ABSTRACT

KNIGHT, JOSEPH. Accuracy Assessment of Thematic Maps Using Inter-Class Spectral Distances. (Under the direction of Dr. Siamak Khorram.)

The goal of this research is to develop a new approach to remote sensing thematic accuracy assessment in which the spectral distances between the classes in a thematic classification are used as inputs to the error estimation process. The conceptual basis for this new approach is that the confusion of relatively spectrally different classes represents a more severe error than confusing relatively spectrally similar classes. Therefore, the accuracy estimate of a classification can be adjusted to take into account the "spectral severities" (or misclassification costs) of the errors in that classification. The benefits of including inter-class spectral distances in the accuracy assessment process are shown in the context of the development of two new accuracy assessment measures called Spectrally Weighted Kappa (SWK) and Spectrally Weighted Fuzzy (SWF). These two new accuracy assessment methods are introduced and tested for their performance relative to current techniques.

The results of this research demonstrate that inter-class spectral distances can be used effectively in accuracy assessment of thematic classifications. The SWK approach can provide information about the spectral costs of errors in a classification that is not as apparent with traditional methods. In addition, SWK provides a quantitative base for establishing weights for Weighted Kappa analysis and allows for the possibility of improving a classification during its development. The SWF method improves upon current fuzzy accuracy assessment techniques by providing a way to establish membership functions

that is based on inter-class spectral distances. We have shown that the SWF method can provide fuzzy membership values that are similar to those that a well-trained human might choose. Therefore, in cases where multiple interpreters would normally have been used to create fuzzy membership values, the SWF method can be employed reduce inter-interpreter bias. In addition, the SWF method provides a quantitative basis for establishment of fuzzy membership values. We expect that these two new accuracy estimation techniques will be of use to the remote sensing research community.

ACCURACY ASSESSMENT OF THEMATIC MAPS USING INTER-CLASS SPECTRAL DISTANCES

by

JOSEPH F. KNIGHT

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

FORESTRY

Raleigh

2002

This dissertation is dedicated to the memory
of my grandmother Rosemary Knight.

BIOGRAPHY

Joseph Knight was born in Fort Wayne, Indiana. He is the oldest child of Joseph Sr. and Lois Knight. Joseph, and his sister Jennifer, were raised in Leo, Indiana and attended Leo High School. It was in high school, during World History class, that Joseph first developed an interest in maps. He also became an avid computer hobbyist while in high school. He had no idea at that time that these interests would converge later in life in the field of remote sensing.

After high school, Joseph attended Purdue University for one year, majoring in Computer Science. Unsatisfied at that time with college life, Joseph joined the Navy. He served his country for five years as a missile system technician aboard the cruiser USS Ticonderoga. His service included participation in the Desert Storm conflict, for which the Ticonderoga was awarded several decorations.

Joseph was honorably discharged in April of 1994 and returned to Purdue University in the fall of that year. It was at Purdue that Joseph became interested in Remote Sensing, after attending a lecture by Dr. Chris Johanssen of Purdue's Laboratory for the Applications of Remote Sensing (LARS). He decided then to pursue a career in that field.

Another turning point in Joseph's life also occurred at Purdue. Joseph participated in a summer study abroad program in Kiev, Ukraine. Another Purdue student in the group was Kirsten Nielsen. Joseph and Kirsten grew close during the summer trip and eventually married in June of 1997.

After graduating from Purdue, Joseph and Kirsten moved to Raleigh, NC to pursue their doctoral studies at North Carolina State. Joseph began a Ph.D. program under the direction of Dr. Siamak Khorram. His next career step is a post-doc with the Environmental Protection Agency.

ACKNOWLEDGEMENTS

I would like to acknowledge the support of my major professor, Dr. Siamak Khorram, and my advisory committee of Dr. Hugh Devine, Dr. John Monahan, and Dr. Joseph Roise. Without their patient help, this project would not have been possible.

I would like to thank the graduate students and staff of the Center for Earth Observation for their help and friendship. The CEO is a wonderful working environment, and leaving will be very difficult.

I would like to thank my family for their support. My parents and sister are a constant source of encouragement. I am one of the first members of my family to attend college, and the first to get a graduate degree, and so my family is very excited about my upcoming graduation. It is their pride and enthusiasm that made these accomplishments possible.

Finally, I would like to thank my wife Kirsten. During our summer together in Kiev, Ukraine, I nicknamed her *solnshka* in Russian, or “little sun”. She has more than lived up to that name during our graduate studies. Having both members of a husband and wife team in Ph.D. programs at the same time can be very difficult. When times get tough, Kirsten is my ray of sunshine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTORY REMARKS

The ability to assess map accuracy has been important for as long as there have been maps. Early map makers were concerned with making maps that allow easy navigation, delineate land ownership, and chart newly explored territory. As civilizations became more technologically advanced, they developed the ability to make ever more sophisticated maps. With the advent of widely available remotely sensed image data, our mapping goals have changed from documenting political boundaries and providing navigation aids to producing high resolution thematic maps. We can now map the entire world using satellite sensors. These sensors are used to produce maps for a wide variety of applications, such as environmental issues like mapping deforestation, socioeconomic issues like forecasting grain yields in third-world countries, and human health issues like monitoring the spread of diseases. These maps are different from previous maps in the way that they are produced. Instead of a laborious on-site mapping effort, maps derived from remotely sensed data are produced from digital images of the study area using automated classification algorithms. The maps can then be referred to as classifications or characterizations of the study area. The quality of these classifications must then be evaluated by comparing them with a set of “ground truth” or reference data. The reference data may be collected by on-site visits or by interpreting aerial photos or other appropriate data.

All classifications contain errors. The truth of this statement becomes apparent when one considers the potential for different forms of reference data

being used to assess the accuracy of one classification. For example, it is possible that a classification could have an overall accuracy estimate of 78% when aerial photos are the reference data and 83% when the classification is compared to ground observations. This raises several questions: Is one of these estimates more correct than the other? What are the contributions of the different sources of error that make up these estimates (the *error budget*)? How useful is the classification for its intended purpose? These are the questions that must be answered by remote sensing researchers before a solid accuracy estimate can be stated. This process is referred to as accuracy assessment.

The goal of this research is to develop a new approach to accuracy assessment in which the quality of a classification derived from multispectral data is weighted by the spectral costs of the various errors in the classification. The specific objectives are discussed in Section 3.0. However some background information and context are needed, which is presented in Section 2.0.

Note: In this dissertation, certain naming conventions will be used. The term “classification” or “map” will refer to a thematic data product derived from remotely sensed data. A “reference site” is a location within the study area where the map and reference class labels are compared. “Reference data” is the agreed-upon true class label for a particular reference site. “Accuracy assessment” will refer to the process of characterizing and examining the impact of errors in a thematic land use and land cover (LU/LC) classification. A person referred to as an “analyst” performs accuracy assessment on a map. An

"interpreter" examines the reference data to determine the correct class label for a reference site.

2.0 BACKGROUND

This section will present a brief review of the accuracy assessment process, from sampling of the reference sites to computation of an accuracy estimate. The evolution of accuracy assessment techniques will be tracked from early qualitative techniques to newer site-specific quantitative methods. An exhaustive treatment of all of the many accuracy assessment measures proposed in the literature is beyond the scope of this review. Rather, we will present an overview of the most commonly used methods.

2.1 DEFINITION OF ACCURACY ASSESSMENT

Accuracy assessment is the process of quantifying data quality so that users may evaluate the utility of a map for their intended applications (Stehman 1998). Accuracy assessment is a critical part of any project in which remotely sensed data are classified to create a thematic map product. Users of these map products must have a robust estimate of the quality of the maps before they make decisions based on the information contained therein.

2.2 Overall Categories of Error Types

There are two main categories of errors in remote sensing accuracy assessment: positional and thematic. Positional errors are errors due to a spatial difference, or misregistration, between different data types (Pontius 2000, Lunetta et al 1991, Janssen and van der Wel, 1994). These data types can be different image data layers or vector layers, but positional error in accuracy

assessment refers to a spatial inconsistency between the map and the reference data. A simple example of how easily positional errors can be introduced into a reference data set is illustrated in the context of collecting ground sampled reference data with the aid of a GPS. If the geographic datum is inconsistent between the map and the reference data in North Carolina – say the map uses NAD83 and the GPS receiver collected coordinates using NAD27 – then there will be a positional error of approximately 1/10 of a mile between the two data sets. Much recent research has been done on quantifying and correcting for positional errors (Seong and Usery 2001, Smith and Atkinson 2001, Greenfield 2001, Gao 2001, Zhou and Li 2000, Pontius 2000, and others).

In contrast, thematic error is confusion between the map class label and the reference or "true" class label. For example, a forest type map may indicate that a certain area is composed of loblolly pine, but when that area is visited by ground sampling crews, they find white oaks. The research presented in this dissertation focuses on quantifying thematic errors. The remainder of this section describes thematic accuracy assessment techniques.

2.3 Early Thematic Accuracy Assessment Techniques

This section describes the early forms of accuracy assessment used before the development of the quantitative site-specific techniques in use today.

2.3.1 Qualitative Accuracy Assessment

Early remote sensing accuracy assessment measures, if they were employed at all, were qualitative rather than quantitative. The analyst examined the classified map and decided whether or not it "looked good" (Aronoff 1982a,b;

Aronoff, 1985, Congalton and Green 1999). This approach did not create a problem in the early days of satellite remote sensing, when the focus was on exploring the capabilities of the new sensors rather than on providing highly accurate maps to third party users. Clearly, quantitatively based accuracy assessment methods would be required to extend the technology to other applications.

2.3.2 Non-Site Specific Methods

As the use of remotely sensed data expanded better accuracy assessment methods were developed. The first of these were non-site specific in nature (van Genderen and Lock 1977, van Genderen et al 1978, Hord and Brooner 1976). Using these methods, the overall areas of the classes on the map were compared with estimates of the areas of the same classes on the ground. For example, in a simple forest vs. non-forest classification, the total number of acres of forest on the classified map would be compared to the measured number of acres of forest on the ground. The same would be done for the non-forest category. This was a step forward from simple qualitative analysis, but non-site specific estimates suffer from a critical limitation: the estimate of the area of a particular class on the classified map may be similar to its actual area on the ground, but the actual spatial overlap of the class on the map and on the ground may be quite low. Due to this problem, non-site specific techniques are not broadly appropriate for estimation of the accuracy of thematic classifications. Development of superior techniques requires sampling of the

study area. At this point a short digression on sampling procedures is appropriate.

2.4 Sampling Procedures

The goal of an accuracy assessment sampling scheme is to adequately represent the true compositions of the thematic classes on the reference data source (or "ground truth" data) in each of the sampled locations (Hay 1979). The reference data class labels are compared with the corresponding location's thematic map class labels to compute estimates of the map's accuracy. The most important criterion of the sampling scheme is that it must be statistically rigorous enough to support inferences made based on the sampled data (Stehman 2001).

An accuracy assessment sampling scheme consists of two parts: the sampling units and the sampling design (Stehman and Czaplewski 1998). The sampling units are the actual locations where the map and the reference data are compared. The sampling design is the method of determining the locations of the sampling units.

2.4.1 Sampling Units

The two basic types of sampling units are points and areal units (Fisher 1997). A point is simply a location which has no area, e.g. a set of latitude and longitude coordinates. Areal units, on the other hand, represent a two-dimensional area on the thematic map and the reference data, e.g. a fixed area, polygon, or pixel that *surrounds* a coordinate location (Stehman and Czaplewski 1998, Curran and Williamson 1986).

Selection of a point sampling unit is advantageous because the sample is considered to be continuous, and so does not require the decision rules involved with using areal units, such as how to determine the precise label for an area when it may be composed of multiple thematic classes. However, the conceptual difficulty of using point samples, and potential problems with precise location of the points on different data layers, have led many researchers to recommend using areal units in accuracy assessment (Fenstermaker 1991, Franklin et al 1991, Janssen and van der Wel 1994, Stenback and Congalton 1990, and Stoms, 1996). In fact, point units are now infrequently considered for accuracy assessment of remotely sensed. The most commonly used sampling units are a single image pixel and a 3x3 image pixel area. (Congalton and Green 1999, Stehman and Czaplewski 1998).

2.4.2 Sampling Design

The sampling design is the method of selecting the sampling units. As mentioned above, the most important criterion of the sampling design is that it be statistically defensible; i.e. it must provide sufficient support for any inferences made from the reference data. To accomplish this, the design must ensure that the inclusion probabilities for all potential sampling units are greater than zero. The inclusion probability represents the likelihood that a particular sampling unit will be included in the sample. By contrast, a design that samples only areas near roads is not statistically defensible because the inclusion probabilities for all areas *not* near roads are zero. Therefore, in that case, the sample cannot be said to represent the entire study area.

The five most common sampling schemes for use in remote sensing reference data collection are: simple random sampling, systematic sampling, stratified random sampling, cluster sampling, and stratified systematic unaligned sampling (Fitzpatrick-Lins 1981, Congalton and Green 1999). These will be briefly described in the following sections.

2.4.2.1 Simple Random Sampling

In a simple random sample, each sampling unit has an equal chance of being selected into the sample. In many cases random map coordinates are chosen to select sampling units into the overall sample. This procedure has the advantages of being easy to implement and, with sufficient sample size, ensures that most all well-represented classes will be adequately sampled. Difficulties occur when there are one or more “rare” classes on the map. In this case, simple random sampling cannot guarantee that there will be sufficient, or even any, sampling units representing those classes. In cases where there are no rare map classes, simple random sampling is a strong choice due to its favorable statistical properties of independence.

2.4.2.2 Systematic Sampling

In systematic sampling, the sampling units are spaced at a regular interval over the study area. There are two ways of going about this approach. First, one can simply lay a grid over the study area and sample at each grid line intersection. This grid is analogous to the “dot grid” area estimation procedure in photogrammetry. The other systematic sampling approach involves randomly selecting the first point and then spacing points at regular intervals after that.

Using either method, systematic sampling has the strength that the study area is sampled more uniformly than with the other designs. A weakness of systematic sampling is, again, the likelihood that rare classes will not be adequately sampled. Also if some phenomenon of interest is regularly spaced in a pattern similar to the layout of the sample units but slightly offset from it, it is possible to exclude each of these phenomena from the sample.

2.4.2.3 Stratified Random Sampling

Using stratified random sampling, the analyst seeks to solve the problem of adequately sampling rare classes by grouping, or stratifying, the thematic classes and then randomly sampling within each of the strata. For example, a thematic map showing land use and land cover (LU/LC) classes would be stratified by LU/LC class. Then, each class is randomly sampled independently of the other classes. The main advantage of stratified random sampling is that one can assign a minimum number of sampling units for each stratum so that all thematic classes can be guaranteed to be well-sampled. A variant of stratified random sampling is equalized random sampling. In this design, the number of sampling units in each stratum is equal. A common problem with stratification is that rare classes, by definition, make up a small percentage of the study area. This problem results in small parts of the study area being intensively sampled while other areas composed of more common classes may be sparsely sampled. This has the effect of magnifying positional errors due to local relief and image misregistration, should those errors occur in the heavily sampled areas (Congalton 1988).

2.4.2.4 Cluster Sampling

Cluster sampling is the process of choosing a number of area samples of a fixed size and then choosing the sampling units for accuracy assessment by exhaustively describing the composition of each cluster. For example, if the sampling unit in a particular project is one pixel, and the cluster size is set to ten pixels, then each cluster would yield ten sampling units. Cluster sampling reduces the number of areas that must be visited or interpreted by grouping the sampling units. However, this clustering may under-represent certain rare classes due to spatial autocorrelation (Cliff and Ord 1973). Another disadvantage of cluster sampling is that the calculation of the standard error is more complex (Czaplewski 1994).

2.4.2.5 Stratified Systematic Unaligned Sampling

Stratified Systematic Unaligned Sampling takes advantage of the strengths of both systematic and stratified sampling schemes. The systematic component guarantees that the entire study area will be adequately sampled. The stratification component ensures that all thematic classes will be represented. However, Stratified Systematic Unaligned Sampling still suffers from the potential that equally spaced phenomena will be either over-sampled or under-sampled.

2.4.3 Sampling Review Conclusion

The most important consideration when choosing a sampling scheme is that it is statistically valid. Satisfying that, the goals and resources of a project must be considered so that a sampling scheme appropriate to the project is

used. After the sampling scheme is designed and implemented and the reference data are collected, the analyst may proceed with computation of the accuracy estimate.

2.5 Recent Thematic Site-Specific Accuracy Assessment Techniques

Recent years have seen the development of site-specific thematic accuracy assessment techniques. With these techniques, a number of randomly determined sample sites are distributed throughout the study area. These sites are then examined on both the classified map and the reference data. If the class labels match, then the site is designated as correctly classified. Using these site-specific techniques, one can calculate an overall percent accuracy estimate for the classification. This section presents several of the more widely used and/or innovative accuracy estimate techniques to provide context for discussing the new techniques presented in this dissertation.

2.5.1 Error Matrix

As discussed previously, the typical accuracy assessment approach involves the selection of a set of statistically determined sample sites that are examined on the chosen reference data (e.g. aerial photographs, ground visits, higher spatial resolution data, etc.) and on the classified map. The classified value on the thematic map is compared with the interpreted class on the reference data on a site-by-site basis. The results of this comparison are conventionally presented in an error matrix (Card 1982). An error matrix is a square n by n matrix, where n is the number of classes. The diagonal elements of the matrix represent correctly classified reference sites. The off-diagonal, or

marginal, cells represent incorrectly classified sites. The error matrix gives the overall percent accuracy of the classification and the omission and commission errors for each class. Other statistics, such as confidence intervals and the Kappa coefficient of agreement can be easily calculated from the error matrix. Due to its utility, the error matrix has been widely used in remote sensing research (Congalton 1999, Khorram et. al. 1999, Lunetta and Elvidge 1998, Congalton 1991, Card 1982, Aronoff 1982a,b).

While it is a very useful tool, the site for site comparison of the error matrix is limited in that the technique is based on the assumption that the reference data interpreter is able to select from the reference data one absolutely correct class for each site. In many cases this is difficult or inappropriate. For example, depending on the acquisition season and scale of the reference data, it can be very difficult to differentiate pasture and cropland cover types. Similarly, site heterogeneity can make choosing only one class problematic. Furthermore, the error matrix design assumes that all off-diagonal (marginal) errors are of equal importance. Clearly, this is rarely the case in a remote sensing classification. For example, confusing deciduous forest with coniferous forest is, from a spectral composition point of view, a much less severe error than confusing deciduous forest with high density urban. The Weighted Kappa analysis technique (discussed below) was introduced to try to solve this problem.

2.5.2 Kappa

Kappa was introduced by Cohen (1960) for use in psychological sampling studies. Kappa is a discrete multivariate technique that tests whether one data

set is significantly different from another. In the case of accuracy assessment, we test whether two error matrices are significantly different (Congalton 1983, Bishop et al 1975). The two error matrices can be from different classifications, as might be the case when conducting change detection, or Kappa may be used on only one error matrix by comparing that error matrix to a hypothetical completely random error matrix. In other words, Kappa's associated test statistic KHAT tests how a classification performed relative to a hypothetical completely randomly determined classification. An important property of Kappa is that it uses the information contained in all of the cells of the error matrix, rather than only the diagonal elements, to estimate the map accuracy. The KHAT statistic ranges from 0 to 1. A KHAT value of 0.75 means that the classification accounts for 75% more of the variation in the data than would a hypothetical completely random classification. A general framework for interpreting KHAT values was introduced by Landis and Koch (1977). They recommended that KHAT values greater than 0.8 represent strong agreement, values between 0.4 and 0.8 represent moderate agreement, and values below 0.4 represent poor agreement.

Since its introduction to remote sensing researchers in 1983 (Congalton 1983), Kappa technique has gained wide acceptance and is now regarded as a standard accuracy assessment technique (Rosenfield and Fitzpatrick-Lins 1986, Hudson and Ramm 1987, Türk 2002). However, some researchers, notably Foody (1992) and Türk (2002), object to the use of Kappa as an accuracy measure because the Kappa chance correction term (defined in Section 4.9.1.5) can be assumed to contain some agreement that is not due to chance, i.e.

"actual" agreement. In spite of this criticism, Kappa is widely recommended and reported in accuracy assessment studies (Congalton 1999, Lunetta and Elvidge 1998, Ma and Redmond 1995, Janssen and van der Wel 1994)

2.5.2.1 Conditional Kappa

Like Kappa, Conditional Kappa tests agreement between the classification in question and a hypothetical random classification (Light 1971). However, Conditional Kappa computes the agreement of the individual categories, or classes, in a classification through the use of a maximum likelihood estimator. The Conditional Kappa algorithm for use in thematic accuracy assessment is defined in Congalton (1999). Conditional Kappa allows the analyst to extract class-specific information from an error matrix, which is not available with the standard Kappa procedure.

2.5.2.2 Weighted Kappa

To account for the different severities of errors in the error matrix marginals, Rosenfield and Fitzpatrick-Lins (1986) proposed the use of the Weighted Kappa statistic in remote sensing thematic accuracy assessment. Weighted Kappa provides the same test statistic, KHAT, as Kappa (on a 0 to 1 scale), but Weighted Kappa provides for the inclusion of marginal error severity weights in the error matrix (also called misclassification costs). These marginal weights are important in that they allow the analyst to assign higher costs to errors that are considered more severe and lower costs to errors that are considered less severe. For example, in a study where the goal is to separate the various forest types from a general "non-forest" class, the analyst might

assign small a small misclassification cost, or weight, to confusions between the different forest types, but a large misclassification cost to confusions between any of the forest types and the non-forest class.

This feature of weighted Kappa represents a very powerful tool for accuracy assessment of thematic data. However, Weighted Kappa has not been well accepted in the literature primarily because of one major problem: defining how to assign appropriate and objective cell weights. Currently weights are assigned subjectively on a study by study basis. The analyst determines the severity of each possible type of error on an ad hoc basis and computes weights for each error type that reflect those severities. This procedure may result in appropriate weights, but it is generally difficult to quantify why the weights were established in a particular way. Additionally, subjectively determined weights are typically not portable from study to study. One of the objectives of this study (see Section 3.0) is to recommend a more objective method of determining weights for use in Weighted Kappa.

2.5.3 GT Index

Unlike Kappa and the other techniques presented in this section, the GT Index is not a measure of classification accuracy. Rather, it is a measure of how well a similar classifier might perform in the future based on its current performance. Introduced by Türk (1979), the GT Index is intended to measure the "diagnostic ability" of a classifier. Türk asserts that some of the reference sites counted as correct in an accuracy assessment are likely to be correct by chance alone. Therefore accuracy estimates that include correct classification by

chance do not necessarily indicate that a classifier will perform well in the future. Türk (2002) presents a thought experiment to illustrate this point: Assume that a classification is created randomly. After accuracy assessment the classification is shown to have a very high overall percent accuracy. Türk asserts that this is misleading because a random classifier has no diagnostic ability; i.e. it will likely not perform acceptably in the future.

The GT Index attempts to solve this problem by providing a measure of the diagnostic ability of a classifier that removes chance agreement. The GT Index is defined as "the proportion of [reference sites] that will *always* be classified correctly" (Türk 1979). In this way, the GT Index draws a distinction between accuracy, an evaluation of present performance, and diagnostic ability, a predictor of future performance.

2.5.4 Tau Coefficient

The Tau coefficient is introduced by Ma and Redmond (1995) and is based on a technique described by Klecka (1980). Like Kappa, Tau is a coefficient of agreement. Tau is similar to Kappa in that it measures the performance of a classification relative to that of a random assignment of pixels. The Tau coefficient seeks to answer the criticism advanced by Foody (1992) (previously discussed) of Kappa's potential for overestimating the proportion of chance agreement by computing the chance agreement with allowance for unequal group probabilities. Ma and Redmond cite three main improvements of Tau over Kappa:

- Tau is easier to understand and interpret.
- Tau and its variance estimate are relatively simple to calculate, as opposed to Kappa's very cumbersome variance estimation.
- Tau compensates for the influence of unequal probabilities of groups on random agreement and the influence of different numbers of groups, depending on which of the two versions of Tau is used.

One significant difference between Tau and Kappa, however, is that the Tau coefficient is based on *a priori* rather than *a posteriori* probabilities. This is a major disadvantage of the Tau coefficient, because the *a priori* class probabilities must be estimated before accuracy assessment can be performed (Smits et al 1999).

2.5.5 Fuzzy Accuracy Assessment

A limitation of the error matrix is that it allows only one reference class for each reference site. This poses a problem in sites where selection of only one class is difficult or inappropriate. If the reference data interpreter is asked to discriminate similar looking classes, there may be some confusion as to which is the proper class. For example, the accuracy assessment of the Federal Region 5 portion of the Multi-Resolution Land Characteristics (MRLC) Consortium's National Land Cover Data (NLCD) required the reference data interpreters to differentiate the agricultural classes of row crops and small grains on winter aerial photos. This is clearly difficult at best. In ambiguous situations such as

these it would be useful to the interpreters to be able to designate more than one class that could be considered correct.

A solution to this problem is suggested by the introduction of fuzzy set theory (Woodcock and Gopal 2000, Gopal and Woodcock 1994, Wang 1990, Zadeh 1965). With this technique, the reference data interpreters provide, not just one correct class, but a range of possibilities. Each reference site is assigned values from a linguistic scale that indicate the reference site's membership in each of the thematic classes. These values are then converted to a numerical scale of the reference points' membership values in each thematic class. The following is a linguistic scale adapted from Gopal and Woodcock (1994):

1. *Absolutely Incorrect*: This class is absolutely unacceptable.
2. *Probably Incorrect*: Not a good answer. There is something about the site that makes selection of this class understandable, but there is clearly a better answer.
3. *Acceptable*: Not necessarily the best class for the site, but it is acceptable; this class does not pose a problem to the user if it is seen on the thematic map.
4. *Probably Correct*: Would be happy to find this class on the thematic map.
5. *Absolutely Correct*: This class is a perfect match for the site.

The linguistic values are then converted to a numerical scale that ranges from 1 to 5. For example, in a thematic classification composed of three classes (A, B, and C), a particular reference site may be assigned a value of 3 (Acceptable) for Class A, 4 (Probably Correct) for Class B, and 1 (Absolutely Wrong) for Class C.

Gopal and Woodcock used these fuzzy membership values as inputs to fuzzy operators that provide the analyst with extensive information about the errors in a classification, such as frequency of matches and mismatches, magnitude of errors, source of errors, and nature of errors. This allows for multiple estimates of the accuracy of a classification depending on the strictness of the rules used. These fuzzy operators are discussed in detail in Section 4.8

Fuzzy accuracy assessment can be of great benefit for thematic map quality assessment because it allows the analyst to derive extensive information about the accuracy of a classification that is not provided by other techniques. Most accuracy measures provide a single accuracy estimate. Fuzzy analysis provides easy-to-interpret information about which classes were confused with which other classes and how often, the severity of errors in the classification, and the effects of choosing different criteria for considering a site to be correct. Fuzzy accuracy assessment has been frequently used in remote sensing studies, and is gaining in popularity (Stehman and Czaplewski 2000, DeGloria 2000, Zhang and Foody 1998, Congalton 1999).

A drawback of this fuzzy method is that, like most accuracy assessment techniques, it relies on subjective interpreter-assigned membership values, which

allows for the possibility of significant inter-interpreter variation in cases where more than one interpreter is used. Furthermore, this method of determining the fuzzy membership values does not have a quantitative base. An objective of this research is to recommend solutions to these problems through the use of a new form of fuzzy accuracy assessment (see Section 3.0)

2.5.6 Accuracy Measures Based on Misclassification Costs

In addition to Weighted Kappa (discussed above) there are other accuracy assessment techniques that seek to provide a weighted measure of the accuracy of a classification. These techniques seek to weight the accuracy estimate by some quantifiable measure.

2.5.6.1 Minimum Accuracy Value

The minimum accuracy value (Aronoff 1985) is, "a way of representing more of the information contained in an accuracy test by indicating not only whether the map passed or failed, but also how well it passed or failed." The minimum accuracy value represents the lowest expected accuracy estimate of a thematic map. This minimum accuracy value is based on the result of a standard accuracy measure, such as overall percent accuracy, an analyst-selected level of "consumer risk," and the cost of misclassifying each of the classes. The consumer risk term represents the probability that a map that is unacceptably accurate will pass the standard accuracy test. These risk values are selected in the same range as the alpha levels of confidence intervals, i.e. 0.05, 0.10, etc. The misclassification costs represent the costs of misclassifying reference sites of each class, as determined by an analyst-selected loss function.

The resulting minimum accuracy value is intended to help the analyst determine whether a map is of sufficient accuracy for its intended purpose by weighting the accuracy estimate to reflect the priorities of the application. The minimum accuracy value represents the lowest expected accuracy of the classification, so it provides a tool to determine whether the worst-case accuracy of the map is appropriate for the application at hand.

The minimum accuracy value is an important addition to the set of accuracy assessment techniques. However, a criticism of the method is that, similar to Weighted Kappa, the analyst-determined weights are typically subjective in nature, and so may not be comparable from study to study.

2.5.6.2 Fuzzy Similarity

Jäger and Benz (2000) present a method for estimating the accuracy of fuzzy classifications accompanied by fuzzy reference data. As discussed above, fuzzy sets can be used to derive fuzzy reference data. Fuzzy sets can also be used to produce fuzzy classifications, where the grade of membership in each fuzzy class is specified for every image pixel. Accuracy assessment of these images is problematic using standard techniques such as overall percent accuracy and Kappa.

Jäger and Benz approach the problem of comparing fuzzy reference data to a fuzzy classification by computing a geometric distance-like measure between the two. They develop fuzzy functions which compare the reference and classified data sets using their fuzzy similarity measure and then “defuzzify” the result of the comparison to provide an interpretable accuracy estimate.

Furthermore, they generalize this approach for use with crisp reference data. They conclude that fuzzy classifications combined with fuzzy reference data provide the most appropriate accuracy estimates. As this approach is very new, it remains to be seen whether this innovative method will be widely used.

2.5.6.3 Economic Cost-Based Accuracy Assessment

Smits et al (2000) present an accuracy assessment protocol for quality assessment of thematic data which is weighted by the economic cost of misclassification. They recommend assessing the objectives of the study and determining a cost table that quantifies each possible misclassification. Following that, multiple classifications of the study area are created using different classifications techniques. Next, each classification is assessed using normalized Kappa values based on the reference data. Finally, they evaluate each classification relative to the sum of the economic costs involved in the errors therein and select the best classifier based on which one incurs the least misclassification cost.

The focus of this new approach is to relate the product of the classification to the needs of the end users – in this case the economic costs of not completely fulfilling the goals of the study. This general approach is not new. A similar accuracy assessment measure could be constructed using Weighted Kappa where the weights are the economic misclassification costs. Like Weighted Kappa, though, the economic cost-based approach suffers from the drawbacks that the misclassification costs may not be comparable from study to study, and may not necessarily have a strong quantitative base.

2.5.6.4 Accuracy Assessment Using Maximum Likelihood Classifier Output

Foody et al (1992) present a method of accuracy assessment that is geared towards improving the classification during its development. This improvement is accomplished by using the output of a maximum likelihood classifier (MLC) to determine areas that are likely to be incorrectly classified and then focusing extra effort on properly assigning those areas. Foody notes that it is common to use an MLC to produce a classified map, but the use of the MLC data often stops there. They assert that the *a posteriori* probabilities produced by the classifier can be used to improve the classification by correcting discrepancies through better training and then repeating the classification. They recommend that the MLC *a posteriori* probabilities be used as an indicator of per-pixel classification quality. These probabilities can be used to identify classes that are atypical of the highest probability class membership. In this way, the classification can be corrected mid-stream or, if correction is not possible, the low confidence areas can be labeled as such to alert map users of possible errors.

2.6 Summary

There have been many accuracy assessment measures proposed for use in evaluating classifications derived from remotely sensed data. Accuracy assessment has progressed from simple visual examination of the thematic map, through non-site-specific methods, to quantitative, categorical, and site-specific accuracy measures. However, none of these analytical techniques has incorporated inter-class spectral distances in accuracy estimates. The most similar study in the literature is the one described in Section 2.5.6.4 that uses the

output of an MLC classifier to revise the classification. However, the MLC technique requires the use of an MLC classifier, and furthermore does not provide single measure of the accuracy of a map.

Drawbacks such as a lack of objectivity and a quantitative basis for weighting schemes, and the potential for inter-interpreter variation in determining fuzzy membership values have led to underutilization of some of these otherwise very powerful techniques. Incorporating inter-class distances to develop new accuracy measures based on these underutilized techniques is expected to provide a new outlook on accuracy assessment that allows more widespread adoption of these powerful techniques. We believe that the incorporation of inter-class spectral distances will strengthen the accuracy assessment tools available to remote sensing researchers. The specific goals of this research and the methods for using inter-class spectral distances in accuracy assessment are described in the following sections.

3.0 OBJECTIVES

The overall objective of this research is to develop a new approach to accuracy assessment in which the spectral distances between the classes in the classification are used as inputs to the error estimation process. The conceptual basis for this new approach is that the confusion of relatively spectrally different classes represents a more severe error than confusing relatively spectrally similar classes. Therefore, the accuracy estimate of a classification can be adjusted to take into account the “spectral severities” (or misclassification costs) of the errors in that classification. The benefits of including inter-class spectral

distances in the accuracy assessment process are shown in the context of the development of two new accuracy assessment measures called Spectrally Weighted Kappa (SWK) and Spectrally Weighted Fuzzy (SWF), which are based on current techniques: Weighted Kappa and fuzzy accuracy assessment.

The SWK combines the strength of Weighted Kappa's variable marginal error weights with inter-class spectral distances. SWK gives the weights an objective and quantitative base and makes the weight determination process much more repeatable than current methods. Furthermore, SWK provides a new perspective on accuracy of classification: one that shows how well the classifier performed in relation to the spectral information it had, rather than basing the accuracy estimate solely on the classification's absolute agreement with the reference data. *The overall hypothesis of the SWK experiments is that examining the difference between the SWK value and a standard Kappa can provide information about the nature of errors in a classification that is not apparent solely from standard accuracy assessment techniques*.

The SWF technique will help to solve the problems with current fuzzy accuracy analysis discussed in Section 2.5.5. As is shown below, the use of multiple reference data interpreters to determine class fuzzy membership values can introduce significant inter-interpreter bias. A more objective method of determining these membership values would be useful. *The overall hypothesis of the SWF experiments is that SWF can create similar membership functions to those that a single well-trained human might create. So in cases where multiple human interpreters would normally be used to determine the fuzzy membership*

values, the SWF approach can be used instead to reduce the potentially significant impact of inter-interpreter bias. Specifically, this approach: 1) reduces biases that may be introduced through the use of multiple reference data interpreters by determining the fuzzy membership values based on inter-class spectral distances, and 2) provides a quantitative basis for the creation of these fuzzy membership functions for accuracy assessment of thematic data derived from remotely sensed data.

4.0 APPROACH

This section describes the procedures for using SWK and SWF analysis techniques and the validation procedures that were undertaken to evaluate the performance of these two newly developed techniques. The general approach in using both SWK and SWF involves determining the spectral distances between each of the classes in the classification, and then incorporating those distances into the accuracy estimates. The inclusion of spectral distances adjusts the accuracy estimates to account for the spectral severities of the various errors present on the map.

The following list gives a general overview of the steps required to accomplish SWK and SWF accuracy analysis. The first four steps in the following general approach are, except for one addition to Step 2 (discussed in Section 4.2), the same whether SWK or SWF analysis is used. The two procedures diverge at Step 5, below.

1) Reference sites are chosen in the study area using a statistically appropriate selection scheme (e.g. simple random, stratified random, etc.)

2) Reference data are collected at each of the reference sites.

3) Samples of each thematic class are delineated on satellite imagery. This is accomplished by using the reference data to assist in locating appropriate areas on the image. The image used should preferably be the same image used to create the classification.

4) The Transformed Divergence (T.D.) algorithm is used to calculate the spectral distance between each of the pairs of image-based thematic

class samples. The T.D. algorithm represents the spectral distance on a transformed scale from 0 to 2000 (Erdas, Inc. 2001). A value of zero means that the classes are inseparable. A value of 2000 means that the classes are completely separable.

SWK:

5) For SWK, the inter-class spectral distances and interpreted reference data are used to construct the weights for Weighted Kappa analysis (Section 4.5).

6) Accuracy analysis is performed using the computed weights (Section 4.6).

SWF:

5) The inter-class spectral distance and interpreted reference data are used to construct the fuzzy class memberships (Section 4.7).

6) Accuracy analysis is performed using the constructed fuzzy membership functions (Section 4.8).

Both:

7) Validation: The performance of the two new techniques is evaluated in relation to currently used accuracy assessment methods. The validation steps are described in detail below (Section 4.9).

4.1 Reference Site Selection

Accuracy assessment requires the comparison of the thematic map data to some set of reference data that provides the class labels that will be considered correct. As discussed in detail in Section 2.0, a probability sampling

scheme should be used for reference site selection to ensure equal inclusion probability for every possible reference site.

The validation portions of this project use reference from the Neuse River Basin in North Carolina. The Neuse River Basin is a "near-lab study site" for the U.S. EPA in Research Triangle Park. The Neuse River Basin (NRB) stretches from north of Durham, NC to Cape Hatteras. The NRB spans two physiognomic regions of North Carolina: Piedmont and Coastal Plain. As such, the NRB contains a wide variety of landscape types; such as the various piedmont and coastal plain vegetation species, bottomlands, urban areas, and agricultural areas; and so is a strong source of data for a project dealing with LU/LC.

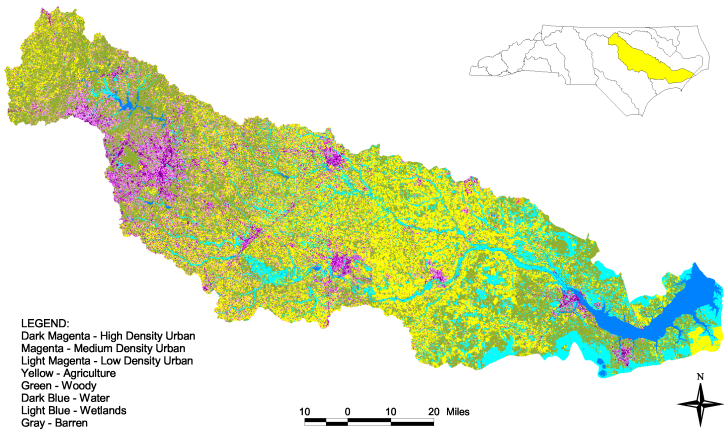

Figure 4. 1 The Neuse River Basin (classification produced by Lunetta, et al 2001)

4.2 Reference Data Collection

Upon selection of the sample sites, the reference LU/LC label must be determined for each site. This is often done through ground visits or by interpreting aerial photographs or other higher resolution data. The objective is to identify the "true" LU/LC at each site. Clearly, there will always be some

subjectivity in this step. Different people may select different reference classes. To further complicate the issue, multiple reference data sources may indicate different "true" classes. This is an issue that has yet to be successfully dealt with in remote sensing research.

There is an additional step in reference data collection for SWF analysis that is necessary for the creation of the fuzzy membership functions. The interpreters choose classes from the reference data in the following manner: If the interpreters are sure of the correct class, they identify that class and give it a membership value of five (5). A membership value of five corresponds to "Absolutely Correct" on the linguistic scale presented above. If there is some doubt as to the proper class, the interpreter chooses a class and assigns it a membership value of four (4). A value of four corresponds to "Probably Correct" on the linguistic scale. As is shown below, these interpreter assigned values are not used to measure the spectral separability of a reference site relative to other sites, and do not affect the outcome of SWF analysis. Rather, these values are used to give the analyst a measure of the confidence of the interpreter in his or her determination of the correct class for each reference site. Clearly, the interpreter's certainty of the "true" class for a particular site does not mean that there are not other classes that may be spectrally similar to that class.

4.3 Class Spectral Samples

The data used to determine the inter-class spectral distances are acquired by delineating on a satellite image areas that are representative of the thematic classes of interest. This is accomplished by using the reference data to assist in

locating appropriate areas on the image. The image used should preferably be the same image used to create the classification. This delineation process is similar to the selection of training samples with which to train a supervised classifier. However, one should not use in the accuracy assessment process the actual training samples used to create the thematic classification or the accuracy assessment could not be considered to be independent of the errors in the classification (Congalton, 1991)

4.4 Transformed Divergence Calculation

The spectral distance is calculated between each possible pair of thematic classes using the representative areas delineated on the image. The Transformed Divergence (T.D.) algorithm (Swain and Davis 1978,), is used for this calculation because: 1) It takes into account the mean, variance, and covariance of the delineated clusters when calculating the distance, whereas other methods (such as Euclidean distance) do not; and 2) it represents the inter-class spectral distances on a transformed scale to enable comparisons from study to study. This scale ranges from 0 to 2000. A T.D. value of zero means that the class pair under consideration is inseparable. A T.D. value of 2000 means that the classes are completely separable. This transformed scale is necessary so that T.D. values from different studies can be compared. T.D. has been used in remote sensing since the 1970s and is a well-accepted algorithm (Erdas, Inc. 2001, Swain and King 1973). The T.D. algorithm is as follows:

$$D_{ij} = \frac{1}{2} tr((C_i - C_j)(C_i^{-1} - C_j^{-1})) + \frac{1}{2} tr((C_i^{-1} - C_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T)$$

Equation 4. 1

$$TD_{ij} = 2\left(1 - \exp\left(\frac{-D_{ij}}{8}\right)\right)$$

Equation 4. 2

where

i and j are the two classes being compared

C_i is the covariance matrix of class I

μ_I is the mean vector of class I

tr is the matrix algebra trace function

T is the matrix algebra transposition function

4.5 Creation of Weights for SWK

The weights for SWK analysis are constructed based on the inter-class spectral distance values calculated above. Transformed Divergence outputs the inter-class spectral distances on the interval (0,2000), where 0 means that the classes are identical. These values are first transformed to a (0,1) scale where 0 means the classes are completely separable. This is accomplished using the following simple formula

$$w_{ij} = 1 - \frac{TD_{ij}}{2000}$$

Equation 4. 3

The resulting weight matrix w is a k by k matrix (where k is the number of thematic classes), composed of values on the interval (0,1), where w_{ij} represents

the weight in the i,jth cell of the weight matrix. These weights are used in place of the typically subjectively determined weights in Weighted Kappa analysis.

4.6 Weighted Kappa Analysis (SWK)

The standard Kappa coefficient describes the accuracy of a classification assuming that all errors are of the same severity. Clearly, this is rarely the case in remote sensing LU/LC classification. Weighted Kappa analysis allows for the use of different error severities for each of the marginals in an error matrix. These severities are represented by weights assigned to the cells in the matrix. SWK assigns these weights based on the spectral distances between the classes in the classification. In this way, confusion between spectrally different classes is given a higher weight than confusion between spectrally similar classes.

The first step in computing Weighted Kappa is to convert the standard error matrix into a matrix of proportions such that

$$p_{ij} = n_{ij} / n$$

Equation 4. 4

where n_{ij} is the original matrix,

p is the proportion matrix,

n is the number of reference sites, and

i and j are the row and column indicators.

The row and column totals for the proportion matrix are defined by

$$p_{i+} = \sum_{j=1}^{k} p_{ij}$$

Equation 4. 5

and

$$p_{+j} = \sum_{i=1}^{k} p_{ij}$$

Equation 4. 6

where k is the number of classes.

Therefore, let

$$p_o^* = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}$$

Equation 4. 7

be the weighted agreement, and

$$p_c^* = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i+} p_{+j}$$

Equation 4. 8

be the weighted chance agreement.

Then Weighted Kappa is defined by

$$\hat{K}_w = \frac{p_o^* - p_c^*}{1 - p_c^*}$$

Equation 4. 9

where: i and *j* are the row and column indicators
k is the number of rows and columns
p_{ij} is the proportion in the i,jth cell of the error matrix ($0 \leq p_{ij} \geq 1$)
w_{ij} is the weight assigned to the i,jth cell in the matrix ($0 \leq w_{ij} \geq 1$).

The large sample variance of Weighted Kappa is estimated by

$$\hat{var}\left(\hat{K}_w\right) = \frac{1}{n\left(1 - p_c^*\right)^4}\left\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[w_{ij}\left(1 - p_c^*\right) - \left(\bar{w}_{i+} + \bar{w}_{+j}\right)\left(1 - p_\circ^*\right)\right]^2 - \left(p_\circ^* p_c^* - 2p_c^* + p_\circ^*\right)^2\right\}$$

Equation 4. 10

where

$$\bar{w}_{i+} = \sum_{j=1}^{k} w_{ij}\, p_{+j}$$

Equation 4. 11

and

$$\bar{w}_{+j} = \sum_{i=1}^{k} w_{ij}\, p_{i+}$$

Equation 4. 12

4.7 Fuzzy Class Memberships Construction for SWF

The fuzzy membership values for SWF analysis are created by starting with the reference class membership value interpreted from the reference data, and then scaling the remaining classes based on their spectral distance from the interpreted class. These remaining classes are assigned scaled fuzzy membership values based on their spectral distance from the class the interpreter selected. Figure 4.2 shows the function used to accomplish this scaling. The X-axis values represent the T.D. spectral distances between classes. The Y-axis determines the multiplier that is used to compute the membership value of a particular class. The fuzzy membership value, f_{ij}, of a class *i* in relation to another class *j* is computed using the following formula:

$$f_{ij} = 4m$$

Equation 4. 13

where m is the membership value read from Figure 4.2.

The number four was used to approximate the range of values present in the linguistic scale presented above. So, the possible fuzzy membership values computed using this method are on the interval {0..4}.

To create the function in Figure 4.2, 100 test sites of various LU/LC types and from geographic different areas and data types were examined manually to determine how the 0-2000 T.D. scale corresponds to the differences between LU/LC classes. The results of these comparisons show that T.D. values below 800 result only from classes that are very spectrally similar. In these cases, the multiplier is fixed at one. Conversely, T.D. values above 1800 represent classes that are very spectrally different. In these cases, the multiplier is fixed at zero. The range between these two extremes is represented by a linear function with a negative slope. The informal manual examination of the aforementioned 100 test points indicates that this function adequately represents fuzzy class membership values such as those that might be created by a human interpreter. An example of this process is provided in the following section.

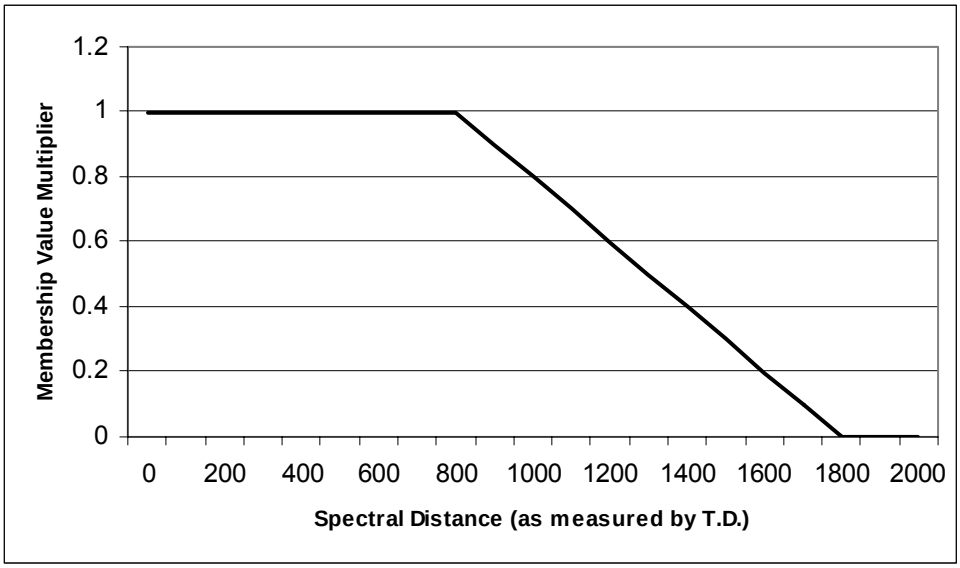

Figure 4. 2 Membership Value Assignment Function

The following example illustrates the process of determining the fuzzy class membership values for a sample site. This example thematic classification is composed of three classes (A, B, and C). The sample site was interpreted as belonging to class A (the "true" class). The calculated T.D. values between class A and classes B and C are shown in Table 4.1.

Table 4. 1 Example site inter-class T.D. values

Class	A	B	C
A	0	1040	1625

Looking up the T.D. values from Table 4.1 on Figure 4.2 gives multiplier values of 0.8 for class A vs. class B, and 0.2 for class A vs. class C. Finally, computing the class membership values using Equation 4.13 gives the results shown in Table 4.2. The spectral similarity between classes A and B results in those classes receiving high membership values for the example site. Since

class C is more spectrally different from the “true” class A, it receives a much lower membership value for the example site.

Table 4. 2 Example site fuzzy membership values

Class	A	B	C
A	4	3.2	0.8

These membership values can then be used in standard fuzzy accuracy assessment. This procedure is discussed in the following section.

4.8 Fuzzy Accuracy Assessment

Fuzzy accuracy assessment is accomplished by comparing the classified, or map, LU/LC class with the LU/LC class fuzzy membership values for each reference sample site. For example, the map class for a particular sample site may be “forest.” A set of interpreted reference fuzzy membership values for that same site derived from the aforementioned linguistic scale could be as follows: agriculture (2), water (1), forest (4), urban (2), shrubland (3), etc. Fuzzy operators are then defined to determine accuracy estimates.

The “Max” operator compares the map class with the fuzzy reference class that has the highest membership value. In the above example the site would be counted as correct according to the Max operator because the map class, forest, matches the reference class that has the highest fuzzy membership, forest (4).

Another operator, “Right,” compares the map class with any reference membership values that equal or exceed a certain pre-defined threshold. The typical threshold value is three (Gopal and Woodcock, 1994). So, in the above

example, the site would be counted as correct according to the Right operator because the map class, forest, has a membership value greater than or equal to three. Note that, if the site had been classified as "shrubland," the site would still have been counted as correct according to the Right operator because the shrubland class has a membership value greater than or equal to three.

Formal definitions of the above concepts are as follows (Gopal and Woodcock, 1994). Let X be a universe of sample sites, with a particular instance of X represented by x. The fuzzy set A of X contains the membership function μ_A that associates each sample site in X with a real number on the interval (0,1). The membership of x in A is represented by $\mu_A(x)$. Therefore:

$$A = \{(x, \mu_A(x)) \mid x \in X\}$$

Equation 4. 14

Let ζ be the set of thematic classes assigned to the sample sites in X. Thematic map accuracy can then be conducted using:

$$A_C = \{(x, \mu_C(x)) \mid x \in S\}$$

Equation 4. 15

where S is a subset, with quantity n, of the sample sites, $S \subset X$. Thus, the two operators, Max and Right are defined as:

$$MAX(x,C) = \begin{cases} 1 & if \mu_C(x) \geq \mu_C(x) \; for \; all \; C' \in \varsigma \\ 0 & otherwise \end{cases}$$

Equation 4. 16

$$RIGHT(x,C) = \begin{cases} 1 & if \mu_C(x) \geq \tau \\ 0 & otherwise \end{cases}$$

Equation 4. 17

where τ is the threshold value, which is defined here as $\tau = 3$.

In this research, the Max and Right values are used to provide accuracy estimates for the test thematic classifications used in the validation section.

4.9 Methods Validation

Thematic classifications must be compared to some kind of objective measure that is accepted as "truth" (i.e. reference data) to determine their quality. However, truth itself is variable. For a given area, there is no one reference data map that can be universally regarded as true. Therefore, when evaluating the performance of a new accuracy assessment method, i.e. a new method of describing the performance of a classifier compared to some true data, there can be no absolute standard against which to compare the old versus the new accuracy assessment schemes. Therefore, testing of accuracy assessment methods is limited to examining the estimates given by the techniques in question for data sets with known errors. In this way, statements can be made

regarding the ability of a particular accuracy assessment method to identify the effects of particular types of errors.

4.9.1 SWK Validation

The hypothesis for these tests is that comparing the accuracy estimates of the SWK and standard Kappa techniques can quantify the effects of certain error types in a thematic classification that may be masked or diminished by current techniques. For example, if a classification contains many errors between classes that are very spectrally similar, a standard Kappa will give each of those errors full weight. This may unfairly penalize the classifier. In contrast SWK will take into account that the errors are between spectrally similar classes and adjust the accuracy estimate accordingly.

4.9.1.1 Error Types Tested

To test this hypothesis, a set of experiments were run that compared the performance of the SWK method with current techniques over a range of data sets. These data sets contained actual errors that one might encounter when performing a thematic land use and land cover (LU/LC) classification, as shown in the following partial list of common errors identified by the remote sensing community (Congalton 1993):

- *Changes in land cover between acquisition dates of the imagery and the reference data* – A frequent problem in remote sensing research, LU/LC change can lead to incorrect estimation of the performance of a classifier

because the interpreted reference data for changed areas may not reflect the composition of the study area at the time the classification was performed.

- *Inappropriate use of remote sensing data source* – This type of error results when the objectives of a study cannot reasonably be achieved with the data types used. A common situation is when two or more LU/LC classes cannot be separated.
- *Inadequate training data* – Errors of this nature occur when the classifier is inadequately trained to discriminate the classes of interest. Inadequate training is a primary cause of misclassification. These errors are different from the "inappropriateness" errors above in that the sensor is capable of discriminating the classes if it is properly trained.

Not included in these tests is a fourth category of errors: positional errors. The SWK technique (like most other accuracy assessment techniques) assumes that the reference data and thematic data for a particular point correspond positionally.

The above list is not meant to be a exhaustive list of possible error sources in thematic classifications; there are several other error types that may occur, such as choosing a sub-optimal classification algorithm, using inappropriately scaled imagery, reference data interpreter error, etc. Rather, the three error types tested were chosen because they represent commonly seen errors and because they can be more objectively quantified.

4.9.1.2 Testing Overview

Several tests were conducted to measure the ability of the SWK technique versus the standard Kappa to detect the effects of the above listed error types. In the first experiment, the two techniques were compared over three geographic areas in the Neuse River Basin using the same image data type to determine the weights. This experiment used three different spectral weighting schemes – one from each area. In a second experiment the two techniques were compared using three image data types in one geographic area. This experiment used three different weighting schemes – one from each image data type. These tests are intended to determine whether the ability of SWK to detect the effects of the above error types is consistent over multiple study areas and data types.

These validation tests use the locations of randomly selected reference sites throughout the Neuse River Basin. The background of these reference sites is that they were originally chosen for use in an EPA LU/LC classification validation study. The sites were chosen using a Systematic Unaligned Random sampling scheme and were visited by EPA field crews during the summers of 1998 or 1999. In the EPA study, LU/LC labels were determined independently by two EPA interpreters using the collected ground sample data as a guide. Lunetta et al (2001) found that the agreement between the two interpreters ranged from 91% to 100% depending on the level of thematic detail that was requested from the interpreters. These high levels of agreement were possible because of the high quality of the ground reference data and proper training of the EPA interpreters.

4.9.1.3 Experiment One: Multiple Geographic Areas

The hypothesis for this experiment is that the SWK technique will perform similarly in relation to the standard Kappa over different local geographic areas within the NRB; i.e. that the impact of local landscape variability will be small enough that it will not significantly change the spectral characteristics of the thematic classes. For this experiment, the Neuse River Basin (see Figure 4.1) was divided into three roughly equal parts (NRB sub-areas) from east to west. From each of these NRB sub-areas, class spectral samples were taken from SPOT multispectral imagery. These samples were then, using the previously described steps, used to create weights for SWK analysis. Next, using these three sets of NRB sub-area weights, SWK analysis was conducted for an example of each of the following sets of error types (as introduced above):

A. Changes in land cover type: Represented by a land cover class conversion from deciduous forest to grass. This test simulates an error caused by an area being logged between the acquisition dates of the imagery and the reference data (a common occurrence). A forest to grass class conversion is common in a recently logged area where sufficient time has passed after the logging to allow some short vegetative cover to re-grow.
B. Inappropriate use of remote sensing data source: Represented by attempting to separate urban grassland from natural grassland. This test simulates an error caused by attempting to map classes that are inseparable on the imagery. In the vast majority of cases, a satellite sensor will not be able to

separate urban grass from natural grass. Note that term “natural grassland” in this case is used because the EPA Neuse River Basin classification uses that name for the class in question. We recognize that there is very little land cover in North Carolina that is actual natural grassland.

C. Inadequate classifier training: Represented by an error in which the goal is to separate bare soil from rooftops (urban), but the classifier is not trained well enough to discriminate these two spectrally similar classes. This was a problem that occurred during the initial investigation stages of a project conducted by Khorram and Knight (2000). The bare soil in question is a sandy soil type. The rooftops are white in color.

4.9.1.4 Experiment Two: Multiple Image Data Types

This experiment tests whether the variance in class spectral values over different image data types affects the relationship between SWK analysis and standard Kappa; i.e. whether it is important to use the same image data type for both the classification and collection of class spectral samples for SWK analysis. For this experiment, three very different image types were analyzed within one NRB sub-area. The image data types are: Landsat 7 Thematic Mapper, SPOT multispectral, and a color infrared digital orthorectified quarter quadrangle (CIR DOQQ). Spectral samples were obtained for the thematic classes using each data type. These samples were then, using the previously described steps, used to create weights for SWK analysis. Next, using these three sets of image data type weights, SWK analysis was conducted for the *Changes in land cover type*

error source as described in Experiment 1, Part A. The image data type tests are referred to as Experiment Two, Parts A, B, and C.

4.9.1.5 Testing Procedure

The following procedure was used to conduct all Parts of Experiments One and Two:

1. For each of these pairs of class confusions an error matrix was constructed using three classes: the two confused classes and a third "placeholder" class to make the error matrix sufficiently large (3x3) to enable the use of Weighted Kappa analysis. These placeholder classes had no effect on the results of these experiments.
2. A stratified random sampling scheme was used to select 100 points from each of the three classes.
3. The reference data were first assumed to be in perfect agreement with the classification for all three classes. Then, the number of errors between the chosen class pair was increased gradually by random number addition to examine the behavior of the SWK value versus the standard Kappa
4. As the number of errors increased, the point at which the difference between the SWK and Kappa becomes statistically significant (according to a Z-test) was recorded. This Z-test formula is:

$$Z = \frac{\left|\hat{K} - \hat{K}_w\right|}{\sqrt{\hat{\mathrm{var}}\left(\hat{K}\right) + \hat{\mathrm{var}}\left(\hat{K}_w\right)}}$$

Equation 4. 18

where

$$\hat{K} = \frac{p_o - p_c}{1 - p_c}$$

Equation 4. 19

where

$$p_c = \sum_{i-1}^{l} p_{i+} p_{+j}$$

Equation 4. 20

and

$$p_o = \sum_{i=1}^{k} p_{ii}$$

Equation 4. 21

Computation of the variance of Kappa, $\hat{\mathrm{var}}\left(\hat{K}\right)$, requires a long series of equations, which can be found in Congalton and Green (1999).

5. This process was repeated for 100 blocks of 100 reps each for each class pair chosen (e.g. natural grassland versus urban grassland). For example, in Experiment 1 Part A, which represented a LU/LC conversion from forest to grass, the number of errors between those two classes was chosen randomly within one of the following sets: 1-10, 11-20, 21-30, etc., to a maximum of 100 errors (100 errors would be 50% of the 200 total reference points in those classes, or equivalent to a random choice between the two classes). From each set (1-10, 11-20, etc.) this error frequency was randomly chosen 100 times. The Z-test was then used to determine if the SWK was significantly different from the standard Kappa

with the chosen frequency of errors. This entire process was then repeated 100 times, thus effectively testing each set of error frequencies 10,000 times.

4.9.1.6 False-Positive Test

An additional test was conducted to determine the likelihood of obtaining a "false positive" result in the significance test between the Weighted Kappa and standard Kappa. This condition is tested by setting up a situation where the SWK and standard Kappa should be equal, which is when the numbers of errors between the class pairs in a classification are exactly inversely proportional to the spectral distances between those classes. *The hypothesis of this test is that the SWK and standard Kappa will not significantly diverge, as measured by the Z test described above, when the numbers of errors between the various class pairs in a classification are exactly inversely proportional to the spectral distances between those classes.* This test is conducted by populating an error matrix of three classes with sets of total numbers of errors (0-10, 11-20, etc.) and then distributing those errors between the classes so that they fit the condition described in the hypothesis. The population of the matrix is then repeated 100 times for each of the sets of error frequencies. The three classes used are the previously discussed natural grassland, forest, and bare soil.

4.9.2 SWF Method Validation

The hypothesis of the SWF tests is that the SWF method can provide fuzzy membership values that are similar to what a well-trained human might

select. The validation of this hypothesis is divided into two parts. First, Section 4.9.2.1 demonstrates that the effects of inter-interpreter variation inherent when multiple interpreters determine fuzzy membership values can be significant. Second, Section 4.9.2.2 tests to determine whether the SWF method can produce membership values similar to those produced by a well-trained human interpreter. The test data sets used in this section contained actual errors that one might encounter when performing a thematic land use and land cover (LU/LC) classification, as described in the SWK testing procedures above.

4.9.2.1 Experiment Three: Multiple Interpreters

The hypothesis for this test is that fuzzy membership functions created by multiple interpreters have the potential for statistically significant inter-interpreter variation. This test presents the results of a large experiment conducted to determine whether multiple interpreters can reliably be used to determine fuzzy membership values for reference sites. This test uses three interpreters to determine membership values for 270 randomly selected points throughout EPA Federal Region 5. This Region consists of the states of Ohio, Indiana, Illinois, Michigan, Wisconsin, and Minnesota. The interpreters were asked to select membership values from the linguistic scale described above for each of 18 LU/LC classes. These membership values were then examined using correlation analysis to determine the amount of similarity between the three interpreters' membership values.

4.9.2.1.1 Testing Context

The aforementioned data was collected as part of the accuracy assessment of the Multi-Resolution Land Characteristics (MRLC) Consortium's National Land Cover Data (NLCD) for Region 5. The MRLC Consortium, including the U.S. Geological Survey (USGS), the U.S. Environmental Protection Agency (EPA), and other Federal agencies, has completed the NLCD program. The dataset provides a consistent and conterminous land cover map of the lower 48 States at approximately an Anderson Level II thematic detail. The program used Landsat Thematic Mapper (TM) 30-meter resolution imagery as the baseline data. The central goal of the program is to provide a regionally consistent land cover product for use in a broad range of applications (Lunetta and Elvidge 1998). Each of the ten Federal regions was mapped independently.

4.9.2.1.2 MRLC Project Accuracy Assessment Sample Determination

A stratified random sampling scheme was used to select 1800 sample sites – 100 for each of the 18 LU/LC classes present in Region 5. For the experiment described in this chapter, 270 (15%) of the 1800 sample sites were chosen using a simple random selection scheme. All of the three interpreters provided fuzzy class membership values for each of these 270 sites. These membership values are the basis for the testing in this experiment.

4.9.2.1.3 MRLC Project Interpreter Training

Comprehensive training procedures were undertaken to ensure that the interpreters were sufficiently knowledgeable to determine the fuzzy membership values. The training consisted of several full-day classroom training sessions

and "on the job" training. The formal classroom training sessions were led by experienced airphoto interpretation and photogrammetry instructors. The training sessions included the following:

- Discussion of color theory and photo interpretation principles and techniques;
- Understanding of the class definitions;
- Interpretation of over 100 sample sites of different classes during the training sessions, followed by interactive discussions about potential discrepancies;
- Creation of sample sites for later reference; and
- Repetition of photo interpretation practice after the sessions.

The focus of this training was on situations that the interpreters would encounter during the project. Each participant was presented with approximately 100 pre-selected sites and was asked to provide their interpretation of the land cover for these sites. Their calls were analyzed and subsequently discussed to minimize any misconceptions.

During the "on the job" portion of the training, the interpreters were randomly assigned 199 sites from the 1800 sample that they interpreted as a group. This approach was used to “calibrate” the interpreters so that there would be as little inter-interpreter variation as possible. Their progress was monitored daily for accuracy and proper methodology. The interpreters kept a log of their calls and the sites for which they were uncertain about the land cover classes.

Problem sites were discussed until the group reached a consensus on the proper class for each site. This calibration procedure provided a solid foundation for further independent photo interpretation work.

4.9.2.1.4 Correlation Analysis

The fuzzy membership values determined by the three interpreters for the 270 sample sites are subjected to correlation analysis using the Pearson correlation coefficient (R). This analysis measures the correlation between the three interpreters' for each of the 18 LU/LC classes, as well as the overall correlation over the entire sample. The results of this work are presented below.

4.9.2.2 Experiment Four: Comparison of SWF versus Human-Derived Memberships

The hypothesis for this experiment is that the SWF technique creates membership values that not statistically different from those created by a well-trained human interpreter. This hypothesis will be tested by comparing Max and Right values computed in a test accuracy assessment for both SWF and human-derived membership values. This test is conducted for each of the following sets of error types (as introduced above): changes in land cover type, inappropriate use of remote sensing technology, inadequate classifier training. The following procedure was used to conduct this test:

1. Five hundred (500) sample sites were selected throughout the Neuse River Basin using a stratified random sampling scheme. The sample was stratified by LU/LC class, with 100 points for each of the five classes used

(forest, natural grassland, urban grassland, bare soil, and rooftops). These classes correspond to those used in the selected error types introduced above.

2. For each of these sample sites, membership values were created in two ways: 1) by a human interpreter; and 2) using the SWF method. The reference data was Landsat Thematic Mapper satellite imagery and Digital Orthorectified Quarter Quadrangles (DOQQs).
3. The reference data for the 500 sample sites were first assumed to be in perfect agreement with the classification for all five classes. Then, the number of errors between the pairs of class confusions was increased by random number addition to examine the behavior of the Max and Right values generated by SWF versus those generated by the human-derived membership values.
4. The values of Max and Right were recorded after each random number addition. These values represent the total number of sites, of the 500 total sites sample, that would have been counted as correct using the applicable Max or Right rule.
5. This process was repeated 100 times for each of the following ranges of error frequencies: 1-10, 11-20, 21-30, etc., to a maximum of 100 errors (100 errors would be 50% of the 200 total reference points in the two confused classes, or equivalent to a random choice between the two classes).

Single factor Analysis of Variance (ANOVA) tests were conducted to measure the correspondence between the Max and Right values determined by SWF and the values determined by the human-derived membership values.

5.0 RESULTS

The following sections present the results of the validation steps described above. The SWK results are presented first, in Section 5.1. The SWF results are presented in Section 5.2.

5.1 SWK Results

To describe the general behavior of SWK versus the standard Kappa, an example of the general behavior of SWK for an arbitrary weighting scheme derived from simulated data is presented in Section 5.1.1. Following that are the results of Experiments One and Two, introduced above.

5.1.1 General Case Behavior of SWK

Using the SWK method of weighting accuracy estimates, the impact of an individual error on the weighed kappa is determined by the spectral distance between the map class and the reference class for that site. A simplified situation is described in the following graphs derived from an arbitrary weighting scheme using simulated data. In general, in cases where the majority of the errors in a classification are between classes that are spectrally similar, the Weighted Kappa will trend higher than kappa because the weights for these errors are smaller (Figure 5.1). If the majority of the errors in a classification are between classes that are spectrally different then the weighed kappa will trend lower than kappa because those errors carry more weight (Figure 5.2.)

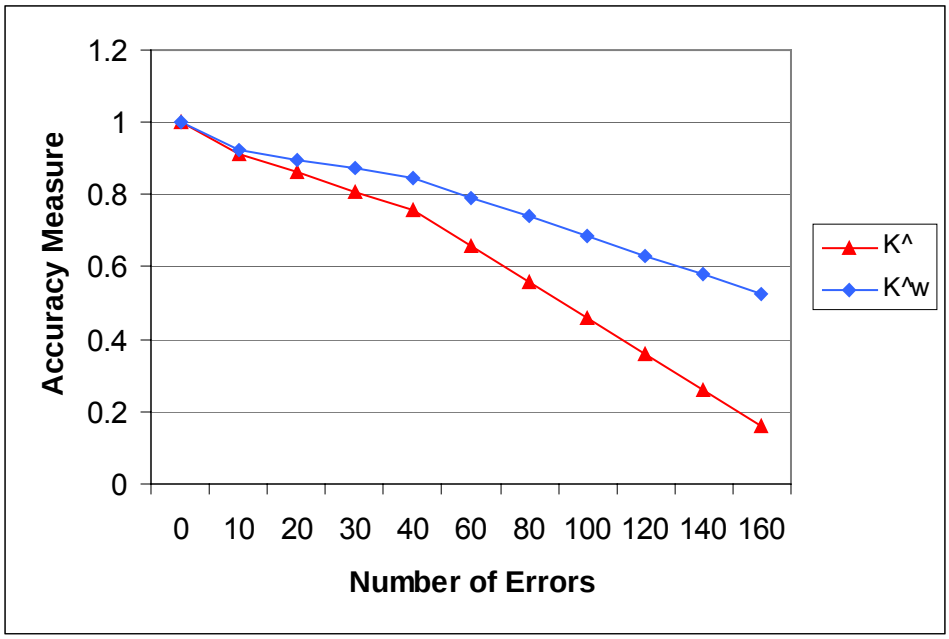

Figure 5. 1 General Case – Errors in Similar Classes

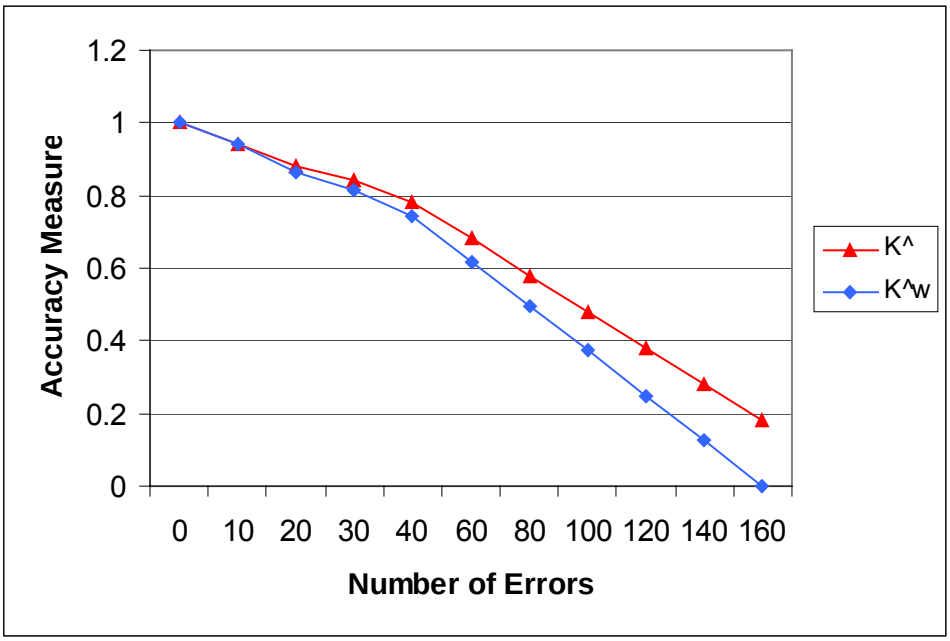

Figure 5. 2 General Case – Errors in Dissimilar Classes

5.1.2 Experiment One Results: Multiple Areas

Experiment One examined the behavior of SWK versus standard Kappa over three geographic areas for three of the most common errors in remote sensing thematic classifications. The results support the hypothesis that SWK is able to emphasize errors that are masked by current techniques such as Kappa.

5.1.2.1 Experiment One, Part A: Forest vs. Grassland

Figures 5.3, 5.4, and 5.5 present the results of the comparison of SWK versus Kappa for a LU/LC class conversion between coniferous forest and grassland for the three NRB sub-areas. The dissimilarity of the spectral samples of the two classes leads the SWK to begin to significantly diverge from the Kappa (as measured by the Z-test) when the number of class errors exceeds 30. As the number of errors increases, the probability of SWK being significantly different from Kappa increases. We reach a very strong probability of detection, 95%, when the number of errors is approximately 60. Sixty errors out of 200 total points is an overall percent accuracy of 70% (140/200), a common value in thematic classifications derived from remotely sensed data. Therefore we can say with confidence that the SWK technique will provide information about the nature of this type of class confusion that is not provided by Kappa when the number of those errors is within bounds commonly seen in remote sensing studies.

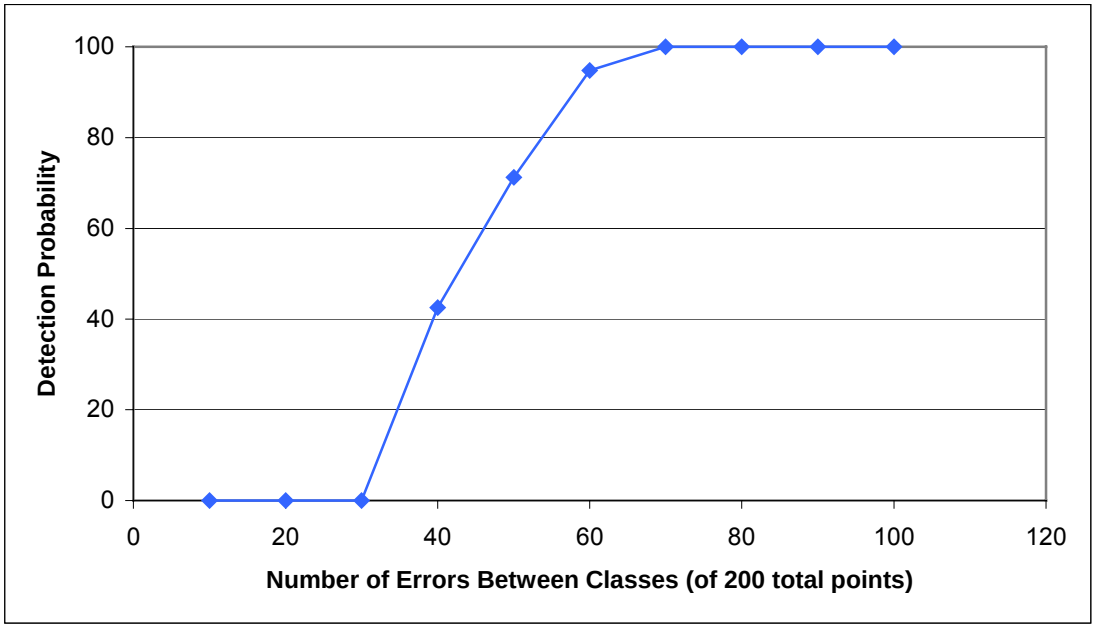

Figure 5. 3 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: NRB Sub-1

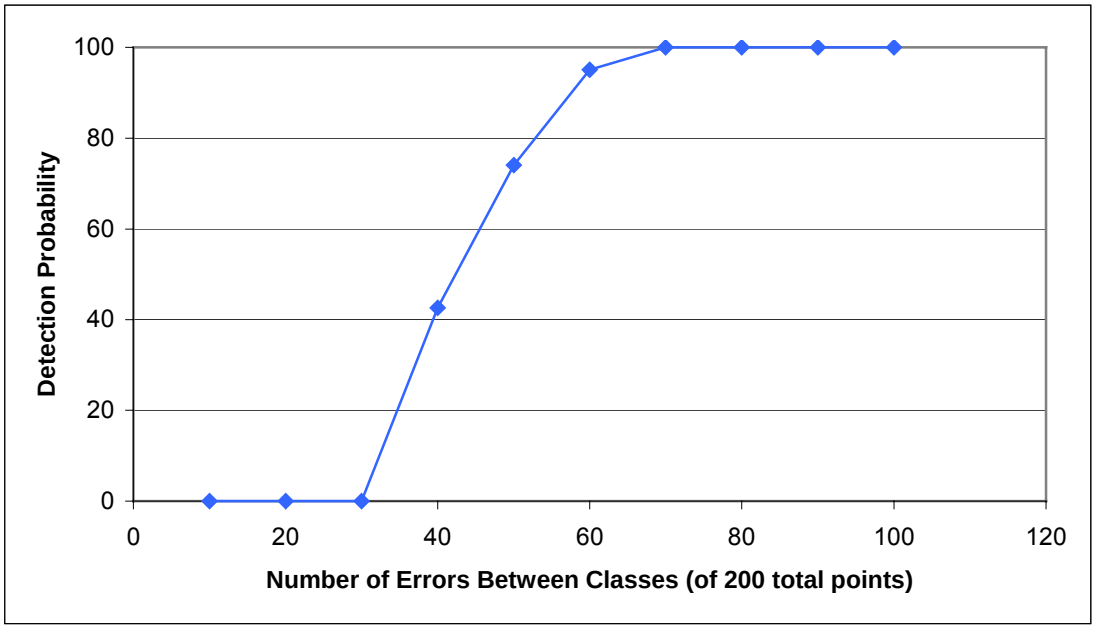

Figure 5. 4 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: NRB Sub-2

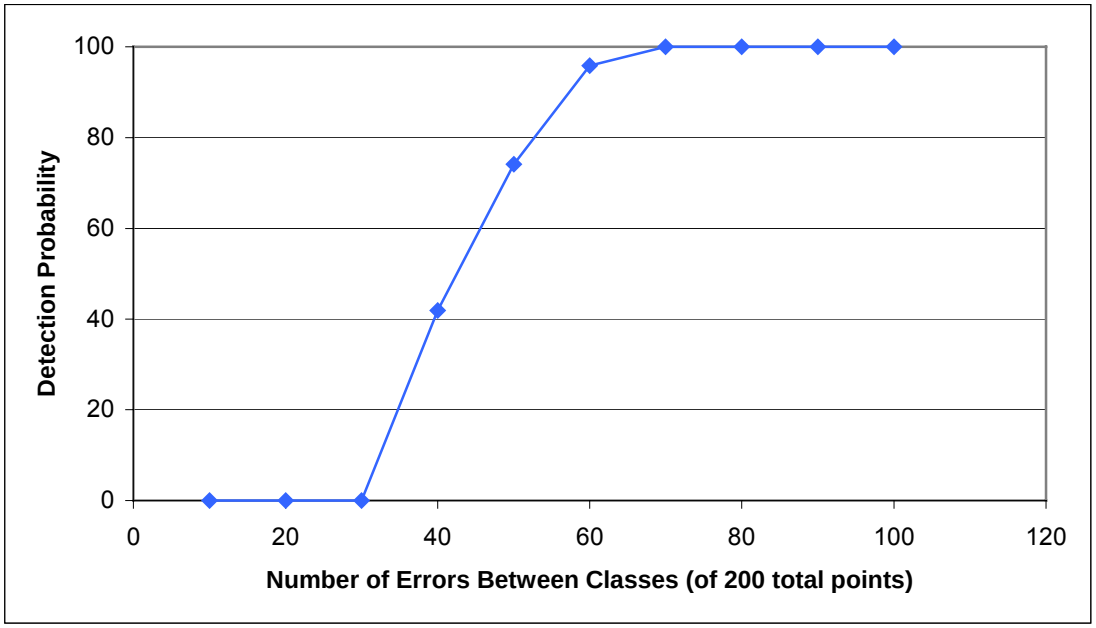

Figure 5. 5 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: NRB Sub-3

5.1.2.2 Experiment One, Part B: White Rooftops vs. Sandy Bare Soil

Figures 5.6, 5.7, and 5.8 present the results of the comparison of SWK versus Kappa for a LU/LC class conversion between rooftops and bare soil for the three NRB sub-areas. These class spectral samples are more similar than those in Part A, and so the SWK to begins to significantly diverge from the Kappa (as measured by the Z-test) when the number of class errors exceeds 25. As before, as the number of errors increases, the probability of SWK being significantly different from Kappa increases. A greater than 95% probability of detection occurs when the number of errors is approximately 58. Fifty-eight errors out of 200 points is an overall percent accuracy of 71% (142/200). Therefore we can again say with confidence that the SWK technique can

highlight this type of class confusion when the number of those errors is within the plausible bounds.

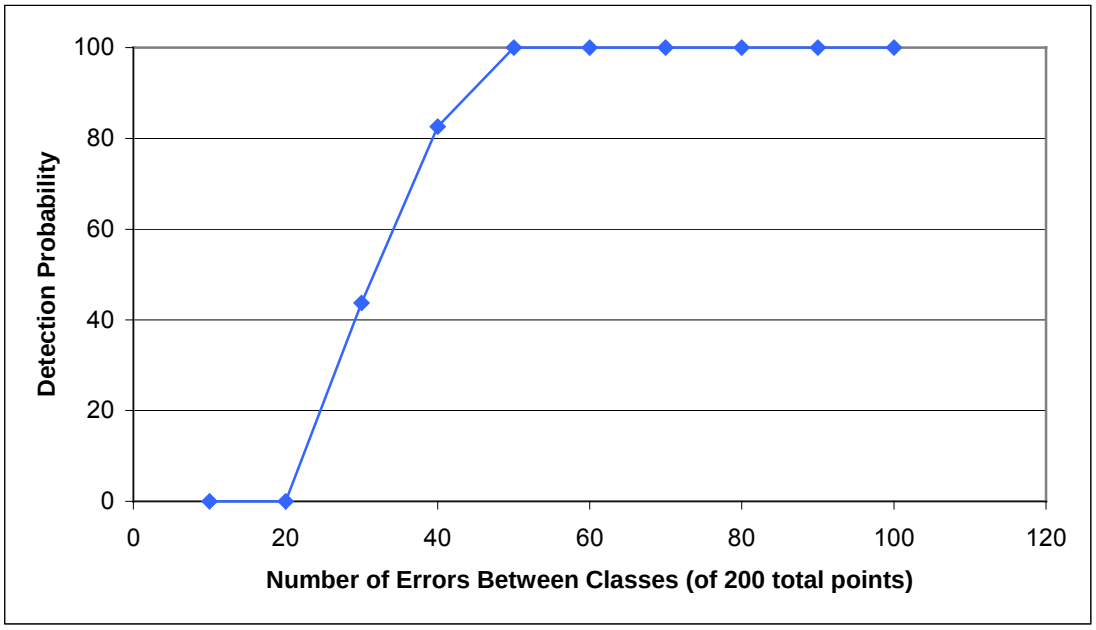

Figure 5. 6 Probability of Detecting a Difference between Kappa and Weighted Kappa - Rooftops vs. Bare Soil: NRB Sub-1

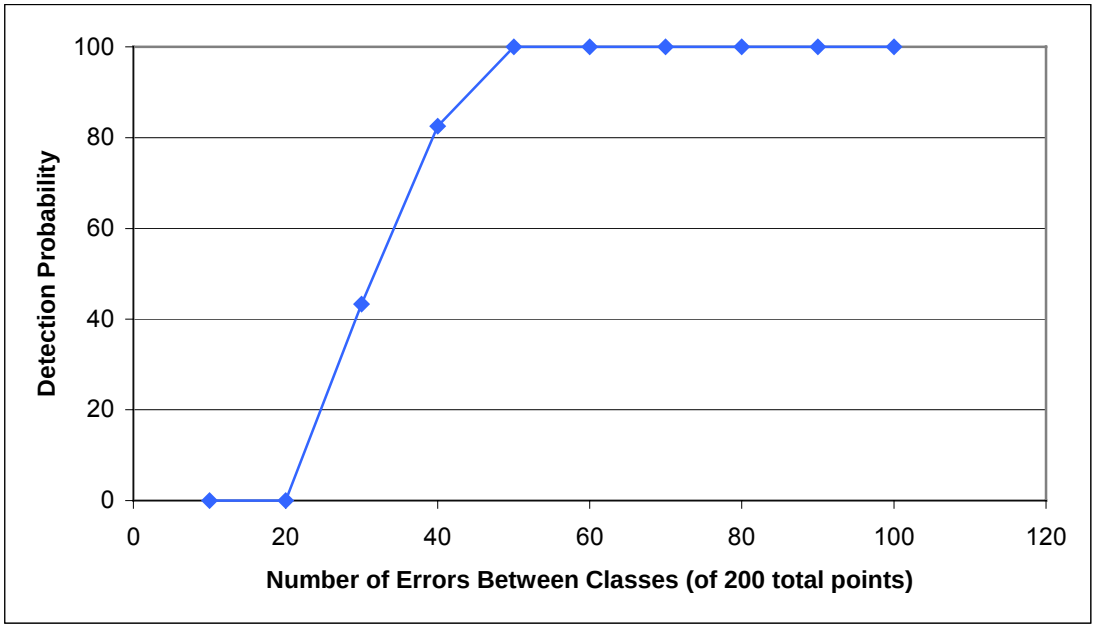

Figure 5. 7 Probability of Detecting a Difference between Kappa and Weighted Kappa - Rooftops vs. Bare Soil: NRB Sub-2

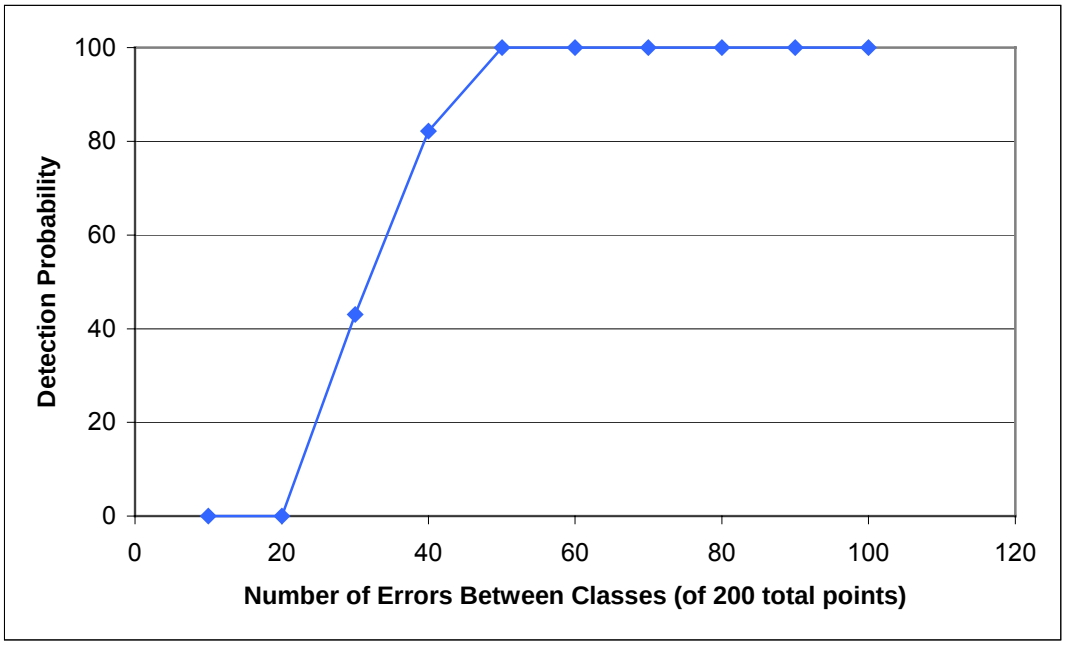

Figure 5. 8 Probability of Detecting a Difference between Kappa and Weighted Kappa - Rooftops vs. Bare Soil: NRB Sub-3

5.1.2.3 Experiment One, Part C: Urban Grass vs. Natural Grass

Figures 5.9, 5.10, and 5.11 present the results of the comparison of SWK versus Kappa for a LU/LC class conversion between the urban grass and natural grass classes for the three NRB sub-areas. Clearly, given similar image acquisition dates, these classes are likely to be nearly identical in spectral composition. This assertion is borne out in the three graphs. The SWK begins to significantly diverge from the Kappa (as measured by the Z-test) when the number of class errors exceeds 10. As before, when the number of errors increases, the probability of SWK differing significantly from Kappa increases. A greater than 95% probability of detection occurs when the number of errors is approximately 38. This value corresponds to an overall percent accuracy of 81% (162/200). With this result, we can say that SWK can provide information, which

Kappa does not, about this type of class confusion for quite low error occurrence rates.

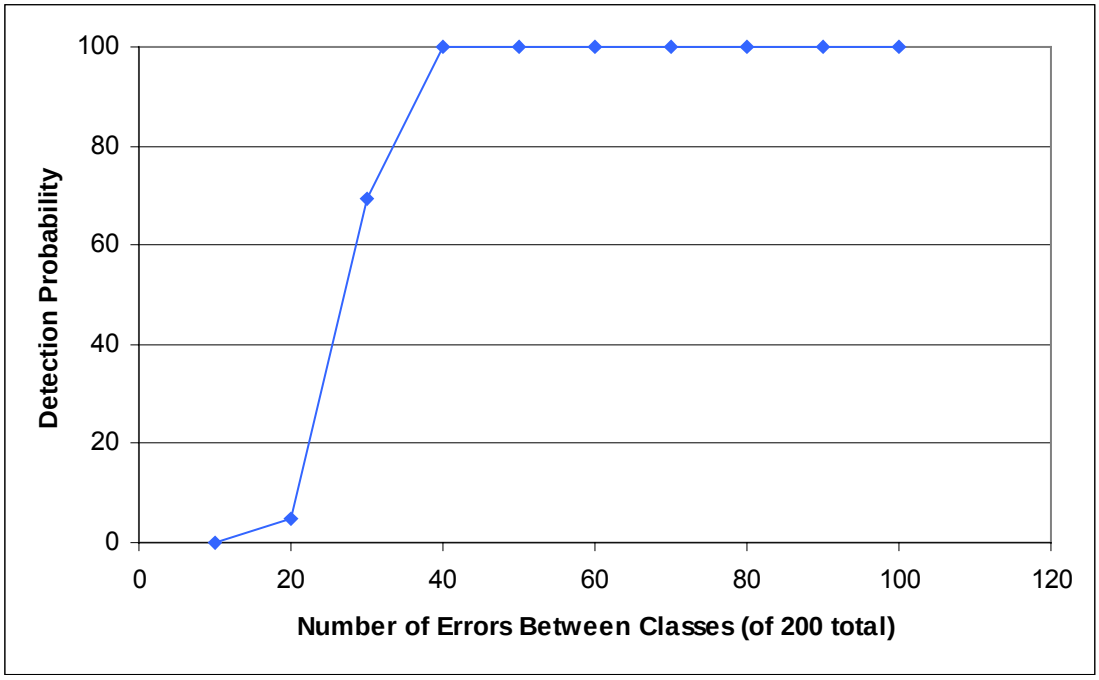

Figure 5. 9 Probability of Detecting a Difference between Kappa and Weighted Kappa - Natural Grass vs. Urban Grass: NRB Sub-1

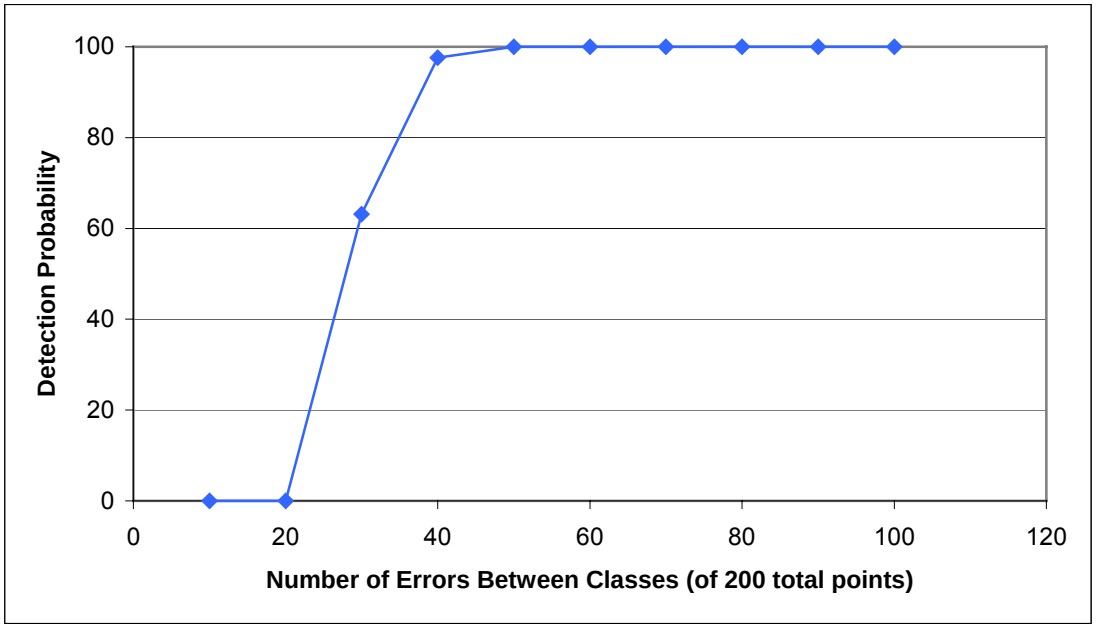

Figure 5. 10 Probability of Detecting a Difference between Kappa and Weighted Kappa - Natural Grass vs. Urban Grass: NRB Sub-2

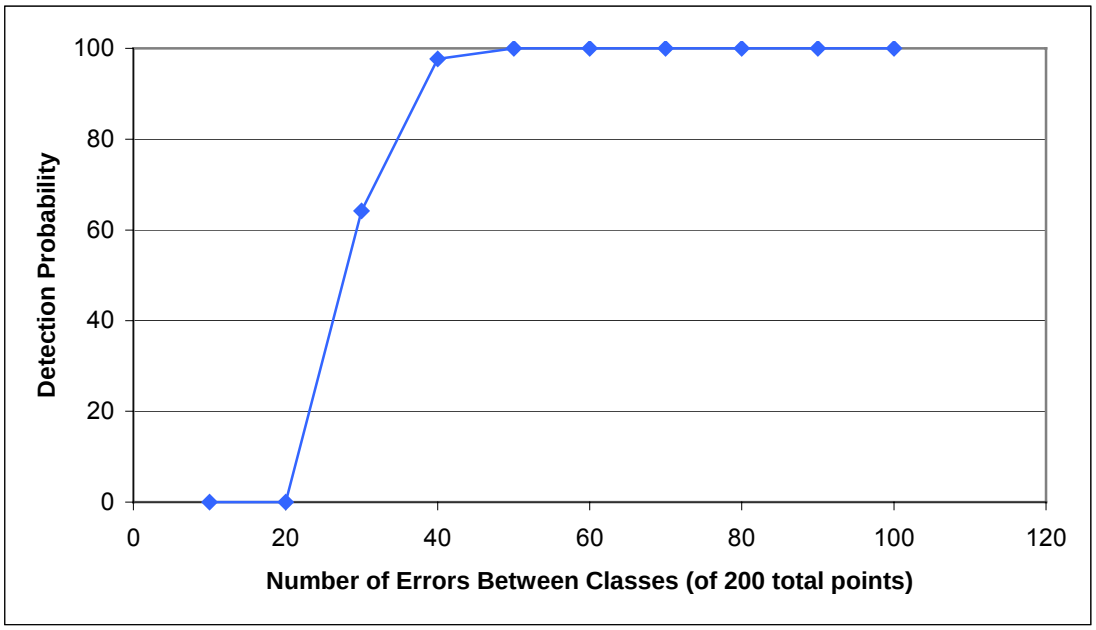

Figure 5. 11 Probability of Detecting a Difference between Kappa and Weighted Kappa - Natural Grass vs. Urban Grass: NRB Sub-3

5.1.2.4 Experiment One, NRB Sub-Area Comparisons

Figures 5.12, 5.13, and 5.14 show the curves for all three NRB sub-areas for each of the Part A, B, and C error types, respectively. For all three error types, the performance of the SWK versus Standard Kappa is consistent over the three geographic regions. Analysis of Variance (ANOVA) tests were conducted to compare the different sub-areas within each error type. In each case, the test failed to reject the null hypothesis of no mean difference between the three sub-areas quite convincingly. In each case, the *p* value was approximately 0.99 for an alpha level of 0.05. Note that due to the scale of the graph, the three curves on Figure 5.13 appear to overlap completely. In fact, they diverge a small amount.

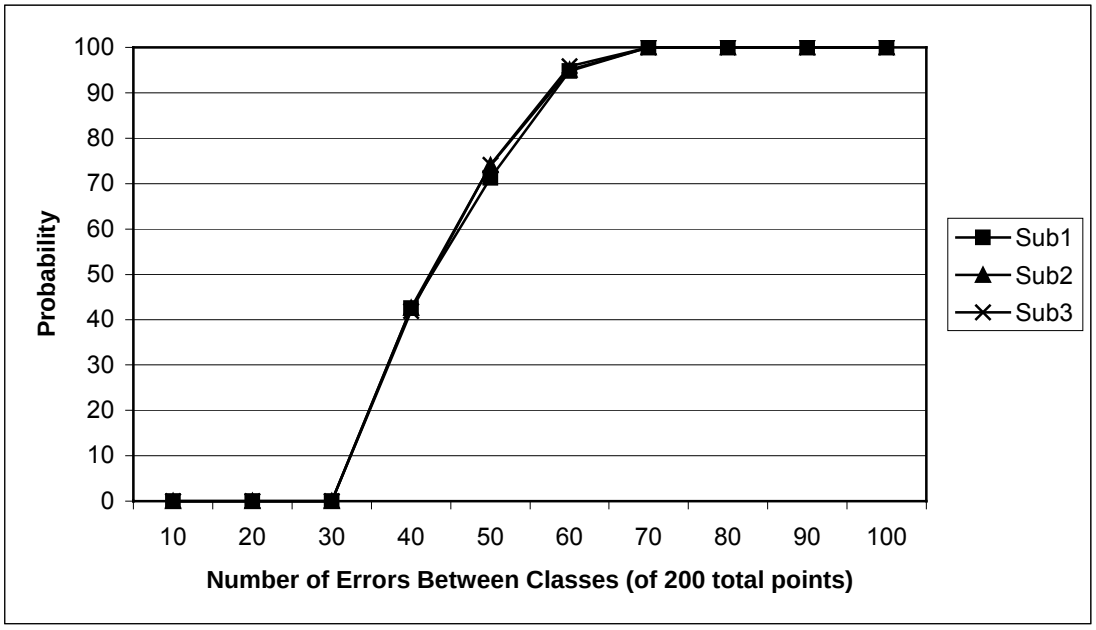

Figure 5. 12 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: All three NRB Sub-Areas

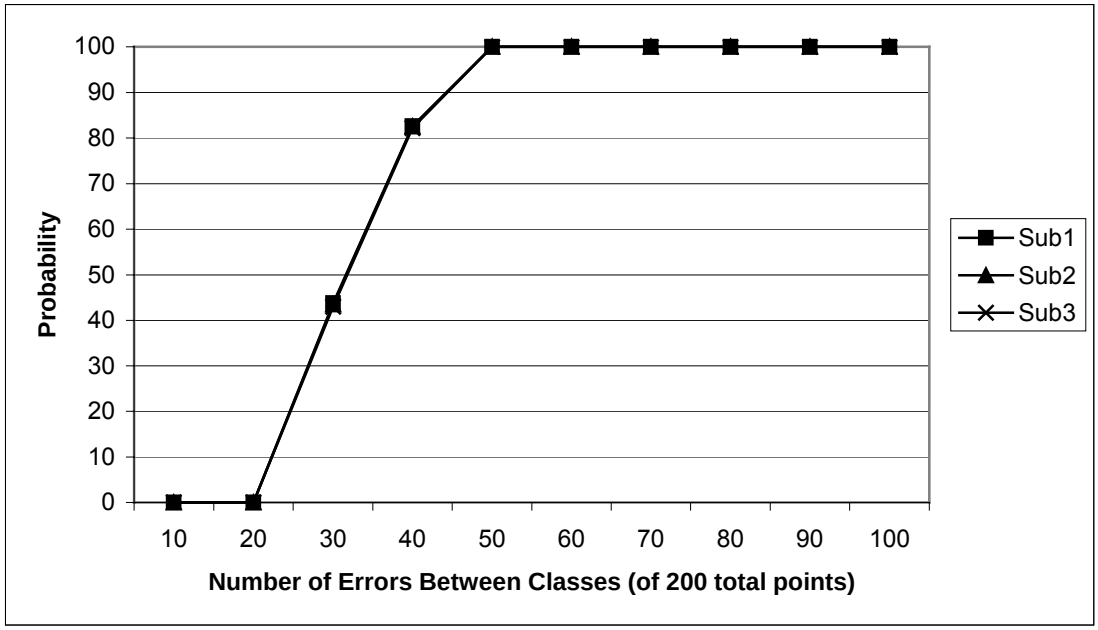

Figure 5. 13 Probability of Detecting a Difference between Kappa and Weighted Kappa - Rooftops vs. Bare Soil: All three NRB Sub-Areas

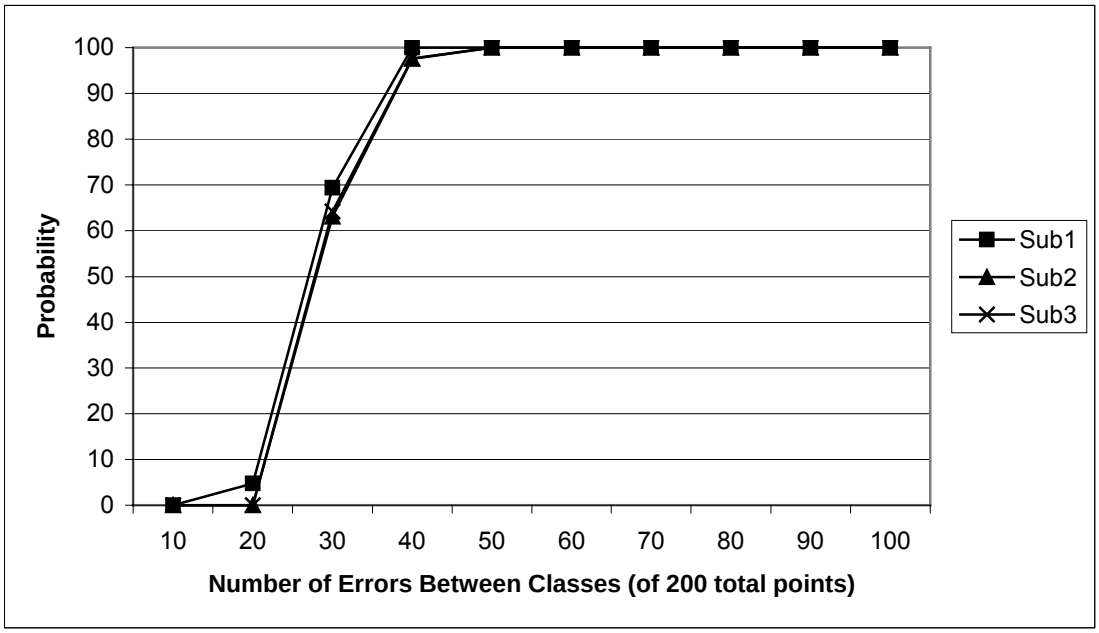

Figure 5. 14 Probability of Detecting a Difference Between Kappa and Weighted Kappa - Natural Grass vs. Urban Grass: All three NRB Sub-Areas

5.1.3 Experiment Two Results: Multiple Image Data Types

Experiment Two examined the behavior of SWK versus standard Kappa over three different image data types for the forest versus grass error type (from Experiment One, Part A) The results support the hypothesis that SWK can detect the effects of errors that are masked by current techniques such as Kappa over different data types, and also indicate that spectral clusters for SWK calculations must be collected from the same image data type as was used in the classification.

Figures 5.15, 5.16, and 5.17 show SWK versus Kappa for the three different image data types: SPOT, Landsat TM, and CIR DOQQ. Note that Figure 5.15, the forest versus grass error type on SPOT data, is the same as Figure 5.3. In each case the SWK technique is able to achieve a 95% error detection probability with a reasonable number of errors. This indicates that SWK is applicable over a range of data types, as well as different geographic areas as shown in Experiment One.

Figure 5.18 shows all three image data type curves. A surprising result is that the SPOT and TM curves reach 100% error detection probability together (at approximately 70 errors), but the CIR DOQQ curve does so with a much lower number of errors (60). This strange result indicates that the forest and grass classes were *more* separable on the three band CIR DOQQ than they were on the five and seven band SPOT and TM images, respectively. It is likely that this difference is due to the CIR DOQQ's higher spatial resolution of one meter versus the 20m and 30m resolutions of the SPOT and TM data.

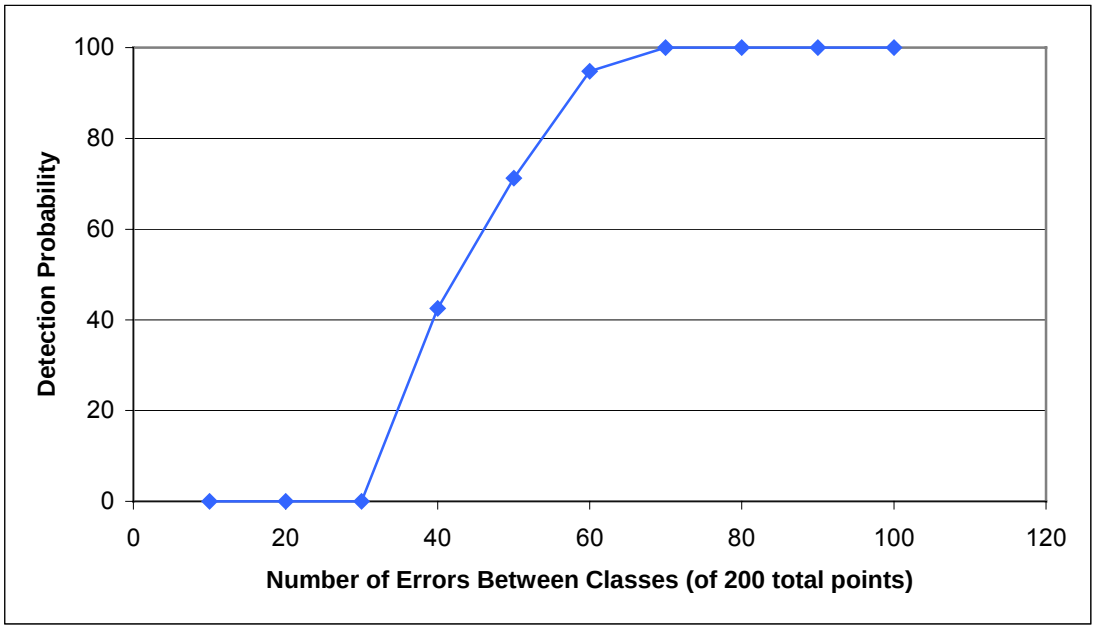

Figure 5. 15 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: SPOT Data (Same as Figure 5.3)

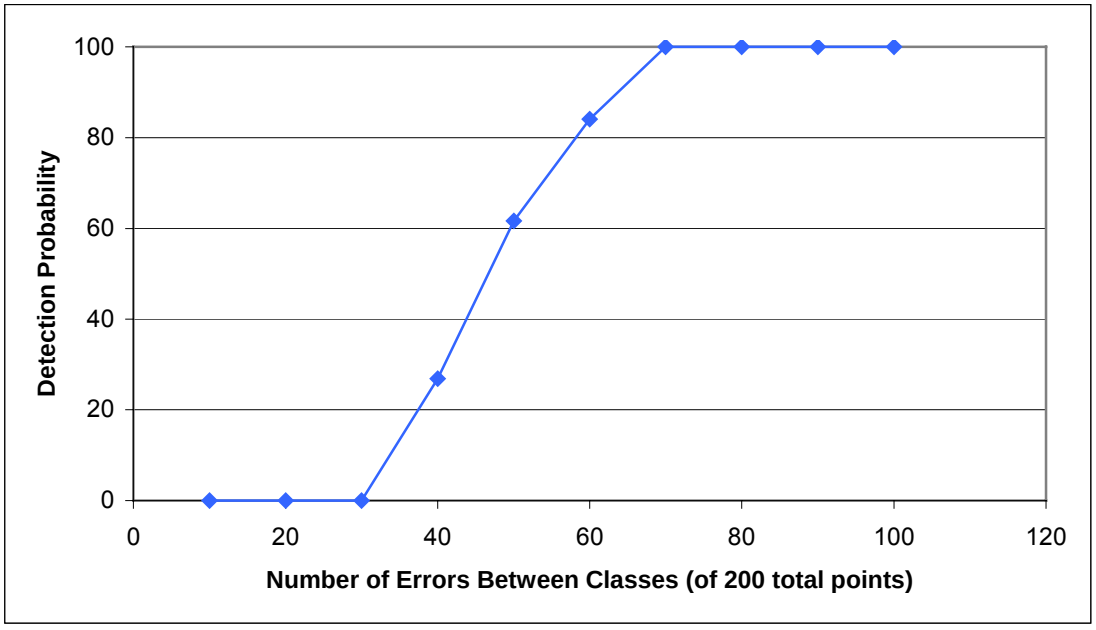

Figure 5. 16 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: Landsat Thematic Mapper

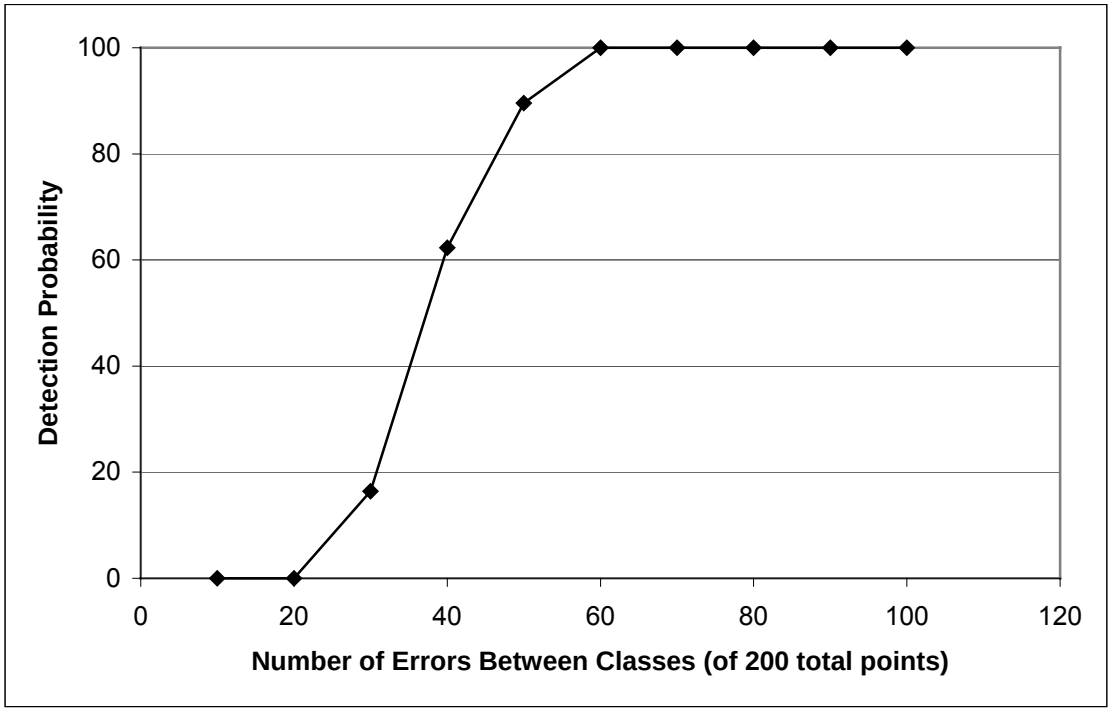

Figure 5. 17 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: CIR DOQQ

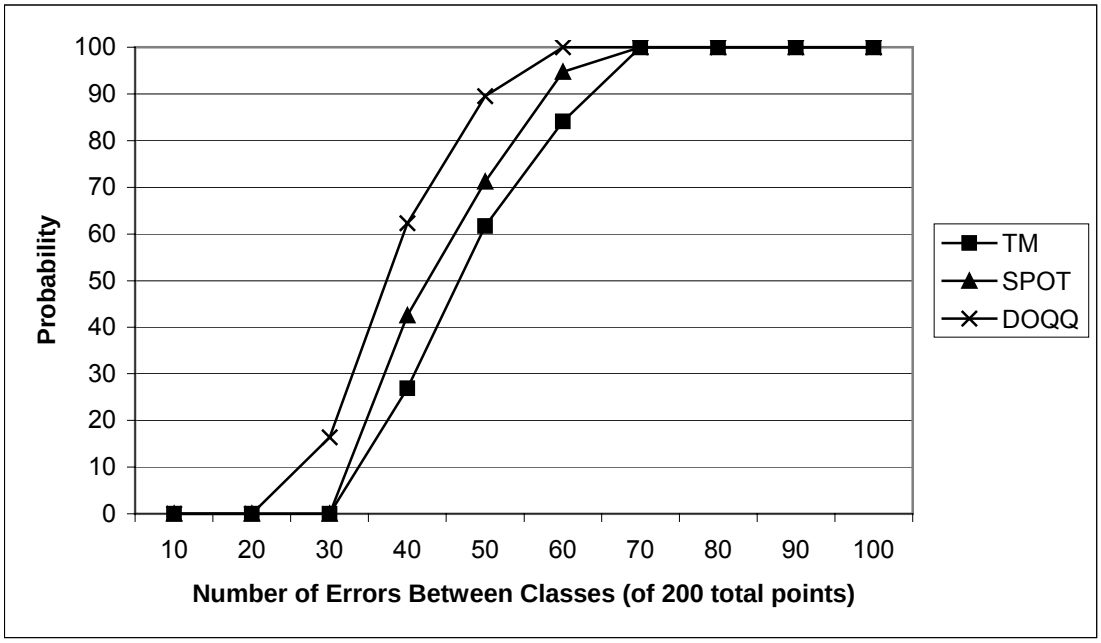

Figure 5. 18 Probability of Detecting a Difference between Kappa and Weighted Kappa - Coniferous Forest vs. Grass: All Image Data Types

Another interpretation of Figure 5.18 shows that it is vitally important that the same data type be used for selecting class spectral samples for use in SWK as that from which the classification was derived. There are large enough differences between the three image curves that one should not use class samples from one image type to assess the accuracy of a classification derived from a different image type.

5.1.4 False-Positive Test Results

Figure 5.19 shows the results of the false-positive test. Figure 5.19 shows that, when the distribution of the total number of errors between classes is inversely proportional to the spectral distances between the classes, the SWK and standard Kappa are very similar. The Z-test described above never indicated a significant difference between the SWK and standard Kappa. In addition, an ANOVA test between the two data sets presented in Figure 5.19 with a null hypothesis of no mean difference resulted in *p*-value (α = 0.05) of 0.94. This result shows that the SWK method does not falsely indicate a significant difference between the SWK and standard Kappa estimates. As a result, it may be possible to use a Z-test that is calibrated for a more stringent alpha level, such as 0.001. However, attempting to recommend different alpha levels based on project goals would add unnecessary complexity to the use of the SWK method, and so the use of the standard 0.05 alpha level Z-test is recommended.

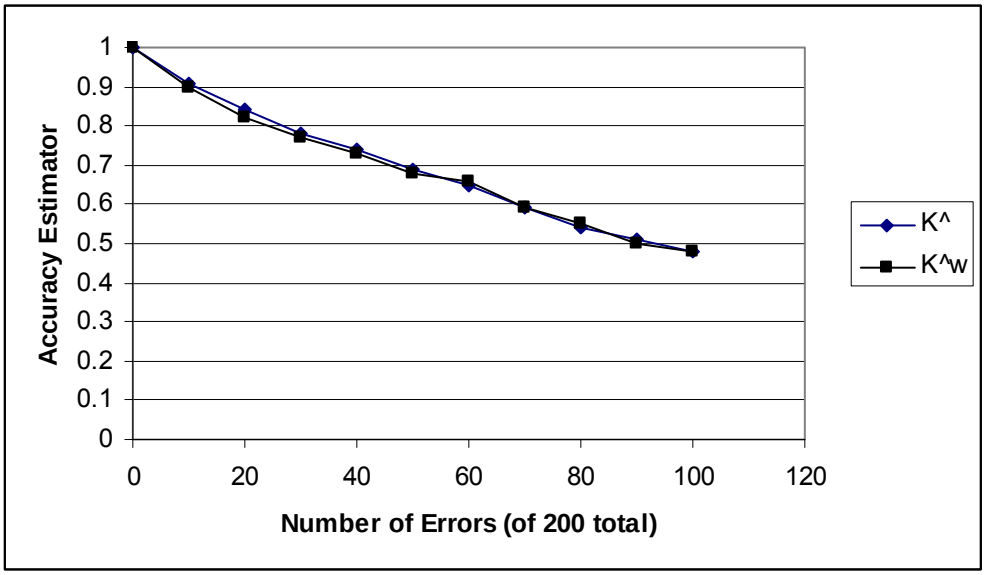

Figure 5. 19 SWK versus standard Kappa in false-positive test.

5.2 SWF Results

The following sections present the results of the analyses described above. First, in Section 5.2.1, are the results of the multiple interpreters test. Following that, in Section 5.2.2 are the results from the SWF versus human-derived fuzzy membership values tests.

5.2.1 Experiment Three: Multiple Interpreters

The interpreters each selected class membership values from the linguistic scale in Section 4.2 for each of the 18 classes in the study for all of the 270 points. The results of the correlation analysis are presented in Tables 5.1 and 5.2.

Table 5.1 gives the average correlation (Pearson correlation coefficient) between the three interpreters' membership values for each class. These values

clearly indicate that, while a few of the classes are highly correlated between the three interpreters, most fall well short of what could be called a strong correlation. Of particular note are the grassland, pasture, and shrubland classes, which have correlation coefficients below 0.5. This is due to the interpreters' difficulty in separating those classes on the reference aerial photographs. The overall average R value is also quite low: 0.61.

Table 5. 1 Average correlation (R value) between the three interpreters for each class. The overall average correlation is 0.61.

Class	Avg. Correl.
Urban: Low Density Residential	0.89
Agriculture: Cropland	0.86
Water	0.84
Barren: Bare Soil/Sand	0.82
Barren: Quarries	0.73
Urban: Commercial	0.69
Transitional	0.69
Forest: Deciduous	0.66
Agriculture: Small Grains	0.60
Forest: Coniferous	0.60
Wetland: Woody	0.58
Urban: High Density Residential	0.54
Wetland: Herbaceous	0.54
Forest: Mixed	0.52
Grassland: Other	0.49
Grassland: Natural	0.42
Agriculture: Pasture/Hay	0.27
Shrubland	0.27

Table 5.2 shows the ten highest correlation values between all 18 LU/LC classes and the three interpreters. The values in parentheses indicate which interpreter was involved in the comparison. Interestingly, of the ten highest correlations, three – including the highest correlation overall – are *between classes* rather than *within classes*. In a study such as this, one would hope to see the highest correlations be between interpreters within the same class. This

would indicate that the interpreters were choosing similar membership values for the classes involved. The values in Table 5.2 show that the interpreters had significant problems agreeing on membership.

Table 5. 2 The ten highest correlations between interpreter and class. The numbers in parentheses are the interpreter designations (1, 2, or 3)

Correl.	Interpreter/Class
0.96	Cropland (2) – Small Grains (2)
0.91	Bare: Soil (1) – Bare: Soil (2)
0.91	Small Grains(1) – Cropland (1)
0.90	L.D. Resid. (2) – L.D. Resid. (1)
0.90	L.D. Resid. (3) – L.D. Resid (2)
0.89	Cropland (2) – Cropland (1)
0.89	Small Grains (1) – Cropland (2)
0.88	L.D. Resid. (3) – L.D. Resid. (1)
0.86	Cropland (3) – Cropland (1)
0.86	Small Grains (2) – Small Grains(1)

The results from Tables 5.2 and 5.3, combined with the very low overall correlation between interpreters, R = 0.61, indicate that there is significant inter-interpreter variation in selecting the class fuzzy membership values. Clearly, in spite of the comprehensive training procedures, the interpreters had trouble choosing similar class fuzzy membership values for these 270 points.

5.2.2 Experiment Four: Comparison of SWF and Manual Memberships

The hypothesis for this experiment is that fuzzy membership functions determined by the SWF method are similar to those determined by a well-trained human interpreter for a set of commonly seen errors in remote sensing thematic classifications. The following graphs show Max and Right values calculated using both methods for each of the three error types.

Figures 5.20, 5.21, and 5.22 show Max and Right for human-derived class fuzzy membership values for the three error types. The graphs show wide variations in the Max and Right values for the three error types. These are due to the effects of the different membership functions chosen. For example, when collecting reference data for a thematic classification derived from satellite data, a reasonable human interpreter would likely give urban grassland a fairly high membership value when examining a natural grassland site (and vice versa) due to the likely spectral similarity of those classes. Figure 5.20 shows the result of giving both grassland classes relatively high membership values. As the number of errors between the two classes increases, the Right operator value decreases more slowly than the Max operator value. This is because, in many of the sample sites, the Right operator counts the site as correct when the Max operator does not. The similarity of the classes results in membership values that fulfill the Right operator's threshold criterion, even though the site doesn't fulfill the Max operator's criterion.

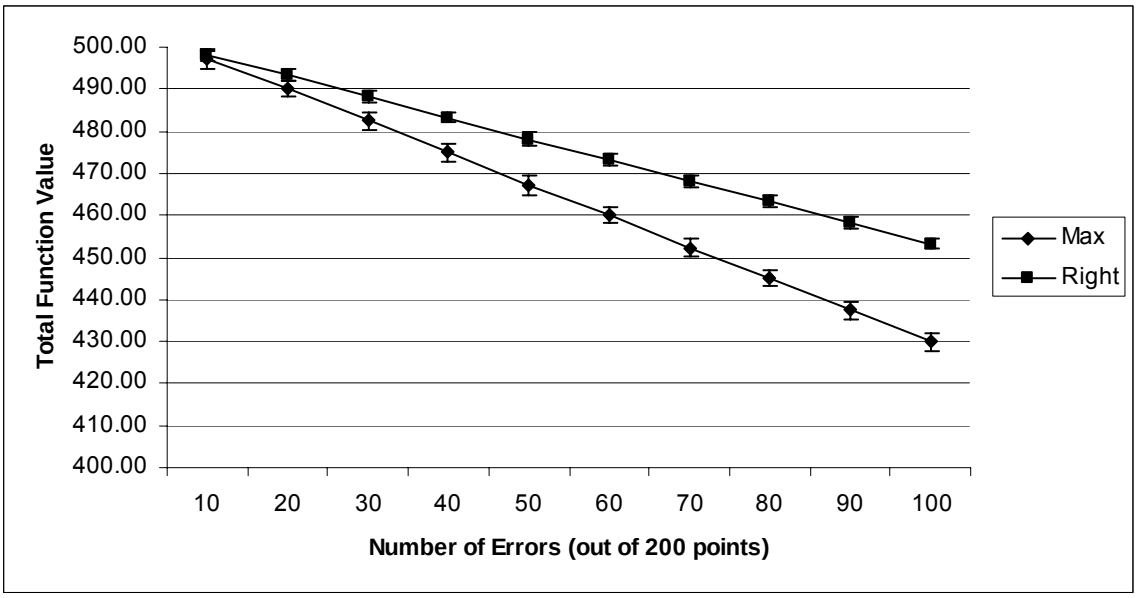

Figure 5. 20 Max and Right for human-derived membership values for the forest versus natural grassland error type (with error bars)

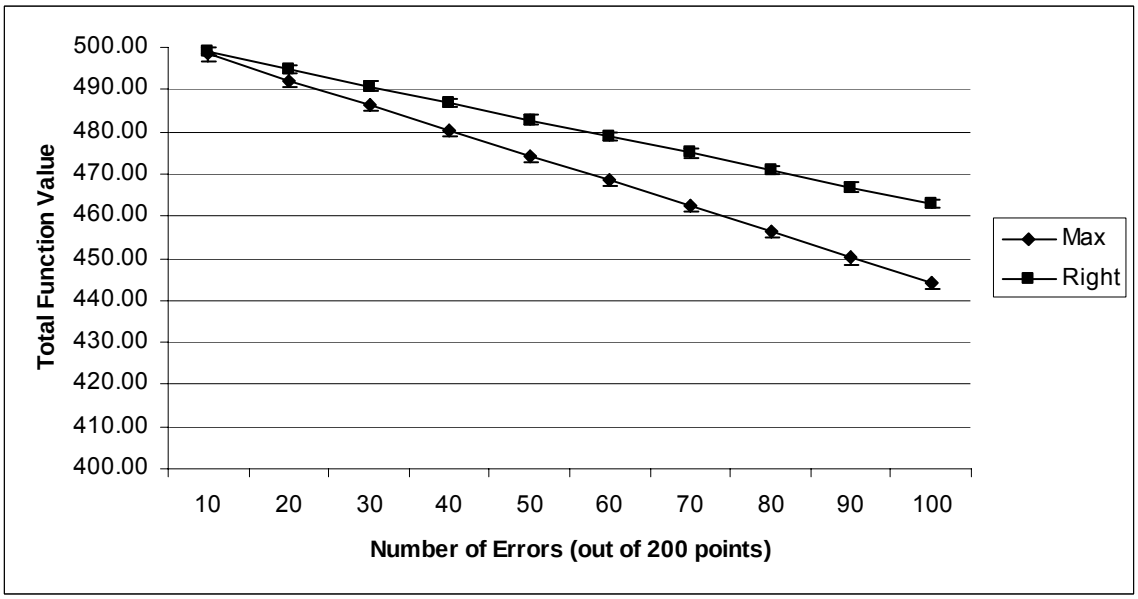

Figure 5. 21 Max and Right for human-derived membership values for the bare soil versus rooftops (urban) error type (with error bars)

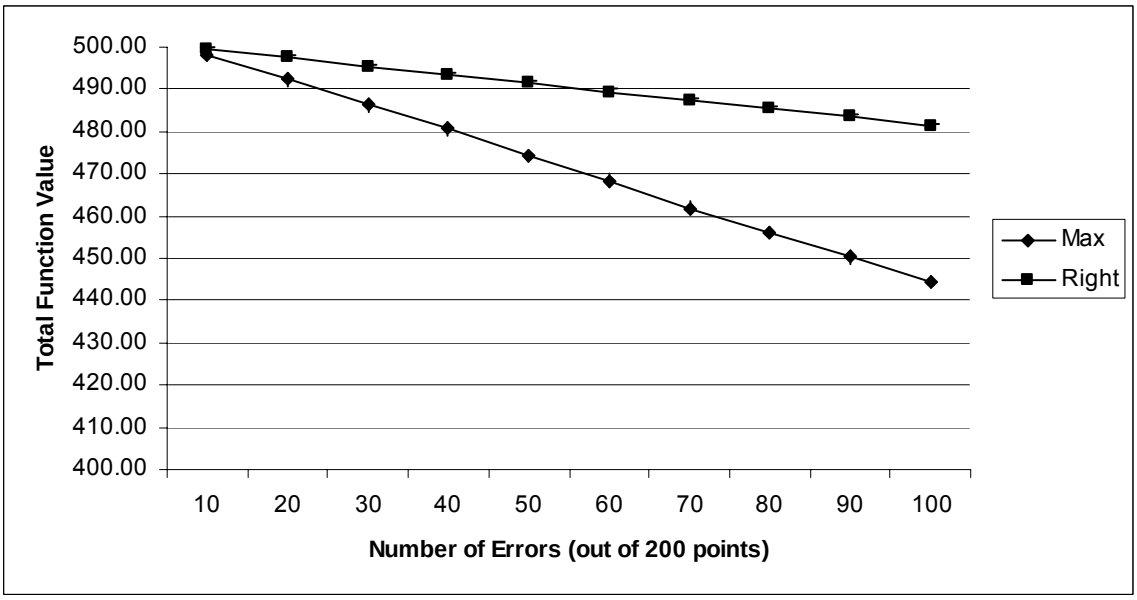

Figure 5. 22 Max and Right for human-derived membership values for the natural grassland versus urban grassland error type (with error bars)

Figures 5.23, 5.24, and 5.25 present the Max and Right values for the same error types as in Figures 5.20, 5.21, and 5.22, but with those values computed by the SWF method instead of from human-derived membership functions. The curves are very similar to those developed by the manual interpretation.

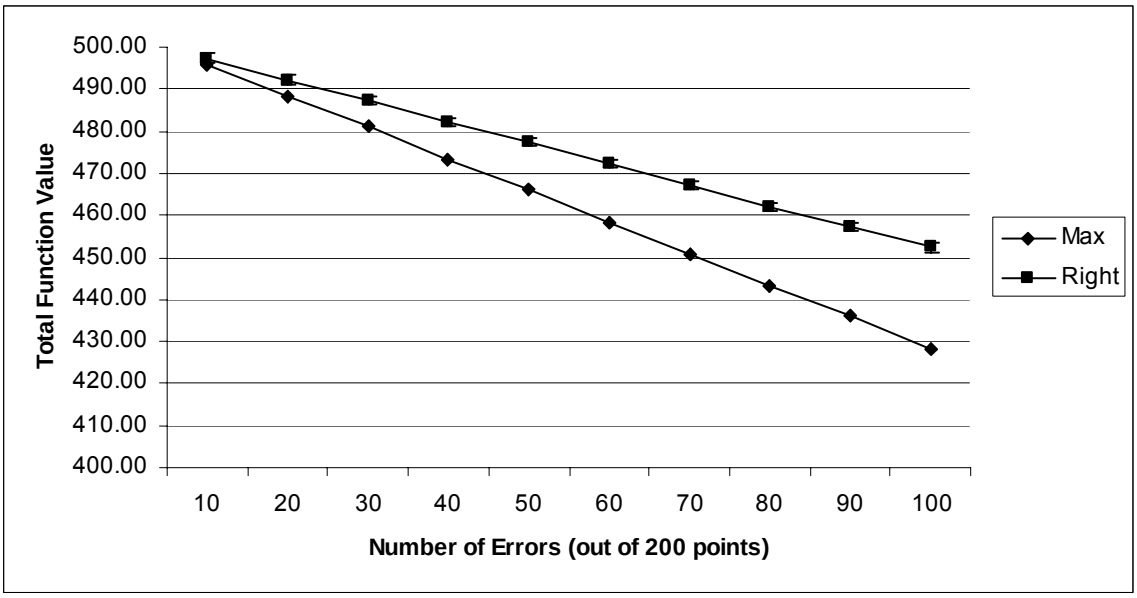

Figure 5. 23 Max and Right for SWF membership values for the forest versus natural grassland error type (with error bars)

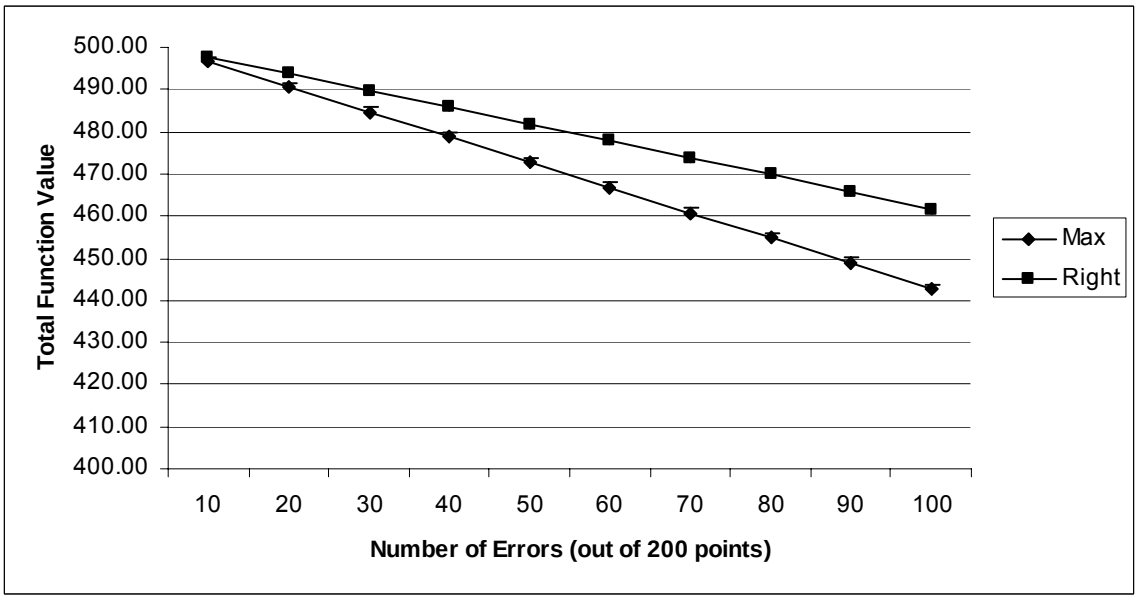

Figure 5. 24 Max and Right for SWF membership values for the bare soil versus rooftops (urban) error type (with error bars)

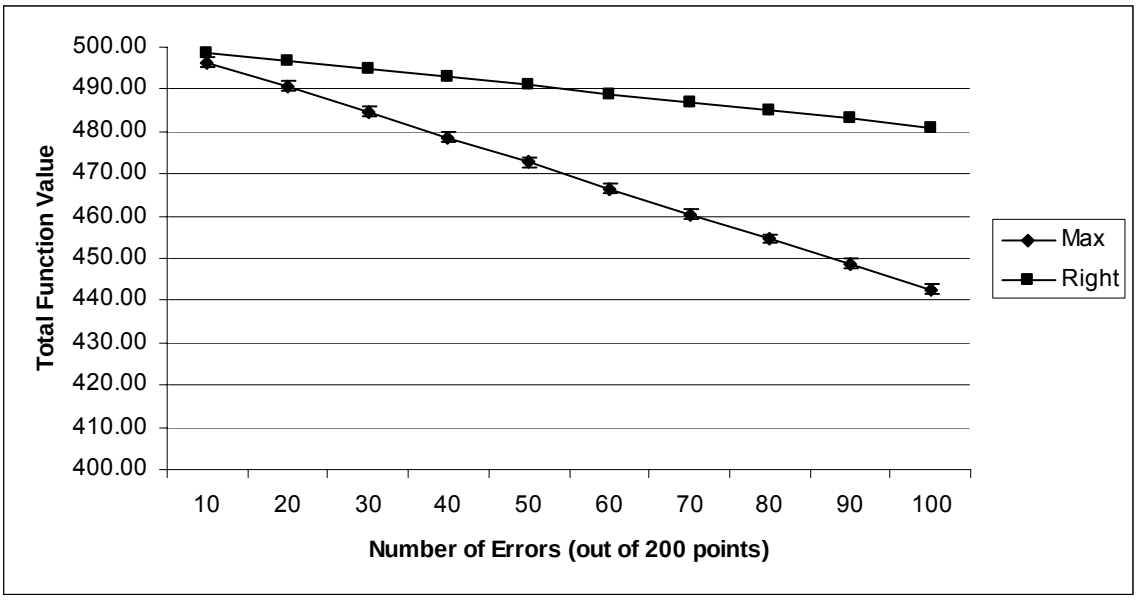

Figure 5. 25 Max and Right for SWF membership values for the natural grassland versus urban grassland error type (with error bars)

Figures 5.26, 5.27, and 5.28 show the human-derived and SWF Max and Right values plotted on the same graph for each error type. In each graph the Max and Right values determined by the SWF method correspond strongly to those from the human-derived membership values.

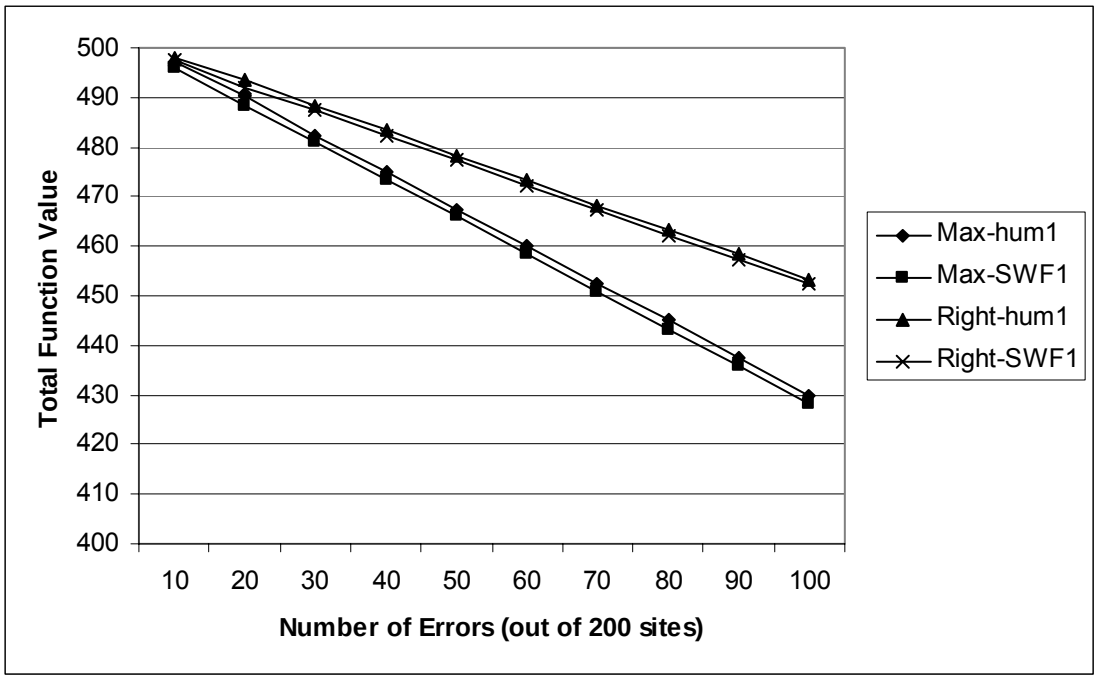

Figure 5. 26 Max and Right for both human derived and SWF membership functions for the forest versus natural grassland error type (error bars removed)

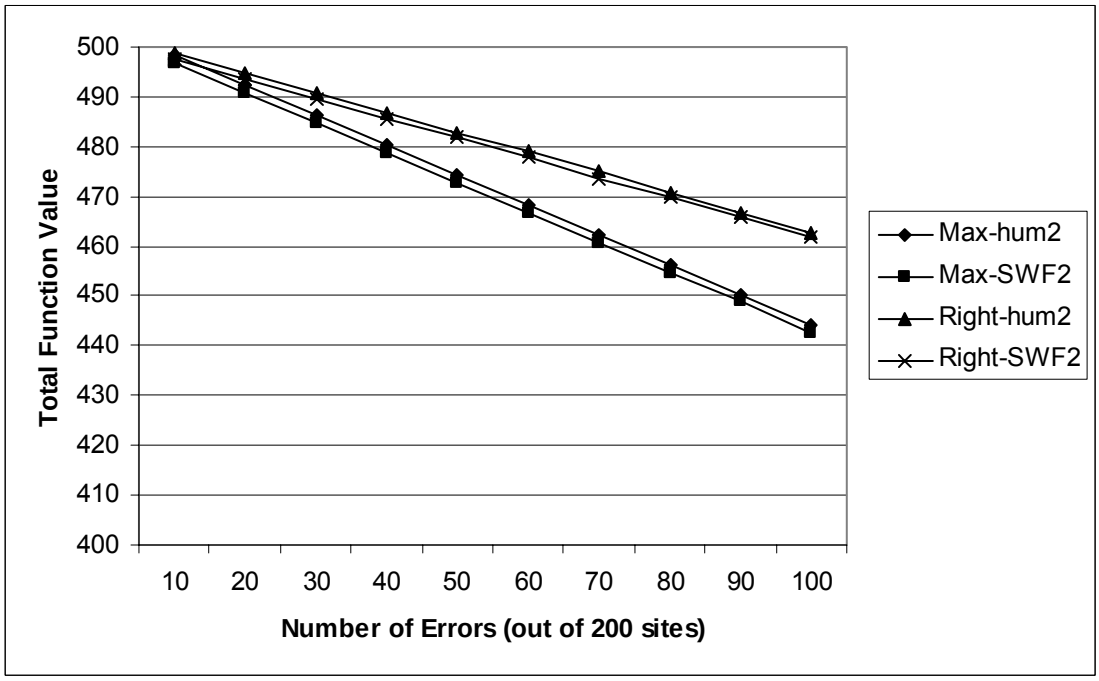

Figure 5. 27 Max and Right for both human derived and SWF membership functions for the bare soil versus rooftops (urban) error type (error bars removed)

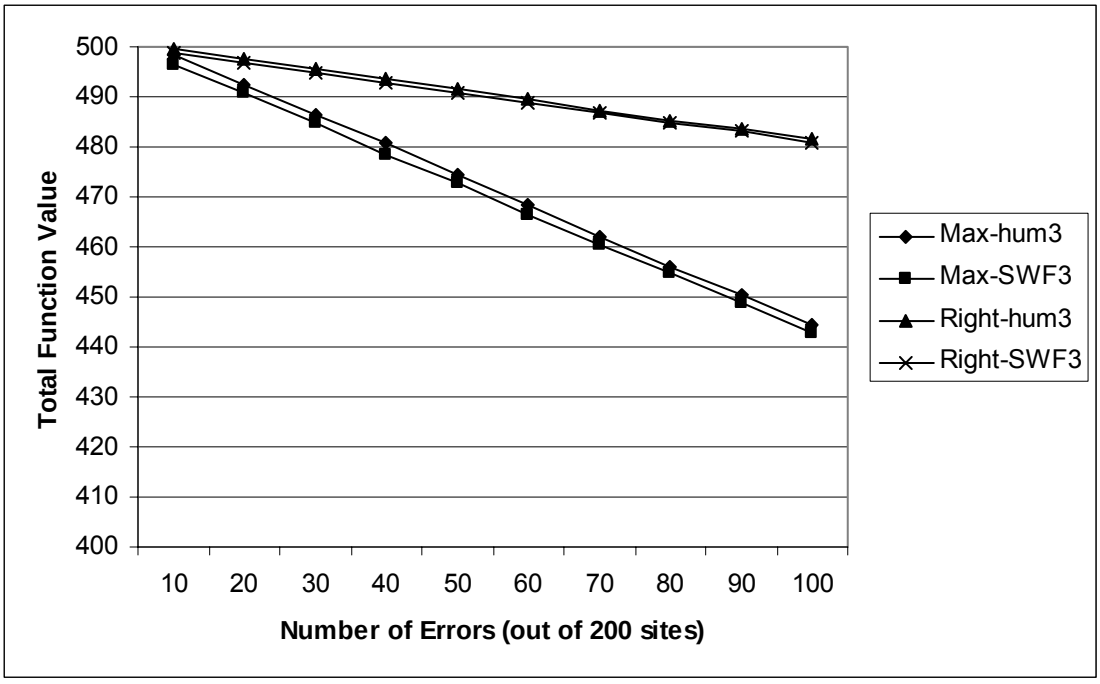

Figure 5. 28 Max and Right for both human derived and SWF membership functions for the natural grassland versus urban grassland error type (error bars removed)

ANOVA comparisons between the SWF and human-derived Max and Right values are presented in Table 5.3. In each case, the F-test failed to reject the null hypothesis of no difference between the SWF and human-derived Max and Right values. In other words, the F-test did not indicate that the human-derived and SWF-derived Max and Right values are statistically different.

Table 5. 3 ANOVA comparisons between Max and Right values

Comparison	ANOVA *p* value
Forest vs. Natural Grassland	
Max vs. Max	0.88
Right vs. Right	0.89
Bare Soil vs. Rooftops (Urban)	
Max vs. Max	0.85
Right vs. Right	0.84
Natural Grassland vs. Urban Grassland	
Max vs. Max	0.84
Right vs. Right	0.84

The results of this experiment support the hypothesis that the SWF technique performs similarly to human-derived membership functions. For each of the three error types tested, use of the SWF technique resulted in Max and Right values that are statistically inseparable from the human-derived values.

6.0 DISCUSSION

Many researchers have contributed to the body of thematic accuracy assessment techniques. Until now, none of these techniques has used the spectral differences between the classes as a modifier of the accuracy estimates. It is expected that techniques such as SWK and SWF will be useful to remote sensing community.

The SWK technique, when compared with a standard Kappa, has been shown to have a greater ability to quantify the effects of the error types examined than other techniques, such as Kappa alone, that do not use the spectral properties of the classes in accuracy assessment. In addition, naturally occurring errors that are not due to misuse or improper training of a classifier, such as those examined in Part A of Experiment One, can be highlighted with the SWK technique. These benefits of SWK are evident when error frequencies are within the range of those frequently seen in remote sensing studies.

The breadth and usefulness of tools for assessing the accuracy of thematic classifications derived from remotely sensed data were increased with the introduction of fuzzy set theory. The new insight provided by the use of fuzzy set techniques allows much more extensive analysis of thematic accuracy than is possible with standard techniques. There are, however, drawbacks to the fuzzy approach as it was initially proposed. Gopal and Woodcock (1994) recommend for further research the creation of, "…a way to standardize results from different interpreters. Some interpreters will be more lenient in the number of multiple set memberships given, and this factor will need to be taken into account." The SWF

method attempts to solve this problem by using an objective method of creating the membership values that is based on inter-class spectral distances. This method reduces the chance that inter-interpreter variation will have a significant affect on the accuracy assessment because the interpreters do not determine the memberships – only the base reference class for a site. In the author's experience, it is much simpler and less time consuming for the reference data interpreters to choose one class for each site than to establish fuzzy membership values for each class. This is especially true in classification systems with as many as fifteen to twenty classes. If the interpreters determine the fuzzy memberships themselves, they must be trained to do so, and must record the values in someway. This reduction in interpretation complexity afforded by the SWF method has the added benefits of reducing both training time and the likelihood that simple data recording errors will affect the accuracy estimate.

Another potential drawback of the fuzzy approach as initially proposed is that the use of reference data interpreters to establish the fuzzy membership values is subjective and is not quantitatively based. This results in membership values that are very heavily dependent on the interpreters, and so may not be reproducible from project to project. The SWF method bases these fuzzy membership values on the spectral distance between the thematic classes of interest.

6.1 Appropriate Treatment of the Classifier in Accuracy Assessment

SWK and SWF focus on the author's concept of "fair" accuracy assessment. In our opinion, it is important that an accuracy assessment scheme

presents the accuracy estimate in such a way as to not unfairly penalize the classifier for confusing spectrally similar classes. It is unreasonable to expect a thematic classifier to be able to discriminate certain class pairs. Some examples of these difficult classes are urban grass versus natural grass (as shown in Experiment One, Part C), row crops versus pasture during the growing season, and residential areas versus forest and grassland (due to trees in yards). As such, it is unfair to weight those errors as highly as errors between classes that are more separable.

Conversely, if the classifier is confusing classes that are very spectrally different, such as water and grass, for example, then these errors should be magnified. Current techniques such as overall percent accuracy, Kappa, and errors of omission and commission do not take into account these spectral factors.

6.2 The Relationship between Kappa and SWK

The SWK will trend higher than the standard Kappa when the majority of the errors in a classification are between classes that are relatively spectrally similar. In contrast, the SWK will trend lower than the standard Kappa when the majority of the errors in the classification are between classes that are relatively spectrally different. As shown in the above experiments, this behavior gives the analyst the potential to find discrepancies in a classification that are not as readily identifiable with other accuracy assessment techniques.

6.3 Importance of Choosing Representative Class Samples

The importance of the class spectral samples to the validity of SWK analysis cannot be overstated. Just as in supervised classification, the quality of the input data determines the quality of the output to a large extent. For example, if the inter-class spectral distances are misrepresented because of poor sample selection, the results of any accuracy analysis derived thereof are biased.

Several factors need to be kept in mind when using SWK and SWF. First, the minimum mapping unit and spectral resolution of the image data used to collect the spectral samples must be appropriate for the goals of the study. For example, too large a minimum mapping unit may result in excessively mixed spectral samples. Second, the potential effects of landscape characteristics must be accounted for. Land cover types must be examined to ensure that adequate spectral samples can be derived. Third, image data characteristics must be evaluated for use with these methods. Care must be taken to ensure that geometric or other error sources do not corrupt the image to such a degree that class samples do not represent the spectral properties of the classes of interest. Fourth, positional errors must be minimized. SWK and SWF require that collected spectral samples match the corresponding locations on the reference data. Fifth, the acquisition dates of the image(s) used to collect the spectral samples must match as closely as possible, with respect to season and year, to the classification image. Where practical, use of the same image for both classification and collection of the spectral samples is preferable.

6.4 Testing Methods

We recognize that the experimental setup used to test the performance of SWK and SWF is not ideal. As the overall objective of this research is to measure the relative abilities of two techniques to assess the agreement of a dataset with the "truth," it would be helpful to have access to such truth. Unfortunately, in remote sensing studies, it is unlikely that one could determine an objective, universally agreed-upon truth. Different reference data sources will show the landscape differently. Also, different reference data interpreters will have different ideas about what they see on the reference data. As such, the experiments conducted herein test the agreement of a well-accepted thematic classification with different simulated reference data sets. These data were designed to contain errors commonly seen in remote sensing research. Then, the abilities of the techniques, SWK and SWF, to identify these errors were tested. It is important to note that every effort was made to ensure that these "created" errors were consistent with confusions found during actual accuracy assessment projects.

The SWF experiments would have been stronger had it been possible to test the technique in different geographic areas and with different image data types. However, the large time investment required to collect two sets of membership values from the 500 reference sites made it impossible to replicate the experiment. Although different geographic areas and image data types were not specifically tested with the SWF technique, the results from the testing of the SWK method indicate that, since both techniques use the Transformed

Divergence algorithm as the basis for their analyses, SWF should also be portable from region to region, but not across data types.

6.5 Similar Covariance Cancellation in Transformed Divergence

A theoretical problem with the SWK technique as presented here involves the use of the Divergence algorithm to calculate the inter-class spectral distances. Referring to Equation 3.1, we see that if the covariance matrices of the classes in question, C_i and $C_{j,}$ are identical then the subtraction of the two matrices will result in a Divergence (D_{ij}) value of zero, irrespective of the means of the classes. However, in informal testing during the completion of this project hundreds of class samples from various areas and data types were taken and compared. Not a single case resulted covariance matrices that were identical, and so Divergence values of zero were not observed. Even homogenous, adjacent samples of the same class had non-zero Divergence values. Thus, though it is possible that two covariance matrices could cancel, it seems that, in practice, this is unlikely to occur.

7.0 CONCLUSIONS

The results of this study demonstrate that inter-class spectral distances can be used effectively in accuracy assessment of thematic classifications derived from remotely sensed data. The SWK approach, when compared with a standard Kappa, can provide information about the spectral costs of errors in a thematic classification that is not as apparent with traditional methods such as Kappa alone. The inter-class spectral weights used in SWK provide a new perspective on the accuracy of a thematic classification that is based on the performance of the classifier relative to the spectral information it had to work with.

The SWK method is expected to benefit remote sensing researchers by providing an accuracy assessment technique that weights the accuracy estimate by the spectral costs of the various errors in the classification. This new approach provides several important additions to current techniques:

- Weighted Kappa has not been well accepted in the literature in large part because of the inherent subjectivity involved in choosing the weights. The SWK technique eliminates this subjectivity by creating the weights in a consistent and repeatable way using the inter-class spectral distances. This new development may lead to more widespread acceptance and use of Weighted Kappa through the SWK method.
- In many cases, journal articles provide only a single statistic, such as overall percent accuracy or Kappa, when discussing the overall accuracy

of a classification. That information, while valuable, is incomplete. As has been shown in this study, the accuracy of a thematic classification may be misestimated due to the confusion of spectrally similar classes. These confusions are given the same weight as confusion of spectrally different classes. The provision of an SWK value in addition to the Kappa value in published studies would give the reader added insight into how well the classifier did with respect to the spectral separability of the classes involved, i.e. how the classifier performed with the information it had.

- A standard error matrix does not provide any information as to the spectral characteristics of the classes in question. One might see grassland and forest classes in an error matrix, but unless one is familiar with the area, one has little idea of the spectral compositions of those classes. If Kappa and SWK statistics are presented with the error matrix, the reader can glean some insight into the spectral costs of the errors in the classification.
- SWK analysis may help to identify problems that the analyst is not aware of. For example, the calculation of a standard Kappa may show that a classification is accurate enough to meet the goals of study for which it was created. However, if the SWK value is significantly different than the Kappa, the creator of the classification may choose to revisit his work and determine why the discrepancy exists. For example, if the SWK value is significantly higher than the standard Kappa, there are many errors in the classification that are due to confusion between spectrally similar classes. The analyst could then choose to revisit the classification scheme and, if

appropriate, combine some of the spectrally similar classes (Chrisman 1982). In this way, SWK analysis can be an aid to improving classifications before they are finalized.

Fuzzy accuracy assessment is not widely used in the literature. Possible reasons for this are the inherent subjectivity, and the risk of significant inter-interpreter variation, in selecting the membership values. The SWF method improves upon the current fuzzy accuracy assessment techniques by providing a way to establish membership functions that is based on inter-class spectral distances. Using these spectral distances instead of asking human interpreters to determine the classes would be particularly advantageous in large accuracy assessment studies where there are hundreds or thousands of reference sites. The best way to control for inter-interpreter variation in such a study is to use only one highly trusted interpreter. In most cases though, it would be unreasonably time-consuming to expect one person to interpret all of the reference sites. The alternative would then be to use multiple interpreters. However, the results of this research indicate that there is significant potential for inter-interpreter bias inherent in fuzzy membership value determination. We have shown that the SWF method can provide fuzzy membership values that are similar to those that a well-trained human might choose. Therefore, in cases where multiple interpreters would normally have been used, the SWF method can reduce this inter-interpreter bias.

In addition, the SWF method provides a quantitative basis for establishment of fuzzy membership values. The T.D. values used in SWF can be compared from study to study. This addresses a concern raised by the researchers who introduced fuzzy accuracy assessment to the remote sensing community, Gopal and Woodcock (1994) when they recommended the creation of, "…a way to standardize results from different [interpreters]. Some [interpreters] will be more lenient in the number of multiple set memberships given, and this factor will need to be taken into account." The SWF method provides a solution to this problem by basing the memberships on inter-class spectral distances. We hope that the introduction of the SWF method will lead to wider use of fuzzy accuracy assessment.

7.1 SWK and SWF Relationships to Current Methods

Current methods such as Kappa, overall percent accuracy are essential to the accuracy assessment process. It is important to note that SWK and SWF are suggested as additions to, not replacements for, currently used methods.

7.2 Future Directions

A potentially very useful extension of the SWK work would be the development of a Spectrally Weighted Conditional Kappa (SWCK). The standard conditional Kappa provides an accuracy estimate for each individual class, rather than just the overall estimate of the standard Kappa. A similar test statistic for SWK would be useful to examine in more depth the errors in a thematic classification.

The SWF work could be improved through the development of a method to allow the reference data interpreters to select more than one class from the reference data. It is often impossible, due to positional errors or heterogeneity, to select only one “true” class. Therefore expanding the SWF method to allow more than one base reference class would be helpful. Additionally, a larger experiment might be undertaken to verify that the SWF method is transportable through different physiographic conditions and classification algorithms.

8.0 REFERENCES

Aronoff, S., 1982a. Classification Accuracy: A User Approach. *Photogrammetric Engineering and Remote Sensing* 48(8):1299-1307.

Aronoff, S., 1982b. The Map Accuracy Report: A User's View. *Photogrammetric Engineering and Remote Sensing* 48(8):1309-1312.

Aronoff, S., 1985. The Minimum Accuracy Value as an Index of Classification Accuracy. *Photogrammetric Engineering and Remote Sensing* 51(1):99-111.

Bishop, Y., S. Fienberg, and P. Holland, 1975. *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA, 575 pp.

Card, D.H., 1982. Using Known Map Category Marginal Frequencies to Improve Estimates of Thematic Map Accuracy. *Photogrammetric Engineering and Remote Sensing* 48(3):431-439.

Chrisman, N. 1982. Beyond Accuracy Assessment: Correction of Misclassification. *Proceedings of the 5th International Conference on Computer-Assisted Cartography*. Crystal City, VA.

Cliff, A.D. and J.K. Ord 1973. *Spatial autocorrelation*. Pion, London, 1973.

Congalton, R.G. and K. Green, 1993. A Practical Looks at the Sources of Confusion in Error Matrix Generation. *Photogrammetric Engineering and Remote Sensing* 59(5):641-644.

Congalton, R.G. and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC Press, Inc. New York, NY.

Congalton, R.G., 1988. A Comparison of Sampling Schemes Used in Generating Error Matrices for Assessing the Accuracy of Map Generated from Remotely Sensed Data. *Photogrammetric Engineering and Remote Sensing* 54(5):593-600.

Congalton, R.G., 1991. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing of Environment* 37:35-46.

Congalton, R.G., R.G. Oderwald, and R.A. Mead, 1983. Assessing Landsat Classification Accuracy Using Discrete Multivariate Analysis Statistical Techniques. *Photogrammetric Engineering and Remote Sensing* 49(12):1671-1678.

Curran, P.J. and H.D. Williamson, 1986. Sample Size for Ground and Remotely Sensed Data. *Remote Sensing of Environment* 20:31-41

Degloria, S.D., M. Laba, S.K. Gregory, J. Braden, D. Ogurcak, E. Hill, E Fegraus, J. Fiore, A Stalter, J. Beecher, R. Elliot, J. Weber, 2000. Conventional and Fuzzy Accuracy Assessment of Land Cover Maps at Regional Scale. *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Amsterdam, The Netherlands.

Erdas, Inc. *The Erdas Imagine Field Guide*. Erdas Press, Atlanta, GA.

Fenstermaker, L., 1991. A Proposed Approach for National to Global Scale Error Assessment. *Proceedings of GIS/LIS 91*, Atlanta, GA, pp. 293-300.

Fisher, P., 1997. The Pixel: A Snare and a Delusion. *Internation Journal of Remote Sensing* 18:679-685.

Foody, G.M., 1992. On the Compensation for Chance Agreement in Image Classification Accuracy Assessment. *Photogrammetric Engineering and Remote Sensing* 58(10):1459-1460.

Foody, G.M., N.A. Campbell, N.M. Trodd, T.F. Wood, 1992. Derivation and Applications of Probabilistic Measures of Class Membership from the Maximum Likelihood Classification. *Photogrammetric Engineering and Remote Sensing* 58(9):1355-1341.

Franklin, S.E., D.R. Peddle, B.A. Wilson, C.F. Blodgett, 1991. Pixel Sampling of Remotely Sensed Digital Imagery. *Comput. Geosci.* 17:759-775.

Gao, J., 2001. Non-Differential GPS as an Alternative Source of Planimetric Control for Rectifying Satellite Imagery. *Photogrammetric Engineering and Remote Sensing* 67(1):49-55.

Gopal, S. and C. Woodcock, 1994. Theory and Methods for Accuracy Assessment of Thematic Maps Using Fuzzy Sets. *Photogrammetric Engineering and Remote Sensing* 60(2):181-188.

Greenfield, J., 2001. Evaluating the Accuracy of Digital Orthophoto Quadrangles (DOQ) in the Context of Parcel Based GIS. *Photogrammetric Engineering and Remote Sensing* 67(2):199-205.

Hay, A.H., 1979. Sampling Designs to Test Land-Use Map Accuracy. *Photogrammetric Engineering and Remote Sensing* 45(4):529-533.

Hord, R.M. and W. Brooner, 176. Land-Use Map Accuracy Criteria. *Photogrammetric Engineering and Remote Sensing* 42(5)671-677.

Hudson, W.D. and C.W. Ramm, 1987. Correct Formulation of the Kappa Coefficient of Agreement. *Photogrammetric Engineering and Remote Sensing* 53(4):421-422.

Jäger, G. and U. Benz, 2000. Measures of Classification Accuracy Based on Fuzzy Similarity. *IEEE Transactions on Geoscience and Remote Sensing* 38(3):1462-1467.

Janssen, L.L.F and F.J.M. van der Wel, 1994. Accuracy Assessment of Satellite Derived Land-Cover Data: A Review. *Photogrammetric Engineering and Remote Sensing* 60(4):419-426.

Khorram, S. and J.F. Knight. Land Cover Classification of the Hominy Creek Watershed. Center for Earth Observation Technical Report 217. June, 2000.

Khorram, S., G. S. Biging, N. R. Chrisman, D. R. Colby, R. G. Congalton, J. E. Dobson,R. L. Ferguson, M. F. Goodchild, J. R. Jensen, and T. H. Mace, 1999, *Monograph*, Accuracy Assessment of Remote Sensing-Derived Change Detection, American Society of Photogrammetry and Remote Sensing (ASPRS), 58p.

Klecka, W.R., 1980. *Discriminant Analysis*. Sage Publications, Beverly Hills.

Landis, J.R. and G.G. Koch, 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159-174.

Light, R.J., 1971. Measure of Response Agreement for Qualitative Data: Some Generalizations and Alternatives. *Psychological Bulletin* 76(5):365-377.

Lunetta, R.S. and C.D. Elvidge (ed.) 1998. *Remote Sensing Change Detection: Environmental Monitoring Methods and Applications*. Ann Arbor Press. Chelsea, MI.

Lunetta, R.S., J.S. Iiames, J.F. Knight, R.G. Congalton, T.H. Mace, 2001. An Assessment of Reference Data Variability Using a "Virtual Field Reference Database." *Photogrammetric Engineering and Remote Sensing* 67(6):707-715.

Lunetta, R.S., R.G. Congalton, L.K Fenstermaker, J.R. Jensen, K.C. McGwire, and L.R. Tinney, 1991. Remote Sensing and Geographic Information System Data Integration: Error Sources and Research Issues. *Photogrammetric Engineering and Remote Sensing* 57(6):677-687.

Ma, Z. and R.L. Redmond, 1995. Tau Coefficients for Accuracy Assessment of Classification of Remote Sensing Data. *Photogrammetric Engineering and Remote Sensing* 61(4):435-439.

Pontius, R.G., 2000. Quantification Error Versus Location Error in Comparison of Categorical Maps. *Photogrammetric Engineering and Remote Sensing* 66(8):1011-1016.

Rosenfield, G.H. and K. Fitzpatrick-Lins, 1986. A Coefficient of Agreement as a Measure of Thematic Classification Accuracy. *Photogrammetric Engineering and Remote Sensing* 52(2)223-227.

Seong, J.C. and E.L. Usery, 2001. Assessing Raster Representation Accuracy Using a Scale Factor Model. *Photogrammetric Engineering and Remote Sensing* 67(10):1185-1191.

Smith, D.P. and S.F. Atkinson, 2001. Accuracy Rectification Using Topographic Map versus GPS Ground Control Points. *Photogrammetric Engineering and Remote Sensing* 67(5):565-570.

Smits, P.C., S.G. Dellepiane, and R.A. Schowengerdt, 1999. Quality Assessment of Image Classification Algorithms for Land-Cover Mapping: A Review and a Proposal for a Cost-Based Approach. *Internation Journal of Remote Sensing* 20(8):1461-1486.

Stehman, S.V., 2001. Statistical Rigor and Practical Utility in Thematic Map Accuracy Assessment. *Photogrammetric Engineering and Remote Sensing* 67(6):727-734.

Stehman, V. and Czaplewski, R.L., 1998. Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remote Sensing of Environment* 64:331-344.

Stenback, J.M. and R.G. Congalton, 1990. Using Thematic Mapper Imagery to Examine Forest Understory. *Photogrammetric Engineering and Remote Sensing* 56:1285-1290.

Stoms, D.M., 1996. Validating Large-Area Land Cover Databases with Maplets. *Geocarto International* 11(2):87-95.

Story, M. and R.G. Congalton, 1986. Accuracy Assessment: A User's Perspective. *Photogrammetric Engineering and Remote Sensing* 52(3):397-399.

Swain, P.H. and R.C. King, 1973. Two Effective Feature Selection Criteria for Multispectral Remote Sensing. *Proceedings of the International Joint Conference on Pattern Recognition*. Washington, D.C., November 1973.

Swain, P.H. and Davis, 1978. *Remote Sensing: The Quantitative Approach*. McGraw-Hill.

Türk, G., 1979. GT-Index: A Measure of Success of Prediction. *Remote Sensing of Environment* 8:65-75.

Türk, G., 2002. Letter to the Editor: Map Evaluation and "Chance Correction." *Photogrammetric Engineering and Remote Sensing* 68(2):123-133.

Van Genderen, J.L. and B.F. Lock, 1977. Testing Land Use Map Accuracy. *Photogrammetric Engineering and Remote Sensing* 43(9):1135-1137.

Van Genderen, J.L., B.F. Lock, and P.A Vass, 1978. Remote Sensing: Statistical Testing of Thematic Map Accuracy. *Remote Sensing of Environment* 7:3-14.

Wang, F., 1990. Improving Remote Sensing Image Analysis Through Fuzzy Information Representation. *Photogrammetric Engineering and Remote Sensing* 56, 1163-1169.

Woodcock, C.E. and S. Gopal, 2000. Fuzzy Set Theory and Thematic Maps: Accuracy Assessment and Area Estimation. *Internation Journal of Geographic Information Systems* 14(2):153-172.

Zadeh, L. 1965. Fuzzy Sets. *Information and Control,* 8, 338-353.

Zhang, J. and G.M. Foody, 1998. A Fuzzy Classification of Sub-Urban Land Cover From Remotely Sensed Imagery. *Internation Journal of Remote Sensing* 19(14):2721-2738.

Zhou, G. and R. Li, 2000. Accuracy Evaluation of Ground Points from IKONOS High-Resolution Satellite Imagery. *Photogrammetric Engineering and Remote Sensing* 66(9):1103-1112.

APPENDICES

APPENDIX A: MRLC/NLCD Region 5 Classification System

The following description is from the data description file that accompanies the Region 5 data distributed by the MRLC Consortium.

The MRLC program utilizes a consistent classification scheme for all EPA Regions at approximately an Anderson Level II thematic detail. While there are 21 classes in the MRLC system, only 15 were mapped in EPA Region 5. The following classification scheme was applied to EPA Region 5 data set:

11 Water

21 Low Intensity Residential

22 High Intensity Residential

23 Commercial / transportation / industrial

31 Bare Rock / Sand

32 Quarries / Strip Mines / Gravel

33 Transitional

41 Deciduous Forest

42 Evergreen Forest

43 Mixed Forest

51 Shrubland

71 Natural Grassland

81 Pasture/Hay

82 Row Crops

83 Small Grains

85 Other Grassland (maintained)

91 Woody Wetlands

92 Emergent Herbaceous Wetlands

The class definitions are as follows:

Water - All areas of open water or permanent ice/snow cover.

11. Open Water - All areas of open water; typically ≥25 % cover of water (per pixel).

Developed - Areas with by a high percentage (≥30 %) of constructed materials (e.g. asphalt, concrete, buildings, etc).

21. Low Intensity Residential - Includes areas with a mixture of constructed materials and vegetation. Constructed materials are 30-80% of cover. Vegetation may are 20-70% of cover. These areas most commonly include single-family housing units. Population densities will be lower than in high intensity residential areas.

22. High Intensity Residential - Includes highly developed areas where people reside in high numbers. Examples include apartment complexes and row houses. Vegetation is ≤20 % cover. Constructed materials are 80-100% of cover.

23. Commercial/Industrial/Transportation - Includes infrastructure (e.g. roads, railroads, etc.) and all highly developed areas not classified as High Intensity Residential.

Barren - Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with little or no "green" vegetation present regardless of its inherent ability to support life. Vegetation, if present, is more widely spaced and scrubby than that in the "green" vegetated categories; lichen cover may be extensive.

31. Bare Rock/Sand/Clay - Prennially barren areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, beaches, and other accumulations of earthen material.

32. Quarries/Strip Mines/Gravel Pits - Areas of extractive mining activities with significant surface expression.

33. Transitional - Areas of sparse vegetative cover ($\leq$ 25% cover) that are dynamically changing from one land cover to another, often because of land use activities. Examples include forest clear cuts, transition between forest and agriculture, the temporary clearing of vegetation, and changes due to natural causes (e.g. fire, flood, etc.).

Forested Upland - Areas with tree cover (woody vegetation generally $\geq$6 m); tree canopy is 25-100% of cover.

41. Deciduous Forest - Areas dominated by trees where ≥75 % of the tree species shed foliage due to seasonal change.

42. Evergreen Forest - Areas dominated by trees where ≥75 % of the tree species maintain their leaves all year. Canopy is never without green foliage.

43. Mixed Forest - Areas dominated by trees where neither deciduous nor evergreen species represent ≥75 % of the cover.

Shrubland - Areas characterized by natural or semi-natural woody vegetation with aerial stems generally ≤6 m, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.

51. Shrubland - Areas dominated by shrubs; shrub canopy accounts for 25-100% of the cover. Shrub cover is generally ≥25% when tree cover is ≤25%. Shrub cover may be ≤25% in cases when cover of other life forms ≤25% and shrubs cover exceeds the cover of the other life forms.

Herbaceous Upland - Upland areas characterized by natural or semi-natural herbaceous vegetation; herbaceous vegetation accounts for 75-100% of the cover.

71. Grasslands/Herbaceous - Areas dominated by upland grasses and forbs. In rare cases, herbaceous cover ≤25%, but exceeds the combined cover of

the woody species present. These areas are not subject to intensive management, but they are often utilized for grazing.

Planted/Cultivated - Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation accounts for 75-100% of the cover.

81. Pasture/Hay - Areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops.

82. Row Crops - Areas used for the production of crops, such as corn, soybeans, vegetables, tobacco, and cotton.

83. Small Grains - Areas used for the production of graminoid crops such as wheat, barley, oats, and rice.

85. Urban/Recreational Grasses - Vegetation (primarily grasses) planted in developed settings for recreation, erosion control, or aesthetic purposes. Examples include parks, lawns, golf courses, airport grasses, and industrial site grasses.

Wetlands - Areas where the soil or substrate is periodically saturated with or covered with water as defined by Cowardin et al.

91. Woody Wetlands - Areas where forest or shrubland vegetation accounts for 25-100% of the cover and the soil or substrate is periodically saturated with or covered with water.

92. Emergent Herbaceous Wetlands - Areas where perennial herbaceous vegetation accounts for 75-100% of the cover and the soil or substrate is periodically saturated with or covered with water.

APPENDIX B: Neuse River Basin Classification System

The classification system for the Neuse River Basin project at EPA (Lunetta et al 2001) is as follows:

110	Urban - High Density
120	Urban - Medium Density
122	Urban - Medium - Agriculture
123	Urban - Medium - Woody
124	Urban - Medium - Herbaceous
125	Urban - Medium - Water
126	Urban - Medium - Wetland
127	Urban - Medium - Barren
130	Urban - Low Density
132	Urban - Low - Agriculture
133	Urban - Low - Woody
134	Urban - Low - Herbaceous
135	Urban - Low - Water
136	Urban - Low - Wetland
137	Urban - Low - Barren
211	Agriculture - Row - Cotton
212	Agriculture - Row - Corn
213	Agriculture - Row - Soybean
214	Agriculture - Row - Tobacco
220	Agriculture - Pasture/Hay
230	Agriculture - Fallow
310	Woody - Deciduous
320	Woody - Evergreen
330	Woody - Mixed
410	Grassland - Natural
420	Grassland - Maintained
500	Water
510	Water - Streams/Rivers
530	Water - Reservoirs
540	Water - Estuaries
550	Water - Ponds
610	Wetlands - Herbaceous
620	Wetlands - Woody
710	Barren - Non-vegetated
720	Barren - Transitional

APPENDIX C: Software Designed For This Project

Custom software was designed and coded by the author that performs not only the accuracy analyses required for this project, but also provides other accuracy estimators commonly used in remote sensing research. The accuracy estimates computed by the program are:

- Standard error matrix
- Overall percent accuracy
- Standard Kappa and variance
- Kappa Z statistic (testing versus a random classification)
- Errors of Commission and Omission
- User's and Producer's Accuracies
- Conditional Kappa and variance
- Weighted Kappa (plus SWK) and variance

As this software is expected to be of use to the remote sensing community, it will be released to the public domain upon completion of this project. Sample output of the software is presented below.

```
Error matrix: matrix_name

               Class 1       Class 2       Class 3       Class 4
Class 1        45            4             12            24
Class 2        6             91            5             8
Class 3        0             8             55            9
Class 4        4             7             3             55

Number of correctly classified points:  246
Overall Percent Accuracy:  73.214
```

```
Overall Kappa:  0.64
Kappa Variance:  0.001014
Kappa Z statistic:  20.109 *

Errors of Commission:
Class 1 :  47.059
Class 2 :  17.273
Class 3 :  23.611
Class 4 :  20.29

Errors of Omission:
Class 1 :  18.182
Class 2 :  17.273
Class 3 :  26.667
Class 4 :  42.708

User's Accuracy:
Class 1 :  52.941
Class 2 :  82.727
Class 3 :  76.389
Class 4 :  79.71

Producer's Accuracy:
Class 1 :  81.818
Class 2 :  82.727
Class 3 :  73.333
Class 4 :  57.292

Conditional Kappas and Variances:
Class 1 :  0.437         0.003001
Class 2 :  0.743         0.002389
Class 3 :  0.696         0.003612
Class 4 :  0.716         0.004059

Weighted Kappa and Variance:  0.692          0.000955
```

APPENDIX D: Glossary

Accuracy assessment – The process of quantifying the quality of a *classification*

Analyst – A person who performs *accuracy assessment*

Areal unit – A polygon or cluster *sampling unit* (i.e. not a *point*)

Band (image) – A sensitivity of an image in a range of the electromagnetic spectrum

Chance correction – Accounting for the possibility that reference sites were classified correctly by chance alone

Class – A category of interest for a particular application

Classification – A map composed of *thematic classes*

Classified map – See "*classification*"

Classifier – A mathematical algorithm that assigns each part of an image to the appropriate *thematic class*

Confidence interval – A statistical estimation of a parameter that provides a variance around the estimation

Data layer – A particular image or a *band* of an image

Diagnostic ability – The projected ability of a *classifier* to produce a classification of specified accuracy in the future (as opposed to *accuracy assessment*, which tests the quality of a completed classification)

DOQQ – Digital Orthorectified Quarter Quadrangle, a USGS product based on 1:40,000 scale aerial photos with a spatial resolution of one-meter

Error budget – Enumeration of the various sources of errors in a *classification* (*thematic*, *positional*, etc.)

Error matrix – An accuracy assessment tool that

Error, commission – Describing a particular *classification* error by denoting the class that was chosen instead of the correct class for the reference site

Error, omission – Describing a particular *classification* error by denoting that the correct class was not chosen for the reference site

Error, positional – Errors resulting from geometric misregistration of image layers

Error, thematic – Errors resulting from confusion of two map *classes* that are not caused by *positional errors*

Fuzzy membership values – The likelihood that a particular *reference site* belongs to each of the *classes* of interest

Fuzzy operators – Fuzzy mathematical constructs used to measure *thematic* accuracy

Fuzzy set theory – A mathematical discipline that is based on gradations of membership in the *classes* of interest

Ground truth – see "*reference data*"

Inclusion probability – The likelihood that a particular part of the study area will be sampled

Inter-class spectral distance – The distance in "spectral space" between two *classes*

Interpreter – A person who interprets *reference data* to determine the correct *classes* for the *reference sites*

Kappa – a discrete multivariate technique that tests whether one error matrix is significantly different from another.

Kappa, Conditional – A Kappa technique that provides accuracy measures for each class in a *classification*

Kappa, Weighted – A Kappa technique that allows for weighting of the various *misclassification costs* of the errors in a *classification*

KHAT – The statistical estimator of Kappa

Landsat Thematic Mapper – A satellite system launched by NASA

Linguistic scale – A word description of class memberships used to assist in assigning *fuzzy membership values*

LU/LC – Land Use and Land Cover

Map class – The *thematic class* shown on the map in a given area

Marginal error – See "*off-diagonal error*"

Maximum likelihood estimator – A classification technique that bases its decisions on gaussian probability of class membership

Misclassification cost – The costs associated with confusing the various *classes* in a *classification*

Multispectral – An image or sensor that has more than one spectral *band*

Neuse River Basin – One of the major hydrologic units in North Carolina

Off-diagonal error – Elements of an *error matrix* that are not correctly classified

Pixel – The smallest unit of measure of an image

Point – A *sampling unit* that is a single image *pixel*

Reference data – The agreed-upon correct class labels for the *reference sites*

Reference site – A location within the study area where *reference data* are collected

Sampling design – The overall design for the collection of *reference data*

Sampling scheme – A statistically valid method of placing reference sites throughout the *study area* for *accuracy assessment*

Sampling units – The individual samples sites used to collect *reference data*

Site-specific – Relating to specific *sample sites* on the ground rather than aggregate areas

Spatial autocorrelation – A concept that describes the likelihood that spatially collocated areas will be composed of similar *thematic* classes

Spectral cost – The *misclassification cost* that is derived from the *spectral distance* between two confused classes

Spectral distances – See "*inter-class spectral distance*"

SPOT – A satellite system launched by the French government

Supervised classification – A classification that was trained using *training samples*

Thematic – Composed of *classes* that represent ground phenomena of interest

Training samples – Class samples used to train a *classifier*