STATISTICAL ANALYSIS TO DETERMINE A SPATIAL RESOLUTION TO IMPROVE IMAGE CLASSIFICATION

By

Nicole M. Rodríguez-Carrión

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

ELECTRICAL ENGINEERING
UNIVERSITY OF PUERTO RICO
MAYAGUEZ CAMPUS
2015

Approved by:

_____	_____
Shawn Hunt, Ph. D. President, Graduate Committee	Date
_____	_____
Luis O. Jiménez, Ph. D. Member, Graduate Committee	Date
_____	_____
Emmanuel Arzuaga, Ph. D. Member, Graduate Committee	Date
_____	_____
Raúl Torres, Ph. D. Chairperson of the Department	Date
_____	_____
Aidsa Santiago, Ph. D. Graduate Studies Representative	Date

Abstract

This study uses hypothesis testing to determine the optimum pixel size to classify hyperspectral images. Pixel size is defined here as the size of the ground area captured in a pixel. Historically, more resolution or smaller pixel sizes, are considered better, but having smaller pixels can cause difficulties in the image classification. If the pixel size is too small, then the variation in pixels belonging to the same class could be vast. By assuming pixels are identically distributed random variables led to a derivation of a hypothesis test that uses the pixels covariance and variance. This new proposed hypothesis method was compared with results from the parametric hypothesis test F-test, and the non-parametric Ansari-Bradley hypothesis test. Promising similar results for synthetic and real hyperspectral images were obtained, validating the usability of the new proposed hypothesis method within the scope of this study.

Resumen

En este estudio se utilizan pruebas de hipótesis para determinar el tamaño de pixel óptimo para clasificar imágenes hiperespectrales. Un tamaño de pixel es definido en este estudio como el tamaño del área capturada en un pixel. Históricamente se conoce que es mejor tener mayor resolución o menor tamaño de pixel, pero el tener pixeles pequeños ocasiona dificultades en la clasificación de imágenes. Si el pixel es muy pequeño, la variación de pixeles que pertenecen a la misma clase podría ser bien grande. Al asumir que los pixeles en la imagen son variables aleatorias idénticamente distribuidas, pero que no son independientes, se pudo derivar una prueba de hipótesis que utiliza la covarianza y varianza entre pixeles. Este nuevo método fue probado al compararlo con la prueba de hipótesis paramétrica llamada "F-test" y con la prueba de hipótesis no paramétrica llamada "Ansari-Bradley". Resultados prometedores y similares se obtuvieron al probar los resultados para imágenes hiperespectrales sintéticas y reales, validando el uso de la nueva prueba de hipótesis bajo las condiciones desarrolladas en este estudio.

Acknowledgments

I would like to thank my academic advisor, Dr. Shawn Hunt, for his encouragement, support, and guidance from the beginning of the pursuit of my graduate degree. He gave me an excellent education since my bachelor's degree and built in me the necessity to study further. His guidance, passion for this area of study, and constant encouragement made possible for this work to be finished successfully. My most sincere gratitude and acknowledgment to the student Humberto Diaz for guiding me through the programming world, for granting me the knowledge and for believing I would pass the same to others.

Special thanks to Eduardo Ortiz for helping me understand that the tools needed to grow within my career were more than academicals, for his support to find my north when I got lost, and for his letters of recommendation that helped me reach my goals. Also, I would like to thank Sandra Montalvo (Sandy) for being the support and light graduate students need at difficult times. Special thanks to all the members of the Graduate Committee for their support and review.

Also, I will like to thank my partner for always being supportive and persuasive to give me the motivation to finish my thesis, and for being strong enough to endure difficult times. Moreover, finally, I would like to thank God for giving me a supporting and loving family, and the strength to fight through every obstacle that may come my way.

Table of Contents

List of Tables

List of Figures

1 INTRODUCTION

1.1 PROBLEM STATEMENT

One of the questions most asked by researchers using remotely sensed data is how large an area should be covered to analyze a natural phenomenon, or at what resolution should the study be conducted. Multilevel, multiscale, and multitemporal data is often available, and a major concern is determining the most appropriate parameters affecting the study. This thesis focuses on finding an optimal selection of the spatial resolution (pixel size) for an image classification problem.

Spatial resolution has a complex effect on image classification. Classification accuracy is not maximized simply by having a high spatial resolution. Woodcock and Strahler [2] found that observed land cover classification accuracies were the result of a tradeoff between two factors. The first factor is the influence of boundary pixels on classification results. With finer spatial resolution, the proportion of pixels falling within the boundary of objects will decrease. Decreasing the boundary objects will result in a higher classification accuracy since boundary pixels have a mix of elements which creates confusion in the classification process. The second factor affecting the classification accuracy is that finer spatial resolution increases the spectral variance of landcover types. Intra-variance or variation inside a class decreases the spectral separability of classes and results in lower classification accuracy.

Many researchers argue that obtaining the optimal spatial resolution depends on the characteristics of the scene and the objective of the analysis. Several methods have been suggested to understand the effects of spatial resolution, and the most popular methods are discussed in the literature review. These methods have not been broadly adopted, due to not having a comprehensive and systematic procedure [3].

The method developed in this thesis focuses on finding the optimal pixel size for individual classes, so statistically based hard classification algorithms have maximum accuracy. In hard image classification algorithms, the pixels are labeled as coming from the one class with the highest maximum posterior probability of membership.

The main assumption used to develop the method here is that the homogeneity of image clusters will increase the accuracy of hard classification by reducing the within-class variance. It is assumed that all classes have a variation that is not modeled well by a single class if the pixel size is small enough, and at some coarser scale, the pixels will become more homogeneous. For example, suppose there is a scene with leaves where small amounts of soil are visible through the leaves. If the pixel scale is tiny, then some of the pixels will be leaves and others soil, and will not be well modeled as a single class. As the pixel scale gets coarser, the pixels will contain both leaves and soil, becoming homogeneous at a scale determined by the scene.

1.2 PURPOSE

This research will benefit all types of hyperspectral classification applications. Greater accuracy during the classification will benefit the use of hyperspectral images for multiple areas such as agriculture, biotechnology, security and defense, and environmental monitoring.

Image remote sensing techniques can be applied to interpret image data acquired from an airborne system (satellite or airplane). One of the highest challenges in developing an autonomous remote sensing monitoring system is to be able to discriminate a pixel correctly as one entity when the individual entity may be smaller than the individual pixels. To find the proper pixel size has shown not to have a one-size fits all solution. Images have unique characteristics that affect the classification accuracy such as the size of spatial features and neighborhood properties, and heterogeneity on spatial patterns.

1.3 MOTIVATION

This study's motivation is to improve a NASA's Puerto Rico Grant Consortium Research named Hyperspectral Imaging for Biodiversity Assessment of Coastal and Terrestrial Ecosystems [4]. This research project consisted of developing an effective biodiversity assessment using hyperspectral image processing techniques. The images were used to study identification, classification, mapping and detection of changes in biodiversity from satellite and airborne systems.

One of the conclusions from the research was that classification accuracy was affected by the pixel size. To access this situation, Lunzer, a graduate student from UPRM, in his thesis approached it by utilizing analysis of variance tests (ANOVA) to provide a method to test if classes in a hyperspectral image are homogeneous. There were various drawbacks such as the high sensitivity of the ANOVA test with real data. Also, the test could not be utilized on images containing only a single spectra, since the test was not sensitive enough to reject the null hypothesis [5].

Typical hypothesis tests are designed for one-dimensional data, where independence assumptions can be made, which is not always the case for multidimensional data. Having a hypothesis test specific for hyperspectral images provides a more reliable and systematic procedure that can be used as the pre-processing procedure to improve classification accuracy. The result provides the researcher with more information about the images, and how to choose an appropriate spatial resolution for a particular problem. Therefore, biodiversity assessment for the NASA's project can be improved as well as any other research that involves image classification.

1.4 OUTLINE

The organization of this thesis is as follows. Chapter 2 contains a relevant literature review of image classification, and techniques available to estimate the optimal resolution to analyze hyperspectral images. In Chapter 3, the theoretical background of spatial resolution and classification is provided. Also, hypothesis testing and sampling statistics theory is explained to provide the background necessary to derive the methods described in Chapter 4.

The methodology defines how random variables are defined in the hyperspectral images to apply hypotheses tests. The multiple statistical tests used presented explaining the advantages and disadvantages between them. Chapter 5 presents the different experiments done to improve the statistical analysis. Each of the experiments has results from synthetic images generated as explained in Section 5.1, and also from real hyperspectral images. Homogeneity results were validated by comparing common hypotheses tests with the one developed. The improvement of the classification accuracy is tested by comparing images classified at the spatial resolution determined by the test, along with a spatial resolution where pixels are less homogeneous.

To be able to understand the statistical analysis made to the images in more detail, a practical example is shown in Chapter 6 where the ideal pixel size for classification is obtained by using the covariance hypothesis test. Chapter 7 encompasses the conclusions from all the experiments realized in Chapter 5. Finally, the recommended future work is included based on the possible improvements that could be made to this research.

2 LITERATURE REVIEW

A literature review was done to find and summarize the studies relevant to the effects of spatial resolution on classification accuracy. The first part of this Chapter summarizes the evolution of the study of the image classification accuracy that started in the early 80's. The main methods currently used to choose an appropriate spatial resolution for remote sensing applications are discussed in the remaining sections.

2.1 HISTORY OF IMAGE CLASSIFICATION ACCURACY

In 1981, Markham and Townshend [6] found that image classification accuracy is affected by two factors. One is the change of the number of mixed pixels located near the boundaries between classes, and the other is the change of spectral variations within classes. Finer spatial resolution (smaller pixels) reduces the mixed pixels and improves the classification accuracy, but it also increases the spectral variation within classes decreasing the classification accuracy. It could be noted that spatial resolution is a crucial factor that needs to be widely studied.

In Bingwen et al. [7] local variance, semivariogram, and wavelet methods were compared on urban, agricultural and forest landscape images to find the optimal spatial resolution. The results varied with method and data source. The local variance gave relatively small values; the semivariogram gave a significant range of values, and the wavelet transform gave values within the bounds of the semivariogram. Also, the local variance is efficient in obtaining an ideal pixel size for small-scale observations, the wavelet transform is more suitable for large-scale areas, and the semivariogram is optimal over a significant range of values.

This problem has been researched for many years [8], but there are still various problems in identifying the optimal pixel size for classification. The results from all the methods give broad and unreliable results that are not consistent between them giving different optimum pixel sizes for the same image.

Local variance has been one of the most popular choices to find the optimal resolution necessary to classify hyperspectral images. Andre´foue¨t [9] presents how spatial resolution affects coral bleaching mapping. Local variance analysis provided information about the optimal resolution required for classification obtaining a range from 40-80 cm. In McCloy [9] the average local variance (ALV) was used to find the pixel size that maximizes the classification accuracy. The graph of ALV versus pixel size was hypothesized to have a peak where the size matches object size. By comparing this method with the semivariogram, similar results were not achieved.

Several research papers show that a consistent, unique solution between methods has been extremely difficult or impossible to obtain. Also, the purpose of these methods is to demonstrate statistical properties of the image and leave the interpretation to the analyst, whom should select an appropriate spatial resolution for their particular study.

2.2 LOCAL VARIANCE

Agronomy researchers were the first ones to show interest in the size of the support. Support is a geostatistical term that is equivalent to the spatial resolution, but normally the size of the support is greater than the spatial resolution due to the sensor point spread function (PSF) [10]. Mercer and Hall [11] and Smith [12] performed studies to crop yields and discovered that the variance between samples decreases as the support increases in size. Woodcock and Strahler [6] noticed that the likelihood that observations close in space are more alike than those apart and relied on this to estimate the spatial dependence using the local variance.

Let x_{ij} be the value of the pixel to be located in the ith row and jth column of an image. The local variance σ_{ij}^2 around x_{ij} can be calculated over a $(2n+1)$ by $(2m+1)$ window as:

$$\sigma_{ij}^2 = \frac{1}{[(2n+1)(2m+1)-1]} \sum_{k=i-n}^{i+n} \sum_{l=j-m}^{j+m} \{x_{kl} - \mu_{ij}\}^2, \quad (1)$$

where μ_{ij} is the mean of the $(2n+1)$ by $(2m+1)$ window centered on the pixel x_{ij}. The mean local variance is the mean of σ_{ij}^2 computed for all $x_{ij}$with the exception of a border equal to either n or m.

Woodcock and Strahler [2] made the assumption that scenes in an image are composed of discrete objects distributed on a continuous background to find the relation between spatial resolution and local variance. Using local variance as a function of spatial resolution, the maximum is an indirect guide to the size of objects in the scene. Spatial resolution will be selected depending on the purpose of the research, and its relation to the object size in the image.

2.3 EXPERIMENTAL VARIOGRAM

In the local variance method, not only the final selection of the spatial resolution depends on the researcher, but also, the approach is strictly empirical. A new method was proposed in [10] by modeling images as realizations of random processes.

A radiation sensed remotely can be modeled spatially as a random function $Z(x)$ defined for positions x in a two dimensional space $\mathbb{R}^2$:

$$Z(x) = m_v + v(x), \quad (2)$$

m_v is the local mean of $Z(x)$ in a region v, and $v(x)$ is a random function with zero mean. The spatial variation in Z allows it to adopt the intrinsic hypothesis of stationarity, such that the expectation exists and does not depend on x. With this characteristic all vectors of spatial separation h, the increment $[Z(x) - Z(x+h)]$, has finite variance which does not depend on x.

$$2\gamma(h) = var[Z(x) - Z(x+h)] = E[[Z(x) - Z(x+h)]^2], \quad (3)$$

where $\gamma(h)$ is the variogram, and (3) is a function that relates the semivariance at lag h, and also summarizes the spatial dependence in Z.

The spatial means $Z_v(x)$ is defined as the integral of $Z(x)$ over an area v. Spatial integration is usually known as regularization, and in the context of remote sensing it is, analogous to increasing the pixel size and coarsening the spatial resolution. Since the punctual variogram is never know a priori, the variogram has to be estimated using the experimental data. The punctual variogram has been computed and used extensively in much research [11] [12].

Having a measured property Z, on observations centered at $x_1, x_2, \ldots$, the method of moments estimator variogram can be computed for $\rho(h)$ pairs of observations:

$$\bar{\gamma}(h) = \frac{1}{2\rho(h)} \sum_{t=1}^{\rho(h)} \{[Z(x_t) - Z(x_t + h)]\}^2 . \quad (4)$$

According to Armstrong in [13] , this experimental estimator may give a poor estimate of the true semi-variogram because the presence of just one outlier can result in an erratic variogram. Therefore, several robust semivariograms were developed to overcome different estimation problems. Some of them are presented in [14] and [15].

Since the variogram is used by punctual support, the measurements in the image are made on pixels on finite area changes have to be made. Taking into consideration that the value of a spatial attribute over an area v is the mean value of all the points within v, the variogram on some support v from the punctual variogram is defined as the following by Journel and Huijbregts [16]:

$$\gamma_v(h) = \bar{\gamma}(v, v_h) - \bar{\gamma}(v, v). \quad (5)$$

Where $\bar{\gamma}(v, v_h)$ is the mean value of the punctual semivariogram between two pixels of size v whose centroids are separated by h, and $\bar{\gamma}(v, v)$ represents the mean value of the punctual semivariogram over the v domain.

Experimentally the variogram is calculated using the estimation in Equation (4), and the spatially dependent components are regularized. Coarser spatial resolutions are found by averaging the pixels values that will be considered a larger pixel value. The semivariance is plotted against the size of support; a typical result is shown in Figure 2-1.

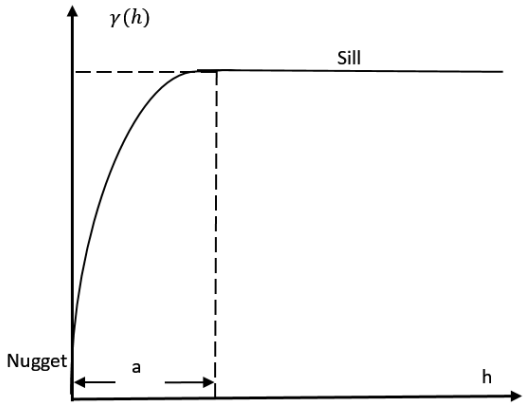


Figure 2-1 A typical semi-variogram function

The graph is used to find at which lag the semivariogram reaches its maximum $\bar{\gamma}_v(v)$. Studies have shown that this value can be used to help ensure that the spatial resolution chosen has the information necessary to shown spatial variation in the data. One issue with this method is being able to choose the optimum spatial resolution based on the spatial variation that the investigator wants to obtain from the image.

Even though variogram estimation seems very straight forward, assuming it is that simple can result in erratic conclusions. Several assumptions made to derive this equation should be verified before using this method. As previously written in this section, the experimental estimator may give a poor estimate of the true semi-variogram leading to false conclusions.

In conclusion, previous literature in the area shows the lack of a reliable method that can be applied to different types of images. The methodology in this thesis demonstrates a viable solution to finding optimum pixel sizes depending on the scene.

3 THEORETICAL BACKGROUND

3.1 REMOTE SENSING

Optical remote sensing makes use of visible, near infrared and short-wave infrared sensors to form images by detecting the solar radiation reflected from targets on the ground [17]. Different materials reflect, absorb, transmit or emit electromagnetic energy at different proportions and different wavelengths. This characteristic allows the targets to be differentiated by using their spectral reflectance signatures in the remotely sensed images.

Remote sensing usually refers to acquiring information about the Earth's surface and atmosphere using airborne or spaceborne platforms. Therefore, the sensors are always looking first through a layer of the atmosphere, which causes wavelength dependent absorption and scattering of radiation. Some of the atmospheric effects can be corrected, but some wavelength bands in the electromagnetic spectrum are almost entirely absorbed by the atmosphere. For this reason, only wavelengths in the microwave, infrared, visible region and part of the near ultraviolet regions are usable for remote sensing. X-rays and gamma rays are also transparent to the atmosphere, but they are not used in remote sensing of the earth.

After the satellite or airborne sensors acquire the electromagnetic radiation, it is transmitted to a ground station where it needs to be pre-processed before using the images for analysis and interpretation. Pre-processing steps are atmospheric corrections (as mentioned before), data error compensations, calibration and map registration. Any remote sensing textbook gives detailed information about this technique such as [18].

Remote sensing images are collected in the form of a data cube where they have two spatial dimensions (X-Y) and a third spectral dimension represented in the Z-direction. The latter consists of a broad number of spectral bands that measure radiation in a particular spectral range. Remote sensing systems are described by the quantity of bands is able to capture.

- Panchromatic system: the sensor is a single channel detector sensitive to radiation in a broad wavelength range.
- Multispectral system: the sensor is a multichannel detector with a few spectral bands, and each channel is sensitive in a narrow wavelength band.
- Hyperspectral system or imaging spectrometer: acquires images in about a hundred or more narrow contiguous bands.

This research is primarily interested in imagery generated by hyperspectral sensors. These consist of a large number of spectral bands (100-300).

3.1.1 SOC700 Hyperspectral Spectrometer

Synthetic images were created along with the acquisition of controlled real world data to be able to research the selection of a spatial resolution to improve classification accuracy. The majority of the hyperspectral images were gathered using a SOC-700 stand mounted Hyperspectral Imager with a spectral range of 430nm to 900nm and 120 spectral bands. Each band consists of 640x640 pixels with a 12-bit dynamic range. The SOC700 utilizes a push broom type sensor and a scanning mirror to generate the along-track dimension. Images gathered using this spectrometer were from areas around Mayagüez, Puerto Rico. Other images were taken from the database located at the Laboratory for Applied Remote Sensing and Image Processing (LARSIP), developed thanks to a NASA EPSCoR grant. The synthetic hyperspectral images were created by using an open source Matlab toolbox called HYDRA, and spectral signatures from the USGS Spectral Database. Additional details are explained in Section 5.1.2.

3.2 SPATIAL RESOLUTION

There are multiple sensor parameters (i.e., spatial resolution, the number of spectral bands, signal-to-noise ratio, spectral resolution, etc.), and the optimal selection of these parameters depends on the

objects under study and instruments used. This thesis focuses on finding the optimal selection of the spatial resolution for a particular problem.

Spatial resolution refers to a parameter that measures the sensor's ability to image or record closely spaced objects so that they are distinguishable as different objects. According to the previous definition, a higher spatial resolution number (1m) means that finer details can be observed in the image compared to 10 m spatial resolution.

There are many definitions of the specific term of spatial resolution because many sensor parameters influence its value. Ideally, a lens would image a point object as a point in the image plane, but diffraction causes the point object to be a bright disc (airy disc). The distribution of energy of this airy pattern is similar to a 2D sinc function, with the main peak followed by some minima and maxima. Two objects can be just resolved if the peak of the airy pattern of one object falls on the first minima of the other, and this is the resolving limit of two objects [19]. Nonetheless, diffraction is not the only influence on the spatial resolution; it is also determined by a combined effect of resolving power of the lens, object contrast, and signal-to-noise ratio of the other components. Therefore defining the spatial resolution as a measure of the smallest object that can be detected in an image is a rather simple definition. For example, it may be possible to identify an object smaller than the resolution of a sensor when there is a significant contrast between the object and its background.

In this thesis, and when sensors using discrete detectors are used to generate an image, the spatial resolution is defined as the projection of the detector element onto the ground through optics [19]. The ground area viewed by a detector is determined mostly by the Instantaneous Field of View (IFOV), which is the angular measurement of the area viewed by a single detector at a given instant in time. When an image is acquired from a remote system as shown in Figure 3-1, the IFOV corresponds to item (B), and it determines the ground resolved area (A) or pixel size from a given altitude at one particular moment in

time. The size of the area viewed (A) is determined by multiplying the IFOV (B) by the distance from ground to sensor (C) [20].

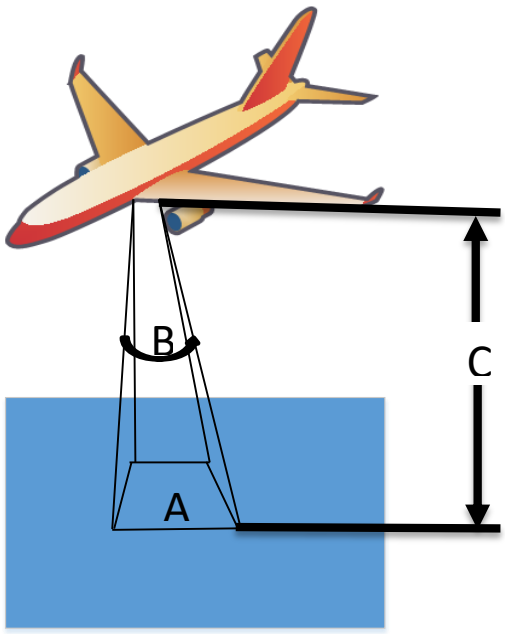


Figure 3-1 Relation between IFOV (B), sensor altitude (C) and ground area (A)

As mentioned in section 3.1.1, images are composed of a matrix of pixels (picture elements). These correspond to the smallest units of an image. Image pixels are normally square and represent a certain area on an image. To be clear on the terminology further explained in this document, low or coarse resolution refers to images where only large features are visible while the fine or high-resolution images refer to images where small objects can be visibly identified.

Woodcock and Strahler [2] found that observed classification accuracies were the result of a tradeoff between two factors. Accuracy is not maximized simply by having a high spatial resolution. The first factor is the influence of boundary pixels on classification results. With finer spatial resolution, the proportion of pixels falling within the boundary of objects will decrease, but will increase spectral variance. The second factor affecting the classification accuracy is that with finer spatial resolution the spectral variance of land covers types associated increases.

3.3 CLASSIFICATION

Classification is a quantitative method to analyze hyperspectral images. It consists of recognizing categories of real-world objects and labeling these entities. There are two approaches to performing classification procedures: supervised and unsupervised classification.

In supervised classification, the classifier algorithm has to be trained to recognize the classes of interest. This is done by using sample pixels that are labeled with their respective classes. Practical Steps to implement a supervised classifier are:

1. Selection of the discrimination rule (parametric or nonparametric)
2. Collection of training samples
3. Use of training data to estimate the parameters of the classifier to be used
4. Use of the trained classifier to label or classify each pixel in the image into one of the classes.
5. Produce tabular summaries or thematic maps to summarize classification results.

In contrast, an unsupervised method does not require defining the training samples, nor defining classes in the image. This method uses clustering to determine the number of distinct categories present in the image and assigns pixels to those categories. Clustering is used to find similarities in characteristics between the inputs.

This thesis is focused on finding an optimal pixel size for individual classes, so its correct classification is maximized. Statistically based classification algorithms will be used. They are based on the assumption that each class has spatial features that can be statistically separated from other classes. Consequently, the spatial features can be modeled such that their means and variances allow this separation [18]. If a pixel is modeled as a single class, standard classification algorithms can be applied to perform pixel-wise classification.

To visualize how spatial resolution affects the classification in images, Figure 3-2 and Figure 3-3 show the results from an image classified at different spatial resolutions. The image in Figure 3-2 was averaged to produce the image shown in Figure 3-3. Neighborhood pixel averaging produced an image ¼ the size of the original image. The classification shows that objects in this image were classified better when the pixel size was increased. For example, the region selected by the black square in both images, in the left the pixels were classified as coming from three classes (shown in different colors) while in Figure 3-3 the pixels in this region are classified as coming from only two classes.

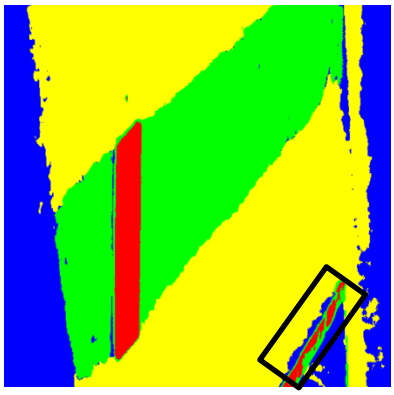


Figure 3-2 Supervised classification (Maximum Likelihood) of the original image

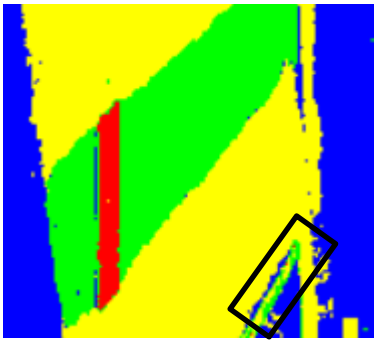


Figure 3-3 Supervised classification (Maximum Likelihood) of ¼ size original image

3.4 HYPOTHESIS TESTING

Inferential statistics enables the ability to measure the behavior of samples with the objective of learning more about the behavior of populations that are normally inaccessible or too large. Hypothesis testing is the method of testing a hypothesis made from a population, based on sample statistics. In an inferential statistical context, hypotheses are formulated as assumptions on

(i) The probability distribution f of one or more random variables $X, Y, \ldots, Z$ in a population Ω, or on

(ii) One or more parameters θ of this distribution function.

Hypothesis testing methodology can be summarized in four steps:

1. A formulation of a hypothesis about what is to be tested, which is called the null hypothesis H_o, and the definition of its alternative hypothesis H_A.
2. A significance level α is selected. This value states the probability at which the null hypothesis is incorrectly rejected.
3. Calculation of a test statistic.
4. Results are analyzed to state a conclusion based on H_0.

The test is performed to determine if the null hypothesis can be rejected. Therefore, not rejecting the null hypothesis does not prove that the null hypothesis is true, but it suggests that the null hypothesis is likely to be true.

3.4.1 Null and Alternative Hypotheses

The null hypothesis (H_0) is a statement about a population parameter, which is assumed to be true. The purpose is to test whether the null hypothesis statement is likely to be true. An alternative hypothesis is stated such as it contradicts the null hypothesis, in case of rejecting the null hypothesis then the alternative hypothesis (H_A) states that the actual population parameter is less than, more than, or not equal to the value indicated in the null hypothesis.

3.4.2 Level of significance

The significance level α is fixed prior to making the test, and it is typically selected such as $\alpha \in [0.01, 0.05]$. Fixing this value prior to performing the statistical test controls the risk of committing a Type I error, which is the probability of rejecting H_0 when is true, defined as $P(H_A|H_0\ true) = \alpha$. Another potential error when performing a hypothesis test is not rejecting H_0 when it's false, which is defined as $P(H_o|H_A\ true) = \beta$. The probability $1 - \beta$, is associated with the power of a statistical test. It is called the power since this is the outcome that is aimed when assuming that the null hypothesis is incorrect. The different possible outcomes when performing a test decision are summarized in the following table.

Table 3-1 Consequences of test decisions

	H_0	H_A
$H_0\ true$	Correct decision $P(H_o \mid H_0\ true) = 1 - \alpha$	Type I error $P(H_A \mid H_0\ true) = \alpha$
$H_A\ true$	Type II error $P(H_o \mid H_A\ true) = \beta$	Correct decision $P(H_A \mid H_A\ true) = 1 - \beta$

3.4.3 Test statistic

After defining the null hypothesis, alternative hypothesis and significance level, a random sample is collected, and the test statistic selected on the hypothesis is calculated. Since the statistic is calculated from a sample, it is necessary to evaluate how likely that sample outcome is if the null hypothesis is true. For example, the test statistic is used to determine how many standard deviations a sample mean is from the population mean. The larger the value of the test statistic, the further a sample mean is from the population mean stated in the null hypothesis. The value of the test statistic is used to make a decision regarding the null hypothesis.

3.4.4 Decision Rules

The decision rules for rejecting the null hypothesis is described in two ways by statisticians, with reference to a p-value or with reference to a region of acceptance [21].

The p-value measures the strength of evidence in support of the null hypothesis. It measures the probability of observing a test statistic as stated in the null hypothesis, assuming the null hypothesis is true. If the p-value is less than the significance level α, the null hypothesis is rejected.

Another method used to define a decision rule is by defining a range of values as the region of acceptance. This region is defined by obtaining the probability of making a Type I error equal to the significance level. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected. Otherwise it is rejected rejected at the α level of significance.

3.5 SAMPLING STATISTICS

The collection of sample data is an arbitrary process where different outcomes are obtained from each experiment. Therefore, the statistic obtained from the sampling data is a random variable and understanding their sampling distributions is important to determine how close the sample statistic is from the population parameter.

When a population parameter θ is unknown, it can be estimated with a random sample $X_1, X_2, \dots, X_n$ of size n. Since population data was randomly sampled, the statistic obtained from the sample is also a random variable with a certain probability distribution. To estimate a parameter θ from the sample it is necessary to model the sample by a probabilistic model that depends on θ.

3.5.1 Sample Mean Distribution

To estimate the population mean μ, the sample mean is defined as

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{6}$$

By definition, when a sample is created by random selections of a population data, the X_i random variables will be independent and identically distributed. Consequently, by combining this property with various expected value properties the theoretical mean and variance of the sample mean can be calculated as,

$$E(\hat{\mu}_x) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}(n\big(E(X)\big) = E(X), \text{and} \tag{7}$$

$$Var(\hat{\mu}_x) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} X_i = \frac{1}{n^2}\big(nVar(X)\big) = \frac{Var(X)}{n}. \tag{8}$$

The last equation indicates that as the sample size increases the variance of the sample mean decreases. Also, if the X_i are normally distributed independent random variables then the sample mean distribution is a normal distribution $\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. The linear combination of independent normal random

variables is a normal distribution, which can be proved by using the characteristic function of the sum of two random variables,

$$\varphi_{X+Y}(t) = E(e^{it(X+Y)}) \text{ [22]}, \tag{9}$$

and the characteristic function of the normal distribution,

$$\varphi(t) = \exp\left(it\mu - \frac{\sigma^2 t^2}{2}\right). \tag{10}$$

3.5.1.1 Central Limit Theorem

In Section 3.5.1 it was established that when there is sample of size n, $X_i \sim N(\mu, \sigma^2)$, for $i = 1,2,\dots,n$, then $\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, but if the random sample does not follow a normal distribution the sample mean distribution can be estimated by using the Central Limit Theorem.

The Central Limit Theorem states that if there is a random sample $X_1, X_2, \dots X_n$ from any distribution, with finite mean μ and finite variance σ^2, and the sample size n is sufficiently large, then:

(i) The sample mean $\hat{\mu}$ follows an approximate normal distribution

(ii) The mean $E(\hat{\mu}_x) = \mu$,

(iii) And the variance of the sample mean is $Var(\hat{\mu}_x) = \frac{\sigma^2}{n}$.

In overview, the Central Limit Theorem (CLT) states that the sampling distribution of the sample mean is approximately normally distributed regardless of the distribution of the random sample [23], if the sample size if sufficiently large. Noting that if it is a random sample then the X_i are independent identically distributed (i.i.d) random variables. The minimum sample size depends on the skewness of the distribution from which the random sample comes [24]:

(i) If the distribution of the sample is symmetric, unimodal or continuous, then a sample size n as small as 4 or 5 could be adequate.

(ii) If the distribution of the sample is skewed, then a sample size n of at least 25 or 30 could yield an adequate approximation.

(iii) If the distribution of the sample is extremely skewed, then a bigger sample size n is needed.

Another version for CLT is that if $X_1, X_2, \dots, X_n$ are independent random variables that are identically distributed with finite means and variances, then the normalized sum $(S_n - n\mu)/\sigma\sqrt{n})$, where $S_n = X_1 + X_2 + \cdots, X_n$, as $n \to \infty$ converges to a Gaussian variable with zero mean and unit variance. In other words, the sum of a sufficient number of independent and identically distributed random variables with finite means and variances will result in a Gaussian random variable [25].

3.5.2 Sample Variance Distribution

The variance is an important population measure because it is a measure of the spread of the data on the mean. The sample variance is defined to be

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_x)^2. \tag{11}$$

A statistic is an unbiased estimate of a parameter when the mean of the sampling distribution of that statistic can be shown to be equal to the parameter being estimated [22]. Since $\hat{\sigma}^2$ is estimating σ^2, if the $E(\hat{\sigma}^2) = E(\sigma^2)$, then $\hat{\sigma}^2$ is an unbiased estimator.

By calculating the expected value of Equation (11), $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$ it is noticed that the sample variance estimation is biased by a factor of $\frac{n-1}{n}$. After correcting for the bias the unbiased sample variance is,

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_x)^2 \tag{12}$$

In the case that X_i are independent observations from a normal distribution, the Cochran's theorem shows that s^2 follows a chi-squared distribution [26]

$$(n-1)(s^2/\sigma^2) \sim \chi^2{}_{n-1}. \tag{13}$$

The Cochran's theorem states that if there is a set $Y_1, \dots, Y_n$ of independent standard normally distributed random variables, and an identity of the form

$$\sum_{i=1}^{n} Y_i^2 = Q_1 + \cdots + Q_k \tag{14}$$

can be written, where each Q_i is a sum of squares of linear combinations of the Ys, then the Q_i are independent and follow a chi-squared distribution with r_i degrees of freedom [27]. Where r_i is the rank of Q_i, and $r_1 + \cdots + r_k = n$. This rank is the number of independent linear combinations included in the sum of squares defining Q_i.

To prove Equation (13), $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_x)^2$, where $X_1, X_2, \dots, X_n$ are independent normally distributed random variables with mean μ and standard deviation σ, then $Y_i = (X_i - \mu)/\sigma$ is a standard normal for each i. By using Cochran's theorem,

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n}\left(\frac{X_i - \hat{\mu}}{\sigma}\right)^2 + n\left(\frac{\hat{\mu} - \mu}{\sigma}\right)^2 \tag{15}$$

since the sample variance can be written as $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \hat{\mu}_x)^2$, it can be substitued in Equation (15) to give

$$\sum_{i=1}^{n} Y_i^2 = \frac{(n-1)S^2}{\sigma^2} + n\frac{(\hat{\mu} - \mu)^2}{\sigma^2}. \tag{16}$$

Equation (16) can be written as,

$$\frac{(n-1)S^2}{\sigma^2} = n\frac{(\hat{\mu} - \mu)^2}{\sigma^2} - \sum_{i=1}^{n}\left(\frac{X_i - \hat{\mu}}{\sigma}\right)^2 = \mathrm{Q}_1 - \mathrm{Q}_2, \tag{17}$$

where Q_1 is of rank 1 and Q_2 is ranked n. Therefore, the theorem states that Q_1 and Q_2 are independent, chi-squared distributions with 1 and n degrees of freedom respectively. The latter proves the result in Equation (13) $(n-1)\frac{s^2}{\sigma^2} \sim {\chi^2}_{n-1}$. Therefore, the mean of the sample variance is, $E(s^2) = E\left(\frac{\sigma^2}{n-1}\,{\chi^2}_{n-1}\right) = \sigma^2$, and its variance is $Var(s^2) = Var\left(\frac{\sigma^2}{n-1}\,{\chi^2}_{n-1}\right) = 2\frac{\sigma^4}{n-1}$.

3.5.3 Sample Covariance Distribution

The covariance is the mean value of the product of deviations of two variates from their respective means. The covariance between X_p and X_q is defined as,

$$C_{pq} = E\{(X_p - \mu_p)(X_q - \mu_q)\}. \tag{18}$$

The covariance matrix of a multivariate random variable is usually not known, and its estimation is needed. The unbiased estimation of the covariance is

$$S = \widehat{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu_x})(Y_i - \widehat{\mu_Y}) \tag{19}$$

If a random vector $X_1, X_2, \dots, X_k$ is said to be k-variate normally distributed then the distribution of $(n-1)S$ is distributed as a Wishart random matrix with $n-1$ degrees of freedom [28].

3.6 HYPOTHESIS TEST: IDENTICALLY DISTRIBUTED RANDOM VARIABLES

A test statistic can be developed to test a null hypothesis indicating that a sum of random variables are dependent and identically distributed. The test statistic can be found by using the property that indicates that the variance of a sum of **m** random variables is the sum of pairwise covariances.

$$Var(\sum_{i=1}^{m} X_i) = Cov(\sum_{j=1}^{m} X_j, \sum_{i=1}^{m} X_i) = \sum_{j=1}^{m}\sum_{i=1}^{m} Cov(X_j, X_i)$$

$$= \sum_{i=j}\sum Cov(X_j, X_i) + \sum_{i\neq j}\sum Cov(X_j, X_i)\,, \tag{20}$$

Since for any i and j the Cov$(X_j, X_i) = Cov(X_i, X_j)$, Equation (20) can be written as,

$$Var(\sum_{i=1}^{m} X_i) = \sum_{i=1}^{m} Var(X_i) + 2\sum_{i\neq j}\sum Cov(X_j, X_i)\,. \quad (21)$$

To use Equation (21) as a test statistic, it can be written as the difference,

$$Var(\sum_{i=1}^{m} X_i) - \sum_{i=1}^{m} Var(X_i) - 2\sum_{i\neq j}\sum Cov(X_j, X_i) = 0. \quad (22)$$

Equation (22) can be used as the statistic in a hypothesis testing problem. For simplicity, a definition of variables could be done by stating that $W = Var(\sum_{i=1}^{m} X_i)$ and Z=$\sum_{i=1}^{m} Var(X_i) + 2\sum_{i\neq j}\sum Cov(X_j, X_i)$, which summarizes as,

$$W - Z = 0\,. \quad (23)$$

The null hypothesis represents the case of the zero difference while the alternative hypothesis indicates that the difference is greater than zero.

3.7 WHITENING TRANSFORM

The test statistic in the previous section needs strong assumptions about the random variables in order to have a less complex methodology for finding the resulting statistics. A pre-processing method can be applied to the images to be able to assume that $X_1, X_2, \ldots X_m$ are uncorrelated random variables, such as $E(X_i, X_j) = E(X_i)E(X_j)$, for all $i \neq j$.

The Whitening Transform is a decorrelation method that transforms an arbitrary set of variables having a known symmetric and positive definite covariance matrix M into a set of new variables whose covariance is the identity matrix. If X is a random column vector with covariance matrix M and mean 0, then M can be written as $M = E[XX^T]$. If M is a symmetric and positive definite matrix, M has a positive

definite symmetric square root $M^{\frac{1}{2}}$, such that $M^{\frac{1}{2}}M^{\frac{1}{2}} = M$. The random vector X is transformed to Y by using $Y = M^{-\frac{1}{2}}$X, which has covariance matrix,

$$Cov(Y) = E[YY^T] = M^{-\frac{1}{2}}E[XX^T]\left(M^{-\frac{1}{2}}\right)^T = M^{-\frac{1}{2}}MM^{-\frac{1}{2}} = I. \quad (24)$$

4 METHODOLOGY

In this chapter, the statistical method developed to find the ideal pixel size or spatial resolution for classification is fully explained. This methodology was developed by an improvement of several statistics methods with different assumptions made according to the hyperspectral image classification problem.

The statistical analysis performed on the images is fully explained in the first sections. It includes the definition of random variables in the image and the downsampling methodology. After defining the variables, a hypothesis test is applied to the image by assuming independence between random variables. Another methodology to improve the results was established by changing independence assumptions made about the image. The last part of the chapter explains the typical parametric and non-parametric hypotheses tests used to determine when the pixels can be well modeled as coming from the same density. This last part is used to provide validation to the results of the hypothesis test developed.

4.1 DOWNSAMPLING

It is assumed that there is a pixel scale where the pixels from a single class become homogeneous. This means that we are trying to determine when statistics of the pixels become homogeneous, or when the pixels can be well modeled as coming from the same density. A downsampling of the image is performed to obtain a data set of multiple spatial resolution images.

In order to have the image at different spatial resolutions the original image is averaged to successively coarser spatial resolution pixels. Let the pixels of the original image be $\{x_{ij}\}$, where i is the row and j the column. The pixels of the lower resolution image are then $\{\tilde{x}_{ij}\}$, where $\tilde{x}_{ij} = \frac{\sum_{l=(i*w)-(w-1)}^{i*w} \sum_{m=(j*w)-(w-1)}^{j*w} x_{lm}}{w \cdot w}$ and the window size w=2,3,4,…,W. Figure 4-1 shows an example where the

window size is w=2, therefore the lower resolution image pixel $\tilde{x}_{11} = \frac{(x_{11}+x_{12}+x_{21}+x_{22})}{4}$, and the next pixel is obtained by shifting a non-overlapping window.

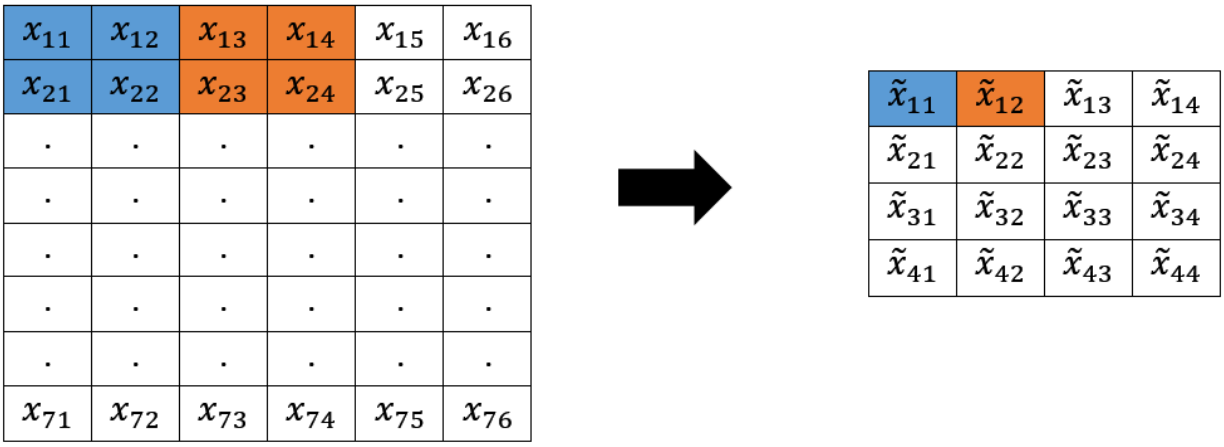


Figure 4-1 Transformation from original image to a lower resolution image by using a 2X2 window

To find the relation between the downsampling window size $w\ x\ w$ and the lower resolution image pixel size $p_{\tilde{x}}$, it can be calculated by multiplying the pixel size of the image before being downsampled p by one dimension of the window size, $p = wp_{\tilde{x}}$.

4.2 Random Variables in the Image

To find a straightforward way of estimating at what resolution the pixels become homogeneous, the image is divided into multiple random variables defined by neighboring pixels. The random variables are defined to exploit the characteristics of the average of m random variables. Each pixel in the downsampled image is obtained by taking an average of m different random variables, where m is determined by the window size. In the example in Figure 4-2, and in the rest of the thesis a window size of w = 2 is used such, such as $m = w^2 = 4$ and $Y_1 = \frac{X_1+X_2+X_3+X_4}{4}$. The new random variable is determined by 4 random variables such as

$$Y_i = \tilde{X} = \frac{X_1+X_2+\cdots X_m}{m} = \left(\frac{1}{m}\right)(X_1 + X_2 + \cdots X_m). \quad (25)$$

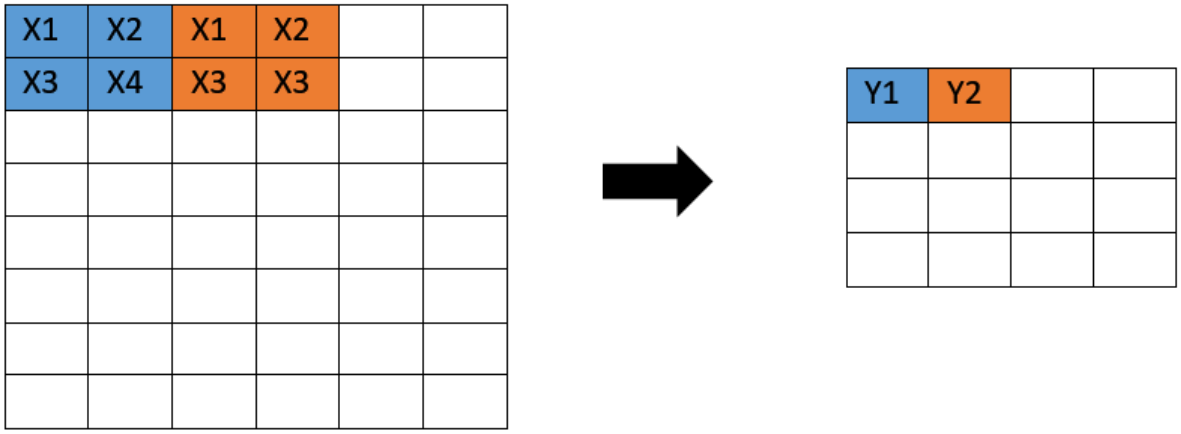


Figure 4-2 Definition of random variables in the original and lower resolution images

4.3 HYPOTHESIS TESTING: VARIANCE

To statistically determine when pixels become homogeneous, or can be modeled from the same density, a hypothesis test can be set up to determine the spatial resolution at which the pixels are identically distributed. Since lower resolutions are obtained by doing a neighboring average of the original image, the variance of an average of identically distributed random variables can be used in the hypothesis test.

Assuming that we have m independent and identically distributed random variables $X_1, X_2, \dots X_m$, then each X_i has the same mean μ and variance σ^2.

Using this the mean and the variance of $\tilde{X}$ is:

$$E(\tilde{X}) = E\left[\left(\frac{1}{m}\right)(X_1 + X_2 + \cdots X_m)\right] = \left(\frac{1}{m}\right)E(X_1 + X_2 + \cdots X_m), \quad (26)$$

$$Var(\tilde{X}) = Var\left[\left(\frac{1}{m}\right)(X_1 + X_2 + \cdots X_m)\right] = \left(\frac{1}{m}\right)^2 Var(X_1 + X_2 + \cdots X_m). \quad (27)$$

To simplify (26), the property of the expected values that says that the expected value of a sum is always the sum of the expected values is used. Therefore, having the previous in consideration and

because the $X_i's$ are identically distributed (they have the same mean and variance) Equation (26) will reduce to:

$$E(\tilde{X}) = \left(\frac{1}{m}\right) m\mu = \mu \tag{28}$$

Equation (27) can also be simplified, since the X_i are independent, the variance of the sum is the sum of the variances:

$$Var(\tilde{X}) = \left(\frac{1}{m}\right)^2 Var(X_1 + X_2 + \cdots X_m) = \left(\frac{1}{m}\right)^2 (m\sigma^2) = \frac{\sigma^2}{m}, \text{ for all m } > 0. \tag{29}$$

Equation (29) can be used to find the dependence of pixel size and variance. To test this hypothesis, the left and right side of Equation (29) is plotted. If X_i are independent and identically distributed, then $Var(\tilde{X}) - \frac{\sigma^2}{m} = 0$.

4.4 HYPOTHESIS TESTING: COVARIANCE METHOD

The variance hypothesis test assumed independence between the pixels. When not assuming independence of the pixels, a new hypothesis test can be set up to determine if spatially the pixels before averaging were identically distributed or not.

The variance of the pixels in the lowered resolution image is:

$$Var(\tilde{x}) = Var\left(\frac{\sum_{l=(i*w)-(w-1)}^{i*w} \sum_{m=(j*w)-(w-1)}^{j*w} x_{lm}}{w \cdot w} \right) = \frac{1}{w^4} Var\left(\sum_{l=(i*w)-(w-1)}^{i*w} \sum_{m=(j*w)-(w-1)}^{j*w} x_{lm} \right). \tag{30}$$

In general, the variance of a sum of variables is the sum of pairwise covariances. To simplify notation, let the pixels in the double summation of Equation (30) be numbered from 1 to $w * w = m$. The variance of a sum can be written as

$$Var\left(\sum_{i=1}^{m} x_i\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} Cov(x_i, x_j) = \sum_{i=1}^{m} Var(x_i) + 2\sum_{\{\{i,j\}:i<j\}} Cov(x_i, x_j). \quad (31)$$

In the case that the pixels in the original image are identically distributed, this can be simplified to

$$Var\left(\sum_{i=1}^{m} x_i\right) = mVar(x_i) + 2\sum_{\{\{i,j\}:i<j\}} Cov(x_i, x_j) \quad (32)$$

Combining (30) with (32),

$$m^2 Var(\tilde{x}) - mVar(x_i) - 2\sum_{\{\{i,j\}:i<j\}} Cov(x_i, x_j) \equiv 0 \quad (33)$$

when the pixels in the original image are identically distributed.

The difference stated in Equation (33) can be used as the statistic in a hypothesis testing problem. The null hypothesis represents the case of the zero difference while the alternative hypothesis indicates that the difference is greater than zero. If the null hypothesis cannot be rejected, then the pixels in the original image are determined to be identically distributed.

4.4.1 Hypothesis Test: Band by Band

The methodology above is applied spatially on the image. In order to include the spectral component of the hyperspectral images, the residual sum of squares (RSS) of the statistic in (33) is applied to each band. If each band is called b and the total of bands are k, the residual sum of squares in the whole image is,

$$RSS = \sum_{b=1}^{k}\left(m^2 Var(\tilde{x}_b) - mVar(x_{i,b}) - 2\sum_{\{\{i,j\}:i<j\}} Cov(x_{i,b}, x_{j,b})\right)^2. \quad (34)$$

4.4.2 Determination of the Statistics of the Test Statistic

The statistics of the test statistic in Equation (33) are necessary to determine a decision rule for the hypothesis test. Since the population variance and mean are unknown most of the time, the sample

mean, variances and covariances equations presented in Sections 3.5.1-3.5.3 need to be used. Noting that since samples are randomly selected, the sample mean, sample variance and sample covariance are random variables. Since many parameters are unknown, it is necessary to make assumptions to be able to calculate the statistics of Equation (33).

A common assumption is that $X_1, X_2, \dots X_m$ are normally distributed random variables. Another assumption made for the purpose of this research is to say they are also identically distributed. Therefore, the statistics of $\widehat{Var}(\sum_{i=1}^{m} X_i)$, can be calculated by using the sample variance of N observations. If $Y = \sum_{i=1}^{m} X_i$, then

$$\widehat{\sigma_Y^2} = \frac{1}{N-1}\sum_{i=1}^{N}(Y - \hat{\mu}_Y)^2, \tag{35}$$

where $\hat{\mu}_Y = \left(\frac{1}{N}\right)\sum_{i=1}^{N} Y_i = \sum_{i=1}^{m} \hat{\mu}_{X_i} = m\hat{\mu}_X$, since it is assumed that each X_i is identically distributed. Also, since they are normally distributed they follow the sample mean distribution $\hat{\mu}_X \sim N(\mu_X, \frac{\sigma_X^2}{N})$. Therefore, $\widehat{\sigma_Y^2}$ is the sum of the squared differences between two normal random variables, which follows a Chi-squared distribution with $N-1$ degrees of freedom and mean and variance $(\sigma_Y^2, 2\frac{\sigma_Y^4}{N-1})$, if Y_i observations are assumed independent.

The Y_i corresponds to the sum of neighboring pixels used in a window size wxw. The observations of these random variables can be assumed independent as the sampling pixels are farther away, which is analogous to defining more than four random variables in an image. In this research four random variables were used through the whole process, therefore this approximation will not apply. Another problem with averaging more than four random variables is the amount of samples, which could put the constraint that the methodology would only be applicable for large images.

The statistics of $\sum_{i=1}^{m} Var(X_i)$ is similar to the statistics of the variance of the sum. Because the X_i are identically distributed, $\sum_{i=1}^{m} Var(X_i) = mVar(X)$, and by using the sample variance equation,

$\widehat{\sigma_x} = \frac{1}{(N-1)}\sum_{i=1}^{N}(X - \hat{\mu}_X)^2$, it also follows a Chi-squared distribution with $\chi^2{}_{N-1}$, and has mean and variance $(\sigma_X^2, 2\frac{\sigma_X^4}{N-1})$.

The statistics of the last term is much more complex since it has the sample covariance function, and several assumptions need to be made. This will be done in the next section.

4.4.2.1 Approximation using the Central Limit Theorem

The test statistic can be written using the sample mean, variance, and covariance,

$$m^2\frac{1}{N-1}\sum_{i=1}^{N}[(Y - \hat{\mu}_Y)^2] - m\frac{1}{N-1}\sum_{i=1}^{N}[(X - \hat{\mu}_X)]^2\text{-}2\frac{1}{N-1}\sum_{k=1}^{N}[\sum_{i<j}(X_{ik} - \widehat{\mu_{x_i}})(X_{j_k} - \widehat{\mu_{X_j}})\,]. \quad (36)$$

As N $\rightarrow\infty$ the central limit theorem can be used to estimate the distribution of Equation (36). The squared differences between two normal random variables can be represented a chi-squared distribution. Therefore, every term in [.] corresponds to a chi-squared random variable, and these random variables are summed N times by the leading summation operator.

A version of the CLT is that if $X_1, X_2, \ldots, X_n$ are independent random variables that are identically distributed with finite means and variances, then the normalized sum $(S_n - n\mu)/\sigma\sqrt{n})$, where $S_n = X_1 + X_2 + \cdots, X_n$, as $n\rightarrow\infty$ converges to a Gaussian variable with zero mean and unit variance [29]. In other words, the sum of a sufficient number of independent and identically distributed random variables with finite means and variances will result in a Gaussian random variable.

The expected value and variance of a chi-squared distribution with N-1 degrees of freedom (or sample size) are,

$$E\left(\chi^2{}_{N-1}\right) = N - 1, \text{ and} \quad (37)$$

$$Var\left(\chi^2{}_{N-1}\right) = 2(N - 1). \quad (38)$$

These moments can be used to convert χ^2 into a standard normal variable by subtracting the expected value and dividing by the standard deviation [30],

$$Z = \frac{[\chi^2 - (N-1)]}{\sqrt{2(N-1)}}. \quad (39)$$

By the CLT, the normalized sum of Z converges to a Gaussian (normal) distribution with zero mean and unit variance. Therefore every sum of the terms in [.], after being converted to standard normal using Equation (39), can be approximated to a normal random variable as N is large. Also, since a linear combination of normal random variables tends to another normal variable, then the test statistic will follow a Gaussian distribution.

4.5 HYPOTHESIS TESTING: WHITENING TRANSFORM

Neighboring pixels are most likely to come from the reflectance of the same object. Therefore, there is a high statistical dependence between pixels in a natural image scene. A pre-processing method could be used to de-correlate pixels in an image and reduce the sensitivity of the sample covariance when there are outliers, or equivalent when the sample size is small.

The whitening transform explained in Section 3.7 converts a random column vector to another vector that has identity covariance matrix [31]. Similarly to the other calculations, the whitening transform is performed band by band. Letting X be each of the two-dimensional images, where the columns are variables and the rows are observations from this variables.

To be able to write the covariance matrix of X as $M = E[XX^T]$, the image X is centralized by subtracting it by its mean. A decorrelation transform can be found by finding the eigenvectors and eigenvalues of M by solving

$$M\Phi = \Phi \Lambda. \quad (40)$$

Λ is a diagonal matrix having the eigenvalues as its diagonal elements. Therefore, the matrix Φ diagonalizes the covariance matrix of X. The diagonalized covariance can be written as

$$\Phi^{\mathrm{T}} M \Phi = \Lambda. \tag{41}$$

To apply the diagonalizing transform to a vector of data, it can be written as

$$y = \Phi^{\mathrm{T}} X. \tag{42}$$

Since the covariance is now a diagonal matrix, the data in y is decorrelated. To whitening the data the diagonal elements of Λ need to be the same, which is obtained by

$\Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}} = I$, which after substituting in (41) the whitened data or image w is,

$$w = \Lambda^{-\frac{1}{2}} \Phi^{\mathrm{T}} X. \tag{43}$$

4.6 Parametric and Non-Parametric Hypothesis Tests

The previous hypotheses methods will be compared with the traditional parametric and non-parametric hypotheses tests that are used to determine when populations can be well modeled as coming from the same density. The drawback of some of these tests is that they are negatively affected by deviations from normality [32]. To test the assumption of normality, histograms and Q-Q plots of the populations can be calculated. A Q-Q plot displays a quantile-quantile plot of two samples. If they come from the same distribution, the plot will be linear.

Since these hypothesis methods are developed to test homogeneity in an area additional steps are needed to apply it to a hyperspectral image where multiple objects are present. Areas, that are desired to be classified as a single class, will be manually extracted from the image, and then the hypothesis will be tested for each area in the image. It is hypothesized that the homogeneity in some parts of the image will increase the accuracy of hard classification by reducing the within-class variance.

4.6.1 F-test

The F-test is designed to test if two populations have the same variance. This test is implemented by comparing the ratio of the two variances. If they are equal, the ratio of the variances will be 1. Therefore, the test statistic is F = s_1^2/s_2^2 where $s_1^2 > s_2^2$. The F hypothesis test for a two-tailed test is defined as:

Ho: ${\sigma_1}^2 = {\sigma_2}^2$, Ha: ${\sigma_1}^2 \neq {\sigma_2}^2$. The hypothesis that the two variances are equal is rejected when:

$$F < F_{\left(1-\frac{\alpha}{2},N_1-1,N_2-1\right)} \text{ or } F < F_{\left(\frac{\alpha}{2},N_1-1,N_2-1\right)}. \quad (44)$$

The assumptions made by this test are:

1. There are two samples from two populations
2. samples are independent
3. populations come from a normal distribution
4. both population variances are unknown

4.6.2 Homogeneity of variance tests

When two or more variances are being compared, a homogeneity of variance (HOV) test can be performed. Most of the time, this test is performed to verify the HOV assumption in an ANOVA test. One major drawback of these tests is that they are not good for detecting small or moderate differences in variances. Also, some tests start from the assumption of normality in the data, but this is not always met. Q-Q plots are used to verify normality. The multiple homogeneity of variance (HOV) tests are explained below [33].

4.6.2.1 Barlett's Test

Barlett's test for k samples is defined as Ho: ${\sigma_1}^2 = {\sigma_2}^2 = \cdots = {\sigma_k}^2$, Ha: ${\sigma_i}^2 \neq {\sigma_j}^2$ for at least one pair (i,j). Its test statistic is

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i-1)ln\, s_i^2}{1+(\frac{1}{3(k-1)})((\sum_{i=1}^{k}(N_i-1))-1(N-k))}. \quad (45)$$

The hypothesis that the variances are equal is rejected when $T > \chi^2_{1-\alpha,k-1}$.

4.6.2.2 Levene's test

Levene's test is an alternative to the Barlett's test that is less sensitive to the normality assumption. It assumes that the samples from the population are independent, and also that the populations under consideration are approximately normally distributed. The Levene test for k samples is defined as Ho: ${\sigma_1}^2 = {\sigma_2}^2 = \cdots = {\sigma_k}^2$, $Ha: {\sigma_i}^2 \neq {\sigma_j}^2$ for at least one pair (i,j). The Levene test statistic is defined as:

$$W = \sum_i n_i(\bar{\bar{z}}_i - \bar{z}..)\text{^}2\,/(g-1))/((\sum_i\sum_j\left(z_{ij} - \bar{\bar{z}}_i\right)^2)/\sum_i(n_i-1)), \quad (46)$$

where $Z_{ij} = \left|y_{ij} - \overline{y_i}\right|$, $\overline{y_i}$ is the mean for the ith treatment, $\overline{z_i} = \sum z_{ij}/n_i$, and $\overline{z..} = \sum\sum z_{ij}/\sum n_i$. The hypothesis that the variances are equal is rejected when $> F_{\alpha,k-1,N-k}$.

Levene test is robust because the true significance level is very close to the nominal for a large variety of distributions. Also, it is not sensitive to symmetric heavy-tailed distributions.

4.6.2.3 Brown-Forsythe Test

The Brown-Forsythe test uses the median instead of the mean. It provides good robustness for non-normal data while retaining good statistical power. Using the mean provides the best hypothesis power for symmetric, moderate-tailed distributions than using another parameter [32]. The Browne-Forsythe test for k samples is defined as:

Ho: ${\sigma_1}^2 = {\sigma_2}^2 = \cdots = {\sigma_k}^2$, $Ha: {\sigma_i}^2 \neq {\sigma_j}^2$ for at least one pair (i,j).

The test statistic used is the one-way ANOVA (analysis of variance) statistic, just as in the Levene test.

$$W = \sum_i n_i(\bar{\bar{z}}_i - \bar{z}..)\text{^}2\ /(g-1))/((\sum_i \sum_j (z_{ij} - \bar{\bar{z}}_i.)^2)/\sum_i (n_i - 1)), \quad (47)$$

where $Z_{ij} = |y_{ij} - \tilde{y}_i|$, where $\tilde{y}_i$ is the median for the ith treatment.

$$\bar{z}_i = \sum z_{ij}/n_i, \text{ and } \overline{z..} = \sum\sum z_{ij}/\sum n_i.$$

The hypothesis that the variances are equal is rejected when $W > F_{\alpha,k-1,N-k}$.

4.6.2.4 O'Brien test

The O'Brien test constructs a dependent variable so that the group means of the new variable equal the group sample variances of the original response. An ANOVA on the O'Brien variable is actually an ANOVA on the group sample variances [34].

4.6.3 Nonparametric tests

Many statistical tests are developed based on the assumption that the population is normally distributed. Nonparametric tests are designed for use when only a few statistical assumptions can be made. In the case of normality, the mean and variance are typically used to describe the center and spread of the population, but they are not robust enough when this assumption is not met. In the non-parametric tests the median is typically used to measure the center of a distribution if it is not heavily influenced by outliers and skewed data. The spread is harder to quantify, but it is usually represented by the interquartile range. This is the difference between the first and third quartiles.

4.6.3.1 Ansari-Bradley

Ansari-Bradley tests for differences in spread. It is a nonparametric alternative to the two-sample F-test for equal variances. It does not require a normality assumption and can be used with samples from distributions that do not have finite variances. One assumption made is that the samples have equal medians. Therefore, it is recommended to subtract the median from the samples to have them centralized around the median.

5 EXPERIMENTS AND RESULTS

Multiple hyperspectral images were used to test all the algorithms described in Sections 4.3, 4.4, and 4.5. First, synthetic images were used to verify the theoretical methodology without having the effect of real life uncertainties. These synthetic images were generated with different signal-to-noise ratios and with different probabilities of coming from two classes. To verify the applicability of the methods to real life, hyperspectral images gathered using the SOC-700 stand mounted were used.

The chapter is organized by experiments realized according to the methodologies explained in Sections 4.3 to 4.5. Synthetic images and real hyperspectral images were used to verify the methods. The images are different in the sense that some images are more homogenous than others in its original spatial resolution.

In the first experiment typical hypotheses tests were applied to determine when pixels in an image can be well modeled as coming from the same density. Secondly, experiments were performed to verify the hyperspectral images applicability to the initial hypothesis test developed in this research, where Ho: $Var(\tilde{X}) = \frac{\sigma^2}{m}$. Most importantly, Section 5.5 contains the modification of the previous hypothesis test, where an applicable statistical methodology was developed to determine a spatial resolution to improve image classification.

Finally, the method was validated by classifying images at different spatial resolutions and verifying that the spatial resolution obtained from the covariance hypothesis test (Section 5.5) improves the classification accuracy.

5.1 GENERATION OF SYNTHETIC IMAGES

5.1.1 Different pixel probability of belonging to a class

Synthetic images were generated by changing the probability of a pixel coming from one of the two classes present in the image. The image is composed with pure pixels where the endmembers used were functions in $\mathbb{Z}^N$. The two endmembers are shown in Figure 5-1, where the first endmember was generated using $\cos\left(\frac{4\pi n}{220}\right) + 1$ and the other $\cos\left(\frac{2\pi n}{220}\right) + 24$, for n bands. The idea behind these endmembers is to be able to test synthetic images were the classes had significantly different intensities.

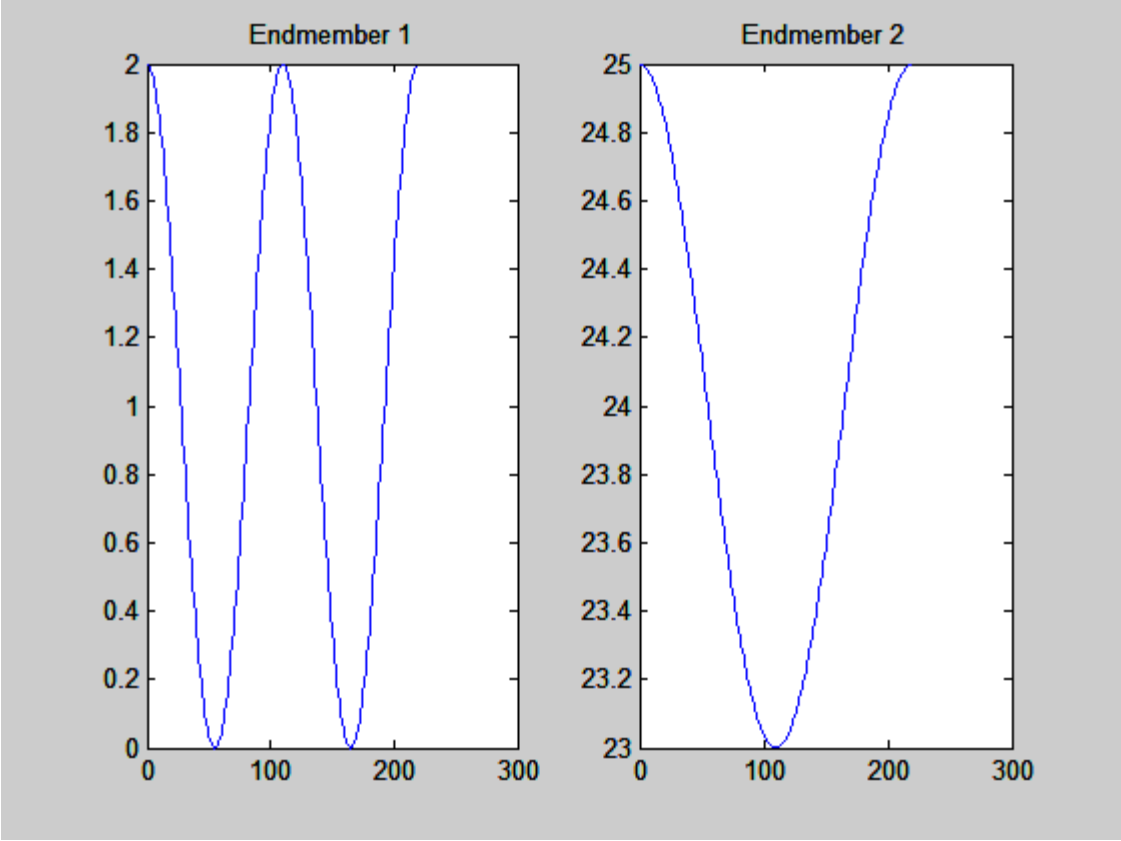


Figure 5-1 Synthetic images endmembers

The classes in the image are defined as class A and class B. Different images were generated by changing the probability of a pixel to belong to a class A. A Bernoulli distribution was used to spatially distribute pixels with success probability p, where in this case success means coming from class A.

Two images were generated, one with p=0.25 and the other with p=0.5. Endmembers one and two from figure 5-1 were spatially distributed using the Bernoulli distribution to generate the two 250x250x220 synthetic hyperspectral images shown in Figure 5-2.

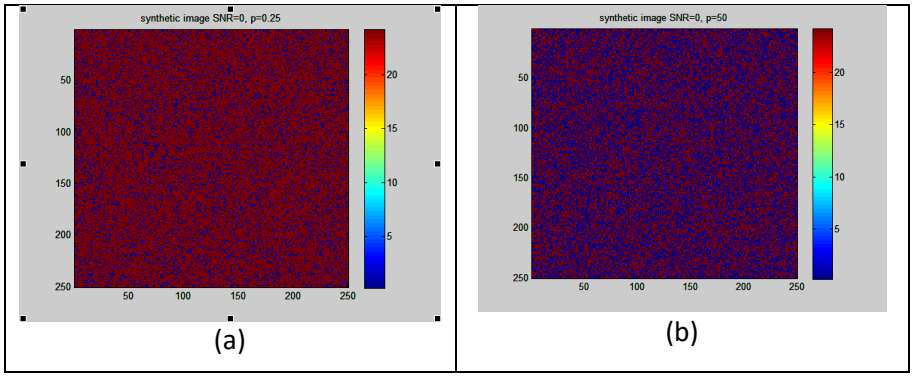


Figure 5-2 Hyperspectral synthetic images with different probabilities of coming from class A: (a) Image generated with p=0.25, (b) Image generated with p=0.5

Since the image (a) shown in Figure 5-2 was generated with p=0.25, it can be seen that the endmember that corresponds to red is more predominant in the image than blue. In contrast image (b) from the same figure has the same quantity of blue than red. To understand the characteristics of the image the histogram and isotropic variogram were calculated and shown in Figure 5-3.

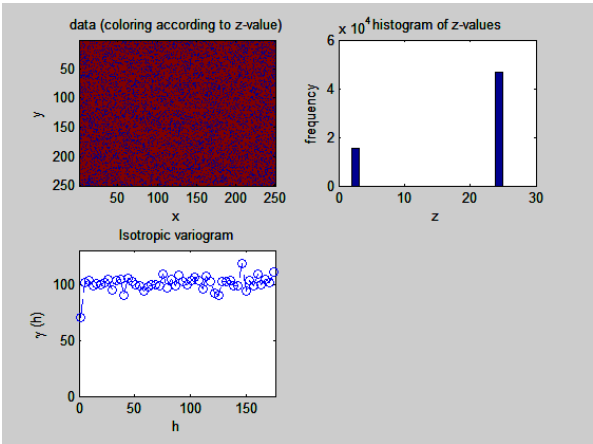


Figure 5-3 Synthetic image generated with p=0.5, with its histogram and isotropic variogram

The isotropic variogram in Figure 5-3 (shown bigger in Figure 5-4) is a graph of semivariance vs. separation distance. When there is autocorrelation present in the image, the semivariance is lower at smaller separation distances. In this case it is lower from 0 to 5 separation distance (h). The sill is the upper

limit of the variogram where all semivariances are invariant with the sample separation distance. Even though in this image the semivariance varies after h=20, the sill can be estimated to be around 100. The range of the variogram model is the separation distance at which samples are spatially autocorrelated, in this case it seems to be around h=8, but the semivariances as the separation distance h increases and does not follow a generalized variogram model as shown in Section 2.3.

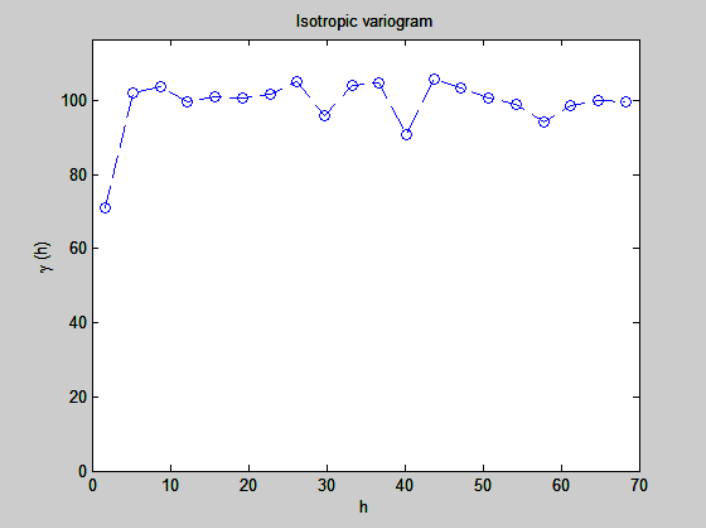


Figure 5-4 Variogram from image in Figure 5-12 (a)

The histogram in Figure 5-3 has two peaks indicating that the reflectance values from the two endmembers are widely separated. Therefore, the previous synthetic images do not simulate real life. Their histograms have two peaks which mean it is bimodal. Having two peaks in a histogram indicates that two processes with different distributions are present in the image. The two distributions in this case represent objects in the images, since these images were generated only with two endmembers. To generate synthetic images closer to real life events spatial and spectral features closer to real events need to be added to the image.

5.1.2 Legendre and Gaussian Abundances Synthetic Images

Spectral features are added to an image by using spectral signatures from the USGS Spectral Library. The spectra of lawn grass and chlorite shown in Figure 5-5 were used, where blue corresponds to

grass, and red to chlorite. The material reflectance were measured from 0.20 to 3.0 μm [35]. The spatial characteristics of the image are computed by abundance image statistics.

The new synthetic hyperspectral images are generated by using a linear combination of pure spectral signatures of endmembers. Let $E = [e_1, e_2]$ be the pure endmember spectral signatures of lawn grass and chlorite respectively. Using a linear model, each hyperspectral signature r at each pixel in the image is formed by the sum of the pixel's signal s and an independent additive noise component n, $r = s + n$. The spatial features of the image are added by using a fractional abundance matrix. Each abundance indicates the fraction of an endmember available at a pixel. The hyperspectral signature r at each pixel in an image with m endmembers can be expressed as,

$$r = s + n = \sum_{i=1}^{m} e_i \phi_i + n_i, \tag{18}$$

where ϕ is the m-dimensional vector of fractional abundances at the given pixel, which given its definition above is subject to constraints: $\phi_i \geq 0$, and $\sum_{i=1}^{m} \phi_i$=1.

The generation of the abundance coefficients is a spatial process performed independently for each desired endmember, and after the abundances matrices are generated normalization conditions are imposed independently for each pixel.

To ensure endmembers are distributed in the image similar to real life the Legendre polynomials and Gaussian fields are used to randomly generate the abundances with different spatial distributions. To reduce the time in developing the algorithms a Matlab Synthesis tools package called HYDRA was used.

HYDRA was developed by the Computational Intelligence group of the Basque Country University. It a set of tools focused on the development of computational methods for the analysis of hyperspectral images, and it contains a set of methods to make synthetic hyperspectral images [36]. HYDRA is an open

source project under the GNU GPL3 license. It provides a GUI (Graphical User Interface), which was used to develop the synthetic images using the USGS spectral signatures.

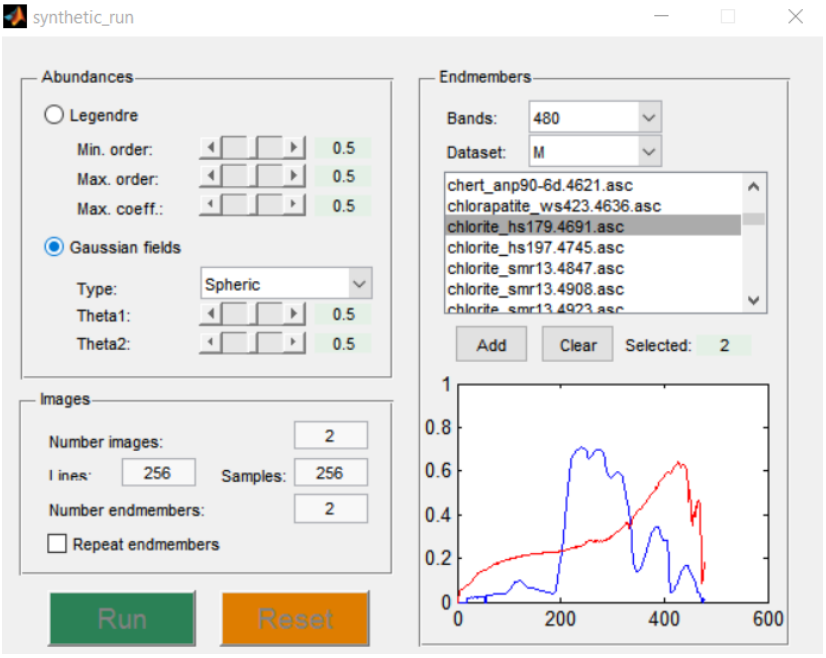


Figure 5-5 Generation of Gaussian Field Synthetic Image using HYDRA GUI

As shown in the endmembers, the final image can have up to 480 bands, but in this research there is no need for so many bands. The final images have 151 bands using bands from 250 to 400.

The images generated using the Legendre and Gaussian abundances methods are shown below. The differences in colors indicate the different intensities of the reflectance coming from the different endmembers present in the image. Even though there are only two endmembers in the image Gaussian additive noise was added to the images causing slight changes in the values.

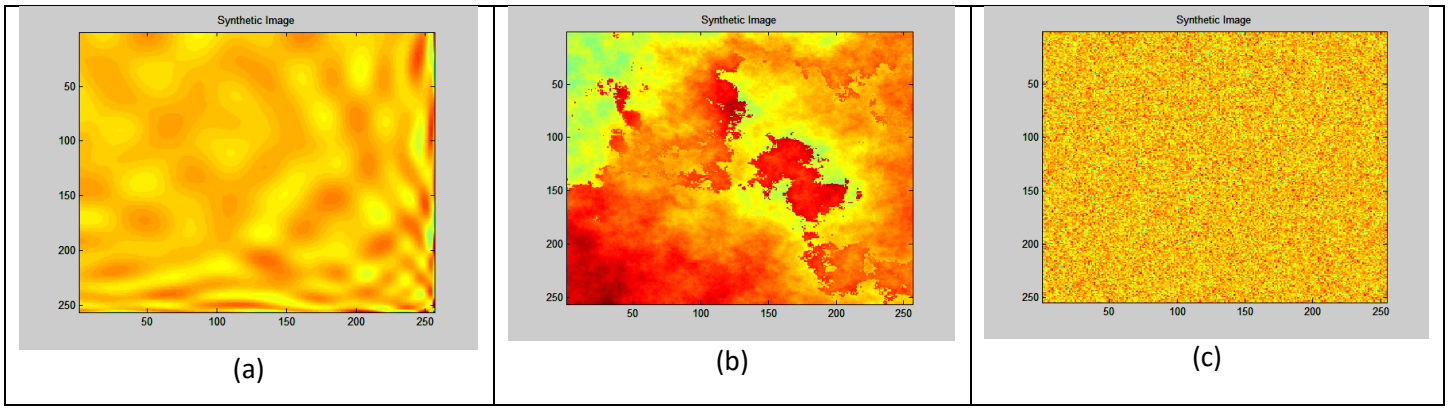


Figure 5-6 Legendre and Gaussian Fields Synthetic Images: (a) Synthetic Image generated using Legendre abundance, (b) Synthetic Image generated using Spherical Gaussian Fields abundance, (c) Synthetic Image generated using Rational Gaussian Fields abundance

The figure below shows the histogram of the Legendre synthetic image. This histogram indicates that the pixel values in the image follow a Gaussian distribution, which is typical in hyperspectral images. Therefore, this images generated by using abundance image statistics have a more realistic characteristic than the previous synthetic images.

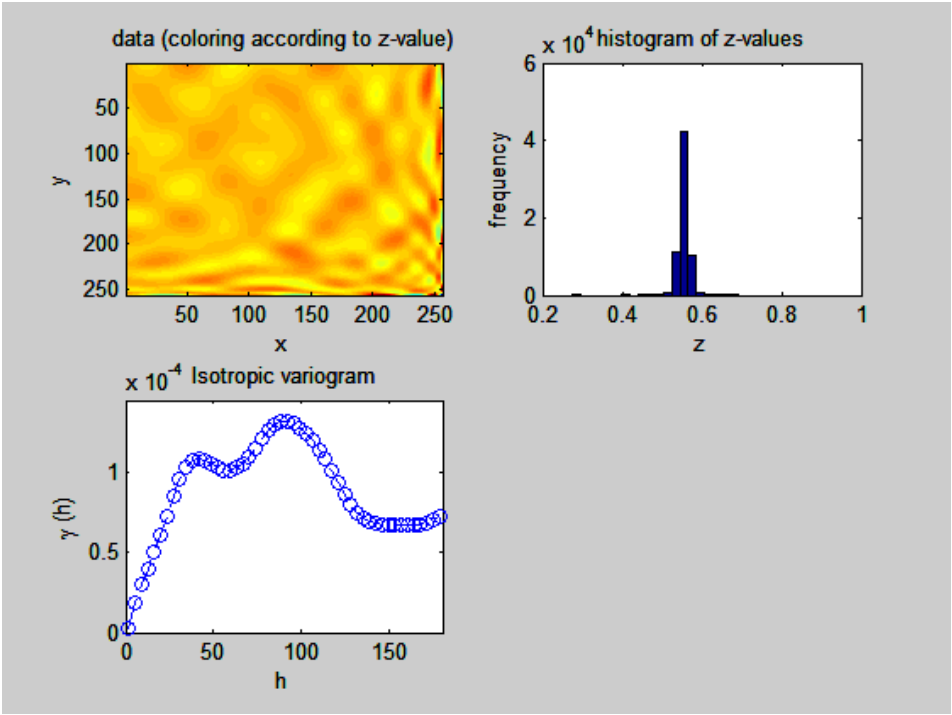


Figure 5-7 Legendre synthetic image, with its histogram and isotropic variogram

Sections in the rest of the chapter show the results from all the experiments realized with all the synthetic images shown above, and also for real images. The typical parametric and non-parametric hypotheses tests are shown first (methodology shown in Section 4.6), and then the results from the two hypotheses tests developed in this research (methodology explained in Sections 4.3 and 4.4).

5.2 PARAMETRIC AND NON-PARAMETRIC TESTS

As explained in Section 4.6 a hypothesis test can be performed to determine when populations can be well modeled as coming from the same density. Samples were taken from the image downsampled using window sizes from n=1 from n=20. Each hypothesis tests if two independent samples come from normal distributions with the same variance. To apply this concept to the hyperspectral images, a random sample vector is taken from the downsampled image at n equal to some factor in the range from 1 to 20. The second sample came from downsampling at n equal to factor+1 and randomly selecting a sample from this new coarser image. Since the hypothesis tests when populations can be well modeled as coming from the same density when the null hypothesis cannot be rejected is an analog to not been able to reject that the images downsampled at n or n+1 the pixels in the image became homogenous. This process was performed from each of the bands in the image.

The different parametric and non-parametric tests explained in Section 4.6 were used to test the hypothesis that populations come from the same density. A significance level of $\alpha = 0.05$, for a 5% of significance was used. Most of the hypothesis test gave results where the null hypothesis was always rejected. The only hypothesis test that gave significant results were the F-test and the Ansari Bradley tests. The Tables 5-1 to 5-5 shows numbers of hyperspectral bands where the null hypothesis was not rejected for the different downsampling window sizes, for each image shown in Figure 5-8.

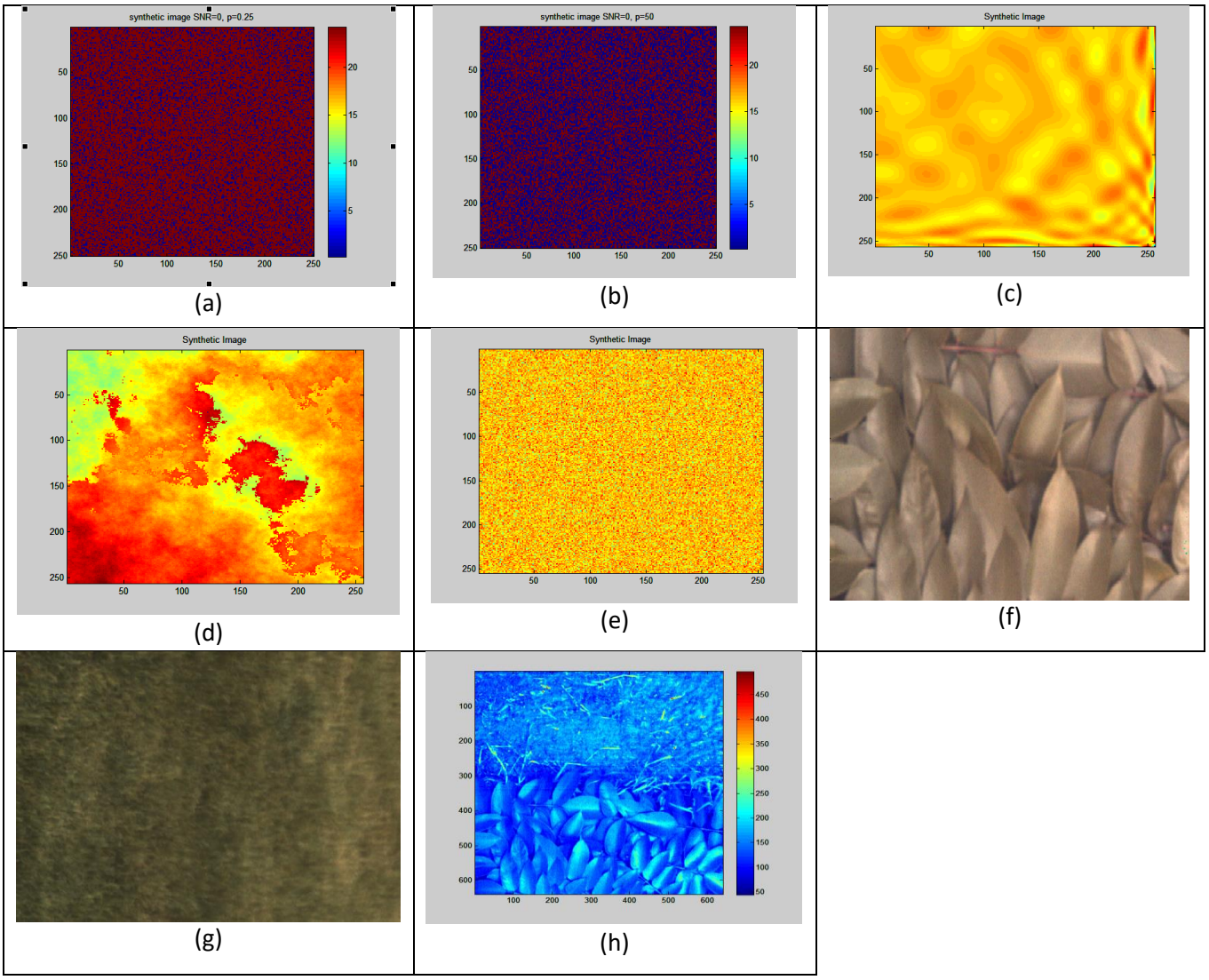


Figure 5-8 Hyperspectral Images used in the experiments, synthetic images from a-e, and real images from f-h: (a) Synthetic Image generated Image generated with p=0.25, (b) Image generated with p=0.5, (c) synthetic image generated using Legendre abundance, (d) Synthetic Image generated using Spherical Gaussian Fields abundance, (e) Synthetic Image generated using Rational Gaussian Fields abundance,(f) leaves real hyperspectral image shown in natural color RGB, (g) lawn grass/soil real hyperspectral image shown in natural color RGB, (h) soil/leaves real hyperspectral image band 45.

The results of the F-test and Ansari-Bradley hypothesis tests are shown below in this section. The windows sizes to be compared with other methods are highlighted in blue. The null hypothesis for the F-test is that the two independent samples come from a normal distribution, with the same variance, while the alternative hypothesis is that they come from normal distributions with different variances. The null hypothesis for the non-parametric test Ansari-

Bradley is that the two independent samples come from the same distribution. The alternative

hypothesis is that samples have same median and shape, but different dispersions (for example variance).

The first hypotheses results shown in Tables 5-1 and 5-2 correspond to the synthetic images generated by selecting the pixel p probability of belonging to a class A that belongs to one of the two endmembers.

Table 5-1 Results from the Ansari-Bradley and F-test for image with p=0.25

	Ansari-Bradley	F-test
Windows size n	Total bands with true hypothesis (Image bands:220)	Total bands with true hypothesis (Image bands:220)
1	220	220
2	220	220
3	220	220
4	220	220
5	220	220
6	220	220
7	220	0
8	220	220
9	220	220
10	0	0
11	0	0
12	0	220
13	0	220
14	0	0
15	220	0
16	0	220
17	0	0
18	0	220
19	0	220
20	0	220

Table 5-2 Results from the Ansari-Bradley and F-test for image with p=0.50

Image bands:220	Ansari-Bradley	F-test
Windows size n	Total bands with true hypothesis (Image bands:220)	Total bands with true hypothesis (Image bands:220)
1	220	220
2	220	220
3	220	220
4	220	220
5	220	220
6	220	220
7	0	0
8	220	220
9	220	220
10	0	0
11	0	0
12	0	220
13	220	220
14	0	0
15	0	0
16	0	0
17	0	0
18	0	220
19	0	220
20	220	220

Both tests gave similar results, were there were 0 bands with true null hypothesis or 220 bands with true null hypothesis. This could be due to how the images were generated. Each pixel has the same probability in all the bands, and the image is noiseless. To understand how these hypotheses act with synthesis images with additive Gaussian noise, the images (c) to (e) from Figure 5-8 were put under test. The results are shown in Tables 5-3 to 5-5.

Table 5-3 Results Ansari-Bradley and F-test for image with Legendre abundances (c)

	Ansari-Bradley (Image bands:151)	F-test (Image bands:151)
Windows size n	Total bands with true hypothesis	Total bands with true hypothesis
1	0	38
2	0	151
3	151	151
4	0	151
5	0	151
6	0	122
7	0	151
8	151	151
9	0	151
10	0	151
11	0	151
12	0	151
13	0	151
14	0	151
15	151	151
16	0	151
17	0	3
18	0	151
19	0	151
20	0	151

Table 5-4 Results Ansari-Bradley and F-test for image with Spherical Gaussian abundances (d)

	Ansari-Bradley	F-test
Windows size n	Total bands with true hypothesis (Image bands:151)	Total bands with true hypothesis (Image bands:151)
1	151	151
2	26	1
3	0	1
4	0	1
5	0	1
6	0	1
7	0	1
8	0	0
9	0	1
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0

Table 5-5 Results Ansari-Bradley and F-test for image with Rational Gaussian abundances (e)

Image bands:151	Ansari-Bradley	F-test
Windows size n	Total bands with true hypothesis. (Image bands:151)	Total bands with true hypothesis (Image bands:151)
1	151	151
2	151	151
3	151	151
4	151	151
5	151	151
6	151	151
7	151	151
8	0	1
9	68	151
10	0	2
11	0	1
12	0	151
13	0	150
14	151	151
15	0	3
16	0	1
17	0	151
18	0	151
19	0	54
20	0	1

The results from the abundance synthetic images (tables above) show more differences between the hypotheses tests. The Ansari-Bradley test rejected the null hypothesis more frequency than the F-test. At first glance, it may seem an assumption that image (c) is more homogenous than (d) and (e), and also (e) is to be more homogenous than (d). The results from the F-test seem to follow this assumption, since for most of the pixel sizes, the null hypothesis could not be rejected. Image (d) gave results where the hypothesis could not be rejected only for finer pixel sizes. Image (e) gave multiple ranges of pixel sizes where the null hypothesis could not be rejected.

These hypotheses tests were also used for the real hyperspectral images shown in Figure 5-8. The results for these images are shown below in Table 5-6.

Table 5-6 Results Ansari-Bradley and F-test for real hyperspectral images (from left to right) shown in Figure 5-8: (f) leaves, (g) grass/soil and (h) dirt/leaves

bands 120	Ansari-Bradley	F-test (Image bands: 120)
n	bands with true hypothesis	bands with true hypothesis
1	114	120
2	7	41
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	0	1
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	12	0
19	0	0
20	0	0

bands 120	Ansari-Bradley	F-test
n	bands with true hypothesis	bands with true hypothesis
1	120	81
2	16	17
3	32	25
4	22	7
5	1	0
6	0	0
7	0	72
8	4	63
9	45	12
10	49	63
11	0	0
12	0	65
13	36	0
14	0	0
15	0	50
16	0	0
17	0	0
18	0	0
19	42	0
20	31	0

bands 120	Ansari-Bradley	F-test
n	bands with true hypothesis	bands with true hypothesis
1	64	64
2	52	52
3	38	38
4	0	0
5	13	13
6	0	0
7	1	1
8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	5	5
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	220	0

The results using the leaves image indicate that pixels in the image cannot be determined to be identically distributed at neither resolution different from its original. Meanwhile, the results using the

grass/soil and dirt/leaves images indicate that at a few spatial resolutions, pixels can be determined to be identically distributed.

The other parametric and non-parametric hypotheses tests were omitted since the null hypothesis was rejected for all window sizes n. Most of the hypothesis tests are dependent on the normality assumption of the data. To verify how the hyperspectral images follow this assumption, Q-Q (quantile to quantile) plots were generated for the hyperspectral synthetic image with p=0.25. Both samples (the one on the left and the one on the right of Figure 5-9) were plotted against a Gaussian distribution shown in dashed lines in red. If the sample comes from the same distribution, the plot will be linear. In this case the plot seems to be linear, but skewed at the first and last quantiles.

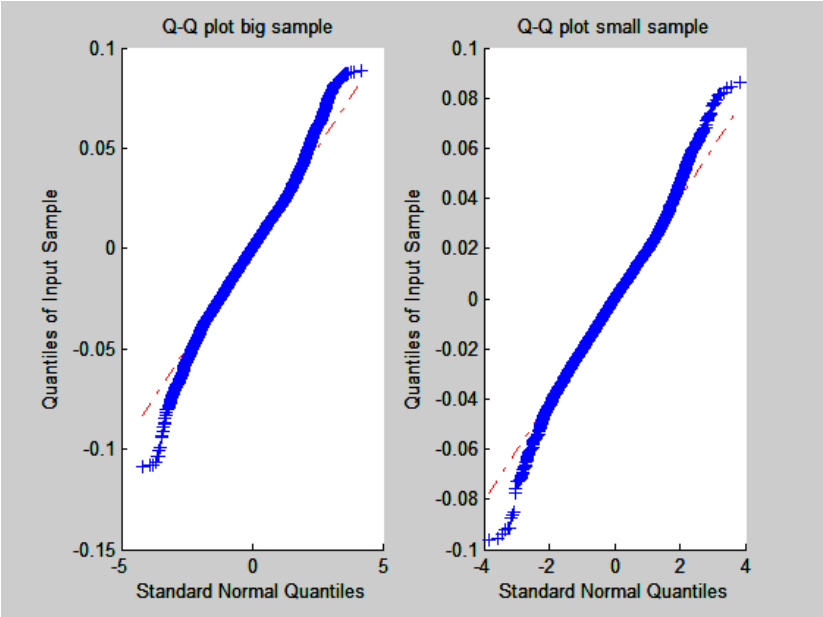


Figure 5-9 Q-Q plots from two samples (shown in blue) from the populations of the synthetic image with p=0.25 against a Gaussian distribution (red).

5.3 SEMIVARIOGRAM

To verify results with another method from the literature, the semivariogram explained in the Literature Review Section 2.3 is used. According to multiple research papers such as [2], [10], spatial resolution used should be finer than the spatial resolution at which the maximum in local variance in the object of interest occurs; and the local variance can be estimated by using the semivariance at a lag of one pixel.

The semivariogram is suited for 2-dimensional data but not n-dimensional. The literature does not provide an alternative for include a spectral dimension. Therefore, the analysis has to be done band by band. Another disadvantage of using the semivariogram is the need to fit a theoretical model which depends on the specific application is being used and involves complex algorithms. A typical semivariogram model fitting used for hyperspectral images is the spherical model, but is not unique. Here, the spherical model was used, and to include the spectral component in the analysis several bands were analyzed independently. The different semivariogram at one lag were plotted for the real images and are shown in Figures 5-10 to 5-12.

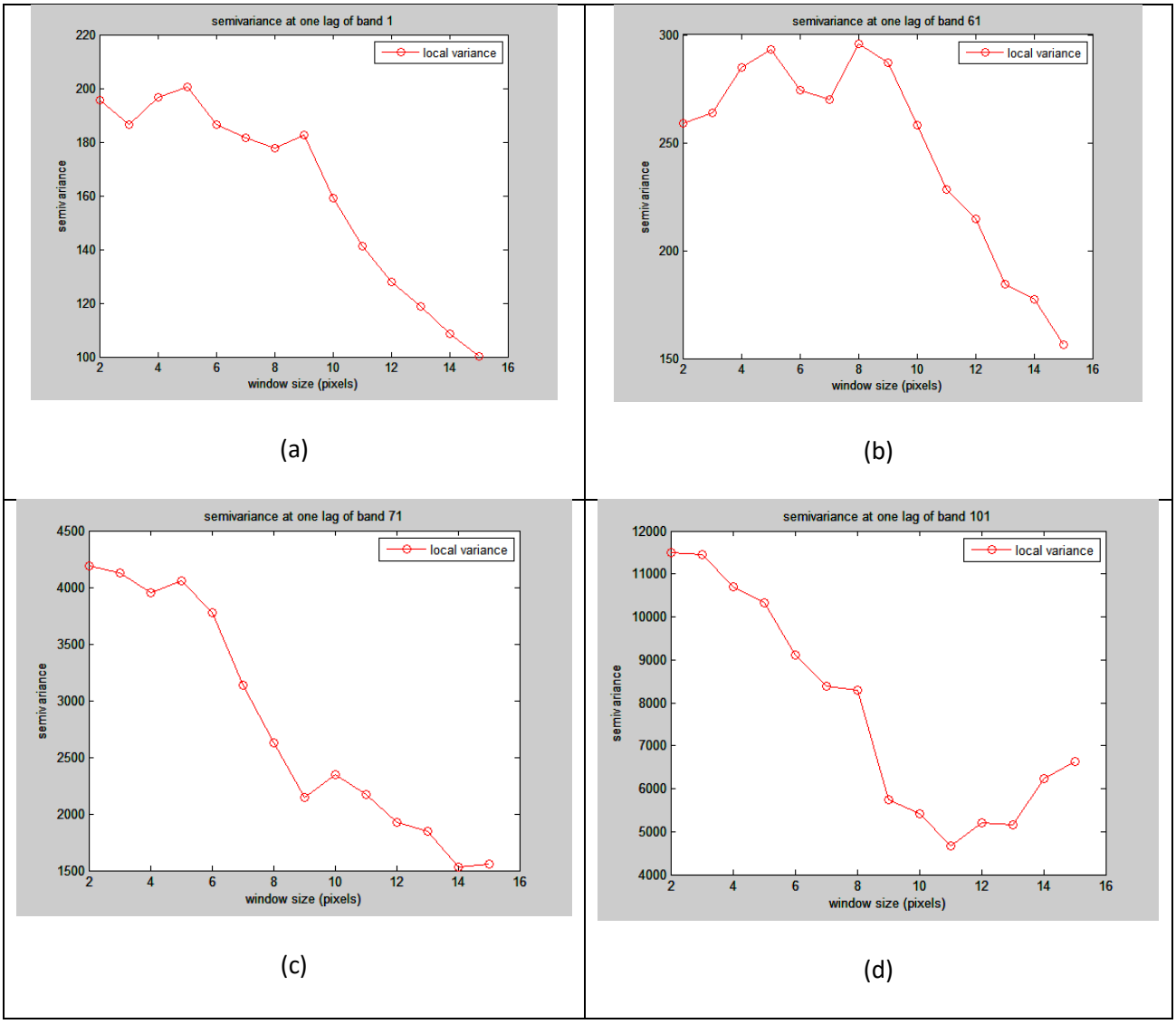


Figure 5-10 Semivariograms from the different spectral bands from the leaves image: (a) band 1, (b) band 61,(c) band 71, (d) band 101

Image classification accuracy can be improved by selecting a finer spatial resolution than the spatial resolution at which the local variance is maximum. Maximum local variance is shown in the plots as the maximum semivariance at a lag of one pixel. According to the semivariograms of the leaves image shown in Figure 5-10, the local variance has local maximums at a window size of n=5 for all bands. Therefore, according to this methodology classification accuracy can be improved by classifying the leaves image at the spatial resolution reached when downsampling at a 5x5 window size.

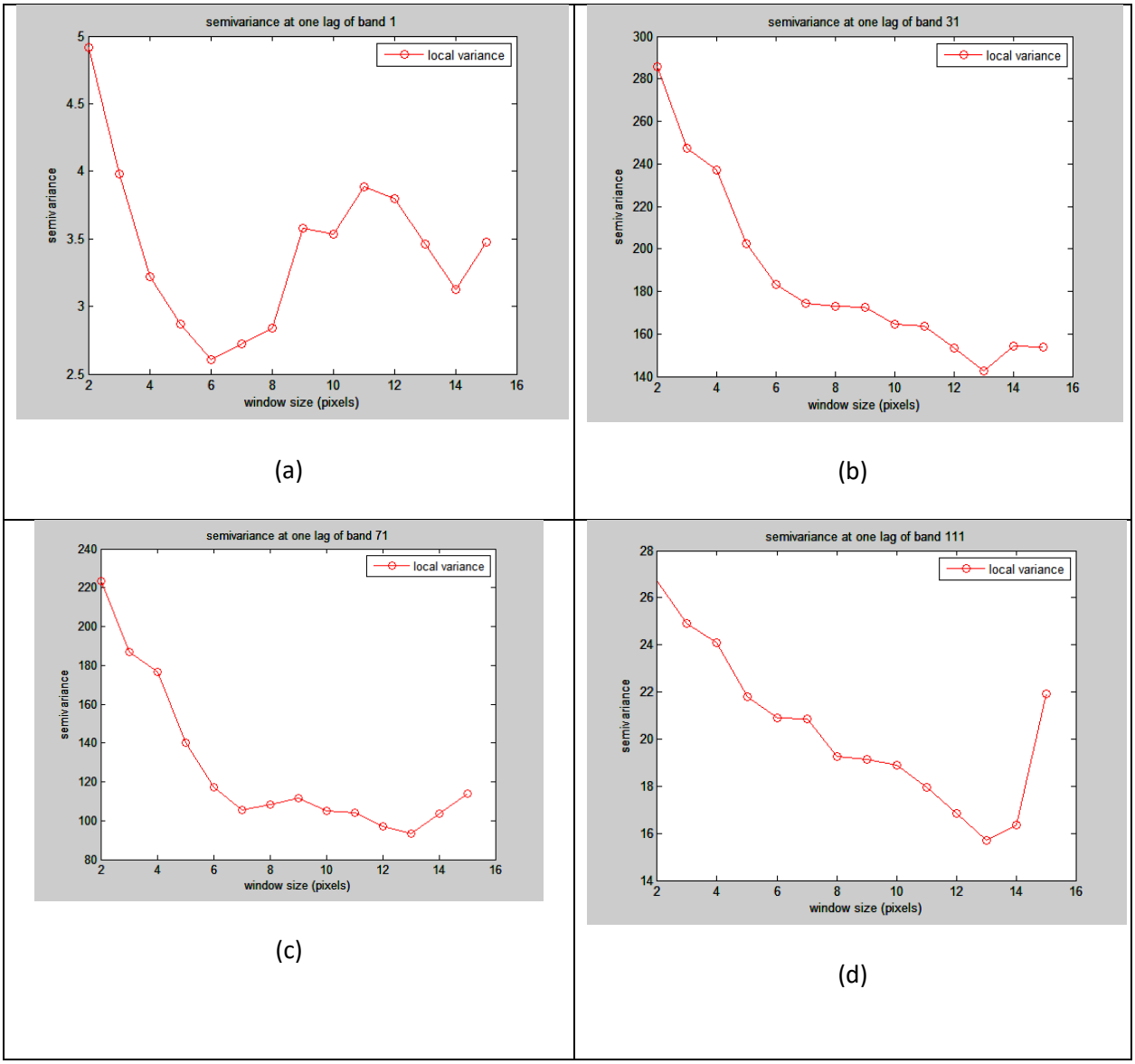


Figure 5-11 Semivariograms from the different spectral bands from the grass/soil image: (a) band 1, (b) band 31,(c) band 71, (d) band 111

For the grass/soil image, the semivariogram has its maximum at n=2 at bands 1,31,71, and 111 giving a consistent result. Hence, the grass/leaves image classification accuracy can be improved by classifying the image at the spatial resolution reached when downsampling at a 2x2 window size.

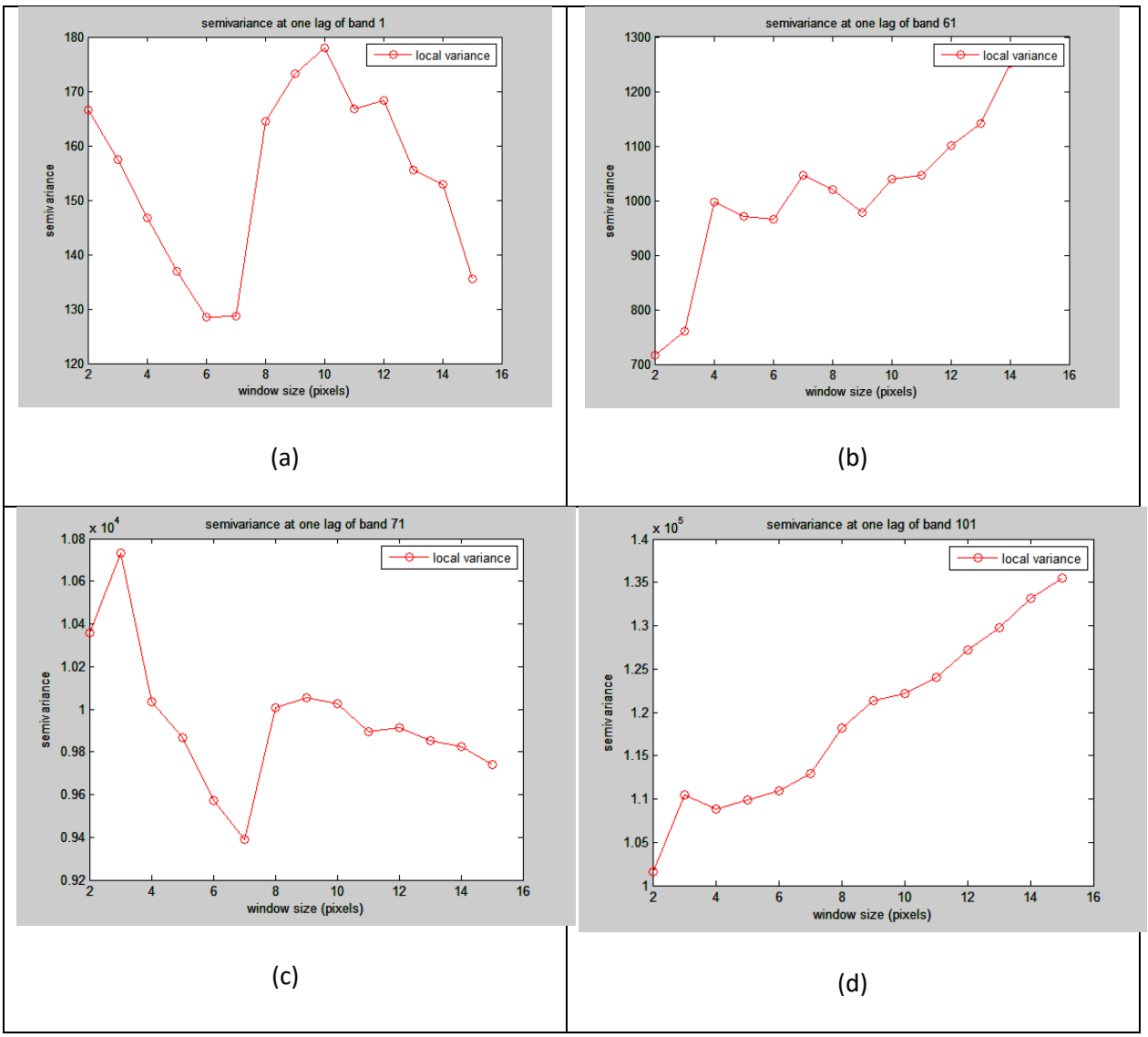


Figure 5-12 Semivariograms from the different spectral bands from the dirt/leaves image: (a) band 1, (b) band 61,(c) band 71, (d) band 101

For the dirt/leaves image, the semivariogram has its maximum at n=9 at band 1, n=15 at band 61, n=3 at band 71, and n=15 at band 101. The most common value between these results is the window size n=15, but the values are not equal and a determination of an optimum pixel size cannot be obtained.

5.4 HYPOTHESIS TESTING: VARIANCE METHOD RESULTS

A hypothesis test was developed to determine if the pixels in an image were identically distributed or not, by using the variance. The method is further explained in Section 4.3.

A hypothesis test with null hypothesis Ho: $Var(\tilde{X}) = \frac{\sigma^2}{n}$, and alternative hypothesis Ha: $Var(\tilde{X}) \neq \frac{\sigma^2}{n}$, were used to test if the pixels X_i are independent and identically distributed random variables. To verify this hypothesis test the left and right side of the Equation (29) from Section 4.3 was plotted for each image.

$$Var(\tilde{X}) = \frac{\sigma^2}{n}, \text{ for all n >0.} \tag{29}$$

If pixels can be determined as coming from independent identically distributed variables, then the residuals from the left and right side of the null hypothesis will be close to zero which is equivalent to not been able to reject the null hypothesis stated above. The hypothesis was tested in the eight images and the results are shown in Figure 5-13 to 5-20.

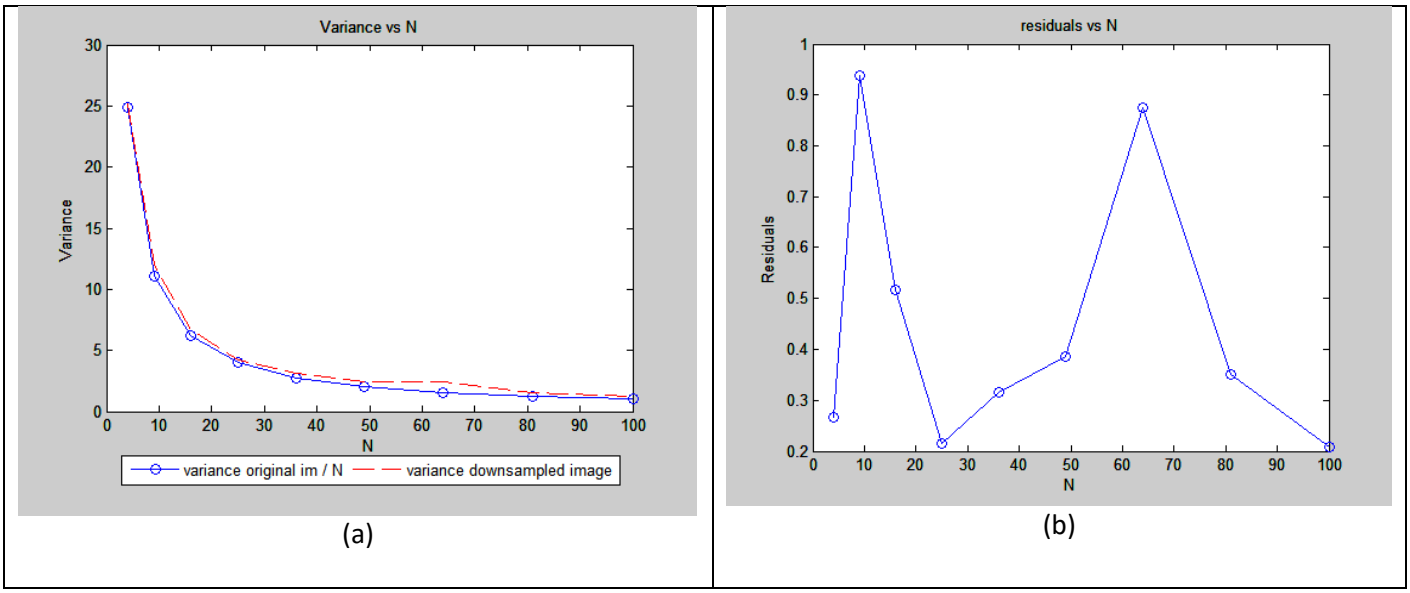


Figure 5-13 Variance method hypothesis test for synthetic image with p=0.25: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

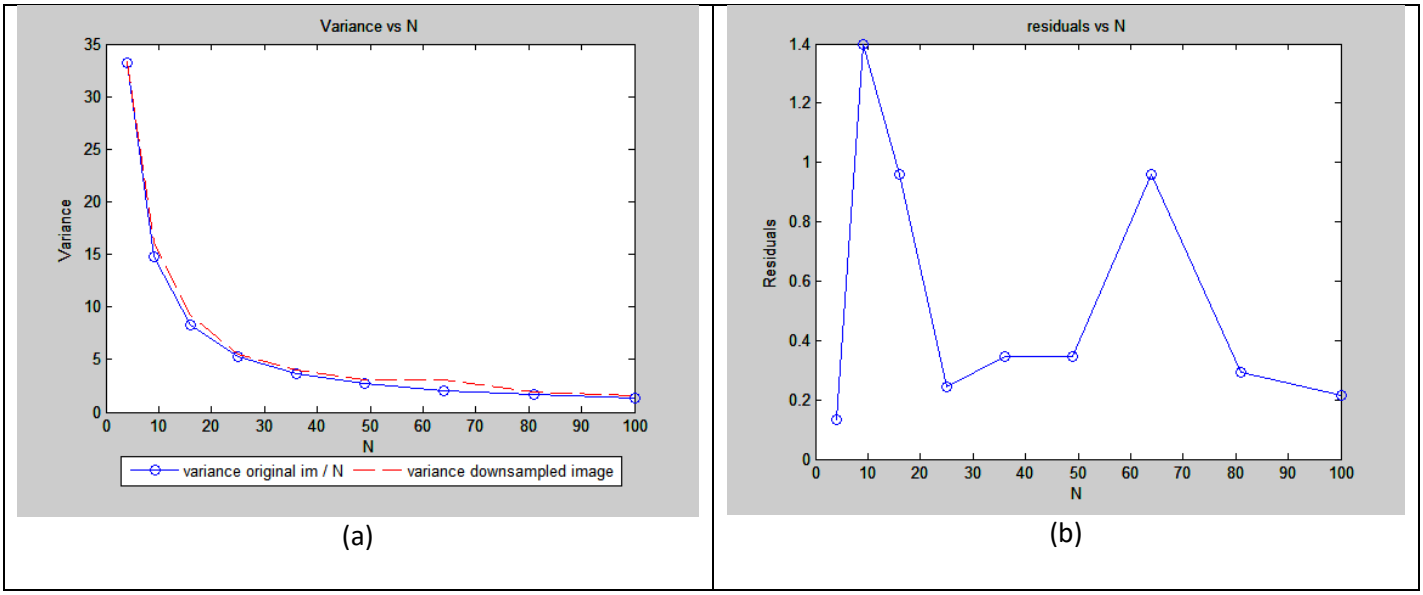


Figure 5-14 Variance method hypothesis test for synthetic image with p=0.50: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

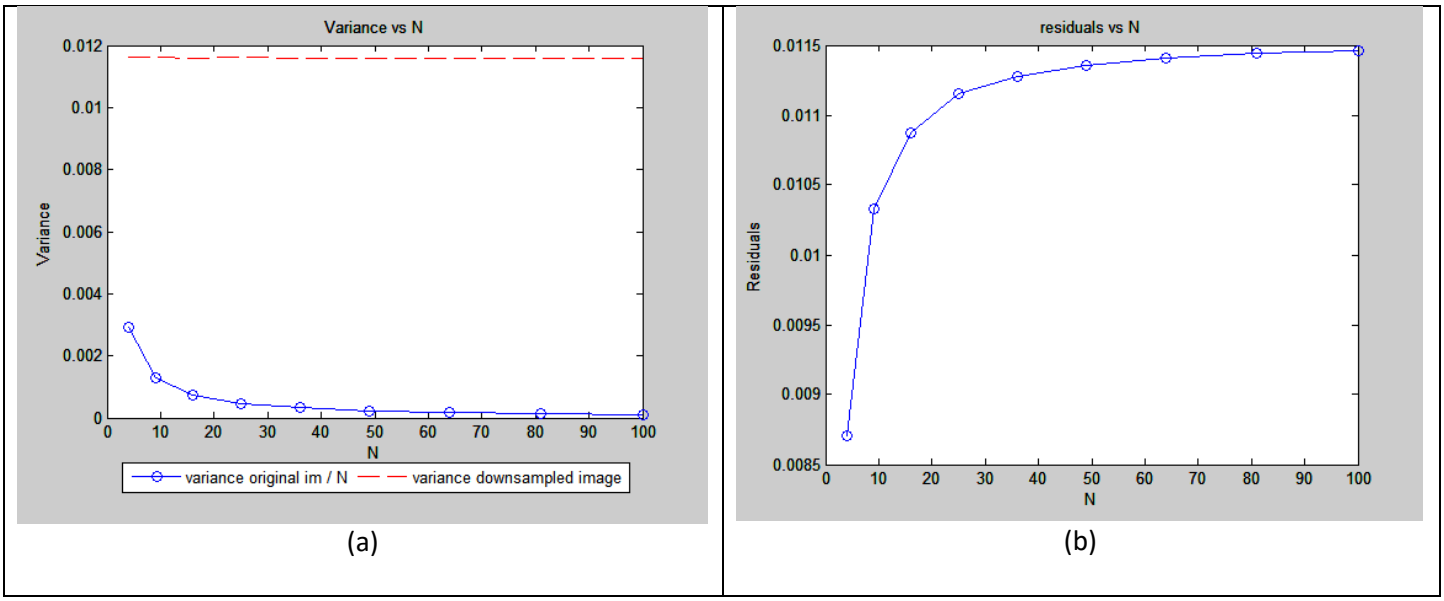


Figure 5-15 Variance method hypothesis test for synthetic image with Legendre abundances: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

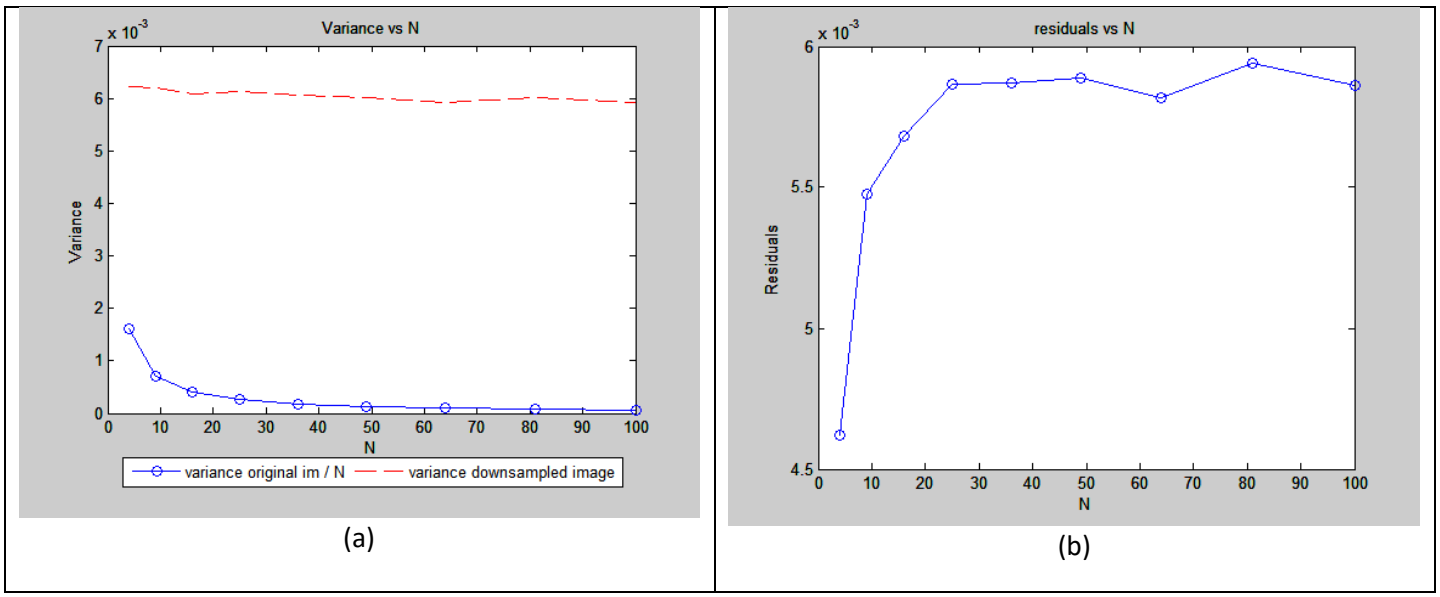


Figure 5-16 Variance method hypothesis test for synthetic image with Spherical Gaussian Fields abundances: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

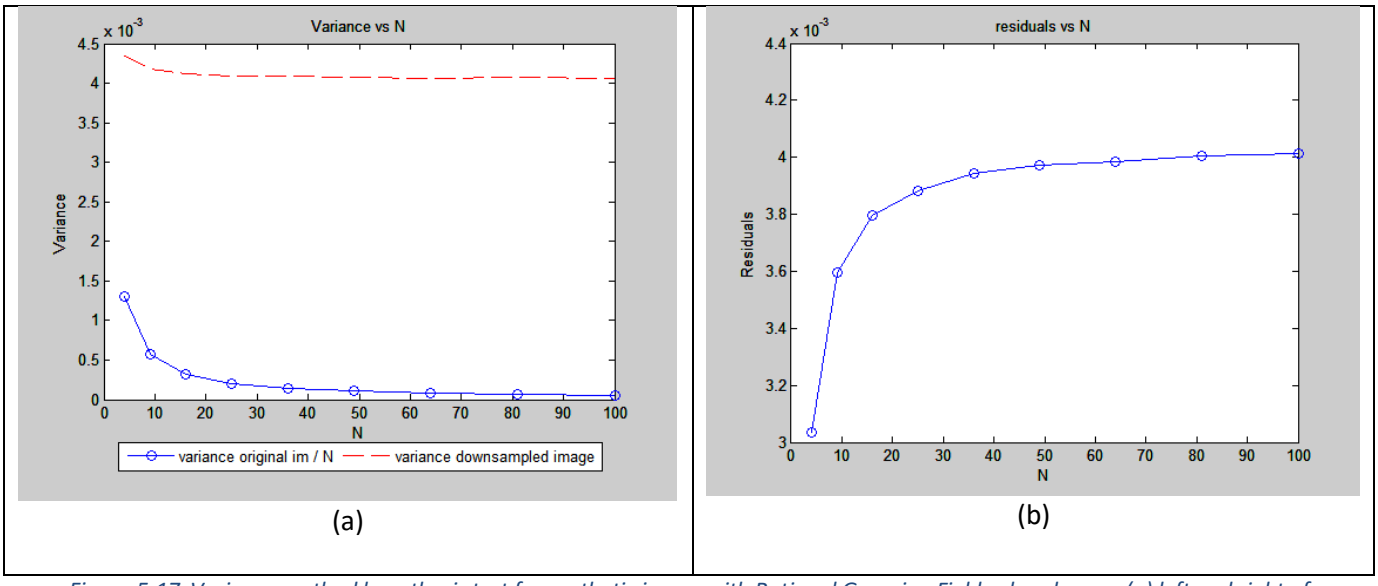


Figure 5-17 Variance method hypothesis test for synthetic image with Rational Gaussian Fields abundances: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

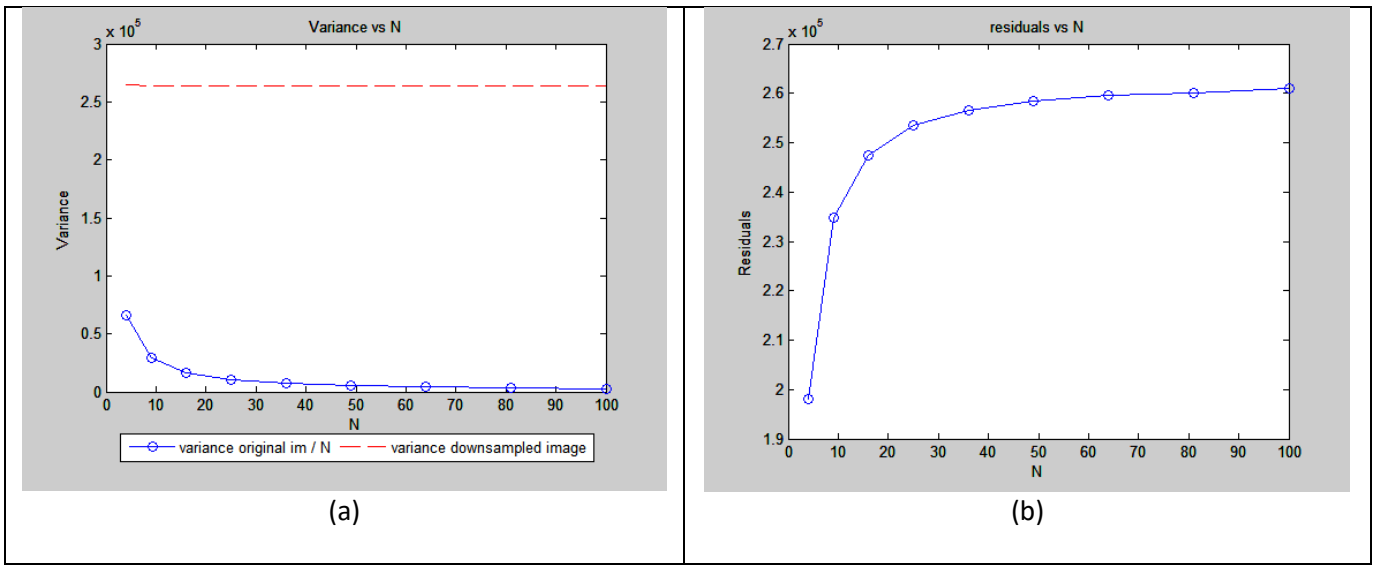


Figure 5-18 Variance method hypothesis test for leaves image: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

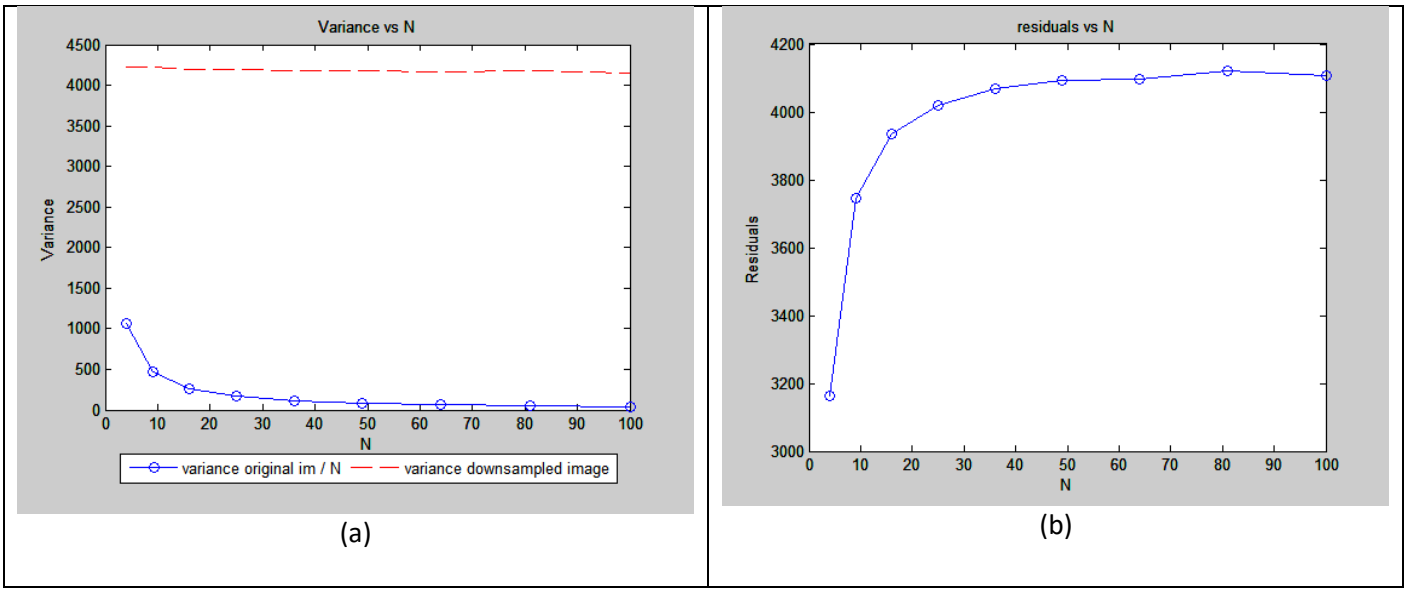


Figure 5-19 Variance method hypothesis test for the grass/soil image: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

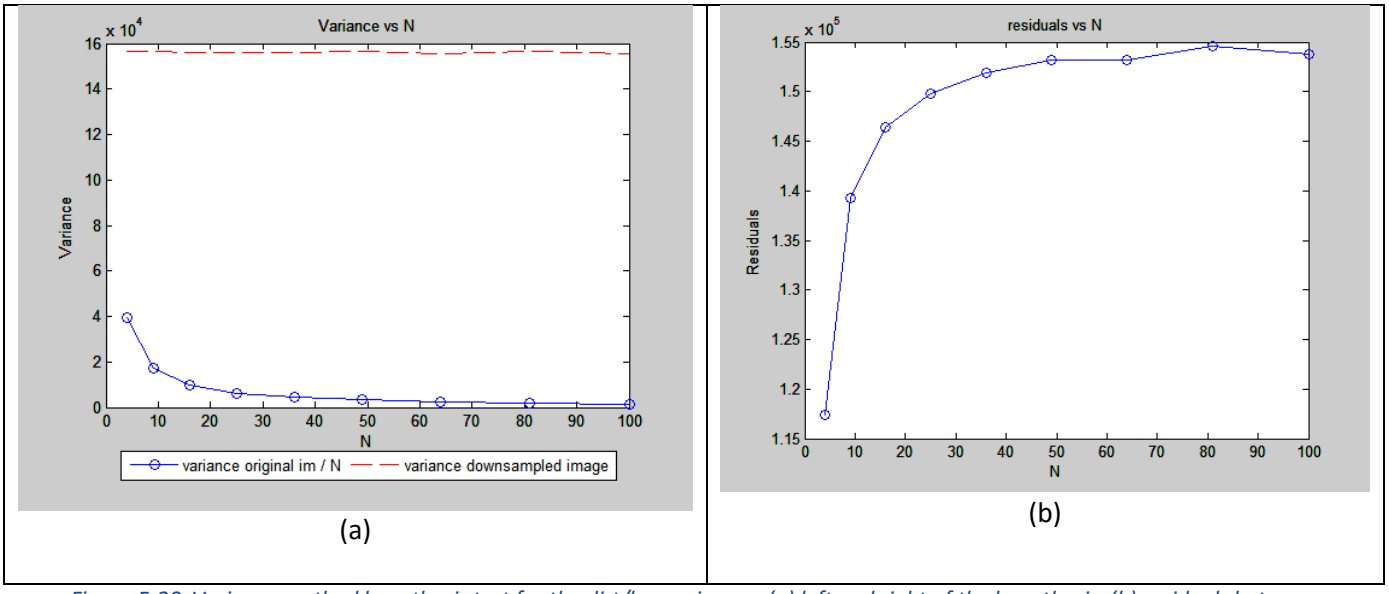


Figure 5-20 Variance method hypothesis test for the dirt/leaves image: (a) left and right of the hypothesis, (b) residuals between the left and right side of the hypothesis

The null hypothesis Ho: $Var(\tilde{X}) = \frac{\sigma^2}{n}$ was rejected for all images except for the first synthetic images, as shown in Figure 5-13 and Figure 5-14. The pixel values in these synthetic images were previously demonstrated (Section 5.1.1) to have a bimodal distribution indicating that the reflectance values from the two endmembers are widely separated. Therefore the pixels in these two images can be easily separated to be determined as coming from independent identically distributed variables.

Since all the other synthetic and real images had residuals (of the left and right side of the null hypothesis equation) on the order of hundreds, the null hypothesis was rejected. All results indicate that this hypothesis is too sensitive to images where the endmembers cannot be widely separated (which is most of the time). A modification to the hypothesis was made by changing the assumption from the pixels X_i being independent and identically distributed random variables, to only being identically distributed.

5.5 HYPOTHESIS TESTING: COVARIANCE METHOD RESULTS

The next hypothesis test developed in this research was tested to validate its usability. This hypothesis test determines if the pixels before averaging were identically distributed or not, by using the variance and covariance. The method is further explained in Section 4.4.

A hypothesis test with null hypothesis Ho: $m^2Var(\tilde{x}) - mVar(x_i) - 2\sum_{\{\{i,j\}:i<j\}} Cov(x_i, x_j) = 0$, and alternative hypothesisHa: $m^2Var(\tilde{x}) - mVar(x_i) - 2\sum_{\{\{i,j\}:i<j\}} Cov(x_i, x_j) \neq 0$, were used to test if the pixels X_i are identically distributed random variables. In overview, the difference stated in Equation (33) from Chapter 4,

$$m^2Var(\tilde{x}) - mVar(x_i) - 2\sum_{\{\{i,j\}:i<j\}} Cov(x_i, x_j) \equiv 0 \tag{33}$$

is used as the statistic in a hypothesis testing problem. If the null hypothesis cannot be rejected, the pixels in the original image are identically distributed.

The methodology explained is applied spatially on the image, to include the spectral component of the hyperspectral images the residual sum of squares (RSS) of the statistic in (33) is applied to each band. Let each band be called b and the total of bands be k such that,

$$\text{RSS} = \sum_{b=1}^{k}\left(n^4 Var(\tilde{x}_b) - mVar\left(x_{i,b}\right) - 2\sum_{\{\{i,j\}:i<j\}} Cov\left(x_{i,b}, x_{j,b}\right)\right)^2 \quad (34)$$

This was performed on each image, and the plot of RSS vs. factor (window size for downsampling) was plotted to visualize the results more easily. The hypothesis was tested using the eight images and the results are shown in Figure 5-21 to 5-28.

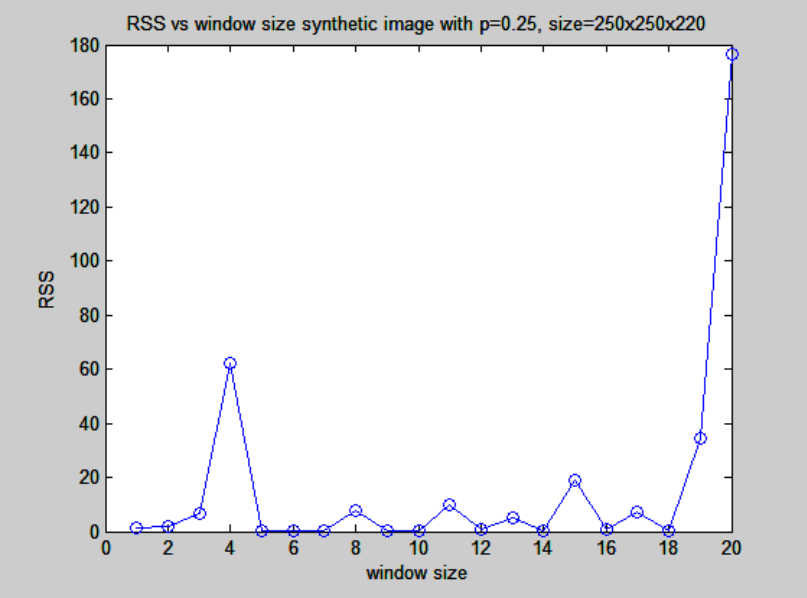


Figure 5-21 Residual Sums of Squares vs. window size of the p=0.25 synthetic image

The p=0.25 synthetic image was tested using the covariance hypothesis method. The RSS vs. window size plot in Figure 5-21 indicate high RSS at windows sizes, n=4,8,11,15,17,19,20. Therefore, the null hypothesis could be rejected indicating pixels are not identically distributed at these spatial resolutions.

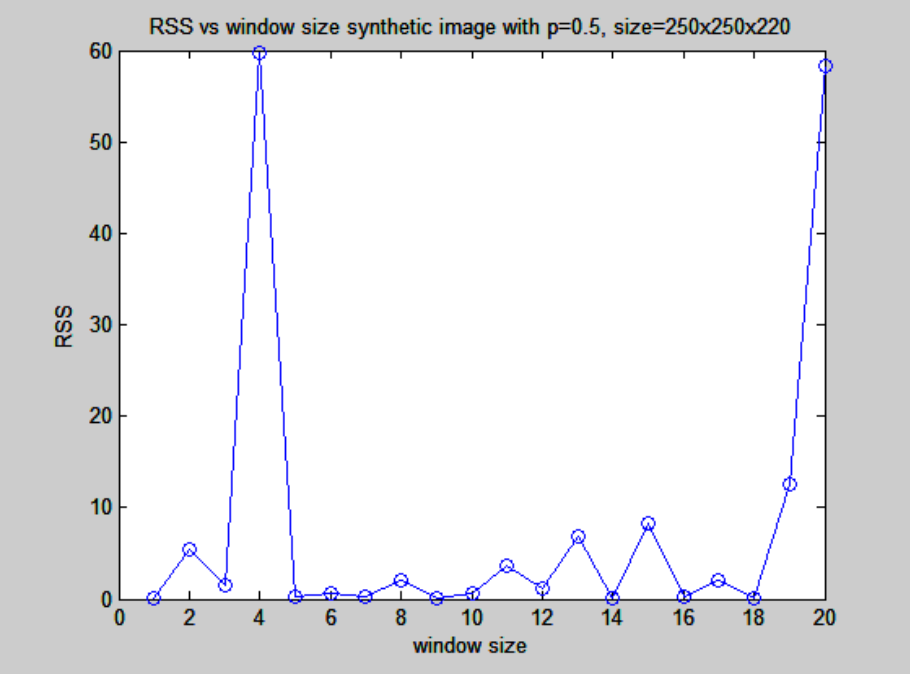


Figure 5-22 Residual Sums of Squares vs. window size of the p=0.50 synthetic image

The RSS vs. window size plot of the p=0.5 synthetic image shown in Figure 5-22 indicates high RSS at windows sizes, n=4,13,15,17,19,20. Therefore, the null hypothesis could be rejected indicating pixels are not identically distributed at these spatial resolutions.

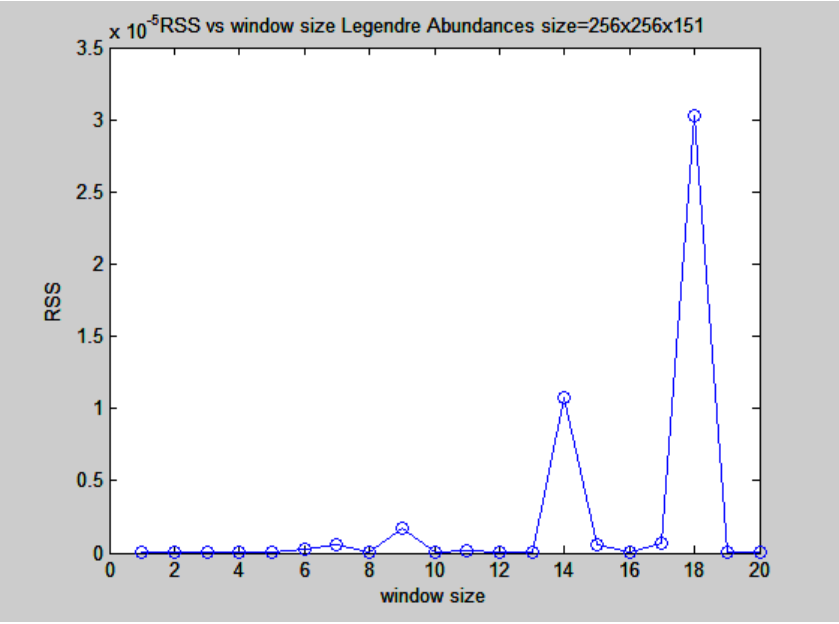


Figure 5-23 Residual Sums of Squares vs. window size of the synthetic image with Legendre abundances

In the Legendre abundances image, the RSS is small for all the window sizes as shown in Figure 5-23. In this case, the null hypothesis could not be rejected for any of the window sizes. Therefore, according to the covariance hypothesis test pixels on the image are identically distributed at all spatial resolutions.

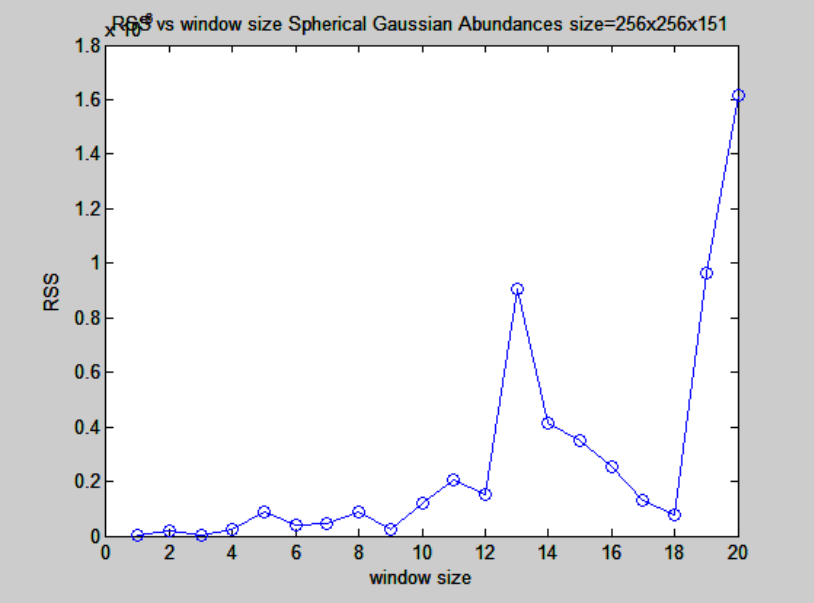


Figure 5-24 Residual Sums of Squares vs. window size of the synthetic image with Spherical Gaussian abundances

Spherical Gaussian abundances image results are similar to the Legendre abundances image. As shown in Figure 5-24 the RSS values are also very small (in the order of $1.0\ x\ 10^{-6}$) for all window sizes. Therefore, the null hypothesis could not be rejected for any of the window sizes and pixels on the image are identically distributed at all spatial resolutions.

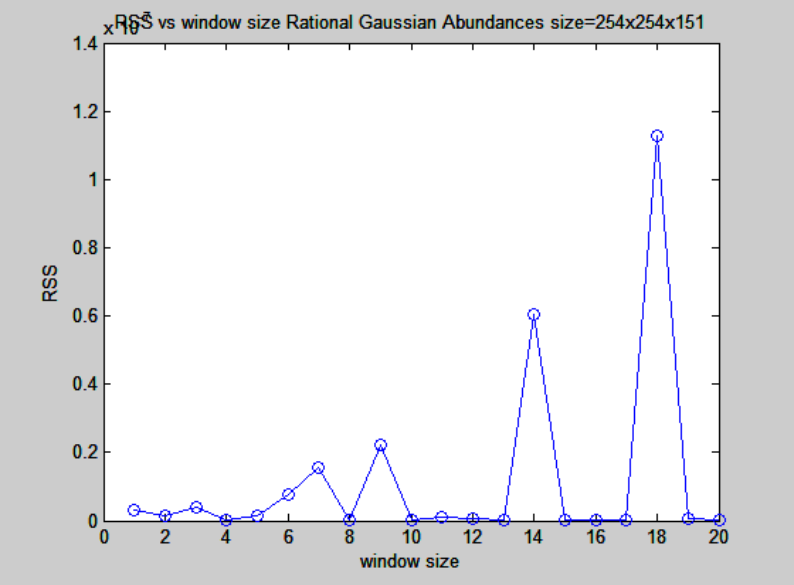


Figure 5-25 Residual Sums of Squares vs. window size of the synthetic image with Rational Gaussian abundances

Rational Gaussian abundances image experiment gave the same results to the previous two. Results shown in Figure 5-25 indicate that the null hypothesis for the covariance method could not be rejected for any of the window sizes. Therefore, pixels in the image are identically distributed at all spatial resolutions.

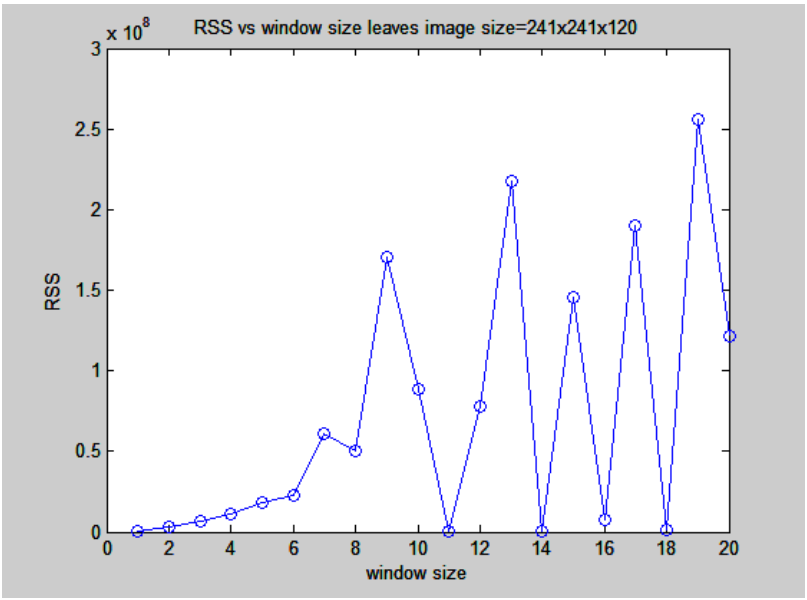


Figure 5-26 Residual Sums of Squares vs. window size of the leaves hyperspectral image

Even though there is only one object in the leaves image, the variance in the image could be high due to the significant difference between the leaves. Leaves could have dissimilarities between healthiness and size resulting in different reflectance values. By observing the results shown in Figure 5-26 the null hypothesis for the covariance method was rejected for all of the window sizes. Therefore, pixels in the image are not identically distributed at all spatial resolutions.

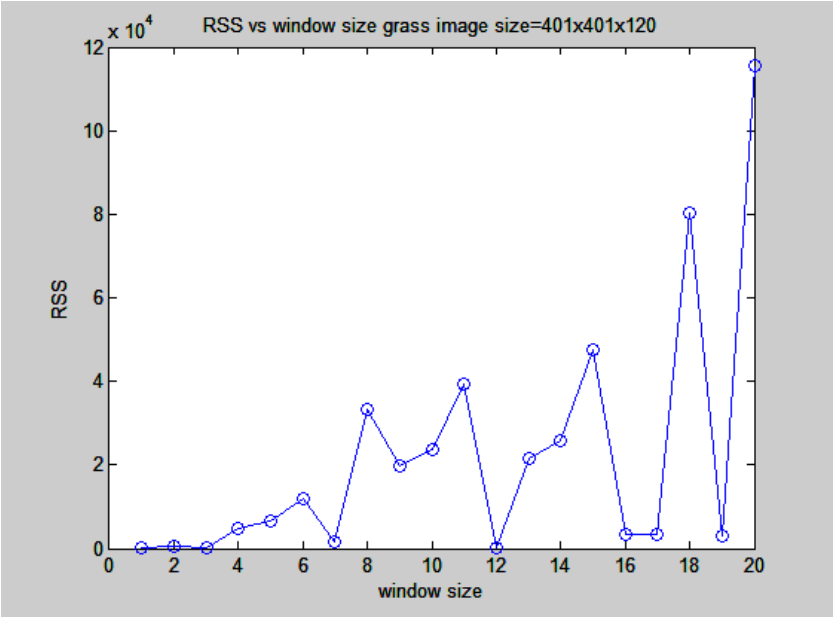


Figure 5-27 Residual Sums of Squares vs window size of the grass/soil hyperspectral image

The grass and soil image has the same characteristics as the leaves image, due to the high heterogeneity present in the grass. The variance in the image could be high due to the significant difference between regions of grass. Similar to leaves, grass could have dissimilarities between healthiness, and size resulting in different reflectance values. By observing the results shown in Figure 5-27 the null hypothesis for the covariance method was rejected for all of the window sizes. Hence, in this image pixels also are not identically distributed at all spatial resolutions.

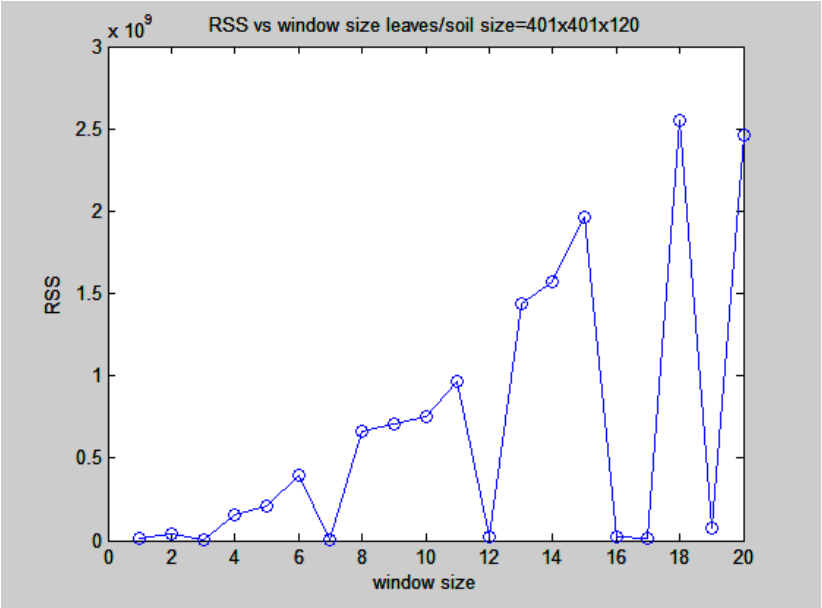


Figure 5-28 Residual Sums of Squares vs. window size of the dirt/leaves hyperspectral image

The dirt/leaves image has a wide separation between dirt and leaves. Therefore, it is expected to be highly heterogeneous. Results shown in Figure 5-28 validates this observation by giving RSS in the order of $1.0\ x\ 10^9$. The null hypothesis was rejected for all of the window sizes, then pixels on the image are not identically distributed at all spatial resolutions.

The covariance hypothesis method tests homogeneity within one class. Therefore the image needs to be divided to be able to apply the hypothesis individually to each class. The image is divided by selecting manually regions that seem homogenous at simple eye. The first spectral band of the resulting dirt and leaves images are shown below.

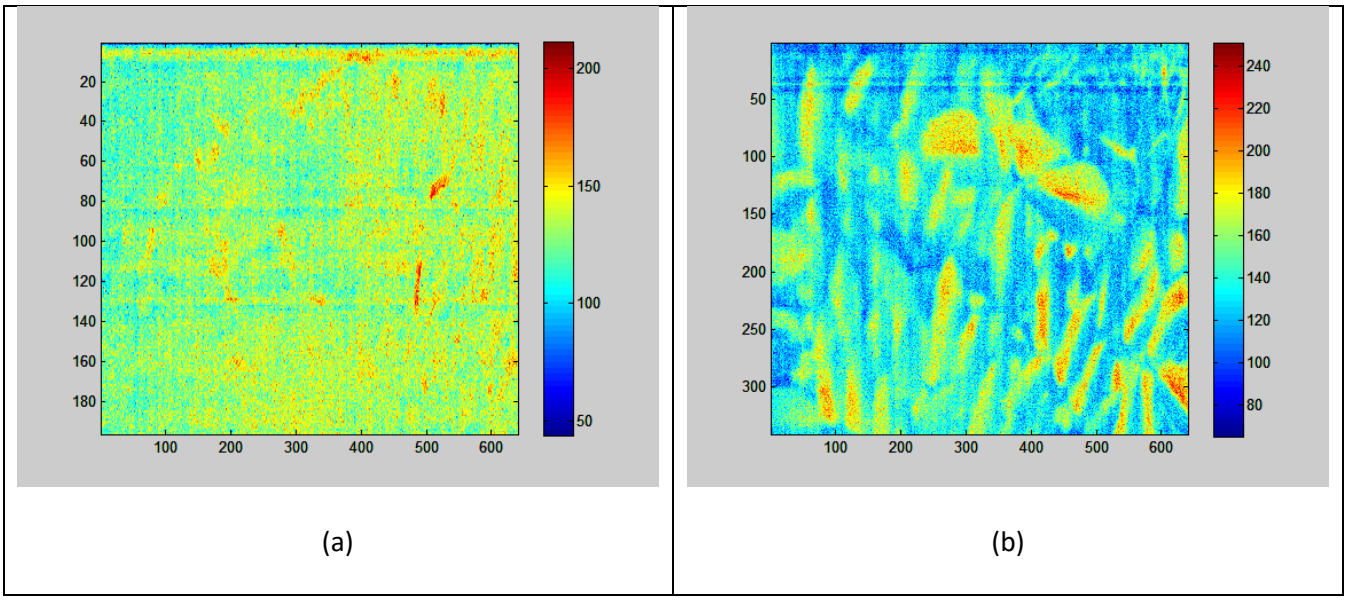


Figure 5-29 Regions extracted from the dirt/leaves image, (a) dirt crop, (b) leaves crop

After separating the two classes on the dirt/leaves image, the covariance hypothesis test was applied to each of the regions. Results from the test are shown in Figure 5-30 and Figure 5-31.

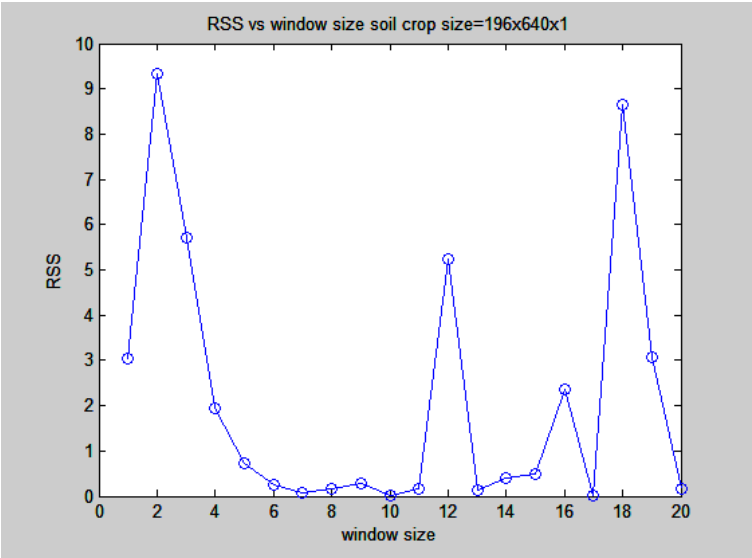


Figure 5-30 Residual Sums of Squares vs. window size of the dirt crop

The RSS vs. window size plot for the dirt crop shown in Figure 5-30 indicates low RSS values at windows sizes, n = 6,7,8,9,1013,17,20. Therefore, the null hypothesis could not be rejected indicating pixels are identically distributed at these spatial resolutions.

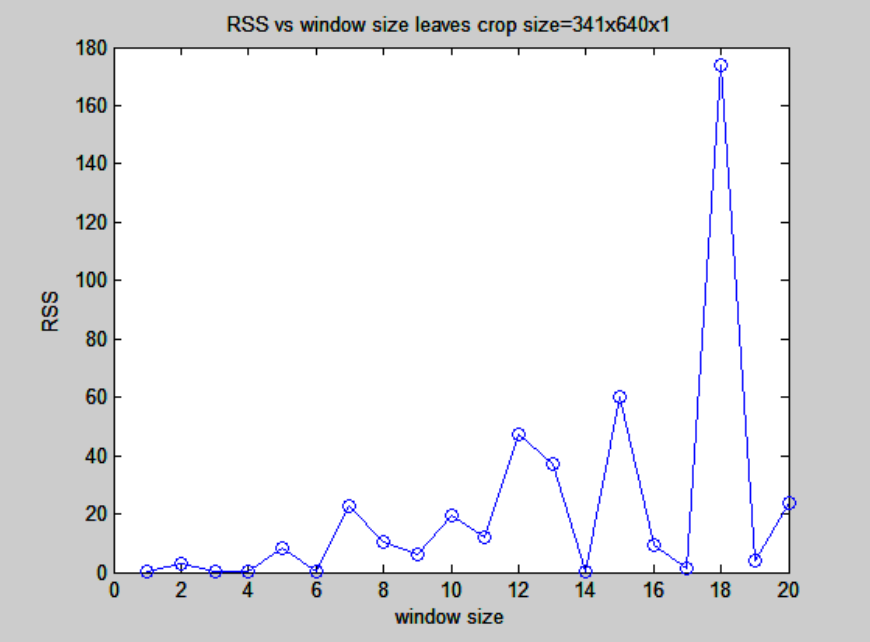


Figure 5-31 Residual Sums of Squares vs. window size of the leaves crop

In the case of the leaves crop, results in Figure 5-31 indicate that at windows sizes, n=2,3,4,6,14,17 the null hypothesis in the covariance method could not be rejected. This determines that pixels in the leaves crop image at these spatial resolutions are identically distributed.

The main reason for testing homogeneity in an image is the assumption that image classification accuracy will improve under this condition. If there are multiple of objects or classes in an image it is assumed that classification accuracy will improve when individual classes are optimally classified. Using the hypothesis covariance test this could be achieved by selecting the pixel size at which the maximum number of classes has a low RSS, or a true null hypothesis. Therefore, to reach the maximum classification accuracy in the dirt/leaves image a window size should be obtained such as the pixels are identically distributed in both classes. The leaves crop image was determined to be homogeneous at n=2,3,4,6,14,17, and the dirt crop is homogenous at n =4,5,6,7,8,9,10,13,17,20. Both results have in common n=4 and n=6, hence any of these spatial resolutions will improve the classification accuracy of the dirt/leaves image. A review of the residual square differences is shown in table 5-7.

Table 5-7 RSS from leaves and dirt at different window sizes

window size	RSS leaves	RSS dirt
1	3.0491	0.0082
2	3.0456	9.3458
3	0.0011	5.712
4	0.4846	1.949
5	8.1173	0.7406
6	0.4982	0.2541
7	22.5543	0.0663
8	10.6777	0.1528
9	6.0803	0.2855
10	19.388	0.0007
11	11.8206	0.1538
12	47.1636	5.2489
13	37.2481	0.1379
14	0.0921	0.4091
15	59.8603	0.5036
16	9.6073	2.3697
17	1.1722	0.0017
18	173.8349	8.6525
19	4.1631	3.0801
20	23.8957	0.1688

5.5.1 Hypothesis Testing After Whitening Transformation

The results from the covariance hypothesis test seem to be consistent with the Parametric and Non-Parametric hypotheses tests, but there are several peaks shown in the plots, especially at high window sizes. Downsampling the image using a window 20x20 reduces the original image by a 1/20 factor. If the original image was 650x650, the new image would be 32 x 32. Reducing the image to a small size reduces drastically the sample size N that is used in the sample variance, mean and covariance.

The variance and covariance estimators are extremely sensitive to outliers [37], which are increased when there is a smaller number of samples or observations. The image can be transformed so that its covariance has unity covariance. This will reduce the effect of the sample covariance by reducing the covariance between neighboring pixels. Shown below are the results obtained by whitening the

leaves/grass image and applying the covariance hypothesis after. Figure 5-32 shows how pixels change with the transformation, and Figure 5-33 shows how the covariance change from the original to the transformed image.

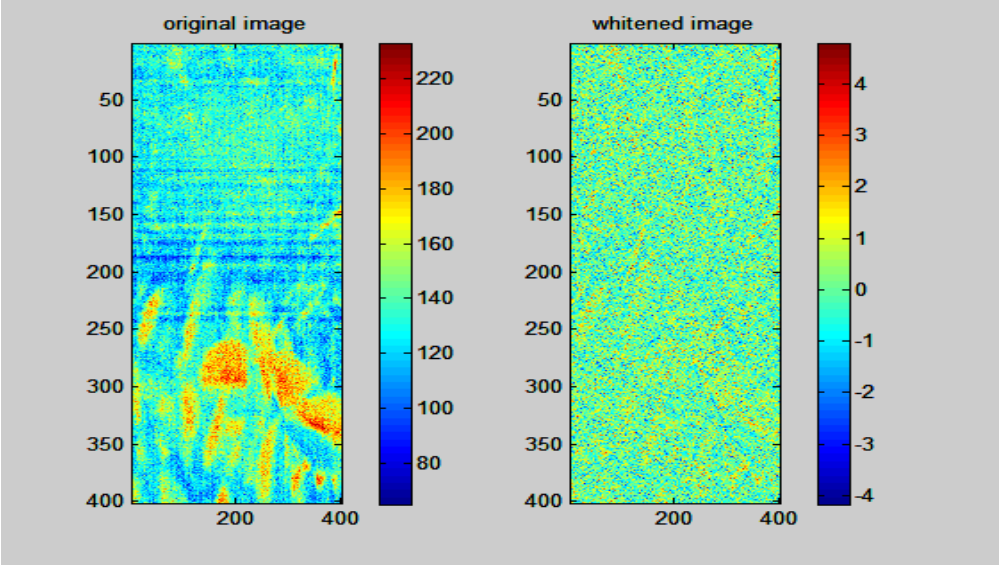


Figure 5-32 Leaves/grass original image and its whitening transform

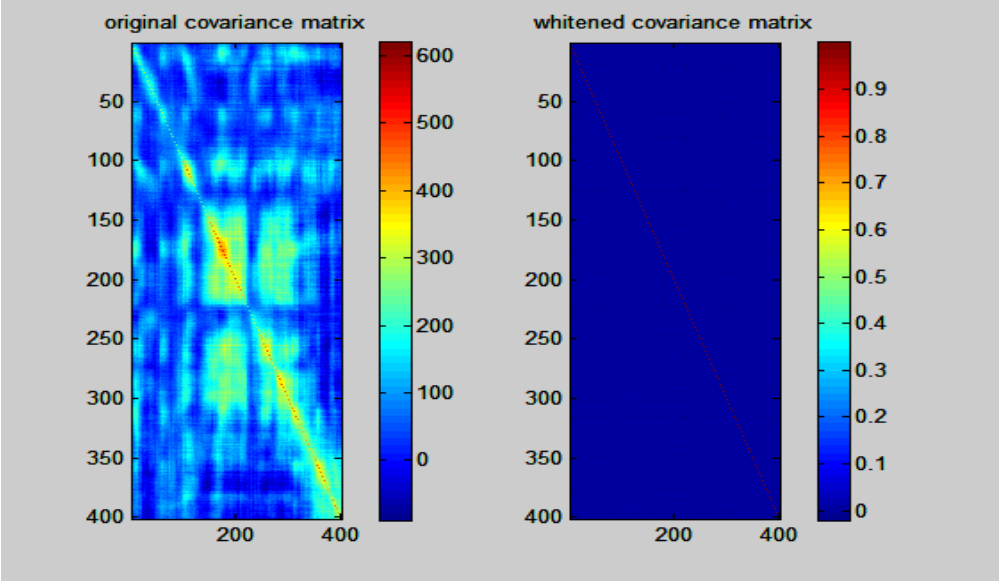


Figure 5-33 Covariance of the original image and the covariance of the whitening transform

After preprocessing the image with the whitening transform, the covariance hypothesis test was applied for the two classes of the leaves/soil image. The RSS vs. windows size plots shown below demonstrates how the peaks shown with the original image are reduced. Also, shows a relation where at bigger pixel sizes there are smaller RSS values.

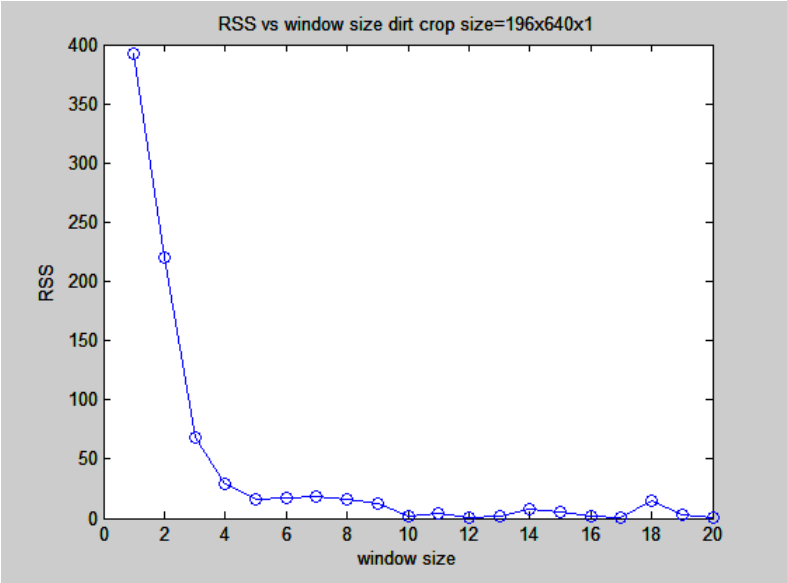


Figure 5-34 RSS vs. window size of the Soil crop after whitening transform

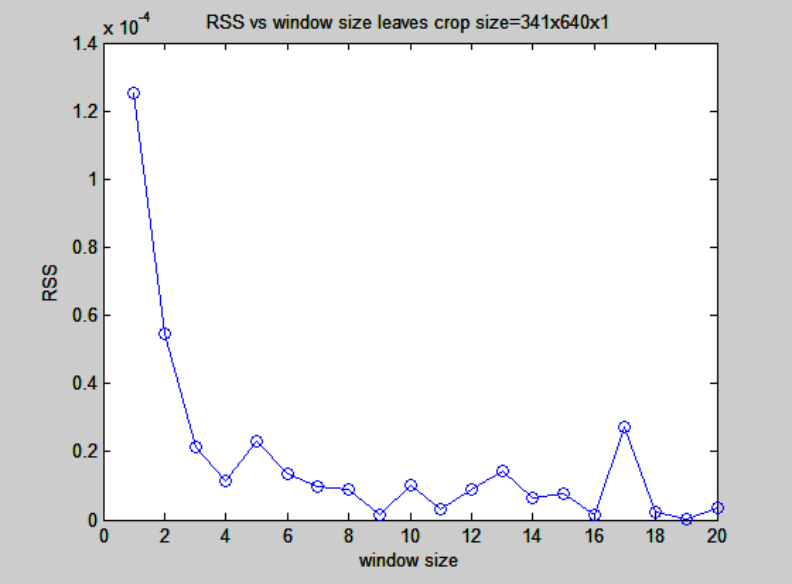


Figure 5-35 RSS vs. window size of the Leaves crop after whitening transform

To compare the results of the hypothesis test before and after pre-processing the image, the RSS vs. window size were plotted at the same logarithm scale (shown below). At high resolution both hypothesis test provide similar information, and the RSS keeps decreasing. At smaller resolution the covariance method without the whitening shows much more erratic behavior, than the whiten one.

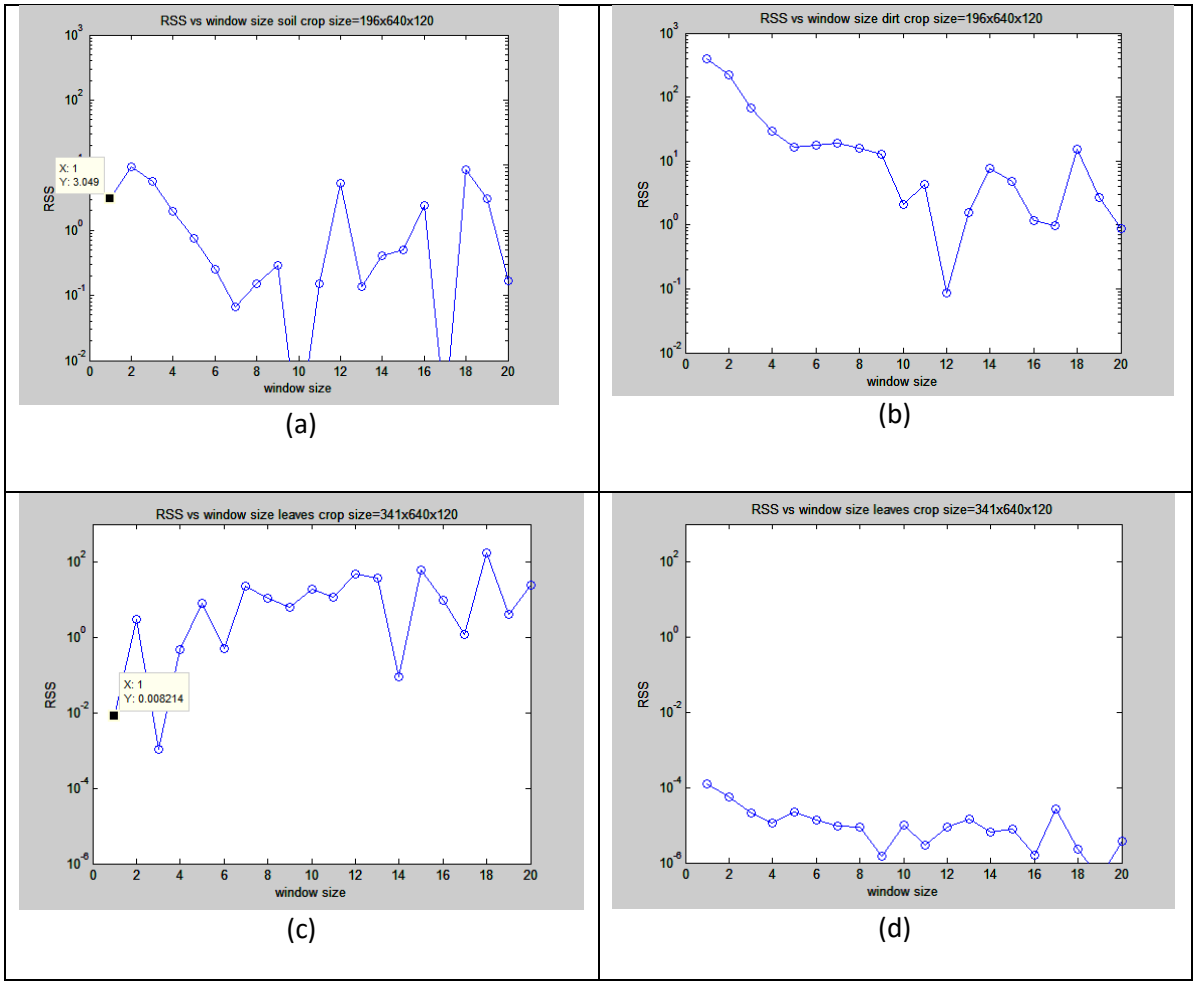


Figure 5-36 RSS vs. window size of the covariance hypothesis test before and after preprocessing: (a) soil crop hypothesis test before pre-processing, (b) soil crop after whitening transform, (c) leaves crop before pre-processing, (d) leaves crop after whitening transform

5.6 IMAGE CLASSIFICATION

The purpose of being able to determine the pixel size where the classes in an image are identically distributed is to obtain the spatial resolution that maximizes the image classification. An increase in accuracy was verified by classifying the leaves/dirt image at the spatial resolutions obtained from the covariance method test (n=6, and n=8). According to the test hypothesis results, the classification accuracy

will be greater at n=6 than at n=8. To validate this, a Maximum Likelihood (ML) classifier was applied at these different resolutions using different regions for training and testing. The results from the pixel labeling and the confusion matrixes are shown below in Figure 3-37 and Tables 5-8 and 5-9.

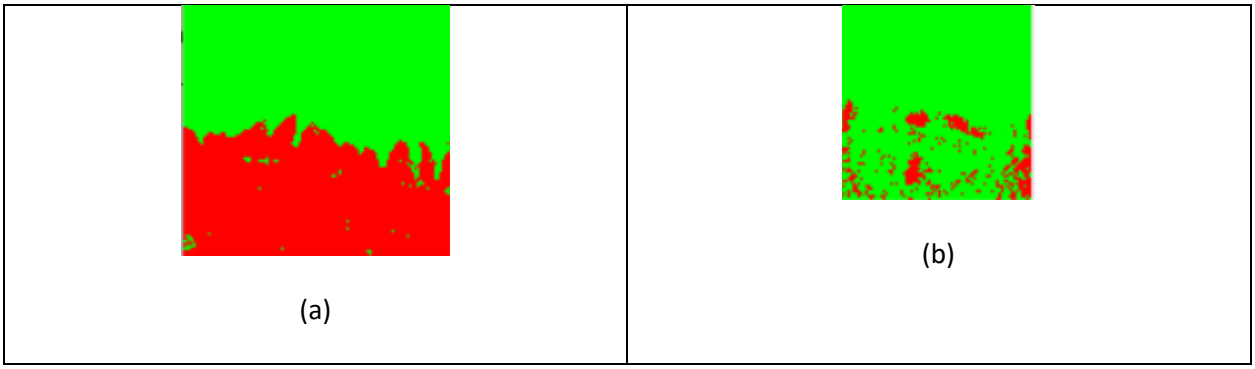


Figure 5-37 Classification labeling using maximum likelihood classifier in the leaves/soil image at different resolutions, red is used for leaves, and green for soil. From left to right: (a) Classification results for image using 6x6 downsampling, (b) classification results for image using 8x8 downsampling.

The classification labeling shows how the 6x6 downsampled image gave accurate results by correctly classifying objects and boundaries between objects. The leaves class is labeled as red, and the dirt as green. The image downsampled with an 8x8 window was incorrectly classified by classifying many leave pixels as dirt (soil). According to the confusion matrices shown in Table 5-8 and Table 5-9, the overall accuracy of the 6x6 downsampled image is 98.82%, with a Kappa Coefficient = 0.9576, and for the 8x8 downsampled image has overall accuracy = 83.87%. The confusion matrices are shown in percentage.

Table 5-8 Confusion matrix ML 6x6 downsampling image (percentage)

Class	Leaves	Soil	Total
Leaves-red	99.30	1.27	17.12
Soil-green	0.70	98.73	82.88
Total	100.00	100.00	100.00

Table 5-9 . Confusion matrix ML 8x8 downsampling image (percentage)

Class	Leaves	Soil	Total
Leaves-red	56.16	0.00	20.67
Soil-green	43.84	100.00	79.33
Total	100.00	100.00	100.00

5.7 RESULTS OVERVIEW

The results in Section 5.2 and Section 5.5 should have the same results, since their purpose is to determine at which spatial resolution the pixels can be statistically identically distributed. Results from the Ansi-Bradley and F-test (Table 5-1 to Table 5-6) were compared with the results of the covariance hypothesis test (Figures 5-21 to Figure 5-28).

5.7.1 Results overview: synthetic image generated by using probability of a pixel coming from one of the two classes p=0.25

The results from the Ansi-Bradley and F-test tests and the covariance hypothesis test were compared for the first image. Comparing Table 5-1 with Figure 5-21 several similarities could be found. The results in Figure 5-21 indicate that at windows sizes, n=3,n=8, and n=13 the null hypothesis in the covariance method test can be rejected, indicating that pixels on this images at these spatial resolutions are not identically distributed. Table 5-1 has the result for the F-test, where the null hypothesis was also rejected at n=7, n=14. These numbers are not equal to the previous result n=8, and n=13, but the Ansi-Bradley and F-test hypothesis tests as explained before was realized by using n, and n+1 as the populations. Therefore, the null hypothesis being rejected at n=7, and n=14 translates to indicate that the populations from n=7 and n=8 do not come from the same density, and also for n=13 and n=14. The hypothesis test for Ansari-Bradley did not reject the hypothesis at n=8, but it rejected it at n=13. Even though the null hypothesis was rejected at other spatial resolution it is ignored, since they are only used to validate this section methodology. Also, the Ansi-Bradley and F-test hypothesis tests are sensitive to

variations in the data and dependent on data normality and the results could possess type I (reject Ho when it is true) or type II errors (accept null hypothesis when it is not true).

5.7.2 Results overview: synthetic image generated by using probability of a pixel coming from one of the two classes p=0.5

The results from the synthetic image generated with p=0.5 gave the same results for the covariance method as the image with p=0.25, with the difference that the residual sum of squares (RSS) at n=3 was greater. Performing the Ansari-Bradley and F-tests hypotheses tests in this image gave consistent results were both tests rejected the null hypothesis at n=7 and n=14. With these two images, the usability of the covariance hypothesis method for hyperspectral images has shown to be effective.

5.7.3 Results overview: synthetic image with Legendre abundances

Comparing Table 5-3 with Figure 5-23 several differences could be found. The Ansari-Bradley test is more sensitive about accepting the null hypothesis, since only window sizes n=3, n=8 and n=15 have true null hypothesis. F-test has considerably different results than the Ansari-Bradley test, by classifying all the spatial resolution as coming from the same distribution. To compare the previous results with the covariance method hypothesis, the RSS vs. window size plot in figure 5-23 explains more in detail why most of the window sizes in the F-test gave true hypothesis results. The RSS vs. window size plot in Figure 5-23 shows that the pixels on this image at most spatial resolutions can be determined as identically distributed, except when n>14. In comparison the F-test and the covariance method has some slight differences, but its tendency in the image are similar, showing homogeneity at most of the pixel sizes.

5.7.4 Results overview: synthetic image with Spherical Gaussian abundances

The Ansari-Bradley and F-test show similar hypotheses results when the image with spherical Gaussian abundances was used. According to these parametric and non-parametric tests the null hypothesis is rejected for all the spatial resolutions, except when n=1, or n=2. This means that these finer resolutions are the best in terms of indicating to have identically distributed pixels. This correlates to Figure 5-24, since the function SSE vs factor keeps incrementing after n=4.

5.7.5 Results overview: synthetic image with Rational Gaussian abundances

Figure 5-25 results show that the null hypothesis from the covariance method cannot be rejected at n=8, n=15, n=16, n=17. In comparison with the Ansari-Bradley and the F-test, the hypothesis can be rejected at n=8, but not at n=7, which indicates same distributions between n=7, and n=8. The hypothesis could also not be rejected at n=14, which indicated same distribution between n=14 and n=15. This two downsampling window sizes resulted in the same in both methods, but n=16 and n=17 from the covariance method gave results where the null hypothesis was rejected in the other two methods.

5.7.6 Results overview: real hyperspectral image with leaves

After extensively verifying the covariance hypothesis test with synthetic images, the real images were put under test. The Figure 5-26 above show the results of the covariance method with the leaves image, where the hypothesis was rejected when the image was downsampled using window sizes of n=9, n=11 and n=13. In the case of the Ansari-Bradley and F-test, at most window sizes the hypothesis was rejected, except for n=1 or its original spatial resolution. The window sizes where the hypothesis was rejected at the covariance test were also rejected by the other hypotheses tests.

5.7.7 Results overview: real hyperspectral image with grass/soil

Another real hyperspectral image was used to compare with the previous results. This image contains grass and soil and can be seen to be, more homogeneous than the leaves image. The covariance method (as shown in Figure 5-27) indicates that the hypothesis could not be rejected at n=2, n=5, and n=10. Also in this image it could be noted that the RSS vs window size function linearly increases after n= 11. The Ansari-Bradley and F-test hypothesis could not be rejected in the case where n=10, n=1, but for n=5 the hypothesis was rejected.

5.7.8 Results overview: real hyperspectral image with dirt/leaves

The last experiment was realized with an image with well spatial separation between the objects in the image. The image contains the top half with only dirt (soil), and the other half is full of leaves. The results from the covariance method shown in Figure 5-28 indicate that the hypothesis could not be

rejected at n= 1,2,3,4,5,8 and n=10. In this image RSS vs window size function is also linearly increasing after n= 11. The Ansari-Bradley and F-test hypothesis could not be rejected in the case where n= 1,2,3, but only in about half of the bands. For n=5, only 13 bands gave a true hypothesis, while for most of the other bands have zero bands with the true hypothesis.

6 APPLICATION EXAMPLE

A practical example is given going through all the steps with an image taken with the SOC700 hyperspectral imager. The image has five different objects submerged 18 inches under water in a tank. The RGB corrected image is shown in Figure 6-1.

Figure 6-1 Tank image shown in natural color RGB

The procedure to statistically find an optimal pixel size to improve the image classification accuracy is summarized below:

1. Select regions of the image that are to be classified. In this case, six regions or classes are chosen as shown in Figure 6-2. Their histograms are presented in Figure 6-3.

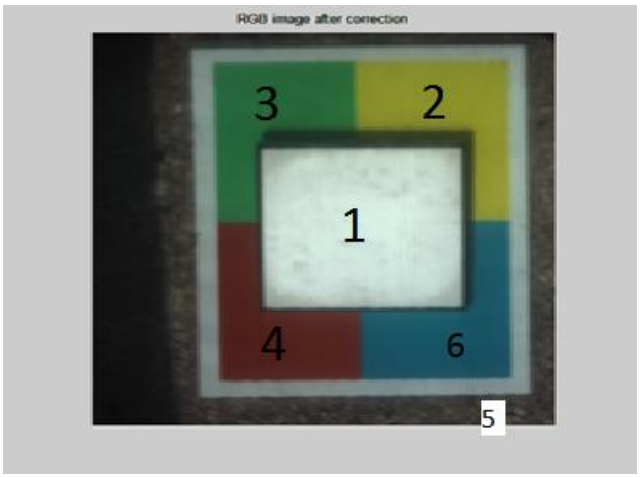
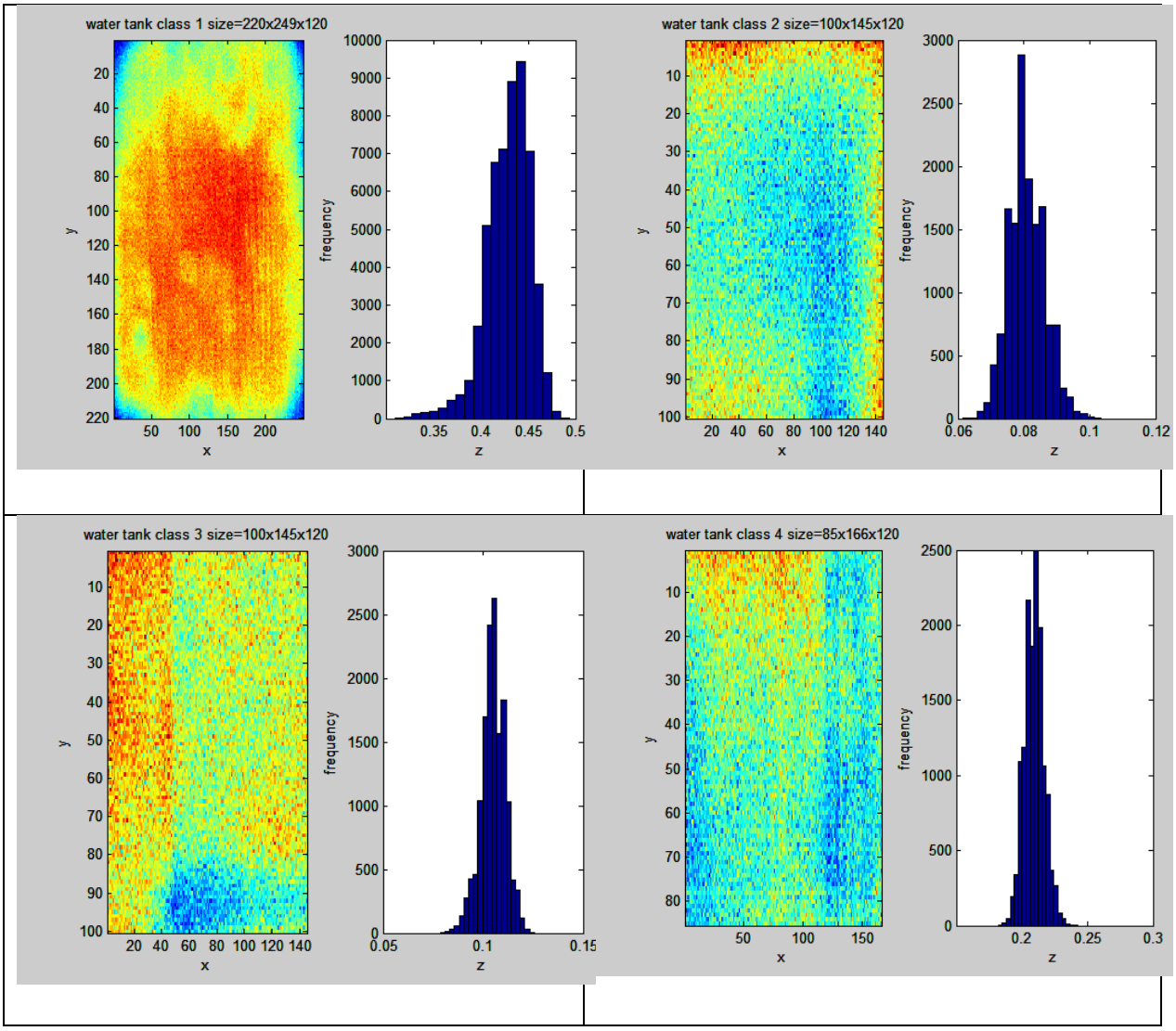


Figure 6-2 Definition classes tank image



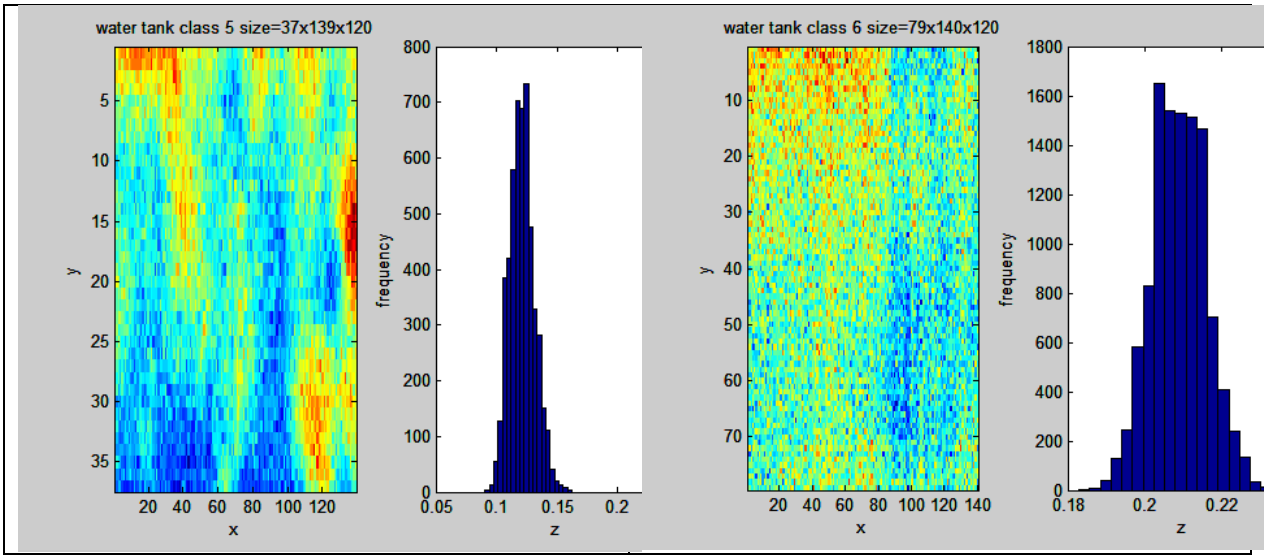


Figure 6-3 Tank six classes with their histograms

2. Pre-process each of image by applying the whitening transform
3. Analyze the residual sums of squares (RSS) at different window sizes, and select the pixel size at which the maximum number of classes has a low RSS. By observing the ranges in the RSS values in this application a $RSS < 0.1$ was selected as the threshold to not reject the null hypothesis of belonging to the same distribution.

Table 6-1 RSS from each of the tank image classes at different window sizes

window (w)	class 1	class 2	class 3	class 4	class 5	class 6
1	0.044729	0.015939	0.018677	0.027251	0.153217	0.038557
2	0.028057	0.006987	0.008518	0.01969	0.218817	0.011608
3	0.031656	0.016104	0.007674	0.013108	0.443555	0.025356
4	0.0389	0.021765	0.006969	0.00376	0.108653	0.003402
5	0.012687	0.003938	0.000482	0.002294	0.084576	0.001386
6	0.020295	0.043134	0.010317	0.011837	0.677991	0.002661
7	0.007788	0.108936	0.013627	0.010514	0.043986	0.002234
8	0.008265	0.094751	0.019936	0.006866	0.184041	0.000389
9	0.018606	0.048136	0.008772	0.002162	2.031752	0.008911

10	0.006132	0.002954	0.001328	0.016942	0.00411	0.000315
11	0.00542	0.043042	0.003877	0.00013	0.01593	0.003377
12	0.031469	0.200918	0.047859	0.003273	0.044367	0.011428
13	0.004957	0.003705	0.00102	0.011306	0.137259	0.035017
14	0.003001	0.04523	0.008559	0.01792	0.498296	0.000307
15	0.015598	0.042946	0.012436	0.000297	0.915019	0.003243
16	0.001098	0.227107	0.033871	0.003685	1.591106	0.000807
17	0.008058	0.003675	0.000884	0.000278	3.234408	0.021025
18	0.016885	0.010669	0.004415	0.0062	6.942972	0.022662
19	0.00239	0.010145	0.001482	0.029004	0.000254	0.066398
20	0.01559	0.003245	0.006452	0.150092	0.000642	0.000307

4. If there are many pixel sizes where the previous criteria fall on, select the lowest spatial resolution if many classes are interfering in the classification, or select the highest spatial resolution if many classes are needed to be classified.

In this particular problem six well spatially separated classes are predominant in the image, and the purpose is to classify all of them. The highest spatial resolution at which the images are homogeneous is the ideal size to use, which according to Table 6-1 is by downsampling the image with a 5x5 window. With the purpose of validating the results the image downsampled with a 6x6 window is also classified. This image is more heterogeneous because the hypothesis test for class 5 determined that pixels within the image are not identically distributed.

The image was classified by using a ML classifier, using training classes shown as polygons, and testing classes as rectangles. The image to the right with 6 colors corresponds to the labeling of the classified image.

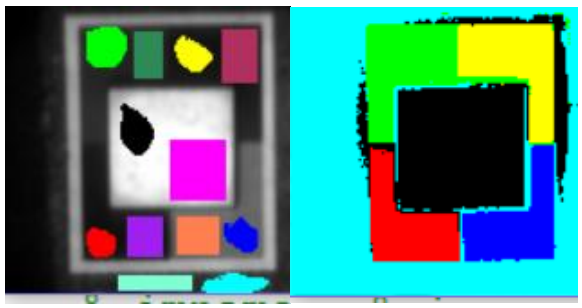


Figure 6-4 Left: Training (polygons) and testing (rectangles) pixels of 5x5 downsampled image, Right: Labeling of tank image after using a Maximum Likelihood Classifier

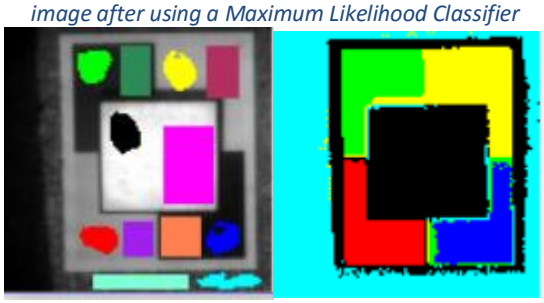


Figure 6-5 Left: Training (polygons) and testing (rectangles) pixels of 6x6 downsampled image , Right: Labeling of tank image after using a Maximum Likelihood Classifier

The results are very similar at both spatial resolutions, but the image downsampled by using a 5x5 window has a higher classification according to the test features, and the confusion matrix. The overall accuracy of the image downsampled with a 5x5 window is of 99.26% and for the 6x6 window it is 96.26%.

Table 6-1 Confusion matrix ML tank image 5x5 downsampling (in percentage)

Class	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Class 1	100	0	0	0	0	0	31.05
Class 2	0	95.83	0	0	0	0	16.93
Class 3	0	4.17	100	0	0	0	13.29
Class 4	0	0	0	100	0	0	13.25
Class 5	0	0	0	0	100	0	14.86
Class 6	0	0	0	0	0	100	10.63
Total	100	100	100	100	100	100	100

Table 6-2 Confusion matrix ML tank image 6x6 downsampling (in percentage)

Class	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Class 1	100	0	0	0	0	0	34.73
Class 2	0	100	10	0	0	0	16.64
Class 3	0	0	90	0	16.02	0	14.40
Class 4	0	0	0	100	0	0	9.41
Class 5	0	0	0	0	83.98	0	12.04
Class 6	0	0	0	0	0	100	12.77
Total	100	100	100	100	100	100	100

Since the image had classes already well spatially and spectrally separated the improvement could not be significant. Even though the classification accuracy in this case was not improved by a lot, it improved by 3%, which could mean a lot if there were a lot of pixels in the image.

7 CONCLUSIONS

This research worked with hypothesis testing to efficiently determine the ideal pixel size for classification. Statistically based classification algorithms are based on the assumption that each class has features that can be statistically separated from other classes. Since their means and variances model this separation, the hypotheses tests developed here used the image statistics. The assumption that pixels in the image are identically distributed random variables, but they are not independent led to a derivation of a hypothesis test using a covariance and variance equation explained in detail in Section 4.4.

The fundamental assumption in this research is that a class or endmember in an image has variations that are not modeled well by a single class if the pixel scale is small enough, but at some coarser scale the pixels will become homogeneous. At this spatial resolution or pixel size, the image classification errors will decrease, resulting in an ideal pixel size for classification.

The experiments produced promising results that were compared by using parametric and non-parametric known hypotheses tests. These tests are used to determine if two populations can be well modeled as coming from the same density. Various hypotheses tests were used with hyperspectral images data. The Ansari-Bradley non-parametric test and the parametric F-test were used as references for comparison as they produced the best results, and were less sensitive to reject the null hypothesis than the other tests. The F-test is designed to test if the samples come from the same density by testing if the two populations have the same variance. It is a parametric test and some of the assumption made to develop the test is that the populations come from a normal distribution (showed in their histograms) and that their variances are unknown. The Ansari-Bradley tests for differences in spread. It is a nonparametric alternative to the two-sample F-test for equal variances. It does not require the normality assumption and can be used with samples from distributions that do not have finite variances. The assumption of variance for these two tests may be the reason for their good performance. At the beginning of the study, the

calculation of sample variances was used as the hypothesis test, but it did not show a good performance because random variables did not appear to be independent. The covariance function had to be used instead because the variance of a sum of variables is the sum of pairwise covariances, and takes into account the correlation between the variables.

All three methods showed promising similar results for synthetic and real hyperspectral images, validating the usability of the new proposed hypothesis method within the test scope. Results show that the covariance hypothesis method is a simple straightforward solution to finding an ideal pixel size for classification. Also, the sensitivity of the covariance to outliers can be addressed by applying a pre-processing step of whitening the image. The biggest advantages of the covariance hypothesis method developed over the other methods shown in this thesis are that this method does not require normality assumption, does not assume that the random variables are independent, and it offers a comprehensive and systematic procedure.

8 FUTURE WORK

Different kind of real hyperspectral images (such as satellite images) should be used to obtain more specific information about the limitations of this method. Also, the images were downsampled by taking a neighborhood average from an original image, which involves the assumption of a square wave response, which is a rough approximation of reality. Real hyperspectral images with different spatial resolutions could be tested.

The statistics of the test statistic was difficult to obtain even after making many convenient assumptions. The dependence of neighboring pixels makes it complicated to obtain a simple solution utilization common methods. More advanced statistical distributions approximations could be applied to obtain the acceptance region for the null hypothesis. For example, the methodology to approximate a linear combination of non-central chi-square random variables explained in research paper [38].

The covariance test statistic developed was obtained by defining the hyperspectral images pixels as multiple one-dimensional random variables. This definition is not unique and could be modified by using multivariate random variables, which can include with major accuracy the characteristics of the image spectral dimension.

9 REFERENCES

[1] N. Rodriguez, S. Hunt, M. Goenaga and M. Velez, "Determining optimum pixel size for classification," in *Conference: Proc SPIE*, Baltimore, MD, 2014.

[2] C. Woodcock and A. H. Strahler, "The factor of scale in remote sensing," *Remote Sensing of Environment*, vol. 21, no. 1, pp. 311-332, 1987.

[3] D. A. Quattrochi and R. E. Pelletier, "Quantitative Methods in Landscape Ecology," in *Remote Sensing for Analysis of Landscapes: An Introduction*, New York, Springer-Verlag, 1991, pp. 51-76.

[4] S. Hunt, "Puerto Rico Space Grant Consortium," 2009-2013. [Online]. Available: http://www.prsgc.upr.edu/index.php?option=com_content&view=article&id=115:hyperspectral-&catid=45:researchawards&Itemid=56. [Accessed 12 May 2015].

[5] J. Lunzer, "Analysis of spectral variability in hyperspectral imagery with application to biodiversity estimation," Unpublished master's thesis, University of Puerto Rico, Mayaguez, Puerto Rico, 2013.

[6] B. Markham and T. J.R., "Land cover classification accuracy as a function of sensor spatial resolution," *Proceedings 15th Int. Symp. on Remote Sensing of Environment*, pp. 1075-1090, 1981.

[7] Q. Bingwen, S. Yinbo, C. Chongcheng and T. Xiaoyang, "Identification of optimal spatial resolution with local variance, semivariogram and wavelet method," *Accuracy 2010 Symposium*, pp. 353-356, 2010.

[8] P. Han, J. Gong, Z. Li and L. Cheng, "The study on the choice of optimal scale in image classification," *Remote Sensing and Spatial Information Sciences*, vol. 37, no. 82, pp. 385-390, 2008.

[9] K. McCloy and P. Becher, "Optimizing image resolution to maximize the accuracy of hard classification," *Photogrammetric: Engineering and Remote Sensing*, vol. 73, no. 8, pp. 893-903, 2007.

[10] P. Atkinson and P. Curran, "Choosing an appropiate spatial resolution for remote sensing applications," *American Society for Photogrammetry and Remote Sensing*, vol. 63, no. 12, pp. 1345-1351, 1997.

[11] P. Curran, "The semivariogram in remote sensing: An introduction," *Remote Sensing of Environment*, vol. 24, no. 3, pp. 493-507, 1988.

[12] I. Clark, "Regularization of a semivariogram," *Computers & Geosciences*, vol. 3, pp. 341-346, 1977.

[13] M. Armstrong, "Basic Linear Geostatistics," *Berlin: Springer-Verlag*, 1998.

[14] J. Chiles and P. Delfiner, Geostatistics: Model spatial uncertainty, New York: John Wiley & Sons, 1999.

[15] I. Clark and W. Harper, Practical Geostatistics 2000, Columbus: Ecosse North America, 2000.

[16] A. Journel and H. C.J., Mining Geostatistics, London: Academic Press, 1978.

[17] R. S., Hyperspectral Remote Sensing and Spectral Signature Applications, New India Pubilshing, 2009.

[18] R. A. Schowengerdt, Remote sensing models and methods for image processing third edition, Burlington, MA: Elsevier, 2007.

[19] G. Joseph, Fundamentals of Remote Sensing Second Edition, India: University Press, 2005.

[20] R. A. Schowengerdt, Remote sensing models and methods for image processing third edition, Burlington, MA: Elsevier, 2007.

[21] "Hypothesis Tests - Statistics and Probability," [Online]. Available: http://www.stattrek.com/hypothesis-test/hypothesis-testing.aspx. [Accessed 1 December 2015].

[22] L. Ludeman, Random Processes, filtering, estimation and detection, Hoboken, NJ: John Wiley and Sons, 2003.

[23] D. Lane, "Sampling Distribution of the Mean," [Online]. Available: http://onlinestatbook.com/2/sampling_distributions/samp_dist_mean.html. [Accessed 10 May 2015].

[24] L. Wasserman, All of Statistics: A Concise Course in Statistical Inference;, New Yotk, NY: Springer-Verlag, 2004.

[25] Z. Michalewicz and D. Fogel, How to Solve It: Modern Heuristics, Springer Science & Business Media, 2013.

[26] K. K., Mathematical Statistics, New York: Chapman and Hall, 2000.

[27] W. G. Cochran, "The distribution of quadratic forms in a normal system, with applications to the analysis of covariance," *Mathematical Proceedings of the Cambridge Philosophical Society* , vol. 30, no. 2, 1934.

[28] R. A. Johnson and W. D.W., Applied Multivariate Statistical Analysis, New Jersey: Person Pentice Hall, 2007.

[29] V. Neuman, "Distribution of the ratio of the mean square succesive difference," *Annals of Mathematical Statistics,* vol. 12, pp. 367-395, 1941.

[30] L. Lapin, Probability and Statistics for Modern Engineering, Belmont: PWS Publishers, 1983.

[31] "Decorrelating and then Whitening data," MIT Media Laboratory, 2015. [Online]. Available: http://courses.media.mit.edu/2010fall/mas622j/whiten.pdf. [Accessed 12 May 2015].

[32] M. Brown and F. A.B, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, pp. 364-367, 1974.

[33] A. Koning, "Homogeneity of Variances," *John Wiley & Sons, Ltd*, 2008.

[34] R. O'Brien, "A general ANOVA method for robust test of additive models for variance," *Journal of the American Statistical Association,* vol. 74, pp. 877-880, 1979.

[35] R. Clark and e. al., "The U.S. Geological Survey, Digital Spectral Librar: Version 1: 0.2 to 3.0 microns," U.S. Geological Survey Open File Report, 1993.

[36] Computational Intelligence Group, "Grupo de Inteligencia Computacional de la Universidad del País Vasco," [Online]. Available: http://www.ehu.eus/ccwintco. [Accessed March 2014].

[37] V. W., "Modern applied statistics with S," *Springer,* 2012.

[38] H. Hyung-Tae and S. Provost, "An Accurate Appoximation to the Distribution of a Linear Combination of Non-Central Chi-Square Random Variables," *REVSTAT Statistical Journal,* vol. 11, no. 3, pp. 231-254, 2013.

[39] M. Ozdogan, Y. Yang, A. G. and C. Cervantes, "Remote sensing of irrgates agriculture: opportunities and challenges," *Remote Sens.*, no. 2, pp. 2274-2304, 2010.

[40] U. N. E. Programme, "What is biodiversity?," 2010. [Online]. Available: http://www.unep.org/wed/2010/english/PDF/BIODIVERSITY_FACTSHEET.pdf. [Accessed 2015].

[41] P. Hsieh, L. L.C. and N. Chen, "Effect of spatial resolution on classification errors of pure and mixed pixels in remote sensing," *IEE Trans. Geosci. Remote Sens.* , no. 39, pp. 2657-2663, 2001.

[42] G. Foody, "Incorporating mixed pixels in the training, allocation, and testing stages of supervised classifications," *Pattern Recognit. Lett.* , no. 17, pp. 1389-1398, 1996.

[43] P. Atkinson and P. Lewis, "Geostatistical classification for remote sensing: an introduction," *Computers and Geosciences,* vol. 26, pp. 361-371, 2000.

[44] J. Gruninger, R. A. J. and .. H. M. L., "The sequential maximum angle convex cone (SMACC)," *SPIE Proceedings, Algorithms for Multispectral, Hyperspectral and Hyperspectral Data,* vol. 5425, no. 1, 2004.

[45] R. A. Schowengerdt, in *Remote sensing models and methods for image processing third edition*, Burlington, MA, Elsevier, 2007, pp. 16-19.

[46] D. Jarvis, P. C. and H. Cooper, Managing Biodiversity in Agricultural Ecosystems, Columbia University Press, 2007.

[47] O. Khalid and S. Romshoo, "Assesing the impact of spatial resolution on the accuracy of land cover classification," *Journal of Himalayan Ecology and Systainable Development,* vol. 9, pp. 33-49, 2014.

[48] F. United Nations, "How to feed the world 2050," 2009.

[49] L. Fabian and D. G., "Defining the spatial resolution requirements for crop identification using optical remote sensing," *Remote Sens.*, vol. 6, pp. 9034-9063, 2014.

[50] G. Metheron, "Principles of Geostatistics," *Economic Geology,* vol. 58, pp. 1246-1266, 196.

[51] N. Bruinsma and A. J., "World agriculture towards 2030/2050: the 2012 revision," June 2012. [Online]. Available: www.fao.org/economic/esa. [Accessed May 2015].

[52] S. Andre´foue¨t, "Choosing the appropriate spatial resolution for monitoring coral bleaching events using remote sensing," *Springer-Verlag: Coral Reefs,* vol. 21, pp. 147-154, 2002.

[53] S. Devlin, G. R and K. J, "Robust Estimation and Outlier Detection with Correlation Coefficients," *Biometrika,* vol. 62, no. 3, pp. 531-545, 1975.