, ND, USA.

Chun, Yongwan. 2008. Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of Geographic Systems* 10, 317-44.

Cliff, Andrew D., and J. Keith Ord. 1981. *Spatial processes: Models and applications*. London: Pion.

———. 1973. *Spatial autocorrelation*. London: Pion.

Congalton, Russell G. 1988. Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed images. *Photogrammetric Engineering and Remote Sensing* 54, (5): 587-92.

Congalton, Russell G., and Kass Green. 2009. *Assessing the accuracy of remotely sensed data principles and practices* CRC Press.

Craig, R. G. 1984. Spatial Structure of Terrain: A Process Signal in Satellite Digital Images.

Craig, Richard G. 1981. Precision in the Evaluation of LANDSAT Autocorrelation: The Terrain Effect.

———. 1979. Autocorrelation in LANDSAT data. *Proceedings of the Thirteenth International Symposium on Remote Sensing of Environment*: 1517-24.

Craig, Richard G., and Mark L. Labovitz. 1980. SOURCES OF VARIATION IN LANDSAT AUTOCORRELATION. *Proceedings of the International Symposium on Remote Sensing of Environment* 3, (23 April 1980 through 30 April 1980): 1755-67.

Cressie, Noel. 1993. *Statistics for spatial data*. New York: Wiley.

Cressie, Noel, and J. Kornak. 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science* 18, (4): 436-56.

Cressie, Noel, Anthony R. Olsen, and Dianne Cook. 1996. Massive data sets: Problems and possibilities, with applications to environmental monitoring. Paper presented at Massive Data Sets Proceeding of a Workshop.

Curran, Paul J. 1988. The semivariogram in remote sensing: An introduction. *Remote Sensing of Environment* 24, (3): 493-507.

Curran, Paul J., and Peter M. Atkinson. 1998. Geostatistics and remote sensing. *Progress in Physical Geography* 22, (1) (Mar 1998): 61-78.

Dacey, Michael F. 1968. Spatial analysis a reader in statistical geography. In Spatial Analysis: A reader in statistical geography, 479-490 Prentice-Hall.

De Cola, Lee. 1989. Fractal analysis of a classified Landsat scene. *Photogrammetric Engineering and Remote Sensing* 55, (5 pt 1): 601-10.

Densham, P. J., and Mark P. Armstrong. 1998. Spatial analysis. In *Parallel processing algorithms for GIS.* London, UK: Taylor Francis.

Derksen, C., M. A. Wulder, E. F. LeDrew, and B. Goodison. 1998. Associations between spatially autocorrelated patterns of SSM/I-derived prairie snow cover and atmospheric circulation. *Hydrological Processes* 12, (15): 2307-16.

Dijkstra, E. 1959. A note on two problems in connection with graphs. *Numerische Mathematik* 1, 101-18.

Diniz-Filho, Jose Alexandre Felizola, and Luis Mauricio Bini. 2005. Modeling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecology and Biogeography* 14, 177-185.

Dray, Stéphane, Pierre Legendre, and Pedro R. Peres-Neto. 2006. Spatial modeling: A comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). *Ecological Modeling* 196, (3-4): 483-93.

Dundar, M. Murat, and David A. Landgrebe. 2004. Toward an optimal supervised classifier for analysis of hyperspectral data. *IEEE Transactions onGeoscience and Remote Sensing*, 42, (1): 271-277.

Durbin, J., and G. S. Watson. 1950. Testing for serial correlation in least squares regression: I. *Biometrika* 37, (3/4): 409-28.

Emerson, C. W., N. S. -N Lam, and Dale A. Quattrochi. 2005. A comparison of local variance, fractal dimension, and Moran's I as aids to multispectral image classification. *International Journal of Remote Sensing* 26, (8): 1575-88.

Fellows, Peter L. 1998. Using spectral and locational information to classify a Landsat TM subset of the 1995 blowdown in the Adirondacks. Syracuse University.

Gasim, Ali Abul. 1988. First order autoregressive models: A method for obtaining eigenvalues for weighting matrices. *Journal of Statistical Planning and Inference* 18, 391-8.

Geary, R. C. 1954. The contiguity ratio and statistical mapping. *The Incorporated Statistician* 5, (3) (Nov.): 115-45.

Getis, Arthur, and Daniel A. Griffith. 2002. Comparative spatial filtering in regression analysis. *Geographical Analysis* 34, (2): 130-40.

Getis, Arthur, and J. Keith Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189-206.

Global Land Cover Facility Earth Science Data Interface: http://glcfapp.umiacs.umd.edu Last accessed 07/17/2010

Gong, P., D. J. Marceau, and P. J. Howarth. 1992. A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data. *Remote Sensing of Environment* 40, (2): 137-51.

Goodchild, Michael F. 1989. Modeling error in objects and fields. In Accuracy of Spatial Databases, 107-113 Taylor & Francis.

Goovaerts, P., G. M. Jacquez, and A. Marcus. 2005. Geostatistical and local cluster analysis of high resolution hyperspectral imagery for detection of anomalies. *Remote Sensing of Environment* 95, (3): 351-67.

Griffith, Daniel A. 2006. Assessing spatial dependence in count data: Winsorized and spatial filter specification alternatives to the auto-Poisson model. *Geographical Analysis* 38, (2): 160-79.

———. 2005. Effective geographic sample size in the presence of spatial autocorrelation, *Annals*, Association of American Geographers, 95: 740-760.

———. 2005. A comparison of six analytical disease mapping techniques as applied to west Nile virus in the coterminous United States. *International Journal of Health Geographics* 4.

———. 2004. Extreme eigenfunctions of adjacency matrices for planar graphs employed in spatial analyses. *Linear Algebra and its Applications* 388, 201-19.

———. 2004. A spatial filtering specification for the autologistic model. *Environment and Planning A* 36, (10): 1791-811.

———. 2003. *Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization*. Berlin; New York: Springer.

———. 2002. A spatial filtering specification for the auto-Poisson model. *Statistics and Probability Letters* 58, (3): 245-51.

———. 2000. Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra and its Applications* 321, (1-3): 95-112.

———. 1996. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying georeferenced data. *The Canadian Geographer* 40, 351-67.

———. 1988. *Advanced spatial statistics*. Dordrecht: Kluwer Academic Press.

———. 1985. An evaluation of correction techniques for boundary effects in spatial statistical analysis: Contemporary methods. *Geographical Analysis,* 17, 81-8.

———. 1983. The boundary value problem in spatial statistical analysis. *J. of Regional Science* , 23: 377-387.

———. 1982. Geometry and spatial interaction. *Annals of the Association of American Geographers* 72, (3) (Sep.): 332-46.

Griffith, Daniel A., and C. Amrhein. 1983. An evaluation of correction techniques for boundary effects in spatial statistical analysis: Traditional methods. *Geographical Analysis,* 15, 352-60.

Griffith, Daniel A., and Carl G. Amrhein. 1997. *Multivariate statistical analysis for geographers*. Upper Saddle River, N.J.: Prentice Hall.

Griffith, Daniel A., and Peter L. Fellows. 1999. Pixels and eigenvectors: Classification of LANDSAT TM imagery using spectral and locational information. In *Spatial accuracy assessment: Land information uncertainty in natural resources.* 455. Chelsea, Mich.: Ann Arbor Press.

Griffith, Daniel A., and Pedro R. Peres-Neto. 2006. Spatial modeling in ecology: The flexibility of eigenfunction spatial analyses. *Ecology* 87, (10): 2603-13.

Haining, Robert P. 2003. *Spatial data analysis: Theory and practice*. Cambridge, UK; New York: Cambridge University Press.

Haralick, R. M. May 1979. Statistical and structural approaches to texture. *Proceedings of the IEEE* 67, (5): 786-804.

Haralick, R. M., K. Shanmugam, and I. Dinstein. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics* 3, (6): 610-21.

Ingram, D. S., and A. L. Actkinson. 1973. *The applicability and effectiveness of cluster analysis.* NASA.

Iyer, P. V. Krishna. 1948. Random association of points on a lattice. *Nature* 162, 333.

Jacob, Benjamin G., Daniel A. Griffith, and Robert J. Novak. 2008. Decomposing malaria mosquito aquatic habitat data into spatial autocorrelation eigenvectors in a SAS/GIS® module. *Transactions in GIS* 12, (3): 341-64.

Jupp, David L. B., Alan H. Strahler, and Curtis E. Woodcock. 1989. Autocorrelation and regularization in digital images - II: Simple image models. *IEEE Transactions on Geoscience and Remote Sensing* 27, (3): 247-58.

———. 1988. Autocorrelation and regularization in digital images-I: Basic Theory. *IEEE Transactions on Geoscience and Remote Sensing* 26, (4): 463-73.

Karakahya, Hakan, Bingul Yazgan, and Okan K. Ersoy. 2003. Artificial neural networks and neural information processing. In *Lecture notes in computer science.* Springer Verlag.

Kosfeld, R., and C. Dreger. 2006. Thresholds for employment and unemployment: A spatial analysis of German regional labour markets, 1992-2000. *Papers in Regional Science* 85, (4): 523-42.

Kyriakidis, Phaedon C., A. M. Shortridge, and Michael F. Goodchild. 1999. Geostatistics for conflation and accuracy assessment of digital elevation models. *International Journal of Geographical Information Science* 13, (7): 677-707.

Lam, N. S. -N, and Lee De Cola. 1993. Fractals in geography. *Fractals in Geography*.

Landgrebe, David A. 2003. *Signal theory methods in multispectral remote sensing* John Wiley and Sons.

Laub, Alan J. 1995. Matrix analysis for scientists and engineers. SIAM.

LeDrew, E. F., H. Holden, M. A. Wulder, C. Derksen, and C. Newman. 2004. A spatial statistical operator applied to multidate satellite imagery for identification of coral reef stress. *Remote Sensing of Environment* 91, (3-4): 271-9.

Luo, Yuancheng, and Ramani Duraiswami. 2008. Canny edge detection on NVIDIA CUDA.

Maillard, P. 2003. Comparing texture analysis methods through classification. *Photogrammetric Engineering and Remote Sensing* 69, (4): 357-67.

Matheron, Georges. 1963. Principles of geostatistics. *Economic Geology* 58, (8) (December 1): 1246-66.

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, (1): 17.

———. 1946. Random associations on lattice. *Nature* 158, 521.

Muller, Jean-Michel. 2006. *Elementary functions algorithms and implementation* BirkHaeuser.

Myint, S. W. 2003. Fractal approaches in texture analysis and classification of remotely sensed data: Comparisons with spatial autocorrelation techniques and simple descriptive statistics. *International Journal of Remote Sensing* 24, (9): 1925-47.

Myint, S. W., E. A. Wentz, and S. J. Purkis. 2007. Employing spatial metrics in urban land-use/land-cover mapping: Comparing the Getis and Geary indices. *Photogrammetric Engineering and Remote Sensing* 73, (12): 1403-15.

National Research Council (U.S.). Panel on Spatial Statistics and Image Processing. 1991. *Spatial statistics and digital image analysis*. Washington, D.C.: National Academy Press.

Ord, J. Keith. 1975. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70, (349): 120-6.

Ord, J. K., and Arthur Getis. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27, (4): 286-306.

Ouma, Y. O., T. G. Ngigi, and R. Tateishi. 2006. On the optimization and selection of wavelet texture for feature extraction from high-resolution satellite imagery with application towards urban-tree delineation. *International Journal of Remote Sensing* 27, (1): 73-104.

Patuelli, R., Daniel A. Griffith, Michael Tiefelsdorf, and P. Nijkamp. 2006. The use of spatial filtering techniques: The spatial and space-time structure of German unemployment data. *Discussion Paper # 06-049/3*, Tinbergen Institute, Amsterdam, The Netherlands.

Pugh, Scott A., and Russell G. Congalton. 2002. Applying spatial autocorrelation analysis to evaluate error in New England forest-cover-type maps derived from Landsat thematic mapper data. *Photogrammetric Engineering and Remote Sensing* 67, (5): 13-620.

Randen, T., and J. H. Husøy. 1999. Filtering for texture classification: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, (4): 291-310.

Read, J. M., and N. S. -N Lam. 2002. Spatial methods for characterizing land cover and detecting land-cover changes for the tropics. *International Journal of Remote Sensing* 23, (12): 2457-74.

Ripley, Brian D. 1981. *Spatial statistics*. John Wiley & Sons.

Rodgers, Josephe Lee, W. Alan Nicewander, and Larry Toothaker. 1984. Linearly independent, orthogonal, and uncorrelated variables. *The American Statistician* 38, (2): 133,134.

Spiegel, M. 1968. *Mathematical handbook of formulas and tables*. Schaum's outline series. McGraw-Hill.

Starks, S. A., J. P. de Figueiredo, and D. L. Van Rooy. 1977. An algorithm for optimal single linear feature extraction from several Gaussian pattern classes. *International Journal of Parallel Programming* 6, (1): 41-54.

Stehman, Stephen V., and W. Scott Overton. 1996. Spatial Sampling. In Practical handbook of spatial statistics, 31-63 CRC Press.

Stein, Alfred, Freek van der Meer, and Ben Gorte. 1999. *Spatial statistics for remote sensing*. Remote sensing and digital image processing; v. 1. Dordrecht ; Boston: Kluwer Academic Publishers.

St-Onge, B. A., and F. Cavayas. 1997. Automated forest structure mapping from high resolution imagery based on directional semivariogram estimates. *Remote Sensing of Environment* 61, (1): 82-95.

Tadjudin, Saldju, and David A. Landgrebe. 2000. Robust parameter estimation for mixture model. *IEEE Transactions Geoscience and Remote Sensing*, 38, (1): 439-45.

Tiefelsdorf, M., and Barry N. Boots. 1997. A note on the extremities of local Moran's *I* and their impact on global Moran's *I*. *Geographical Analysis* 29, (3): 248-57.

Tiefelsdorf, Michael. 2007. Controlling for migration effects in ecological disease mapping of prostate cancer. *Stochastic Environmental Research and Risk Assessment (SERRA)* 21, 615-624.

———. 2000. *Modeling spatial processes: The identification and analysis of spatial relationships in regression residuals by means of Moran's I.* Lecture notes in earth sciences. Springer Verlag.

Tiefelsdorf, Michael, and Barry N. Boots. 1995. The exact distribution of Moran's I. *Environment and Planning A* 27, (6): 985 - 999.

Tiefelsdorf, Michael, Barry N. Boots, and Daniel A. Griffith. 1999. A variance-stabilizing coding scheme for spatial link matrices. *Environment & Planning A* 31, (1): p165.

Tiefelsdorf, Michael, and Daniel A. Griffith. 2007. Semiparametric filtering of spatial autocorrelation: The eigenvector approach. *Environment and Planning A* 39, (5): 1193-221.

Tuceryan, Mihran, and Anlin Jain. 1993. Texture analysis. In *Handbook of pattern recognition and computer vision.*, eds. C. H. Chen, L. F. Pau and P. S. P. Wang, 235-235-276 World Scientific Publishing Company.

Turton, I. 2000. Parallel processing in geography. In *GeoComputation.*, eds. S. Openshaw, T. Harris and R. AbrahartGordon and Breach.

Warner, T. A., and M. C. Shank. 1997. Spatial autocorrelation analysis of hyperspectral imagery for feature selection. *Remote Sensing of Environment* 60, (1): 58-70.

Woodcock, Curtis E., and Alan H. Strahler. 1987. The factor of scale in remote sensing. *Remote Sensing of Environment* 21, (3): 311-32.

Wulder, M. A., and Barry N. Boots. 1998. Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the Getis statistic. *International Journal of Remote Sensing* 19, (11): 2223-31.

Xia, Zong-Gou, and Keith Clarke. 1997. Approaches to scaling of geo-spatial data. In *Scale in remote sensing and GIS.*, eds. Dale A. Quattrochi, Michael F. Goodchild, 309 CRC Press.

VITA

Melissa Joy (Tolene) Rura was born in Chicago Heights, Illinois on September 25th, 1978 to Jerome and Susan Tolene. She is the third child of a family of seven. Her childhood followed the maintenance of the track, which ran the legendary City of New Orleans train. She received a Bachelor of Science in mathematics and geography from Murray State University in May 2000, where she was the Max G. Carmen scholarship winner and a member of Phi Mu Alpha honors math fraternity. After graduation, as a recipient of the Congress / Bundestag exchange for young professionals fellowship, Melissa studied at Ludwig Maximilian Universitaet and worked at the Bavarian geological survey in Munich. After a 6 month return to the USA to work for the Mid American Remote Sensing Center in Murray, Kentucky as a research assistant, Melissa moved to Russia where she married her husband and directed an after school program for Russian children. In 2003, she returned again to the US to have her first child and to begin a master's degree in Civil Engineering with a specialty in Geomatics from Purdue University, which she completed in December 2004. After the birth of her second child Melissa returned to academics pursuing a Ph.D. in Geographical Information Science from the University of Texas at Dallas, for which this dissertation is written. Melissa is the recipient of the Science, Mathematics And Research for Transformation (SMART) scholarship and a member of the AAG, ASPRS, and SPIE.