Second Edition

Assessing the Accuracy of Remotely Sensed Data

Principles and Practices

Reference Data

		Urban	Farm	Forest	Water	Total
Map	Urban	93	3	2	2	100
Data	Farm	10	83	4	3	100
	Forest	2	3	90	5	100
	Water	1	0	0	99	100
	Total	106	89	96	109	400

Overall Accuracy = (93+83+93+100)/400 = 91%

Producer's Accuracy			*User's Accuracy*		
Urban	*= 93/106 =*	*88%*	*Urban*	*= 93/100 =*	*93%*
Farm	*= 83/89 =*	*93%*	*Farm*	*= 83/100 =*	*83%*
Forest	*= 90/96 =*	*94%*	*Forest*	*= 90/100 =*	*90%*
Water	*= 99/109 =*	*91%*	*Water*	*= 99/100 =*	*99%*

Russell G. Congalton
Kass Green

CRC Press
Taylor & Francis Group

Second Edition

Assessing the Accuracy of Remotely Sensed Data

Principles and Practices

Second Edition

Assessing the Accuracy of Remotely Sensed Data

Principles and Practices

Russell G. Congalton
Kass Green

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-5512-2 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Library of Congress Cataloging-in-Publication Data

Congalton, Russell G., 1957-
Assessing the accuracy of remotely sensed data : principles and practices / Russell G. Congalton and Kass Green. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4200-5512-2 (alk. paper)
1. Remote sensing--Evaluation. I. Green, Kass. II. Title.

G70.4.C647 2008
621.36'78--dc22 2008030228

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Dedication

The second edition of this book is dedicated to our families, including our spouses Jeanie Congalton and Gene Forsburg: our children Ashton, Emma, and Brandon Congalton, and our parents Bob and Janet Congalton, John and Jean Samson, Mary Green, William and Carol McDevitt, and Frank and Janet Forsburg. Together they have made us much more than we would be alone. We are forever grateful for their love, support, and companionship.

Table of Contents

Preface

The field of assessing the accuracy of maps derived from remotely sensed data has continued to develop and mature since the first edition of this book was published in 1999. The original eight chapters have been expanded to eleven. Of most significance is a new chapter that covers positional accuracy. The accuracy of any spatial data set is a combination of both the positional accuracy and the thematic accuracy. Therefore, a complete presentation of how to assess the positional accuracy of a map has been added along with a discussion of the impact of positional accuracy on thematic accuracy. The use of fuzzy accuracy assessment has increased since the first edition, and we have included an entire chapter on this important process. Also, the chapter on assessing the accuracy of a map of change detection has been expanded with a more thorough discussion of the special sampling issues that must be considered to effectively assess the change. Finally, a new case study has been presented that is up-to-date and reflects the complications and issues one would face when conducting an accuracy assessment today.

Acknowledgments

There are a large number of people to thank for their help on this book. The first edition of this book was inspired by Dr. John Lyon and dedicated to Dr. Roy Mead. In addition, we are especially thankful to Dr. John Jensen, Dr. Greg Biging, Dr. Tom Lillesand, Dr. Jim Smith, Mr. Ross Lunetta, Mr. Mike Renslow, Dr. George Lee, and Dr. Jim Campbell for their positive feedback and support. Along the way, many colleagues at Pacific Meridian Resources, Space Imaging Solutions, Fugro-Earthdata, Dewberry, Tukman Geospatial and Sanborn Solutions as well as numerous graduate students in the Department of Natural Resources at the University of New Hampshire helped make this book better.

We would especially like to thank Mr. Mike Palmer for the many hours of discussion on change detection accuracy assessment that added significantly to this second edition. Many professionals have contributed to our new chapter on positional accuracy assessment. We would like to especially thank Drs. George Lee, Greg Biging, and Dave Maune, who carefully reviewed multiple drafts of Chapter 3 and provided invaluable new materials, insight and clarification. We would also like to thank Dewberry, the NOAA Coastal Services Center, and Sanborn Solutions for allowing us to use samples of their real world accuracy data in our examples. Additionally, we thank Mark Tukman of Tukman Geospatial and Chad Lopez of Fugro–Earthdata for their contribution and insight. We also gratefully acknowledge all of our friends and colleagues in the geospatial community who inspired and encouraged us on many occasions. Finally, we would like to thank our families for the time they managed without us while we worked on this book.

About the Authors

Russell G. Congalton is professor of remote sensing and GIS in the Department of Natural Resources and the Environment, University of New Hampshire. He is responsible for teaching courses in photogrammetry and photo interpretation, digital image processing, and geographic information systems. Russ has authored or coauthored more than 150 papers and conference proceedings. He is the author of eight book chapters and is coeditor of a book on spatial uncertainty in natural resource databases titled *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing.* Russ served as president of the American Society for Photogrammetry and Remote Sensing (ASPRS) in 2004–2005 and as the National Workshop Director for ASPRS from 1997–2008. In January 2008 he was appointed editor-in-chief of *Photogrammetric Engineering and Remote Sensing.*

Dr. Congalton received a B.S. (in natural resource management) from Rutgers University in 1979. He earned an M.S. (1981) and a Ph.D. (1984) in remote sensing and forest biometrics from Virginia Tech. In addition to his academic position, Russ served as chief scientist of Pacific Meridian Resources from its founding in 1988 until 2000, and then as chief scientist of Space Imaging Solutions from 2000–2004. Currently he serves as senior technical advisor with the Solutions Group of Sanborn, the oldest mapping company in the U.S.

Kass Green is the current president of the American Society of Photogrammetry and Remote Sensing (ASPRS 2008–09) and the president of Kass Green and Associates, where she consults on geospatial strategy, technology and policy issues to private, educational, and public organizations. Several years ago, Kass retired as president of Space Imaging Solutions. Prior to joining Space Imaging (now GeoEye), Kass was the president of Pacific Meridian Resources, a geospatial services company she cofounded in 1988 and sold to Space Imaging in 2000.

Kass received her B.S. degree in forestry from the University of California at Berkeley and her M.S. degree in resource policy and management from the University of Michigan; she advanced to her Ph.D. candidacy at the University of California at Berkeley.

1 Introduction

WHY MAP?

The earth's resources are scarce. As we continue to add more people to the earth, the scarcity of resources increases, as does their value. From land use conversion throughout the world, to fragmentation of tropical bird habitat, to acid rain deposition in Eastern Europe, to polar bear habitat loss in the Arctic, to the droughts in Africa, to wars, people have significantly affected the ecosystems of the world. The ever-increasing world population and need for resources continues to cause the price of resources to increase and to intensify conflicts over resource allocation.

As resources become more valuable, the need for timely and accurate information about the type, quantity, and extent of resources multiplies. Allocating and managing the earth's resources requires accurate knowledge about the distribution of resources across space and time. To efficiently plan emergency response, we need to know the location of roads relative to fire and police stations, hospitals, and emergency shelters. To improve the habitat of endangered species, we need to know what the species habitat requirements are, where that habitat exists, where the animals live, and how changes to the habitat and surrounding environments will affect species distribution, population, and viability. To plan for future developments, we need to know where people will work, live, shop, and go to school. Because each decision (including the decision to do nothing) impacts the (1) status and location of resources and (2) the relative wealth of individuals and organizations who derive value from those resources, knowing the location of resources and how they interact spatially is critical to effectively managing those resources and ourselves over time.

WHY ASSESS THE ACCURACY OF A MAP?

Thus, decisions about resources require maps, and effective decisions require accurate maps or at least maps of known accuracy. For centuries, maps have provided important information concerning the distribution of resources across the earth. Maps help us to measure the extent and distribution of resources, analyze resource interactions, identify suitable locations for specific actions (e.g., development or preservation), and plan future events. If our decisions based on map information are to have expected results, then the accuracy of the maps must be known. Otherwise, implementing such decisions will result in surprises, and these surprises may be unacceptable.

For example, suppose that you wish to have a picnic in a forest on the edge of a lake. If you have a map that displays forest, crops, urban, water, and barren land cover types, you can plan the location of your picnic. If you do not know the accuracy of the map, but the map is 100% accurate, you will be able to travel to your forest lakeside location, and in fact, find yourself in a nice picnic spot. However, if the maps are not spatially accurate, you may find that your picnic location falls in the middle of the lake rather than on the shore; and if the maps are not labeled correctly (i.e., thematically accurate), you may find yourself in a city next to a fountain, or in an agricultural field next to an irrigation ditch. However, if you know the accuracy of the map, you can incorporate the known expectations of accuracy into your planning and create contingency plans in situations when the accuracy is low. This type of knowledge is critical when we move from our lighthearted picnic example to more critical decisions such as endangered species preservation, resource allocation, peace-keeping actions, and emergency response.

There are many reasons for performing an accuracy assessment. Perhaps the simplest reason is curiosity—the desire to know how good a map you have made. In addition to the satisfaction gained from this knowledge, we also need or want to increase the quality of the map information by identifying and correcting the sources of errors. Third, analysts often need to compare various techniques, algorithms, analysts, or interpreters to test which is best. Also, if the information derived from the remotely sensed data is to be used in some decision-making process (i.e., GIS analysis), then it is critical that some measure of its quality be known. Finally, it is more and more common that some measure of accuracy be included in the contract requirements of many mapping projects. Therefore, a valid accuracy is not only useful, but may be required.

Accuracy assessment determines the quality of the map created from remotely sensed data. Accuracy assessment can be qualitative or quantitative, expensive or inexpensive, quick or time consuming, well designed and efficient or haphazard. The goal of *quantitative* accuracy assessment is the identification and measurement of map errors.

The purpose of this book is to present the theory and principles of quantitative accuracy assessment and to instruct readers how to adequately design and implement an accuracy assessment. Throughout the book, we emphasize that no single recipe exists for accuracy assessment. Just as there is no one way to produce a map, there is no one way to assess the accuracy of a map. Instead, this book will teach you to consider every aspect of a mapping project and to design and implement the best possible assessment given the strengths and limitations of each mapping project you conduct, fund, or rely on.

TYPES OF MAP ACCURACY ASSESSMENT

There are two types of map accuracy assessment: positional and thematic. Positional accuracy deals with the accuracy of the location of map features, and measures how far a spatial feature on a map is from its true or reference location on the ground (Bolstad, 2005). Thematic accuracy deals with the labels or attributes of the features of a map, and measures whether the mapped feature labels are different from the true feature label. For example, in the picnic example, the earth's surface was classified as forest, water, crops, urban, or barren. We are interested in both the accuracy of the location of the features so we can locate our picnic spot in a forest on the shore of a lake, and in the thematic accuracy so we truly end up in a forest and not in a city, desert, or agricultural field that was erroneously mapped as a forest.

The accuracy of any map or spatial data set is a function of both positional accuracy and thematic accuracy, and this book considers both. However, because thematic accuracy is much more complex than positional accuracy, the book devotes considerably more attention to thematic accuracy assessment.

CRITICAL STEPS IN ACCURACY ASSESSMENT

As previously stated, there is no single procedure for conducting either a positional or a thematic accuracy assessment. However, all accuracy assessments include three fundamental steps:

1. Designing the accuracy assessment sample
2. Collecting data for each sample
3. Analyzing the results

Each step must be rigorously planned and implemented. First, the accuracy assessment sampling procedures are designed, and the sample areas on the map are selected. We use sampling because time and funding limitations preclude the assessment of every spatial unit on the map. Next, information is collected from both the map and the reference data for each sample site. Thus, two types of information are collected from each sample:

- *Reference accuracy assessment sample data:* The position or class label of the accuracy assessment site, which is derived from data collected that are assumed to be correct.
- *Map accuracy assessment sample data:* The position or class label of the accuracy assessment site, which is derived from the map or image being assessed.

Third, the map and reference information are compared, and the results of the comparison are analyzed for statistical significance and for reasonableness. In summary, effective accuracy assessment requires (1) design and implementation of unbiased sampling procedures, (2) consistent and accurate collection of sample data, and (3) rigorous comparative analysis of the sample map and reference data.

Because there is no single procedure for designing and implementing accuracy assessments, there are a number of important questions to ask and considerations to think about when conducting a valid assessment. This book addresses the most important ones, including the following:

1. Questions concerning the design of an accuracy assessment sample:
    - What are the map classes to be assessed and how are they distributed across the landscape?
    - What is the appropriate sampling unit?
    - How many samples should be taken?
    - How should the samples be chosen?

2. Questions concerning how the reference data should be collected:
    - What should be the source of the reference data?
    - How should the reference data be collected?
    - When should the reference data be collected?
    - How do I ensure consistency and objectivity in my data collection?
3. Questions concerning how the analysis should be conducted:
    - What are the different analysis techniques for continuous as compared to discontinuous map data?
    - What is an error matrix and how should it be used?
    - What are the statistical properties associated with the error matrix and what analysis techniques are applicable?
    - What is fuzzy accuracy and how can you conduct a fuzzy accuracy assessment?
    - How is accuracy assessment conducted on change detection maps?
    - How is accuracy assessment conducted on maps created from multiple layers of data?

ORGANIZATION OF THE BOOK

The organization of this book takes you through each of the previously mentioned fundamental accuracy assessment steps as follows:

- The next three chapters (Chapters 2 through 4) introduce the basic concepts of positional and thematic accuracy assessment. Chapter 2 begins with a review of the history and basic assumptions of map making and accuracy assessment. Chapter 3 introduces the reader to positional accuracy assessment, while Chapter 4 introduces the concepts of thematic accuracy, including the error matrix.
- Chapter 5 reviews sample design considerations.
- Chapter 6 is devoted to factors that must be taken into account during the collection of reference data.
- Chapters 7 through 9 detail thematic accuracy assessment analysis, which is much more complex than positional accuracy assessment analysis. The basic analysis techniques that can be applied to an error matrix are discussed in Chapter 7. Chapter 8 discusses the causes of differences in the error matrix, whether from map errors or from other nonerror sources. Chapter 9 presents a solution to some of the nonerror differences in the error matrix by suggesting the use of fuzzy accuracy assessment.
- Chapter 10 presents a case study that reviews all the design, data collection and analysis methods presented in Chapters 3 through 9.
- Chapter 11 delves into more advanced topics in accuracy assessment, including change detection accuracy assessment and multilayer accuracy assessment.

2 The History of Map Accuracy Assessment

HOW MAPS ARE MADE

Before the invention of aircraft, maps were created from human observations made on the earth's surface using survey equipment and the most basic, yet sophisticated, remote sensing devices: the human eyes and the analytical capabilities of the human brain. By the early sixteenth century Portuguese navigators were able to map the coast of Africa (see Figure 2.1) by relying on measurements taken at sea from astrolabes, quadrants, cross-staffs, and other early navigation tools. During their exploration of the American Northwest, Lewis and Clark were able to produce the remarkably detailed map in Figure 2.2. Indian pundits secretly mapped the Himalayas to high precision in the mid-1800s by pretending to be Buddhist pilgrims (Hopkirk, 1992), keeping count of their paces using holy beads and concealing compasses and other instruments in their clothing and walking sticks. However, all of these maps were not without error, and when observations on the earth's surface were unobtainable, map makers often interpolated between field observations with questionable results, as illustrated in one of the earliest, and obviously incorrect, maps of California, displayed in Figure 2.3.

One of the most notorious examples of reliance on an erroneous map created from field observations was the disastrous Donner party in 1846, which chose to follow Hasting's cutoff rather than the established Oregon–California trail during their migration from the Midwest. As a result, they added miles to their journey (the positional accuracy was in error), were forced to cross unexpected steep mountains and expanses of waterless desert (the thematic accuracy was in error), as shown in Figure 2.4, and ended up attempting to cross the Sierra Nevada mountains in late fall rather than during the summer. The group ended up stranded in 20 ft of snow just below the summit for the entire winter, and lost almost half of their party to starvation, hypothermia, and cannibalism (Stewart, 1960). Regardless of how the map is made, not knowing the accuracy of maps can have catastrophic results!

Today, most map makers use remote sensing[†] rather than field observations as the main source of spatial information. While field observations are still important, they are ancillary to the remote sensing data, providing information at sample locations instead of a total enumeration of the area to be mapped. Since the first aerial photograph was captured from a balloon in 1858, data collected using remote sensing has supplanted ground observations for map making. Satellites and aircraft offer humans

[†] Remote sensing is defined as the collection and interpretation of information about an object from a distant vantage point. Remote sensing systems involve the measurement of electromagnetic energy reflected or emitted from an object, and include instruments on aircraft and satellites.

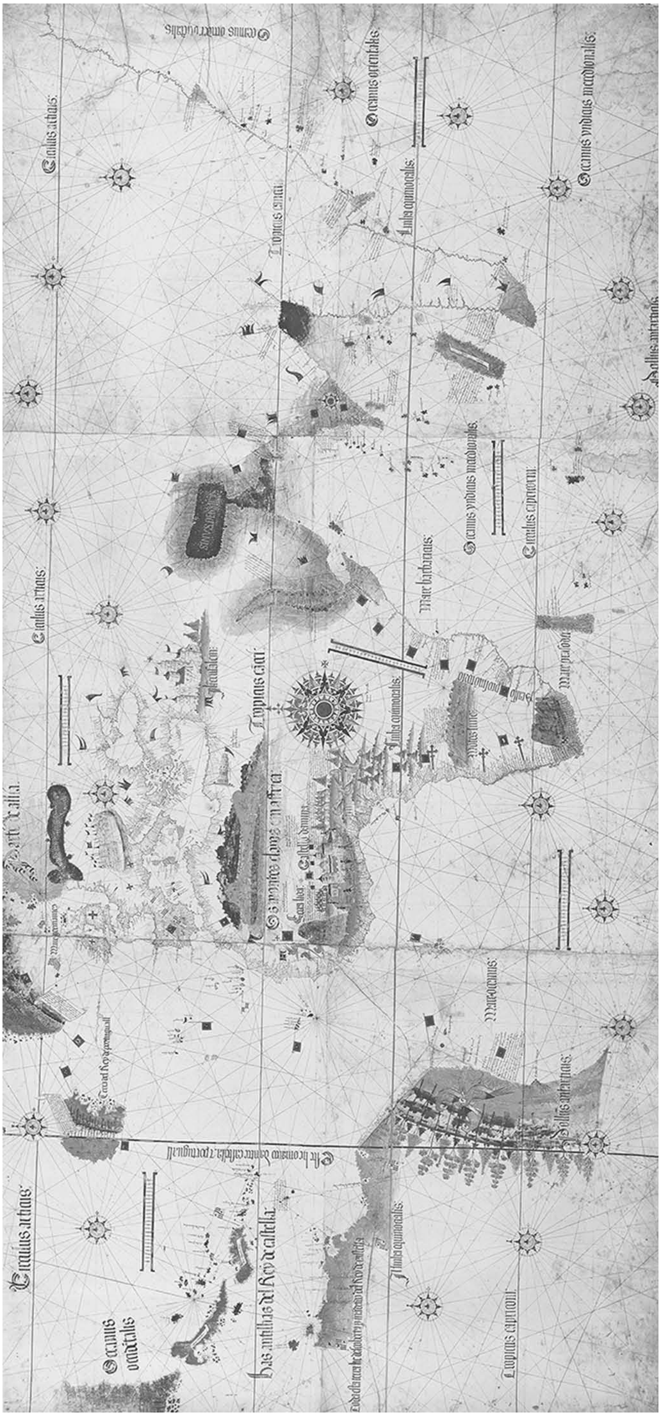


FIGURE 2.1 ***(Color version follows page 112)*** The Cantino World Map, which is a map of the known coastlines of the world, created by sixteenth century navigators.

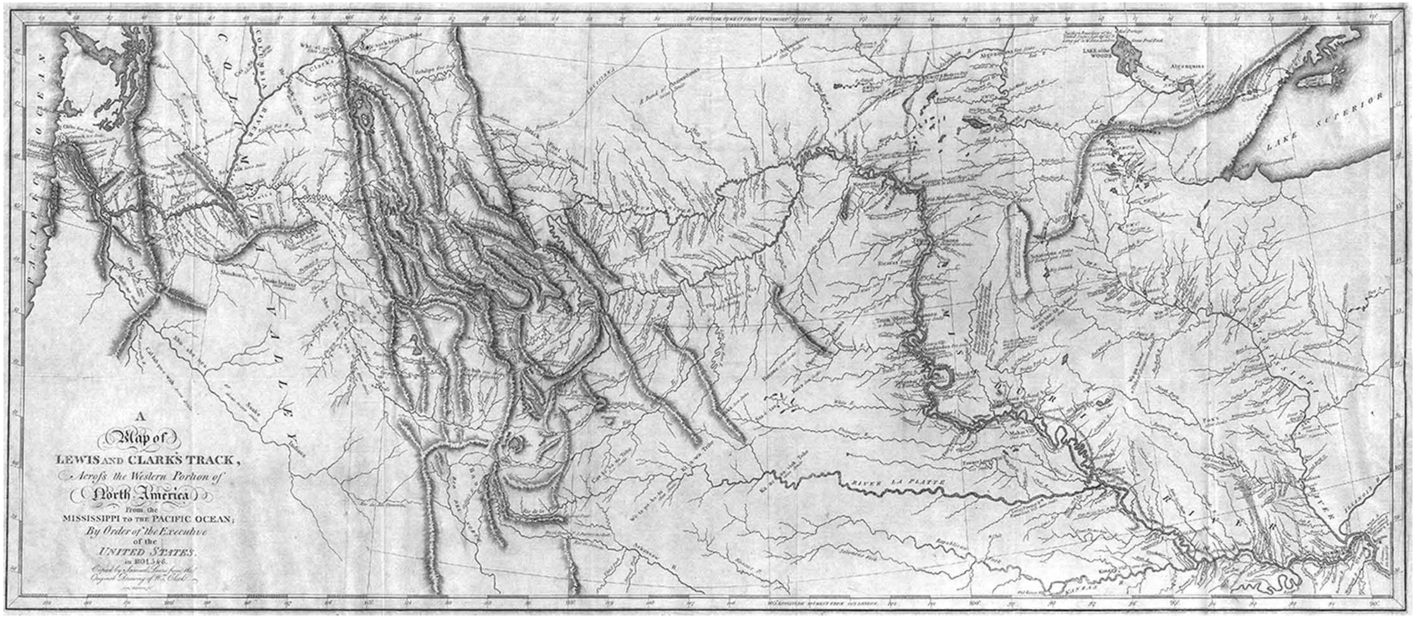


FIGURE 2.2 Map of the American Northwest created by Lewis and Clark. (From Lewis, Meriwether, William Clark, Nicholas Biddle, Paul Allen. 1814. *Maps of Lewis and Clark's Track across the Western Portion of North America*. Bradford and Inskeep, Philadelphia.)

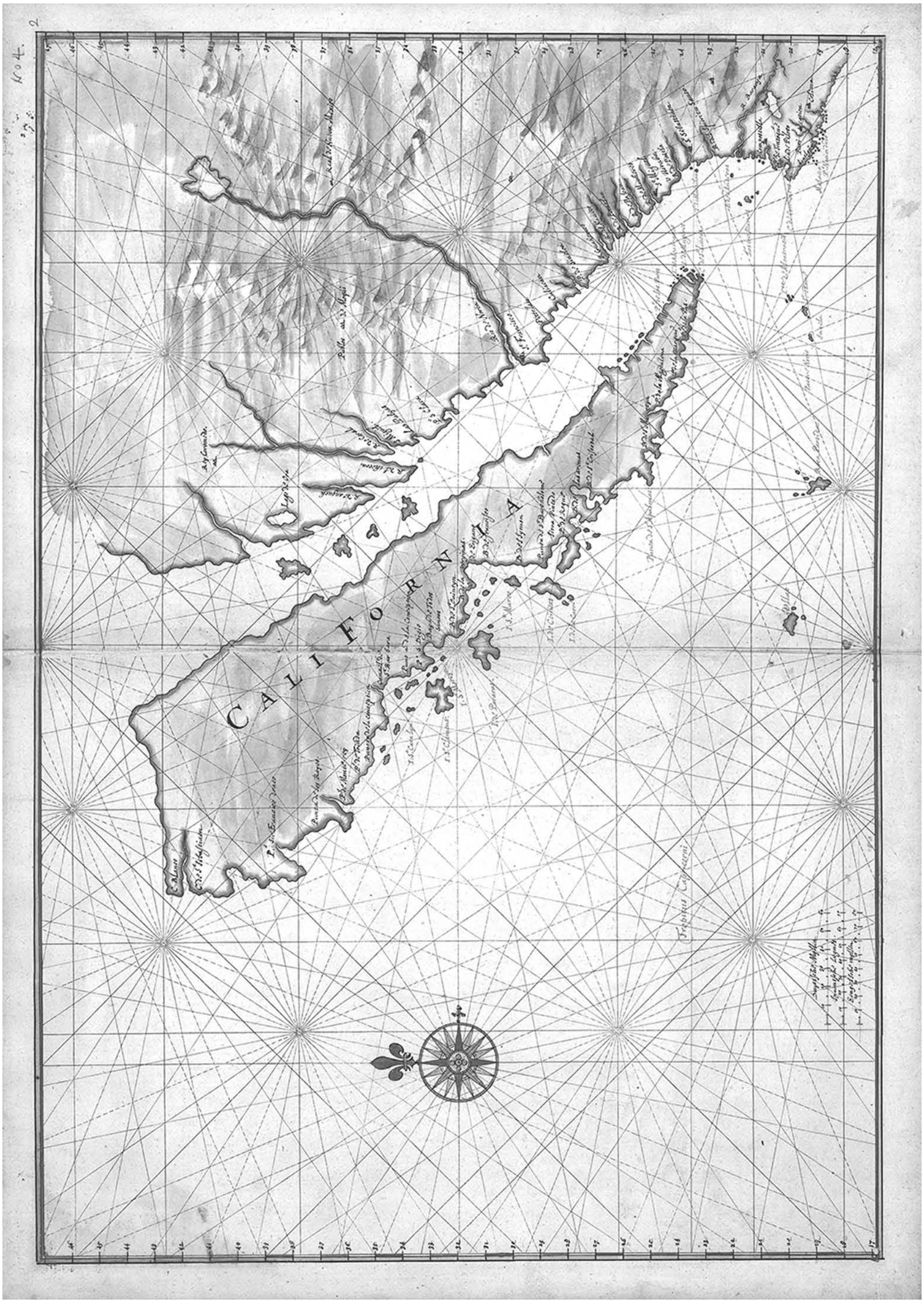


FIGURE 2.3 ***(Color version follows page 112)*** A seventeenth century map of California.

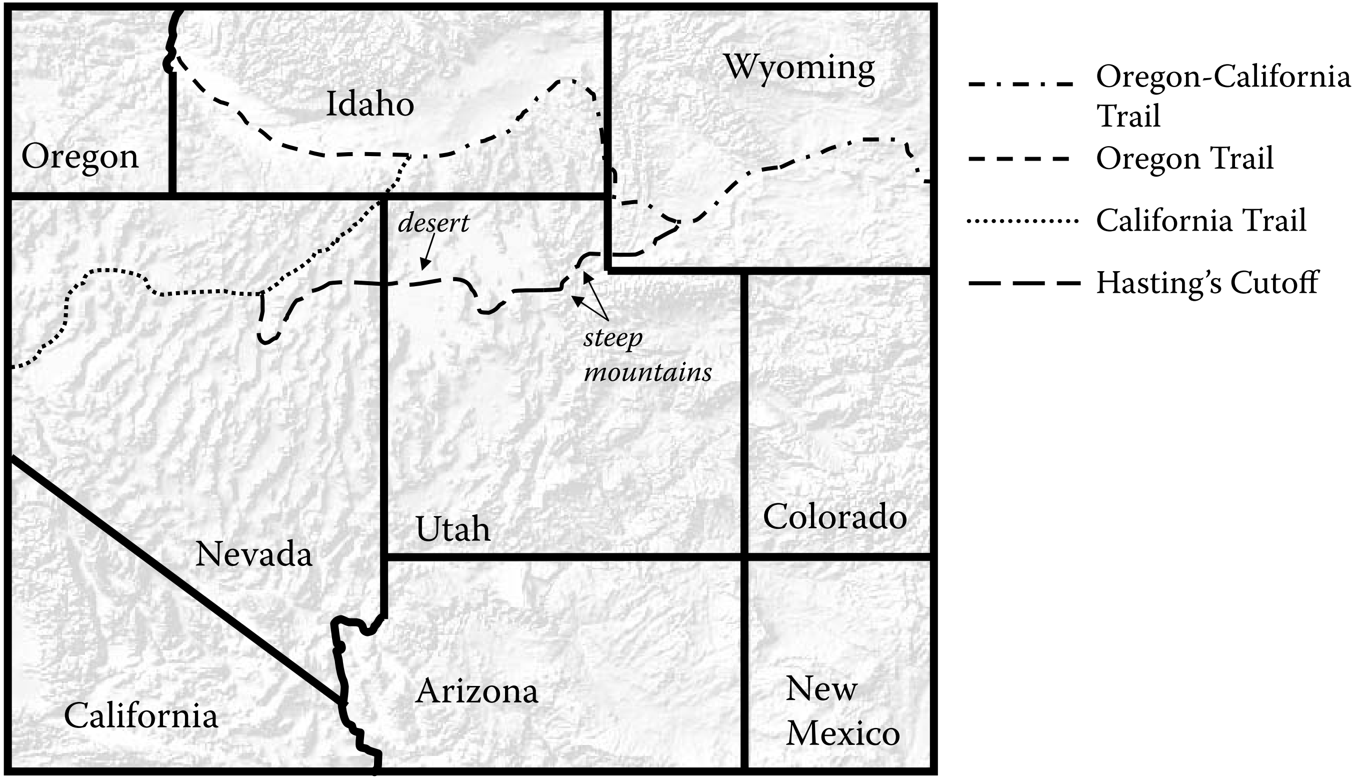


FIGURE 2.4 ***(Color version follows page 112)*** Hasting's Cutoff versus the safer California and Oregon trails used in 1846 by emigrants.

a view of their surroundings that humans cannot obtain on their own. Well before the first human went aloft in a balloon in 1783, humans had long been fantasizing about flight. Once humans invented successful flying machines, it was an easy step to put cameras in flying machines so that the pilot's perspective could be shared with those on the ground.

We use remotely sensed data to make maps because it:

- Is significantly less expensive and more efficient than creating maps from observations on the earth's surface,
- Offers a perspective from above (the "bird's-eye or synoptic view"), improving our understanding of spatial relationships, and
- Permits capturing imagery and information in electromagnetic wavelengths that humans cannot sense, such as the infrared portions of the electromagnetic spectrum.

Remotely sensed imagery is irresistible because it provides a view that can be readily understood, is inimitably useful, and yet is impossible to obtain without the use of technology. The innovation of air and space remote sensing has fundamentally changed the way we conduct war, manage inventory and resources, perform research, and respond to disasters.

Map making with remotely sensed data requires:

1. Precise linkage of the distances in the remote sensing imagery to distances on the ground so that spatial features can be accurately located, and
2. Understanding what causes variation in the features to be mapped and understanding how the remotely sensed data and ancillary information respond to those variations, so that the spatial features can be labeled.

Remotely sensed data provide an excellent basis for making maps because (1) remote sensing instruments and platforms are highly calibrated, and (2) a high correlation exists between variation in remotely sensed data and variation across the earth's surface.

However, there is never a complete one-to-one correlation between variation in remotely sensed data and variation on the earth's surface. Aircraft movement, topography, lens distortions, clouds, shadows, and a myriad of other factors can combine to weaken the relationship between the imagery and the earth's surface. Thus, much judgment, analysis, and interpretation are required to turn remotely sensed data into maps, and as a result, errors can occur during the many steps throughout any mapping project. As illustrated in Figure 2.5, the possible sources of error are multiple and compounding. Error can derive from the acquisition of imagery, to its rectification and classification, through its presentation as a map, and the application of the map in a decision-making process. Also, of course, error can also occur in the accuracy assessment itself. Accuracy assessment estimates, identifies, and characterizes the impact that arises from all of the sources of error.

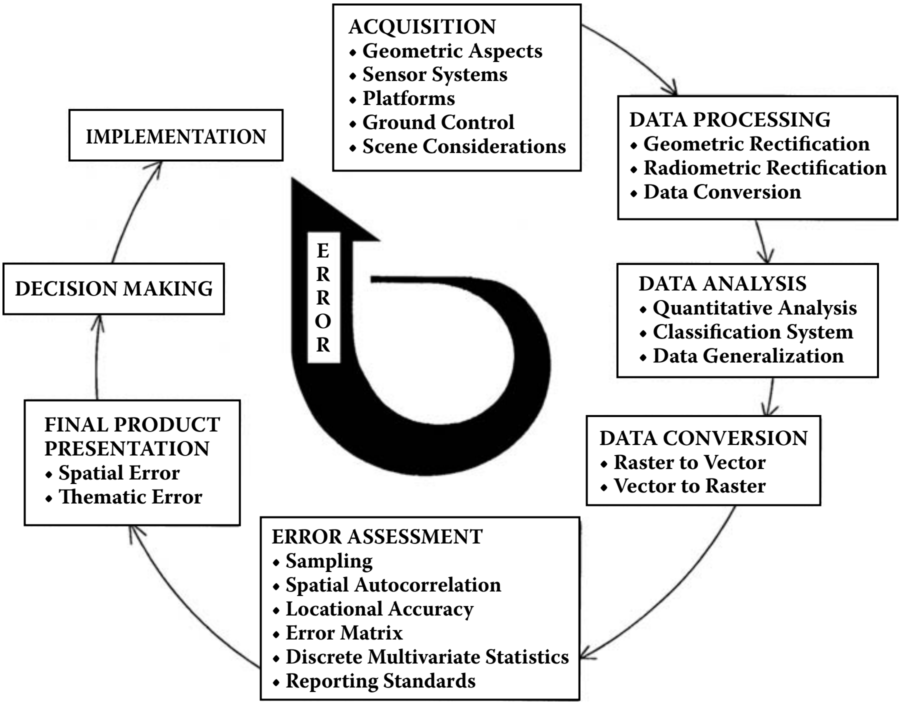


FIGURE 2.5 Sources of error in remotely sensed data. (Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Lunetta, R., R. Congalton, L. Fenstermaker, J. Jensen, K. McGwire, and L. Tinney. 1981. Remote sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering and Remote Sensing.* 57(6): 677–687.)

HISTORY OF ACCURACY ASSESSMENT

The widespread acceptance and use of remotely sensed data have been and will continue to be dependent on the quality of the map information derived from it. As we learned in the previous section, the history of using remotely sensed data for mapping and monitoring the earth is a relatively short one. Aerial photography (analog or film-based remote sensing) has been used as an effective mapping tool only since the early 1900s. Digital image scanners and cameras on satellites and aircraft have an even shorter history beginning in only the mid-1970s. The following two sections briefly review the history of positional and thematic accuracy assessment of maps created from remotely sensed data.

Positional Accuracy Assessment

Photogrammetry, the science of determining the physical dimensions of objects from measurements on aerial photographs or imagery, was first implemented in 1849 using terrestrial photographs taken on the earth's surface (McGlone, 2004). Aerial photogrammetry, which utilizes images taken from aerial or satellite platforms, followed soon after the first photographs were taken from aircraft. Adoption of aerial photographs to create maps exploded with:

- The need to rebuild Europe following World War I,
- Development of roll film by George Eastman (founder of Kodak),
- Reduction of camera lens distortion,
- Improvements in camera bodies including increased sturdiness, permanently mounted lenses, techniques for holding the film flat, and inclusion of a mechanism for aligning the camera axis,
- Employment of fiducial marks for the definition of the image plane,
- Development of analytical photogrammetry equations, and
- Invention of the stereo plotter. (Ferris State University, 2007)

From the very first days of aerial photogrammetry, positional accuracy has been assessed by comparing the coordinates of sample points on a map against the coordinates of the same points derived from a ground survey or some other independent source deemed to be more accurate than the map. In the early twentieth century, mapping scientists focused on map production and attempted to characterize each different contributor to positional error. Now, positional error assessment is more user-focused, emphasizing the estimation of the total error, regardless of the source.

In 1937, the American Society of Photogrammetry (now the American Society for Photogrammetry and Remote Sensing or ASPRS) established a committee to draft spatial accuracy standards. Soon after, the U.S. Bureau of the Budget published the *United States National Map Accuracy Standards* (NMAS) in 1941. The current version of the National Map Accuracy Standards was published in 1947 (U.S. Bureau of the Budget, 1947) and is included in the following text:

1. "Horizontal accuracy. For maps on publication scales larger than 1:20,000, not more than 10% of the points tested shall be in error by more than 1/30th inch, measured on the publication scale; for maps on publication scales of 1:20,000 or smaller, 1/50th inch. These limits of accuracy shall apply to positions of well-defined points only. Well-defined points are those that are easily visible or recoverable on the ground, such as the following: monuments or markers, such as bench marks, property boundary monuments; intersections of roads and railroads; corners of large buildings or structures (or center points of small buildings). In general, what is well defined will also be determined by what is plottable on the scale of the map within 1/100th inch. Thus, while the intersection of two roads or property lines meeting at right angles would come within a sensible interpretation, identification of the intersection of such lines meeting at an acute angle would not be practicable within 1/100th inch. Similarly, features not identifiable upon the ground within close limits are not to be considered as test points within the limits quoted, even though their positions may be scaled closely upon the map. This class would cover timber lines and soil boundaries.
2. Vertical accuracy, as applied to contour maps on all publication scales, shall be such that not more than 10% of the elevations tested shall be in error by more than one-half the contour interval. In checking elevations

taken from the map, the apparent vertical error may be decreased by assuming a horizontal displacement within the permissible horizontal error for a map of that scale.

3. The accuracy of any map may be tested by comparing the positions of points whose locations or elevations are shown upon it with corresponding positions as determined by surveys of a higher accuracy. Tests shall be made by the producing agency, which shall also determine which of its maps are to be tested, and the extent of such testing.
4. Published maps meeting these accuracy requirements shall note this fact in their legends, as follows: "This map complies with National Map Accuracy Standards."
5. Published maps whose errors exceed those aforestated shall omit from their legends all mention of standard accuracy.
6. When a published map is a considerable enlargement of a map drawing (manuscript) or of a published map, that fact shall be stated in the legend. For example, 'This map is an enlargement of a 1:20,000-scale map drawing,' or 'This map is an enlargement of a 1:24,000-scale published map.'
7. To facilitate ready interchange and use of basic information for map construction among all federal map-making agencies, manuscript maps and published maps, wherever economically feasible and consistent with the use to which the map is to be put, shall conform to latitude and longitude boundaries, being 15 minutes of latitude and longitude, or 7.5 minutes, or 3.75 minutes in size."

Establishment of the standards was a critical step in implementing consistency in positional accuracy across the United States. However, NMAS focuses on errors measured at the map instead of ground scale, which became problematic over the years as maps migrated from paper to digital formats that can be printed at variable map scales. Additionally, the standards state the requirements for spatial accuracy, but only briefly discuss procedures for collecting samples to determine whether or not those standards have been met. Thus, while the accuracy percentage was standardized, the procedures for measuring accuracy were not.

In the 1960s a precursor to the present-day National GeoSpatial-Intellegence Agency (NGA), the Aeronautical Chart and Information Center, printed a report entitled *Principles of Error Theory and Cartographic Applications* (Greenwalt and Schultz, 1962, 1968) that meticulously laid the statistical foundation for estimating the distribution of positional map error from a sample of reference points. The basic concepts of the report derive from the probability theories developed in the 1800s to predict the probable distribution of artillery shells fired at a target. Relying on the root-mean-square error (RMSE)[†] as the parameter to be estimated in characterizing positional map accuracy, the report became, and has remained, the foundation for all other publications that stipulate the calculation of map error from a set of sample points

[†] RMSE is the square root of the average squared differences between accuracy assessment sample map and reference locations. The equations for calculating RMSE are presented in Chapter 3.

(ASPRS, 1990; DMA, 1991; FGDC, 1998; MPLMIC, 1999; Bolstad, 2005; Maune, 2007). However, unlike later publications, the report focused only on how to calculate error and did not address how the sample points should be chosen or measured.

In the late 1970s, the American Society for Photogrammetry and Remote Sensing's (ASPRS) Specifications and Standards Committee started a review of the 1947 standards with the goal of updating them to include standards for both hardcopy and digital maps. The result was the 1990 publication of *ASPRS Interim Accuracy Standards for Large-Scale Maps* (ASPRS, 1990), which stipulated that accuracy be reported at ground scale rather than map scale, thereby allowing the consideration of digital as well as hardcopy maps. The standards established the maximum RMSE (measured at ground distances) permissible for map scales from 1:60 to 1:20,000. It also cited Greenwalt and Schultz (1962, 1968) in establishing RMSE as the pivotal map accuracy parameter. Finally, it provided guidance on how accuracy sample points should be identified, measured, and distributed across the map and how these points should be collected.

Soon after the release of the ASPRS Standards, the Ad Hoc Map Accuracy Standards Working Group of the Subcommittee on Base Cartographic Data of the Federal Geographic Data Committee (FGDC) produced the *U.S. National Cartographic Standards for Spatial Accuracy* (NCSSA) (FGDC, 1998) to create positional accuracy standards for medium- and small-scale maps.

Following public review, the NCSSA was significantly modified so as to adopt positional accuracy assessment procedures in lieu of accuracy assessment standards. The result was the 1998 publication of FDGC *National Standard for Spatial Data Accuracy* (NSSDA) (FGDC, 1998), which relies heavily on the ASPRS standards and "implements a statistical and testing methodology for estimating the positional accuracy of points on maps and in digital geospatial data, with respect to georeferenced ground positions of higher accuracy." The standard explicitly does not establish threshold standards (as did the NMAS and ASPRS), but encourages map users to establish and publish their standards, which it was recognized would vary depending on the user's requirements.

Also relying on Greenwalt and Schultz (1962, 1968), the NSSDA specifies that positional accuracy be characterized using RMSE, requires that accuracy be reported in ground distance units at the "95% confidence level,"[†] and provides guidance on how samples are to be selected. NSSDA continues to be the accepted standard on positional accuracy assessment. It is often used in conjunction with the ASPRS large-scale map standards, with NSSDA providing standardized processes for assessing positional accuracy and the ASPRS (1990) standards setting the maximum errors allowable for different map scales.

More recently, three new guidelines have been established for assessing digital elevation data. All three call for the stratification of positional accuracy assessment samples into land cover types. Two of the guidelines also mandate that accuracy be reported at the "95th percentile error" in addition to the NSSDA statistic.

[†] Confusion exists in the mapping field between the terms "95% precision level" and "95% confidence level." Chapter 3 examines the difference in detail.

Thematic Accuracy Assessment

Unlike positional accuracy, there is no government standard for assessing and reporting thematic accuracy. This omission is partially due to the inherent complexity of thematic accuracy, but primarily to the fact that when maps were made from aerial photographs, thematic accuracy was generally assumed to be at acceptable levels. It was the development and use of digital remote sensing devices that had the most profound impact on thematic accuracy assessment of maps created from all remotely sensed data.

Spurr, in his excellent book *Aerial Photographs in Forestry* (1948), presents the early prevailing opinion about assessing the accuracy of photo interpretation. He states, "Once the map has been prepared from the photographs, it must be checked on the ground. If preliminary reconnaissance has been carried out, and a map prepared carefully from good quality photographs, ground checking may be confined to those stands whose classification could not be agreed upon in the office, and to those stands passed through en route to these doubtful stands." In other words, a qualitative visual check to see if the map looks right has traditionally been the recommended course of action for assessing photo interpretation.

However, in the 1950s some researchers saw the need for quantitative assessment of photo interpretation in order to promote their discipline as a science (Sammi, 1950; Katz 1952; Young, 1955; Colwell, 1955). In a panel discussion entitled "Reliability of Measured Values" held at the 18th Annual Meeting of the American Society of Photogrammetry, Mr. Amrom Katz (1952), the panel chair, made a very compelling plea for the use of statistics in photogrammetry. Other panel discussions were held, and talks were presented that culminated with a paper by Young and Stoeckler (1956). In this paper, these authors actually propose techniques for a quantitative evaluation of photo interpretation, including the use of an error matrix to compare field and photo classifications, and a discussion of the boundary error problem.

Unfortunately, these techniques never received widespread attention or acceptance. The *Manual of Photo Interpretation* published by the American Society of Photogrammetry (1960) does mention the need to train and test photo interpreters. However, it contains no description of the quantitative techniques proposed by those brave few in the 1950s.

There is no doubt that photo interpretation has become a time-honored skill, and the prevailing opinion for decades was that a quantitative thematic accuracy assessment was unnecessary. In speaking with some of the old-time photo interpreters, they remember those times when quantitative assessment was an issue. In fact, they mostly agree with the need to perform such an assessment and are usually the first to point out the limitations of photo interpretation. However, it was mostly agreed that the results of any photo interpretation grouped areas that were similar and that there was more variation between these polygons or vegetation types or forest stands than between them. Hence, with this goal achieved, no quantitative assessment was necessary. Therefore, the quantitative assessment of photo interpretation is typically not a requirement of any project. Rather the assumption that the map was correct or at least good enough prevailed. Then along came digital remote sensing, and some of these fundamental assumptions about photo interpretation needed to be further scrutinized and adapted.

As in the early days of aerial photography, the launch of Landsat 1 in 1972 resulted in a great burst of exuberant effort as researchers and scientists charged ahead trying to develop the field of digital remote sensing. In those early days, much progress was made and there was not much time to sit back and evaluate how they were doing. This "can do" mentality is common in many developing technologies. The GIS (geographic information system) community has experienced a similar development pattern. However, as a technology matures, more effort is dedicated to data quality and error/accuracy issues. By the early 1980s, some researchers began to consider and realistically evaluate where they were going and, to some extent, how they were doing with respect to the quality of maps derived from digital remotely sensed data.

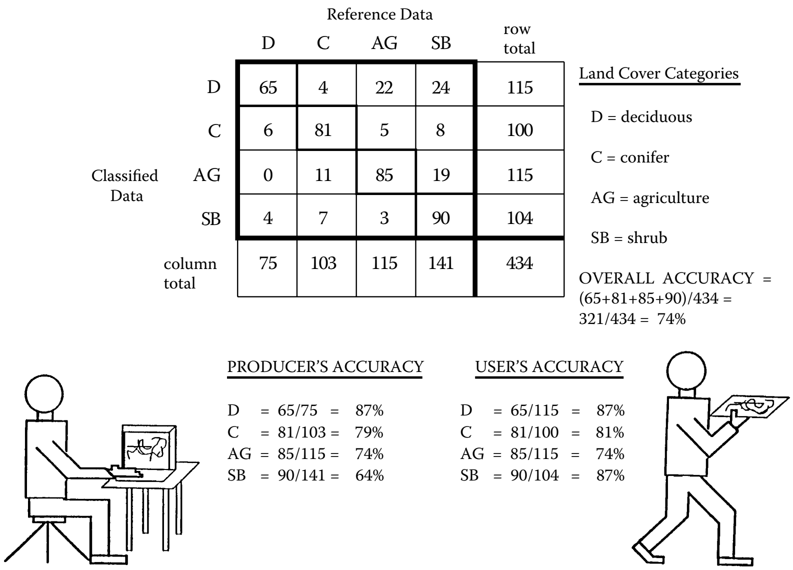
The history of assessing the thematic accuracy of maps derived from remotely sensed data is relatively brief, beginning around 1975. Researchers, notably Hord and Brooner (1976), van Genderen and Lock (1977), and Ginevan (1979), proposed criteria and basic techniques for testing overall map accuracy. In the early 1980s, more in-depth studies were conducted and new techniques proposed (Aronoff, 1982, 1985; Rosenfield et al., 1982; Congalton and Mead, 1983; Congalton et al. 1983). Finally, from the late 1980s up to the present time, a great deal of work has been conducted on thematic accuracy assessment. More and more researchers, scientists, and users are discovering the need to adequately assess the thematic accuracy of maps created from remotely sensed data.

The history of digital accuracy assessment can be effectively divided into four parts or epochs. Initially, no real accuracy assessment was performed but rather an "it looks good" mentality prevailed. This approach is typical of a new, emerging technology in which everything is changing so quickly that there is not time to sit back and assess how good you are doing. Despite the maturing of the technology over the last 25 years, some remote sensing analysts and map users are still stuck in this mentality.

The second epoch is called the age of non-site-specific assessment. During this period, total acreages by map class were compared between reference estimates and the map without regard for location. It did not matter if you knew where it was; rather, just the total amounts were compared. While total acreage is useful, it is far more important to know where a specific land cover or vegetation type exists. Therefore, this second epoch was relatively short-lived and quickly led to the age of site-specific assessments.

In a site-specific assessment, actual locations on the ground are compared to the same location on the map and a measure of overall accuracy (i.e., percentage correct) presented. This method far exceeded the non-site-specific assessment, but lacked information about individual land cover/vegetation categories. Only overall map accuracy was assessed. Site-specific assessment techniques were the dominant method until the late 1980s.

Finally, the fourth and current age of accuracy assessment could be called the age of the error matrix. An error matrix compares information from reference sites to information on the map for a number of sample areas. The matrix is a square array of numbers set out in rows and columns which express the labels of samples assigned to a particular category in one classification relative to the labels of samples assigned



		Reference Data				
		D	C	AG	SB	row total
Classified Data	D	65	4	22	24	115
	C	6	81	5	8	100
	AG	0	11	85	19	115
	SB	4	7	3	90	104
	column total	75	103	115	141	434

Land Cover Categories

D = deciduous

C = conifer

AG = agriculture

SB = shrub

OVERALL ACCURACY = (65+81+85+90)/434 = 321/434 = 74%

PRODUCER'S ACCURACY

D = 65/75 = 87%
C = 81/103 = 79%
AG = 85/115 = 74%
SB = 90/141 = 64%

USER'S ACCURACY

D = 65/115 = 87%
C = 81/100 = 81%
AG = 85/115 = 74%
SB = 90/104 = 87%

FIGURE 2.6 Example error matrix.

to a particular category in another classification (Figure 2.6). One of the classifications, usually the columns, is assumed to be correct and is termed the reference data. The rows are usually used to display the map labels or classified data generated from the remotely sensed image. Thus, two labels from each sample are compared to one another:

- Reference data labels: The class label or value of the accuracy assessment site, which is derived from data collected that is assumed to be correct; and
- Classified data or map labels: The class label or value of the accuracy assessment site derived from the map.

Error matrices are very effective representations of map accuracy because the individual accuracies of each map category are plainly described along with both the errors of inclusion (commission errors) and errors of exclusion (omission errors) present in the map. A commission error occurs when an area is included in an incorrect category. An omission error occurs when an area is excluded from the category to which it belongs. Every error on the map is an omission from the correct category and a commission to an incorrect category.

In addition to clearly showing errors of omission and commission, the error matrix can be used to compute not only overall accuracy, but also producer's accuracy, and user's accuracy, which were introduced to the remote sensing community by Story and Congalton (1986). Overall accuracy is simply the sum of the major diagonal (i.e., the correctly classified sample units) divided by the total number of sample units in the error matrix. This value is the most commonly reported accuracy assessment statistic and was part of the older, site-specific assessment. Producer's and user's accuracies are ways of representing individual category accuracies instead of just the overall classification accuracy (see Chapter 4 for more details on the error matrix).

Proper use of the error matrix includes correctly sampling the map and rigorously analyzing the matrix results. The techniques and considerations involved in the building and analyzing of an error matrix are the main themes of this book.

3 Positional Accuracy

Critical to any accuracy assessment project is the determination of precisely the same location on both the reference data and on the imagery or map being assessed. If this correspondence is not achieved, then the resulting poor positional accuracy may cause a false thematic error to occur. For example, it is possible to be in the correct location and mislabel (incorrectly measure or classify) the attribute. It is also possible to correctly label the attribute, but be in the wrong location. In either case, error is introduced into the map or spatial data set. These two factors are not independent of each other, and great care needs to be taken to not only assess each of these factors but also control them to minimize the errors.

As we learned in Chapter 1, accuracy assessment is characterized by two measures: positional and thematic accuracy. This chapter reviews the concepts of positional accuracy and is organized into the following sections. The first section introduces positional accuracy and briefly reviews the causes of positional accuracy error. The second section compares and contrasts the seven common standards for positional accuracy. The next section reviews basic statistics and positional accuracy sample design and collection within the overall framework set forth in the most commonly used standard, the *National Standard for Spatial Data Accuracy* (NSSDA) (FGDC, 1998). The fourth section explains how to analyze the accuracy assessment sample data to estimate positional accuracy under each standard. Finally, the last section compares the standards to one another and outlines a recommendation for positional accuracy that incorporates the concepts of existing standards, yet avoids some of the assumptions required by them.

A major goal of this chapter is to bring clarity to the language and equations of positional accuracy assessment. Since the development of the first standards in 1942, each new standard has introduced new concepts and interpreted old concepts in new ways. As a result, the language of positional accuracy assessment is often confusing, and the equations that comprise the accuracy assessment standards are, unfortunately, sometimes incorrect.

WHAT IS POSITIONAL ACCURACY?

The *Glossary of the Mapping Sciences* (ASPRS and ASCE, 1994) defines positional accuracy as "the degree of compliance with which the coordinates of points determined from a map agree with the coordinates determined by survey or other independent means accepted as accurate." All locations on maps and georeferenced images are expressed by a set of values: x- and y-coordinates for horizontal location. Many data sets also include elevations, which are represented by the letter z.

Positional accuracy uses sampling to estimate the discrepancy between a map or image feature's coordinates or elevations and their "true" location on the earth's surface. Positional accuracy can refer to either horizontal (planimetric) or vertical (elevational) accuracy, and this chapter discusses both.

Several factors can affect the positional accuracy of a map or georeferenced image. For example, the sensor lens may be distorted, or the aircraft carrying the sensor may suddenly tilt or yaw, changing the relationship of the sensor's image plane to the ground. However, the most important cause of positional error arises from the impact of topography on remotely sensed imagery. Because the sensor image plane is flat and the earth has relief such as hills and ravines, the scale of the remotely sensed imagery relative to the earth varies with topographic changes, requiring that some sort of adjustment be made to "terrain-correct" the image. This correction is a complex process that is highly prone to error.

Figure 3.1 presents an example of horizontal positional inaccuracy in which an inaccurate road layer is displayed over the top of an ortho-corrected digital image. The reference data, which have been "accepted as accurate," are survey points indicated by a crosshair on the figure. As you can see, the road layer does not exactly align with the points (i.e., there are positional errors)—the roads are shifted to the north and west of their "true" location, as determined by the survey points. While we can clearly see that

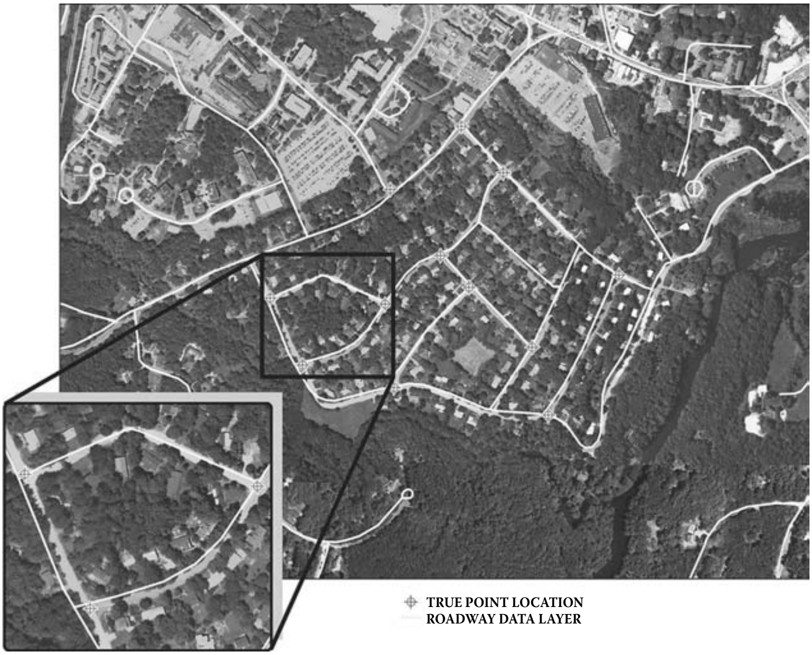
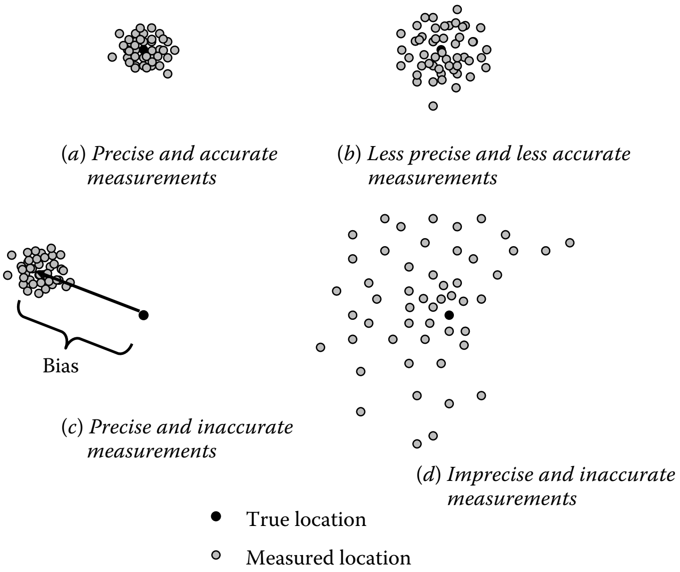


FIGURE 3.1 Illustration of positional errors in a road map (in white) compared to the image, which is assumed to be accurate.



(*a*) *Precise and accurate measurements*

(*b*) *Less precise and less accurate measurements*

(*c*) *Precise and inaccurate measurements*

(*d*) *Imprecise and inaccurate measurements*

FIGURE 3.2 Illustrations of precision versus accuracy.

the position of the road is inaccurate, we need to use quantitative accuracy assessment to estimate the mean error in the accuracy of the position of the road layer.

In statistics and accuracy assessment, there are two terms that are commonly used and confused that need clarification. Accuracy and precision are often thought of as synonymous, but actually have very different meanings. Accuracy refers to the bias of an estimator. It measures how close an estimated or calculated value is to its true value. Precision refers to the variability in an estimator. It quantifies how repeated measures of the same estimator will vary. Inaccurate measurements can be very precise, and accurate measurements can be imprecise. Figure 3.2 illustrates the concepts of accuracy and precision with an example of multiple measurements made of one location.

In positional accuracy assessment, we are interested in characterizing the accuracy of a geospatial data set. We take samples to determine if a bias (systematic inaccuracy) exists in the data set, and we estimate the magnitude and precision of the bias. We also strive to ensure that our measurements of each sample's reference and geospatial data set's locations are themselves accurate, and we must take enough samples so that our estimate of the bias (if it exists) is precise.

WHAT ARE THE COMMON STANDARDS FOR POSITIONAL ACCURACY?

The *National Standard for Spatial Data Accuracy* (NSSDA) (FGDC, 1998) is the most widely used positional accuracy standard. However, new standards have been developed and several earlier standards are still in use. In addition, it is not uncommon to have two standards applied to the same project. This section compares and contrasts the seven primary positional accuracy standards:

1. *United States National Map Accuracy Standards* (NMAS) (U.S. Bureau of the Budget, 1947),
2. *Principles of Error Theory and Cartographic Applications* (Greenwalt and Schultz, 1962, 1968), which is cited by all subsequent standards,
3. *ASPRS Interim Accuracy Standards for Large-Scale Maps* (ASPRS, 1989),
4. The Federal Geographic Data Committee's *National Standard for Spatial Data Accuracy* (FGDC, 1998),
5. The Federal Emergency Management Agency's (FEMA) *Guidelines and Specifications for Flood Hazard Mapping Partners* (FEMA, 2003),
6. *ASPRS Guidelines, Vertical Accuracy Reporting for Lidar Data* (ASPRS, 2004), and
7. The National Digital Elevation Program (NDEP) *Guidelines for Digital Elevation Data* (NDEP, 2004).

National Map Accuracy Standards

NMAS (reproduced in its entirety in Chapter 2) stipulates that:

- For horizontal accuracy, not more than 10% of the points tested may be in error by more than 1/30th of an inch (at map scale) for maps larger than 1:20,000 scale, or by more than 1/50th of an inch for maps of 1:20,000 scale or smaller, and
- For vertical accuracy, not more than 10% of the elevation tested may be in error by more than one half the contour interval.

The standard is very straightforward and simple and does not require any assumptions about the distribution of error. Using what is later termed the "percentile method," NMAS merely states that no more than 10% of the samples may exceed the maximum error allowed. However, because it relies on map versus ground units, and because it provides no guidance for creating statistically valid bounds on the estimated error, NMAS is rarely used today.

Principles of Error Theory and Cartographic Applications

The *Principles of Error Theory and Cartographic Applications* (Greenwalt and Schultz, 1962, 1968) report (hereafter referred to as Greenwalt and Schultz) approaches positional accuracy from a diametrically opposite standpoint compared to NMAS by proposing equations that should be applied to estimate the maximum error interval that would occur at various probabilities. The report interprets NMAS' (1947) "10% of the points taken" to limit the size of errors to that within "which 90% of the well defined points will not exceed" (Greenwalt and Schultz, 1962, 1968), which it terms the "map accuracy standard" (MAS). The report uses probability theory to develop equations for calculating one-dimensional elevation (z) "map accuracy standard" and two-dimensional (x and y) "circular map accuracy standard" (CMAS) statistics by assuming that map errors are normally distributed.[†] MAS is the estimated interval around the mean vertical error, and CMAS is the estimated interval around the horizontal mean error within which 90% of the errors are predicted to occur.

While seemingly similar, the two standards—NMAS and Greenwalt and Schultz—are very different. NMAS *stipulates* the maximum size of error that 10% of the samples may not exceed. The Greenwalt and Schultz standard does not stipulate a maximum error. Rather, it *calculates* the probable maximum error interval around the mean error from the sample data.

Additionally, Greenwalt and Schultz does not specify 90% as the only probability level to be employed. Instead, it shows how to estimate the distribution of errors under various probability levels and provides tables for converting from one probability level to another.

ASPRS Interim Accuracy Standards for Large-Scale Maps

Similar to NMAS, ASPRS (ASPRS, 1989) standards stipulate a maximum distance beyond which errors may not exceed. However, ASPRS differs from NMAS in stating how to determine if the errors have exceeded the maximum acceptable error. Rather than stipulating that no more than 10% of the errors may exceed the stipulated maximum, ASPRS states that the mean error estimated from the samples may not exceed the stipulated maximum distance. Most importantly, the ASPRS standards migrate the units of measurement of error from map units to ground units. The ASPRS standards also restate the Greenwalt and Schultz CMAS equations, but do not imply that the equations should necessarily be used.

National Standard for Spatial Data Accuracy

As mentioned in Chapter 2, the *National Standard for Spatial Data Accuracy* (FGDC, 1998) established much-needed guidelines for measuring, analyzing, and reporting positional accuracy of both maps and georeferenced imagery such as orthophotos or orthoimages. While developed for federal agencies, the NSSDA standards have been widely accepted by many local and state government agencies, as well as by the private sector. Because of the importance of the NSSDA in establishing positional accuracy assessment procedures, we highly recommend that the reader download the NSSDA at http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3.

NSSDA explicitly rejects setting a maximum allowable error at any scale and suggests instead that the maximal allowable error threshold be determined as needed. Instead, accuracy is to be reported "in ground distances at the 95% confidence level," which is interpreted as allowing "one point to fail the threshold given in the product

[†] We will examine the implications of the assumption of normality later in this chapter.

specification”[†] when a sample of 20 points is used. It is not unusual for positional accuracy projects to use the equations of NSSDA to calculate accuracy statistics *and* to require that those statistics not exceed the distances established in the ASPRS (1989) standards. Similar to the ASPRS standards, NSSDA relies on ground rather than map units and uses the mean error as an accuracy statistic. NSSDA increases the probability level to 95%, an increase of 5% above the Greenwalt and Schultz-interpreted NMAS level of 90%. NSSDA also incorporates the approach of Greenwalt and Schultz by referencing its equations and defining accuracy as a measure of the maximum error expected at a specific probability level. However, as we will learn later, NSSDA incorrectly implements the Greenwalt and Schultz equations.

Guidelines and Specifications for Flood Hazard Mapping Partners

FEMA’s *Guidelines and Specifications for Flood Hazard Mapping Partners* (FEMA, 2003) adds a new dimension to positional accuracy assessment by requiring that a minimum of 20 samples be collected for each major vegetation type of which there may be a minimum of 3, resulting in a minimum of 60 total sites sampled. The vegetation types specified are:

1. Bare-earth and low grass,
2. High grass, weeks, and crops,
3. Brush lands and low trees,
4. Forested, fully covered by trees,
5. Urban areas,
6. Sawgrass, and
7. Mangrove.

ASPRS Guidelines: Vertical Accuracy Reporting for Lidar Data

ASPRS Guidelines for Reporting Vertical Accuracy of Lidar Data (ASPRS, 2004) ratify the FEMA guidance to stratify the landscape into different land cover classes. The ASPRS classes differ slightly from the FEMA classes and are:

1. Open terrain,
2. Tall weeds and crops,
3. Brush lands and low trees,
4. Forested areas fully covered by trees, and
5. Urban areas with dense human-made structures.

The ASPRS guidelines also call for vertical accuracy to be reported in three different ways depending on the ground cover of the area being mapped or imaged:

[†] While NSSDA assumes that the two quotes in this sentence refer to the same statistic, they, in fact, imply two different statistics. The first quote refers to a “confidence level,” which in statistics is the measure of reliability of the parameter being estimated, in this case the RMSE. The second quote refers to the estimated distribution of errors. The difference between these two statistics will be discussed in more detail later in this chapter.

1. "Fundamental vertical accuracy" is computed only from samples measured in open terrain and relies on the NSSDA equations for calculating accuracy.
2. "Supplemental accuracy" is measured from samples taken in nonopen terrain cover types and is determined using the "95th percentile error" method, which is defined as the "absolute value in a data set of errors. It is determined by dividing the distribution of the individual sample errors in the data set into 100 groups of equal frequency. " By definition, 95% of the sampled errors will be less than the 95th percentile value.
3. "Consolidated vertical accuracy" is a combination of the samples from both open terrain and other ground cover classes and is reported as a 95th percentile error.

Guidelines for Digital Elevation Data

The National Digital Elevation Program (NDEP) *Guidelines for Digital Elevation Data* (NDEP, 2004) essentially mirror the ASPRS (2004) lidar guidelines for vertical accuracy reporting in calling for the computation of Fundamental Vertical Accuracy, Supplemental Vertical Accuracy, and Consolidated Vertical Accuracy. Both documents also mandate that errors higher than the 95th percentile be documented in the metadata. NDEP refers to this aspect of its standard as its "truth in advertising approach."

POSITIONAL ACCURACY ASSESSMENT DESIGN AND SAMPLE SELECTION

Positional accuracy assessment requires the appropriate selection of samples to estimate the statistical parameters of the population of errors (e_i) occurring in the spatial data being assessed. Parameters such as the mean (μ), standard deviation (σ), and standard error (σ_μ) characterize the distribution of the population of errors and the reliability of estimators. The mean (μ) is the expected value of a random variable. In the case of positional accuracy, the mean is the expected error, which is usually estimated by the root-mean-square error, or RMSE. The standard deviation (σ) is the square root of the population variance. The variance measures how much the variables of a population deviate from the population mean. The standard error (σ_μ) is the square root of the variance of the estimate of the mean. It measures how estimates of the population mean will deviate from the true mean and is used to create a confidence interval around an estimate of the mean. Equations for calculating these variables and their estimators are presented in the following text.

Estimating positional error parameters requires the comparison of coordinates and/or elevations of identical sample locations from:

- The spatial data set to be assessed (map or imagery) and
- The reference data, which must be an "independent source of higher accuracy" (FGDC, 1998).

We rely on samples because measuring every point in the geospatial data set being assessed would be prohibitively expensive, and sampling can provide highly reliable estimates of the error population's parameters.

NSSDA (FGDC, 1998) outlines several requirements that govern positional accuracy sampling design and collection. They are:

Data independence. To ensure the objectivity and rigor of the assessment, it is critically important that the reference data be independent from the data being tested. In other words, the reference data cannot have been relied upon during the creation of the map or image being assessed. Thus, control points or digital elevation models used to create the spatial products being tested are unsuitable sources of reference data.

Source of reference data. The source of the reference points depends on a number of factors. In some cases, a map of larger scale than the map or image being assessed may provide sufficiently detailed reference coordinates. This is especially true if the map/image to be tested is small in scale and covers a large area. In other cases, such as engineering site drawings, much more precision is required for the reference data points; a field survey or use of a high-precision GPS may be required. NSSDA (FGDC, 1998) stipulates that the reference source data "be of the highest accuracy feasible and practicable." Other handbooks suggest that the reference data be from one to three times more accurate than the anticipated accuracy of the data being tested (Ager, 2004; MPLMIC, 1999; NDEP, 2004; ASPRS, 2004).

Number of samples. The NSSDA (FGDC, 1998) requires a minimum of 20 sample points. Other standards require a minimum of 20 samples per ground cover class and suggest that at least 30 sample points per class are preferred (NDEP, 2004; ASPRS 2004). For statistical rigor, more than 20 sample locations should be chosen. Fewer than 20 points do not provide sufficient samples for a statistically valid estimate. If the population of errors is normally distributed, as illustrated in Figure 3.3, then taking more than 30 samples results in the effort or cost required to collect additional samples exceeding the additional sample's contribution to the precision of the accuracy estimate. If the distribution of the population of errors is skewed or flat, then the sample size should be increased. Lopez et al. (2005) argue that at least 100 samples points are required to achieve a 95% confidence level. However, collecting reference samples, especially ground survey locations, can be extremely expensive and most positional accuracy assessments rely on the NSSDA minimum of 20 samples.

As an alternative, the number of samples required to meet a specified probability level can be calculated as long as reliable approximations of the mean and standard deviation are available. The equations for this calculation can be found in Appendix 3.1 of this chapter.

Identification of samples. The samples must consist of "well-defined points" that "represent a feature for which the horizontal position is known to a high degree of accuracy and position with respect to the geodetic datum" (FGDC, 1998). What constitutes a "well-defined point" will vary with the scale of

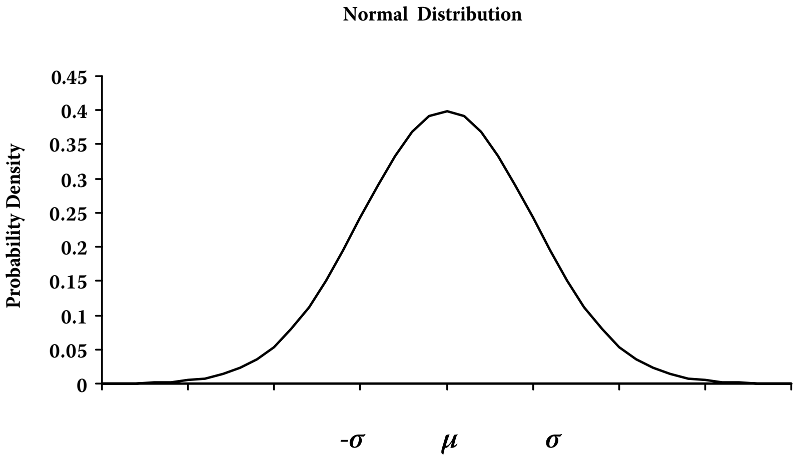


FIGURE 3.3 Shape of the normal distribution.

the map or imagery being assessed. Each point must be clearly identifiable on the spatial data being assessed and in the reference data set. Any error in locating the test points can significantly impact the positional accuracy results. NSSDA (FGDC, 1998) suggests that, "For graphic maps and vector data, suitable well-defined points represent right-angle intersections of roads, railroads, or other linear mapped features, such as canals, ditches, trails, fence lines, and pipelines. For orthoimagery, suitable well-defined points may represent features such as small isolated shrubs or bushes, in addition to right-angle intersections of linear features. For map products at scales of 1:5000 or larger, such as engineering plats or property maps, suitable well-defined points may represent additional features such as utility access covers, and intersections of sidewalks, curbs, or gutters."

Distribution of samples. The sample points must also be well distributed across the project area, and represent the full variety of topography, as topography has the largest impact on positional accuracy. Several options are available for distributing samples across the map or image being assessed:

- The points may be randomly selected using a random number generator. However, the sample points must be identifiable on the imagery or map being assessed, as well as on the reference data. Only a subset of the total population of map or image points will be identifiable. Additionally, locating random points in the field can be problematic if the points fall on private property or inaccessible terrain.
- NSSDA suggests that samples "may be distributed more densely in the vicinity of important features and more sparsely in areas that are of little or no interest." However, emphasis on "important features" will most likely result in a biased sample that may produce biased estimates of the error population parameters.

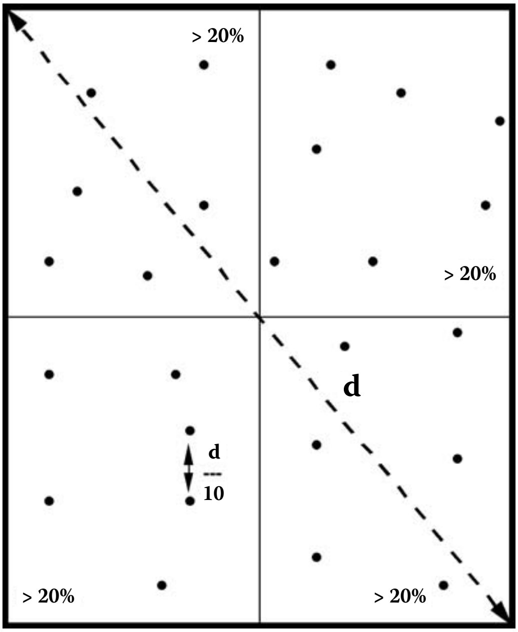


FIGURE 3.4 ASPRS (1989) suggested distribution of positional accuracy assessment sample locations.

- Figure 3.4 illustrates the ASPRS-suggested (1989) systematic sampling method, which ensures that the sample points are well distributed throughout the map or image being assessed. To implement the ASPRS sample distribution, first, the map or image is divided into quadrants. Next, a minimum of 20% of the sample points are allocated to each quadrant. To ensure adequate spacing between the sample points, no two points should be closer than *d*/10 distance from each other, where *d* is the diagonal dimension of the map or image. This spacing will minimize spatial autocorrelation (a topic that will be discussed in detail in later chapters). In addition, using the ASPRS systematic sample distribution requires assuming that the sample distribution is not correlated with map or image error. This is a reasonable assumption because most positional error is correlated with topography, and topography is rarely distributed on a grid pattern.

In summary, the design and collection of positional accuracy assessment requires the simultaneous consideration of several factors. Often there is a trade-off between well-distributed sample points and easily identifiable sample points. It is not uncommon for some of the desired sample points to fall on private land which may be inaccessible if a ground survey is being used as the reference data. Often, easily identifiable points are concentrated in small areas or are not evenly distributed throughout the map. Care must be taken to obtain the best possible combination of good test points that are appropriately distributed throughout the map or image being assessed.

HOW IS POSITIONAL ACCURACY ANALYZED?

Analyzing positional accuracy involves using sample data to estimate the fit of the spatial data layer (map or image) being assessed to the reference layer, which is assumed to be correct. The accuracy of the fit is depicted by characterizing the distribution of error using the error population's estimated mean, standard deviation, and standard error. Because much confusion exists between the commonly used accuracy standards, we will begin this section with a review of the basic statistics and then move on to the specific equations for depicting positional accuracy.

Review of Basic Statistics

The concepts in this section may be found in any standard statistics textbook. Documents directly relied upon for this text include *Principles of Error Theory and Cartographic Applications* (Greenwalt and Schultz, 1962, 1968), *Biostatistical Analysis (*Zar, 1974), and *Analysis and Adjustment of Survey Measurements (*Mikhail and Gracie, 1981).

This section first provides the equations for calculating and estimating the parameters of a population of values. Next, it discusses the assumptions and equations required to estimate the dispersal of values around the mean. Finally, it provides the equations for calculating a confidence interval around the estimate of the mean.

Parameters and Statistics

The arithmetic mean (μ) of a population of random variables *(*X_i*)* is the expected value of any random variable and is calculated by

$$\mu_{X_i} = \sum_i^N X_i/N \tag{3.1}$$

where

X_i = the value of the *i*th individual in the population, and
N = the total number of individuals in the population.

The mean is estimated from a sample by the variable $\overline{X}$ and is calculated by

$$\overline{X} = \sum_i^n x_i/n \tag{3.2}$$

where

x_i = the value of the *i*th sample unit chosen from the population, and
n = the total number of sample units chosen.

The standard deviation (σ) is the square root of the population variance, which measures how much the variables of a population deviate from their expected value (i.e., the population mean). The standard deviation is calculated by

$$\sigma = \sqrt{\sum_i^N (X_i - \mu)^2/(N-1)} \tag{3.3}$$

where X_i, μ, and N are defined as before.

The standard deviation is estimated from a sample by the variable S and is calculated by

$$S = \sqrt{\sum_{i}^{n} (x_i - \bar{X})^2/(n-1)} \tag{3.4}$$

where x_i, $\bar{X}$, and n are defined as before.

A final key parameter in statistics is the standard error $(\sigma_{\bar{X}})$, which helps characterize the spread in the distribution of the possible means, which could be derived from a single *sample* of a population (rather than the entire population itself). According to the central limit theorem, the standard error,[†] which is the square root of the variance of the population of estimated means, is a valuable parameter because it allows us to estimate our confidence in our estimate of the mean. There is a population of possible estimated means (instead of just one) because there are many possible values of $\bar{X}$, each resulting from a different selection of samples of size n from the population.

The standard error is calculated by

$$\sigma_{\bar{X}} = \sigma/\sqrt{n} \tag{3.5}$$

where σ and n are defined as before.

The standard error is estimated from a sample by the variable $S_{\bar{X}}$ and is calculated by

$$S_{\bar{X}} = S/\sqrt{n} \tag{3.6}$$

where S and n are defined as before.

Estimating the Dispersal of Variables

Assuming that the frequency of the values of variables is normally distributed about the mean as depicted in Figure 3.3, the normal or Gaussian distribution can be used to approximate the distribution of population variables. Additionally, the standard normal distribution can be used to estimate an interval of X_i at specified probabilities within which the mean of the population (μ) will fall. To do so, the distribution of the population variables must be standardized by transforming the scale of the standard normal distribution to the scale of the population being studied.

Figure 3.3 illustrates the shape of the normal distribution. All normal distributions are shaped like the curve in Figure 3.3, with the area underneath the curve equal to 1. The *standard* normal distribution represents the distribution of the standard normal

[†] The term *standard error* is unfortunately used to denote different parameters in different professions. While most statistics texts define the standard error as the square root of the variance of the population of means $(\sigma_{\bar{X}})$, many mapping texts define the standard error as the square root of the variance of the population signified by σ, which statisticians call the *standard deviation*.

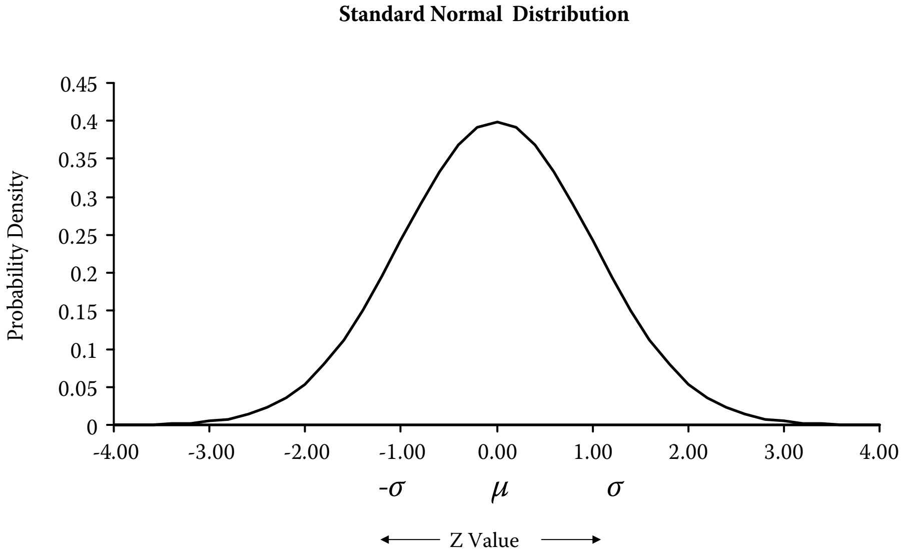


FIGURE 3.5 The standard normal distribution.

variable (Z_i) and is unique because it has a mean of 0 and a standard deviation of 1, as illustrated in Figure 3.5.

The standard normal variable, Z_i, is defined as

$$Z_i = (X_i - \mu)/\sigma \tag{3.7}$$

where

Z_i is the value from the x-axis of the standard normal distribution at the ith probability level,

X_i is the corresponding value from the x-axis of the population of interest, and

μ and σ are defined as before.

Using algebra, we can transform the x-axis scale of the normal distribution to that of our population by solving for values of X_i such that

$$Z_i * \sigma = (X_i - \mu) \text{ and}$$

$$X_i = \mathrm{Z}_i * \sigma - \mu \tag{3.8}$$

With this formula, we could transform every Z_i value of the standard normal distribution into an X_i value of our population. More commonly, the transformation is used to calculate an interval at a specified probability level within which values of X_i will occur such that:

$$X_i < \mu < X_\iota, \text{ or using Equation 3.8, the interval becomes}$$

$$[\mu - Z_i * \sigma, \mu + Z_i * \sigma] \tag{3.9}$$

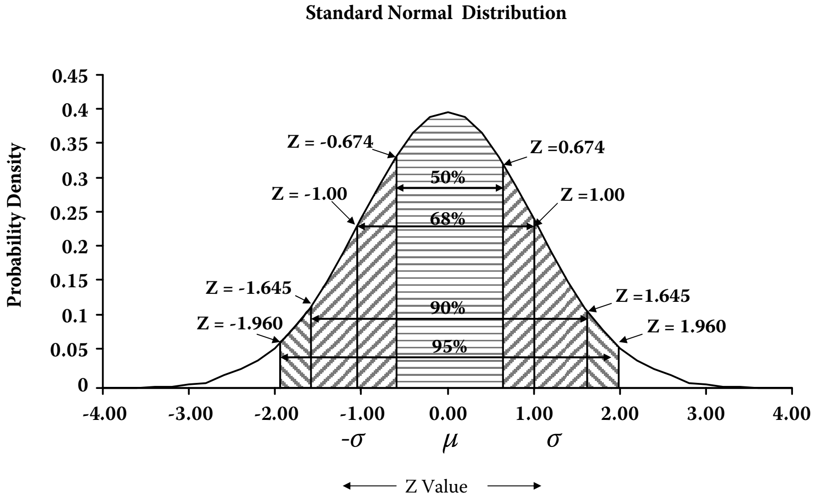


FIGURE 3.6 Probability areas and corresponding Z_i values of the standard normal distribution.

Transforming the values of the normal standard distribution into that of our population of interest requires that the distributions of variables of the normal distribution and of the population of interest be almost identical. This is not a big leap of faith, as the normal distribution characterizes a multitude of natural phenomena ranging from organism population dynamics to human polling behavior. However, it is always important to fully understand whether or not the population you are studying is actually normally distributed or not. Equation 3.9 expresses the dispersal around the mean of the variable X_i at the stipulated probability level if and only if, the population of X_i's is normally distributed. Figure 3.6 illustrates the portions of the normal distribution and the corresponding Z_i values that match various levels of probability.

To summarize, determining the interval at a specific probability within which the mean (μ) of our population of interest will fall requires simply:

1. Assuming that the population is normally distributed,
2. Looking up the Z_i value for the specified probability level in a standard normal table (which may be found in the back of any statistics text or by searching on the Internet),
3. Multiplying the Z_i value times the standard deviation (σ) of the population of interest, and
4. Adding and subtracting the resulting $Z_i*\sigma$ value from the mean (μ).

For example, the interval within which 90% of the values of a normally distributed population with a mean (μ) of 20 and a standard deviation (σ) of 4 can be determined by:

1. Looking up the Z_i value for 90% probability in a Z table or from Figure 3.6. At 90% probability, Z_i is equal to 1.645.

2. Calculating $Z_i * \sigma$ by multiplying 1.645 times the standard deviation of 4, which equals 6.58.
3. Adding and subtracting 6.58 from the mean to determine the interval at 90% probability:

$$= 20 - 6.58, \quad 20 + 6.58$$

which results in the interval ranging from 13.42 to 26.58.

Therefore, we know that 90% of the values of our population will fall within a range between 13.42 and 26.58. Figure 3.7 shows how the *x*-axis scale of the standard normal distribution transforms to that of our example.

Usually we do not know the true mean and the standard deviation of the population. However, because $\bar{X}$ and S are unbiased estimators of μ and σ, we can use the sample estimates of the mean ($\bar{X}$) and the standard deviation (S) to calculate the interval, which becomes

$$\bar{X} - Z_i * S < \mu < \bar{X} + Z_i * S \tag{3.10}$$

Estimating the Reliability of the Estimate of the Mean

Often we want to understand how reliable our estimate of the mean is. To do so requires using the sample data to develop a "confidence interval" around the estimated mean. Estimating the confidence interval once again employs the standard normal variable (Z_i), which, for the population of sample means is defined as

$$Z_i = (\bar{X}_i - \mu)/\sigma_{\bar{X}} \tag{3.11}$$

and can be estimated by

$$Z_i = (\bar{X}_i - \bar{X})/S_{\bar{X}} \tag{3.12}$$

where $\bar{X}_i$ is the value from the population of estimated means that corresponds to the Z_i value from the normal distribution and $\mu, \sigma, \bar{X},$ and $S_{\bar{X}}$ are defined as before.

The confidence interval on the estimate of the mean is calculated as

$$\bar{X} - Z_i(S_{\bar{X}}) < \mu < \bar{X} + Z_i(S_{\bar{X}}) \tag{3.13}$$

when sample sizes are large, and by

$$\bar{X} - t_i(S_{\bar{X}}) < \mu < \bar{X} + t_i(S_{\bar{X}}) \tag{3.14}$$

when sample sizes are small, where t_i is the value from the *x*-axis of the Student's *t* distribution[†] at the *i*th probability level.

[†] The Student's *t* distribution should be used instead of the *Z* distribution when sample sizes are below 30. Similar to the *Z* distribution, the values of the Student's *t* distribution can be found at the back of any statistics text or on the Web.

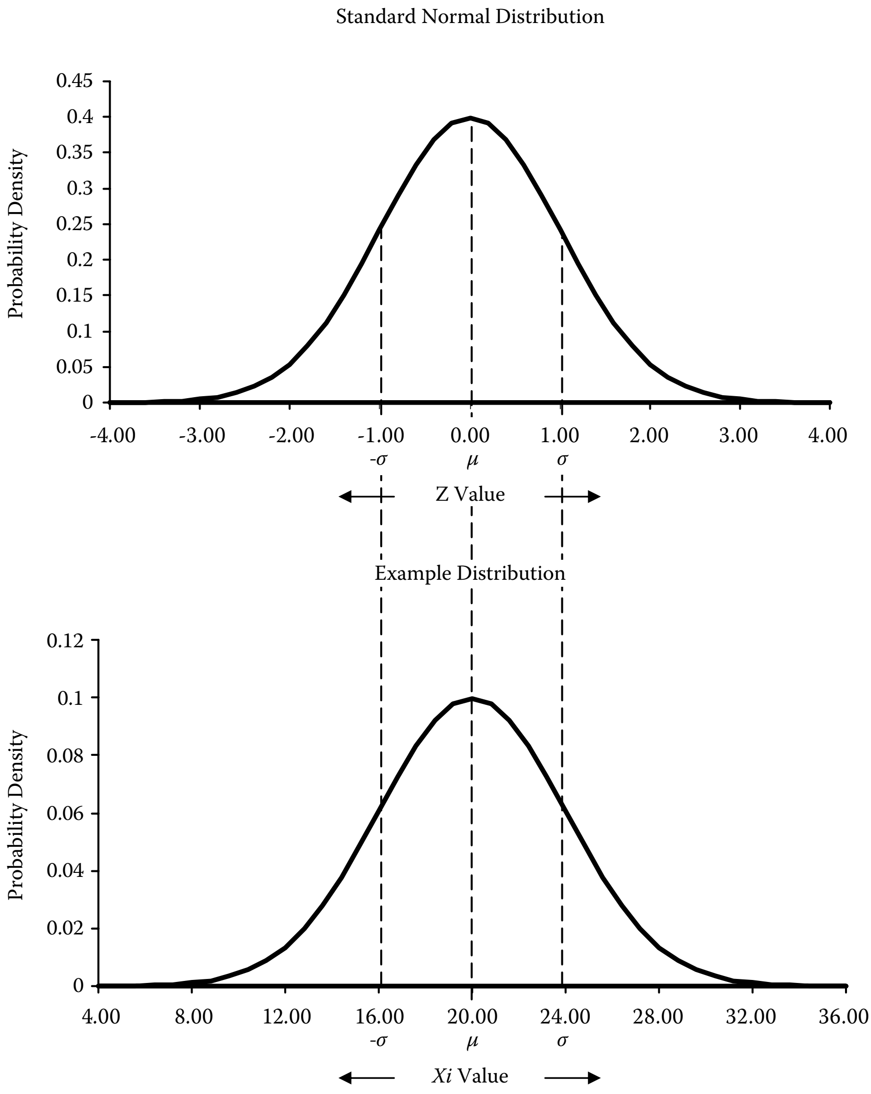


FIGURE 3.7 Transformation of the standard normal distribution x-axis scale to the x-axis scale of the example.

There is a subtle but very important distinction between estimating the *dispersal* interval of population values around the mean at a specified probability as calculated in Equation 3.10, and the calculation of a *confidence interval* around the estimate of the mean as shown in Equations 3.13 and 3.14. The former expresses the dispersal of a population of values around the mean at specified probabilities. The latter expresses the reliability of the estimate of the mean at specified probabilities.

An interesting aspect of the population of sample means is that it will be normally distributed even when the underlying population of variables is not. This important concept, which is derived from the *central limit theorem*, tells us that when the sample size is large enough and the samples are chosen without bias, then the distribution of the population of means *will* be normally distributed even when the population distribution from which the samples were chosen to estimate the means *is not* normally distributed. The central limit theorem permits us to state our confidence in our estimate of $\bar{X}$ regardless of the distribution of X_i, which allows us to rely on, rather than assume, that the shape of the standard normal distribution is the shape of the distribution of sample means.

Statistics in Positional Accuracy Assessment

In positional accuracy assessment, the NSSDA-specified and accepted measure of accuracy is the mean square root of squared differences between the map and the reference points. This term is called the root-mean-square error, or RMSE. RMSE is estimated from a sample of map and reference points. The mean square root of the square of the differences is used instead of the mean of the simple arithmetic differences to compensate for the fact that the errors can have both positive and negative values. An alternative estimator that would also deal with negative values would be to take the absolute value of the arithmetic mean of the errors.

The estimate of the standard deviation (S) of the squared differences is also an important parameter in many positional accuracy assessment standards (Greenwalt and Schultz, 1962, 1968; ASPRS, 1989). This chapter also suggests the use of the estimated standard error of the RMSE (S_{RMSE}) to build a confidence interval around the estimate of RMSE.

All positional accuracy parameters are estimated by comparing reference coordinates or elevations to the map or image coordinates or elevations of the data set being assessed at each sample location. Unfortunately, positional accuracy standards often confuse the estimate of the mean error (RMSE) with the estimate of the standard deviation (S), and the estimate of the standard deviation with that of the standard error (S_{RMSE}). In addition, terms that are commonly used in other disciplines are often applied slightly differently in positional accuracy assessment, which also adds to the confusion.

For example, statisticians and mapping professionals use the term "root-mean-square error," or RMSE, to refer to different error population parameters. "Error" in positional accuracy is the difference between the reference location and that of the geospatial data set being assessed. It is a measure of accuracy and measures the magnitude of an inaccurately estimated or calculated value. "Error" in statistics is the difference between an observed value and its statistical estimator, and is a measure of precision.

As a result, the equations for RMSE differ between the two applications:

- Mapping professionals define RMSE as the square root of the mean squared differences between the sample reference sample locations and the corresponding locations on the geospatial data set being assessed. The equation for calculating RMSE in mapping applications is

$$\mathrm{RMSE} = \sqrt{\sum_{i}^{n} (e_i)^2/n} \tag{3.15}$$

where

$$e_i = e_{ri} - e_{mi} \tag{3.16}$$

and

e_{ri} equals the reference elevation at the *i*th sample point,
e_{mi} equals the map or image elevation at the *i*th sample point, and
n is the number of samples.

- Statisticians define RMSE as the square root of the mean squared differences between a statistical estimator of a parameter and the value actually observed. The equation for calculating RMSE in statistics is

$$\sqrt{\sum_{i}^{n} (e_i - \bar{e})^2/(n-1)} \tag{3.17}$$

where $\bar{e}$ is the unbiased estimator of the mean or average difference, which is calculated by

$$\bar{e} = \sqrt{\sum_{i}^{n} (e_i)^2/n} \tag{3.18}$$

The only time that the mapping and the statistical RMSE are equal to one another is when the average error $(\bar{e})$ equals zero, which is a condition that is rare and which should always be tested for by calculating $(\bar{e})$ and determining if it is significantly different from zero. Unfortunately, many mapping standards make the assumption that $(\bar{e})$ equals zero, which results in misleading characterizations of map error.

This use of the same term to mean different things has, understandably, led to much confusion. The RMSE used in mapping is the square root of the estimated mean of the squares of the geospatial data set's positional errors. The RMSE used in statistics is the square root of the variance of the errors and characterizes how errors differ from the mean error. Because mapping professionals rely on statistics to characterize the frequency distribution of positional errors, it is critical that confusion be minimized and that statistics be properly applied to mapping applications.

This chapter attempts to eliminate the confusion surrounding positional accuracy assessment analysis by:

1. Detailing what equations should be used to characterize positional accuracy, and
2. Correcting the mistakes in currently used standards.

First, one-dimensional vertical accuracy assessment is discussed. Next, two-dimensional horizontal accuracy assessment is reviewed.

Vertical Accuracy

Statistical Parameters

The mean vertical positional error (μ_v)[†] is depicted in mapping applications by the vertical root-mean-square error (RMSE_v) of the sample of vertical errors (e_{vi}) and is estimated by

$$\text{RMSE}_v = \sqrt{\sum_i^n (e_{vi})^2/n} \tag{3.19}$$

where

$$e_{vi} = v_{ri} - v_{mi} \text{ and} \tag{3.20}$$

v_{ri} equals the reference elevation at the *i*th sample point,
v_{mi} equals the map or image elevation at the *i*th sample point, and
n is the number of samples.

An alternative estimator is the arithmetic mean of the absolute error values and is calculated by

$$|\bar{e}_v| = \sum_i^n |e_{vi}|/n \tag{3.21}$$

The standard deviation (σ_v) of the population of vertical errors is estimated by

$$S_v = \sqrt{\sum_i^n (e_{vi} - \text{RMSE}_v)^2/(n-1)} \tag{3.22}$$

and the standard error of estimates of RMSE_v is estimated by

$$S_{\text{RMSE}_v} = S_v/\sqrt{n} \tag{3.23}$$

Assuming that the vertical errors are normally distributed, the estimated interval of errors at a specific probability can be expressed as

$$\text{RMSE}_v \pm Z_i(S_v) \tag{3.24}$$

At a 95% probability level, the equation becomes

$$\text{RMSE}_v \pm 1.96\,(S_v) \tag{3.25}$$

If RMSE equals zero, then the factor $\pm Z_i(S_v)$ will express the interval of error at the probability level specified by the Z_i variable and the interval at 95% will equal ±1.96 (S_v). A 90% interval, with RMSE_v to zero, will be 1.645 (S_v).

[†] Some mapping texts use the subscript "*z*" to denote vertical error. Because this text (and most statistics texts) uses the variable Z_i to denote the standard normal variable, we use the subscript "*v*" to denote vertical error.

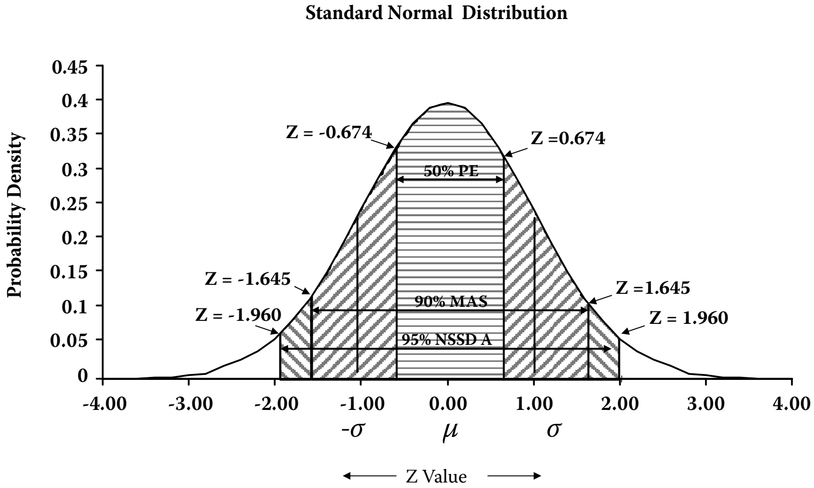


FIGURE 3.8 Areas and Z_i values of the standard normal distribution corresponding to the probability levels of various map accuracy standards.

The *Principles of Error Theory and Cartographic Applications* report is the first report to propose use of the $Z_i(S_v)$ interval as a standard in estimating positional accuracy (Greenwalt and Schultz, 1962, 1968). The report relies on estimating the interval $Z_i\,(S_v)$ at various probability levels where it is referred to as the probable error at 50% and the map accuracy standard at 90% (Greenwalt and Schultz, 1962, 1968). Figure 3.8 illustrates the portions of the normal distribution that correspond to the probable error at 50%, the map accuracy standard at 90%, and the NSSDA standard at 95%.

The Greenwalt and Schultz (1962, 1968) equations estimate the interval of errors around the mean error at different probability levels. Derived from the military science of ballistics, the equations result in an estimate of the probable dispersal of error around the mean error (RMSE_z) at specified probabilities.[†] Use of the interval as an accuracy standard was confirmed by subsequent Defense Mapping Agency reports (DMA, 1991), and is used in ASPRS's large-scale mapping standards (ASPRS, 1989), as well as NSSDA (FGDC, 1998).

Note that the $Z_i(S_v)$ interval is not a confidence interval around the estimate of RMSE_v, nor is it the range of expected errors at a given probability. Rather, it is an estimate of the maximum interval of error that will exist at a specified probability assuming that mean error equals zero and the errors are normally distributed. Unfortunately, spatial errors are often biased and interrelated, bringing the assumption of normality into question.

[†] Greenwalt and Schultz (1962, 1968) define the vertical map accuracy standard as "the size of error which 90% of the elevations will not exceed." However, the interval $Z_i\,(S_z)$ meets this definition only when RMSE_v equals zero. The estimated size of elevation errors which will not be exceeded at a probability level specified by Z_i is $\text{RMSE}_v \pm Z_i\,(S_z)$.

To measure the reliability (or our confidence in) the estimate of RMSE$_v$, we calculate a confidence interval around RMSE$_v$, by converting the general confidence interval equation (Equation 3.13)

$$\bar{X} - Z_i(S_{\bar{X}}) < \mu < \bar{X} + Z_i(S_{\bar{X}})$$

to our mapping application terminology such that

$$\text{RMSE}_v - Z_i(S_{\text{RMSE}}) < \mu < \text{RMSE}_v + Z_i(S_{\text{RMSE}}) \tag{3.26}$$

for large sample sizes, and

$$\text{RMSE}_v - t_i(S_{\text{RMSE}}) < \mu < \text{RMSE}_v + t_i(S_{\text{RMSE}}) \tag{3.27}$$

for small samples sizes, where all variables are defined as before.

In most situations, if we have more than 30 samples, at a 95% confidence level the equation becomes:

$$\text{RMSE}_v - 1.96(S_{\text{RMSE}}) < \mu < \text{RMSE}_v + 1.96(S_{\text{RMSE}}). \tag{3.28}$$

That means that we are 95% certain that the interval contains the true, but unknown, population average error.

Table 3.1 displays the map and reference elevations for a hypothetical digital elevation data set. The errors at each sample point are calculated as well as the estimated RMSE$_v$, $|\bar{e}_v|$, S_v, S_{RMSE_v}, NSSDA accuracy statistic, MAS, and a 95% confidence interval around the estimate of RMSE$_v$. All the equations for these calculations are shown in Table 3.2.

NSSDA

The NSSDA (FGDC, 1998) requires that accuracy be reported at the 95% level, which is defined by NSSDA as meaning "that 95% of the positions in the data set will have an error with respect to true ground positions that is equal to or smaller than the reported accuracy." NSSDA references the Greenwalt and Schultz (1962, 1968) equations, but *mistakenly* stipulates that the vertical accuracy interval at the 95% probability be computed by multiplying the appropriate Z_i statistic times the *estimated mean (*RMSE$_v$*)* instead of the *estimated standard deviation* (S_v). The resulting NSSDA equation for calculating the NSSDA vertical accuracy statistic is[†]

$$\textit{NSSDA Vertical Accuracy}_v = 1.96\ (\text{RMSE}_v) \tag{3.29}$$

rather than the Greenwalt and Schultz (1962, 1968) equation, which is

$$\textit{Accuracy}_v = 1.96\ (S_v). \tag{3.30}$$

Estimating the interval within which 95% of the errors will fall requires assuming that our errors are normally distributed and converting the scale of the standard

[†] 1.96 is the standard normal distribution Z statistic (the value from the x-axis of the standard normal distribution) for an interval with a probability of 95%.

TABLE 3.1
Vertical Accuracy Example

Point ID	v_{ri} Reference	v_{mi} Map	Error = e_{vi} = Reference – Map ($v_{ri} - v_{mi}$) = e_{vi}	Error Squared $(v_{ri} - v_{mi})^2 = e_{vi}^2$	Absolute Error	(Absolute e_{vi} – $\text{RMSE}_v)^2$
1202	2362.2075	2361.3100	–0.8975	0.8055	0.8975	0.3502
1230	2421.5855	2420.9000	–0.6855	0.4699	0.6855	0.1442
1229	2701.6110	2701.1700	–0.4410	0.1945	0.4410	0.0183
125	705.3117	705.0190	–0.2927	0.0857	0.2927	0.0002
316	1009.2344	1009.0300	–0.2044	0.0418	0.2044	0.0103
369	920.0574	919.8740	–0.1834	0.0336	0.1834	0.0150
292	586.3659	586.2400	–0.1259	0.0159	0.1259	0.0323
143	761.4684	761.3910	–0.0774	0.0060	0.0774	0.0521
132	712.1791	712.1320	–0.0471	0.0022	0.0471	0.0669
1005	1190.4284	1190.4000	–0.0284	0.0008	0.0284	0.0769
274	809.0433	809.0500	0.0067	0.0000	0.0067	0.0894
112	387.2611	387.2960	0.0349	0.0012	0.0349	0.0734
339	965.6910	965.7480	0.0570	0.0032	0.0570	0.0619
130	1059.1342	1059.2300	0.0958	0.0092	0.0958	0.0441
113	428.7700	428.9630	0.1930	0.0372	0.1930	0.0127
122	1012.0117	1012.3100	0.2983	0.0890	0.2983	0.0001
136	308.7100	309.0110	0.3010	0.0906	0.3010	0.0000
104	529.4721	529.8260	0.3539	0.1252	0.3539	0.0023
101	427.1653	427.5840	0.4187	0.1753	0.4187	0.0128
1221	2690.1380	2689.5200	–0.6180	0.3819	0.6180	0.0975
129	483.4317	483.0480	–0.3837	0.1472	0.3837	0.0061
128	492.7014	492.5810	–0.1204	0.0145	0.1204	0.0344
114	799.9452	799.8560	–0.0892	0.0080	0.0892	0.0469
367	1273.0857	1273.0300	–0.0557	0.0031	0.0557	0.0625
108	1235.0128	1235.0300	0.0172	0.0003	0.0172	0.0833
325	1040.9078	1040.9700	0.0622	0.0039	0.0622	0.0593
250	211.4375	211.5230	0.0855	0.0073	0.0855	0.0485
1010	1189.4876	1189.6200	0.1324	0.0175	0.1324	0.0300
Sum				**2.77**	**6.31**	**1.60**

normal distribution to that of the error population using the estimated standard deviation as detailed in Equation 3.10 below.

$$[\bar{X} - Z_i(S), \quad \bar{X} + Z_i(S)]$$

If the mean error is equal to zero, then the interval becomes the Greenwalt and Schultz (1962, 1968) statistic of Z_i (S). While the NSSDA standard is applied ubiquitously, it is valid only when RMSE$_v$ = S_v. If S_v is less than RMSE$_v$ then the NSSDA statistic will overestimate the error interval, and if S_v is greater than RMSE$_v$, the NSSDA statistic will underestimate the error interval.

TABLE 3.2
Vertical Accuracy Example Equations and Statistics

Definition	Equation	Value
Estimated root-mean-square error of the population of vertical errors	$\mathrm{RSME}_v = \sqrt{\Sigma_i^n (e_{vi})^2/n}$	0.320
Estimated absolute arithmetic mean of the population of vertical errors	$\lvert\bar{e}_v\rvert = \Sigma_i^n \lvert e_{vi}\rvert/n$	0.234
Estimated variance of the population of vertical errors	$S_v^2 = \Sigma_1^n (e_{vi} - \mathrm{RMSE}_v)/(n-1)$	0.059
Estimated standard deviation of the population of vertical errors	$S_v = \sqrt{\Sigma_i^n (e_{vi} - \mathrm{RMSE}_v)^2/(n-1)}$	0.244
Estimated standard error of the population of RMSEs	$S_{\mathrm{RMSE}_v} = S_v/\sqrt{n}$	0.047
Greenwalt and Schultz MAS standard normal interval of e_{vi} at 90% probability	$1.645 * S_v$	0.401
Greenwalt and Schultz standard normal interval of e_{vi} at 95% probability	$1.96 * S_v$	0.478
NSSDA statistic	$1.96 * \mathrm{RMSE}_v$	0.628
95% confidence interval around the estimate of RMSE	$\mathrm{RMSE} \pm 1.96 * S_{\mathrm{RMSE}}$	RMSE ± 0.092 which results in a range from 0.228 to 0.412

FEMA, ASPRS LIDAR, and NDEP Standards

The most recent standards and guidelines for elevation data all require that the landscape be stratified by vegetative cover class, that a minimum of 60 samples be chosen, and that the different types of cover classes be evaluated using different methods. While open terrain is evaluated using NSSDA equations, other cover classes are evaluated using the 95th percentile method, as is the consolidated vertical accuracy, which combines open terrain samples with other ground cover types.

Figure 3.9 shows an example of positional accuracy analysis with each sample point sorted by ground cover type and charted by the error measured between each points reference and map data. Both the NSSDA statistic at ±0.82 ft, and the consolidated vertical accuracy at ±0.91 ft are displayed.

Horizontal Accuracy

Statistical Parameters

Horizontal accuracy is more complex than vertical accuracy because the error is distributed in two dimensions (both the x and y dimensions), requiring the calculation of the radial error and reliance on the bivariate normal distribution to estimate

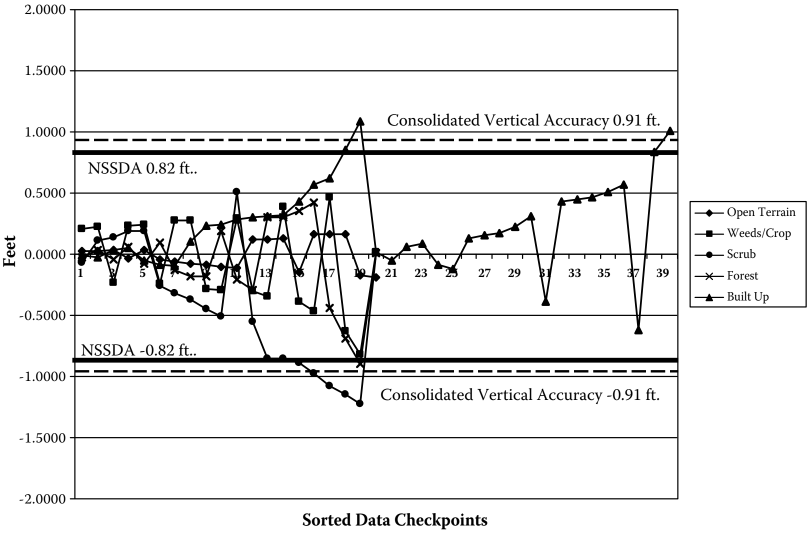


FIGURE 3.9 Application of 95th percentile criteria versus NSSDA for reporting positional accuracy. (Courtesy of Dewberry.)

probabilities. To calculate the horizontal root-mean-square error (RMSE$_h$),[†] first, the x-coordinate from the reference data is recorded followed by the x-coordinate from the spatial data set being assessed. Then the difference between the two locations is computed, followed by a squaring of this difference. The same process is used for the y-coordinate. Each test point then has an associated error distance, e_i, defined by the following equation:

$$e_h = \sqrt{(x_{ri} - x_{mi})^2 + (y_{ri} - y_{mi})^2} \tag{3.31a}$$

and

$$e_h^2 = (x_{ri} - x_{mi})^2 + (y_{ri} - y_{mi})^2 \tag{3.31b}$$

where x_r and y_r are the reference coordinates and x_m and y_m are the map or image coordinates for the ith sample point in the spatial data set being assessed.

The equation for the average horizontal error or horizontal root-mean-square error (RMSE$_h$) is calculated from the errors of the individual test sample points using the following equation:

$$\mathrm{RMSE}_h = \sqrt{\sum_i^n ((x_{ri} - x_{mi})^2 + (y_{ri} - y_{mi})^2)/n} = \sqrt{\left(\mathrm{RMSE}_x^2 + \mathrm{RMSE}_y^2\right)}/n \tag{3.32}$$

[†] Greenwalt and Schultz (1962, 1968) refer to horizontal error as *circular error,* which they designate with the subscript "c." NSSDA (FGDC, 1998) refers to horizontal error as radial, designated by the subscript "r." Because the errors are usually elliptical rather than circular, and because we have already designated the subscript "r" to indicate a reference value of an accuracy assessment sample, this text uses the subscript "h" to designate horizontal error.

or

$$\mathrm{RMSE}_h = \sqrt{\frac{\sum_i^n e_{hi}^2}{n}} \tag{3.33}$$

where e_{hi} is defined in the preceding equation and n is the number of test sample points.

An alternative estimator is the arithmetic mean of the absolute error values and is calculated by

$$|\bar{e}| = \sum_1^n |e_{hi}|/n \tag{3.34}$$

Once RMSE$_h$ has been estimated, the standard deviation (S_h) of the population of horizontal errors can also be approximated from the samples by calculating the average standard deviation, using the Greenwalt and Schultz (1962, 1968) equation below:

$$S_h = (S_x + S_y)/2 \tag{3.35}$$

where

$$S_x = \sqrt{\sum_i^n ((x_{ri} - x_{mi}) - \mathrm{RMSE}_x)^2/n - 1} \tag{3.36}$$

and

$$S_y = \sqrt{\sum_i^n ((y_{ri} - y_{mi}) - \mathrm{RMSE}_y)^2/n - 1} \tag{3.37}$$

The estimated standard error of the population of RMSE$_h$'s is

$$S_{\mathrm{RMSE}_h} = S_h/\sqrt{n} \tag{3.38}$$

Assuming that the errors are normally distributed, the estimated interval of errors at a specified probability can be expressed as

$$\mathrm{RMSE}_h \pm Z_i(S_h) \tag{3.39}$$

If RMSE$_h$ is equal to zero, then the error interval becomes the Greenwalt and Schultz (1962, 1968) specified Z_i (S).

The confidence interval around the estimate of the mean horizontal error can be calculated as follows:

$$\mathrm{RMSE}_h - Z_i(S_{\mathrm{RMSE}_h}) < \mu < \mathrm{RMSE}_h + Z_i(S_{\mathrm{RMSE}_h}) \tag{3.40}$$

for large sample sizes, and

$$\mathrm{RMSE}_h - t_i(S_{\mathrm{RMSE}_h}) < \mu < \mathrm{RMSE}_h + t_i(S_{\mathrm{RMSE}_h}) \tag{3.41}$$

for small samples sizes.

Because horizontal error is measured in two dimensions, the *bivariate* standard normal distribution must be used to characterize the distribution of errors. Figure 3.10

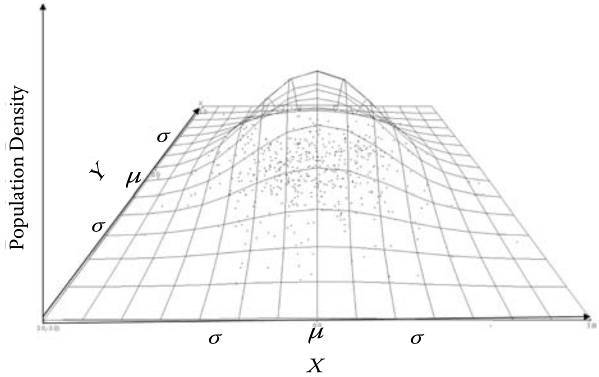


FIGURE 3.10 Three-dimensional representation of the standard normal bivariate distribution.

provides a three-dimensional illustration of the bivariate normal distribution. Figure 3.11 is an overhead view of the bivariate standard normal probability distribution for the commonly used map standards of the circular error probable (CEP) at 50%, the circular map accuracy standard (CMAS) at 90%, and NSSDA at 95%.

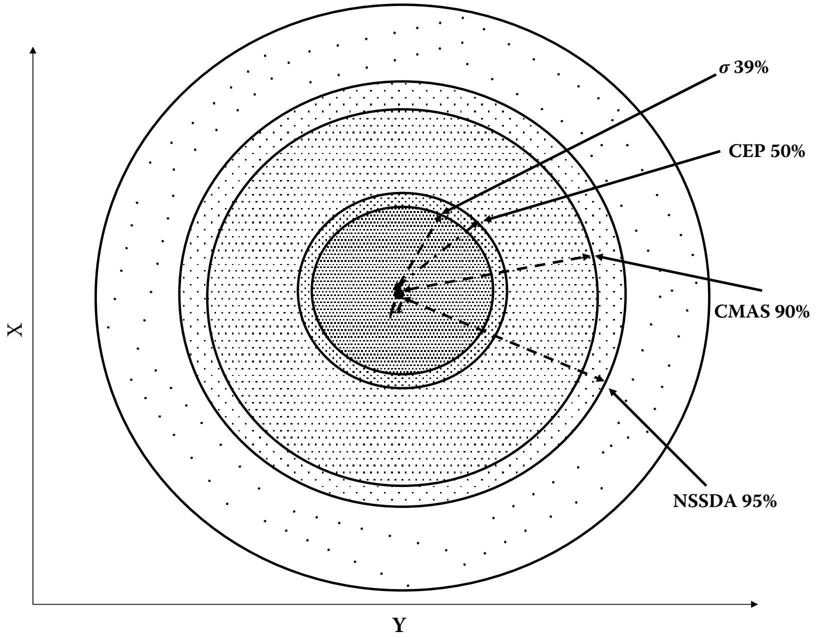


FIGURE 3.11 Two-dimensional representation of the normal standard bivariate or circular distribution with the probabilities of common horizontal map standards (From Greenwalt, C. and M. Schultz. 1962, 1968. *Principles of Error Theory and Cartographic Applications.* United States Air Force. Aeronautical Chart and Information Center. ACIC Technical Report Number 96. St. Louis, MO. 60 pages plus appendices).

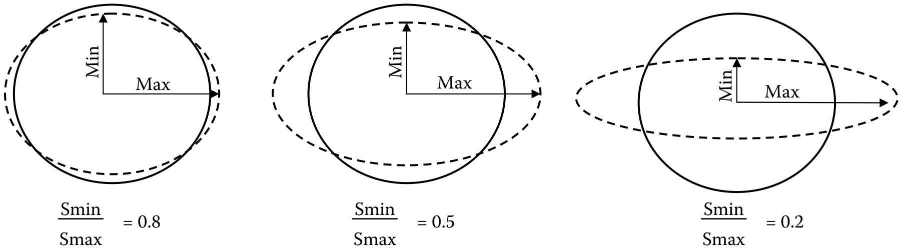


FIGURE 3.12 Comparison of circular to elliptical distributions for various ratios of S_{min}/S_{max} (From Defense Mapping Agency. 1991. *Error Theory as Applied to Mapping, Charting, and Geodesy.* Defense Mapping Agency Technical Report 8400.1. Fairfax, Virginia).

Relying on the bivariate standard normal distribution to characterize the distribution of horizontal errors requires that we assume that the horizontal errors are distributed in a circle with S_x equal to S_y. We can test for circularity by calculating the ratio of the S_{min} to S_{max} (where S_{min} is the lower of S_x or S_y, and S_{max} is the larger of S_x or S_y). Figure 3.12 shows how differences in S_x and S_y affect the shape of the distribution of errors. If the ratio of S_{min} to S_{max} is 0.2 or greater, Greenwalt and Schultz (1962, 1968) state that the circular distribution can be assumed.

As with vertical accuracy, many standards rely on the Z_i (S_h) as the statistic to estimate horizontal accuracy (DMA, 1991; NSSDA, 1998). The statistic estimates the maximum interval of error on either side of RMSE_h that will exist at a specified probability. The bivariate standard normal distribution Z_i statistic at 95% probability is 2.4477 (Greenwalt and Schultz, 1962, 1968) and the resulting interval of errors at 95% probability is

$$2.4477\ ((S_x + S_y)/2) \tag{3.42}$$

or

$$2.4477\ S_h \tag{3.43}$$

The interval of errors within which 95% of the errors will occur (assuming the errors are normally distributed) is

$$[\text{RMSE}_h - 2.447 * S_h,\ \text{RMSE}_h - 2.447 * S_h] \tag{3.44}$$

If RMSE_h is equal to zero, the estimated interval reduces to the Greenwalt and Schultz (1962, 1968) and ASPRS (1989) accuracy statistic of $2.447(S_h)$.

Because the distribution of RMSE_h is one-dimensional (even though the distribution of errors is two-dimensional), a confidence interval on RMSE_h at the 95% level is expressed by

$$\text{RMSE}_h - 1.96 S_{\text{RMSE}_h} > \mu > \text{RMSE}_h - 1.96 S_{\text{RMSE}_h} \tag{3.45}$$

for large samples and

$$\text{RMSE}_h - t_{95\%,\ n-1\ degrees\ of freedom}\ S_{\text{RMSE}_h} > \mu > \text{RMSE}_h + t_{95\%,\ n-1\ degrees\ of freedom}\ S_{\text{RMSE}_h} \tag{3.46}$$

for small samples.

NSSDA

As with vertical accuracy, the NSSDA accuracy statistic incorrectly applies the RMSE_h rather than S_h to calculate the NSSDA accuracy statistic under two different conditions described in the following text: when the RMSE_y and RMSE_x are equal and when they are not equal.

When the errors are circular. NSSDA defines errors as circular if $\mathrm{RMSE}_y = \mathrm{RMSE}_x$ (rather than when $S_x = S_y$). Under NSSDA

$$\text{if } \mathrm{RMSE}_y = \mathrm{RMSE}_x, \text{ then}$$

$$\mathrm{RMSE}_h = \sqrt{(2\mathrm{RMSE}x)^2} = \sqrt{(2\mathrm{RMSE}_y)^2}$$

$$= 1.4142 * \mathrm{RMSE}_x = 1.4142 * \mathrm{RMSE}_y \tag{3.47}$$

Applying the circular error normal distribution Z statistic at 95% probability of 2.4477 results in

NSSDA Horizontal Accuracy = 2.4477 RMSE_h or

$$= 2.4477 * \mathrm{RMSE}_h/1.4142$$

$$= 1.7308 * \mathrm{RMSE}_h \tag{3.48}$$

Most organizations use this simplified equation regardless of whether the errors are distributed circularly or not.[†] However, as with elevational accuracy, the NSSDA horizontal accuracy value has no statistical basis. The population parameter that should be used to determine the interval of error at a specific probability level is the standard deviation of the horizontal errors (S_h), and not RMSE_h (Ager, 2004).

When the errors are not circular. If $\mathrm{RMSE}_y \neq \mathrm{RMSE}_x$, then the NSSDA stipulates that the NSSDA accuracy statistic is

$$= 2.4477 \; ((\mathrm{RMSE}_x + \mathrm{RMSE}_y)/2) \tag{3.49}$$

Tables 3.3 and 3.4 present sample reference and map coordinates from our earlier example and calculate the RMSE_h, S_h, S_{RMSE_h}, $Zi * S$ at 95% probability, as well as the circular map accuracy standard (CMAS) accuracy interval at 90%, the NSSDA statistic, and a 95% confidence interval around RMSE_h.

There is one final, very important issue that must be understood when implementing positional accuracy. A different RMSE value is often calculated as part of the spatial data set registration process. We will call this RMSE_{reg}. The calculation of RMSE_{reg} during the registration process is a test of the goodness of fit of the registered data set to its control points. Because of its lack of independence from the data set being assessed, RMSE_{reg} is not a valid measure of positional accuracy, and will almost always be lower (i.e., better) than RMSE_z or RMSE_h. Independent positional

[†] Circularity is defined by Greenwalt and Schultz as S_{min}/S_{max} greater than or equal to 0.2. However, NSSDA restricts application of the circular distribution to those situations in which S_{min}/S_{max} is between 0.6 and 1.0.

TABLE 3.3
Horizontal Accuracy Example

Point ID	x_{ri} Reference	x_{mi} Map	Error in x Dimension = Reference − Map $(x_{ri} - x_{mi}) = e_{xi}$	Absolute Error $\lvert e_x \rvert$	Error in x Dimension Squared $(x_{ri} - x_{mi})^2 = e_{xi}^2$	(Absolute e_{xi} − RMSE$_x$)^2	y_{ri} Reference	y_{mi} (Map)	Error in y Dimension = Reference − Map $(y_{ri} - y_{mi}) = e_{yi}$	Absolute Error $\lvert e_y \rvert$	Error in y Dimension Squared $(y_{ri} - y_{mi})^2 = e_{yi}^2$	(Absolute y_i − RMSE$_y$)^2	Sum of Squared Errors $e_{xi}^2 + e_{yi}^2$	Sum of Absolute Error
107	6463928.4275	6463928.2891	0.1384	0.1384	0.0192	0.0388	1740487.9905	1740488.2089	−0.2184	0.2184	0.0477	0.0912	0.0669	0.3569
108	6478942.9446	6478942.9707	−0.0261	0.0261	0.0007	0.0956	1757945.7986	1757945.5996	0.1991	0.1991	0.0396	0.1033	0.0403	0.2251
110	6498179.1383	6498179.2172	−0.0789	0.0789	0.0062	0.0657	1736983.2778	1736983.7799	−0.5021	0.5021	0.2521	0.0003	0.2583	0.5810
111	6500864.5792	6500866.2526	−1.6734	1.6734	2.8004	1.7906	1758833.2498	1758830.8834	2.3664	2.3664	5.5999	3.4075	8.4003	4.0398
116	6527762.0733	6527762.1410	−0.0677	0.0677	0.0046	0.0716	1731210.4027	1731210.7259	−0.3232	0.3232	0.1045	0.0389	0.1091	0.3910
117	6539890.0536	6539890.2650	−0.2113	0.2113	0.0447	0.0154	1755842.1176	1755841.9103	0.2073	0.2073	0.0430	0.0981	0.0876	0.4186
122	6452053.8265	6452053.8601	−0.0336	0.0336	0.0011	0.0910	1728034.3838	1728034.6916	−0.3078	0.3078	0.0948	0.0452	0.0959	0.3414
123	6435447.0261	6435446.7694	0.2567	0.2567	0.0659	0.0062	1737489.6870	1737489.9830	−0.2960	0.2960	0.0876	0.0504	0.1535	0.5527
124	6445012.8528	6445012.7143	0.1385	0.1385	0.0192	0.0387	1757524.8057	1757524.7919	0.0138	0.0138	0.0002	0.2567	0.0194	0.1524
206	6523662.6628	6523662.7526	−0.0898	0.0898	0.0081	0.0603	1753217.8809	1753218.0854	−0.2045	0.2045	0.0418	0.0998	0.0499	0.2944
216	6503988.9073	6503989.0881	−0.1808	0.1808	0.0327	0.0239	1728652.7232	1728653.2982	−0.5750	0.5750	0.3306	0.0030	0.3633	0.7558
222	6497217.5322	6497217.6331	−0.1009	0.1009	0.0102	0.0549	1751316.3332	1751316.3331	0.0001	0.0001	0.0000	0.2708	0.0102	0.1010
227	6532154.2998	6532154.2726	0.0272	0.0272	0.0007	0.0949	1740450.9200	1740451.2630	−0.3430	0.3430	0.1177	0.0315	0.1184	0.3702
228	6514726.6170	6514726.6231	−0.0061	0.0061	0.0000	0.1084	1748724.1696	1748724.4427	−0.2731	0.2731	0.0746	0.0612	0.0746	0.2792
229	6480333.2958	6480333.3200	−0.0242	0.0242	0.0006	0.0968	1742388.0615	1742388.2686	−0.2071	0.2071	0.0429	0.0982	0.0435	0.2313
283	6510536.2705	6510536.4059	−0.1354	0.1354	0.0183	0.0400	1757706.5081	1757706.5519	−0.0438	0.0438	0.0019	0.2272	0.0202	0.1791
200	6509030.6018	6509030.5422	0.0596	0.0596	0.0036	0.0760	1746587.3294	1746587.4407	−0.1113	0.1113	0.0124	0.1674	0.0159	0.1709
112	6502026.5461	6502026.5552	−0.0091	0.0091	0.0001	0.1064	1779378.9142	1779378.7511	0.1631	0.1631	0.0266	0.1277	0.0267	0.1723

(*Continued*)

TABLE 3.3 (CONTINUED)
Horizontal Accuracy Example

Point ID	x_{ri} Reference	x_{mi} Map	Error in x Dimension = Reference – Map $(x_{ri} - x_{mi}) = e_{xi}$	Absolute Error $\|e_x\|$	Error in x Dimension Squared $(x_{ri} - x_{mi})^2 = e_{xi}^2$	(Absolute e_{xi} $-\text{RMSE}_x)^2$	y_{ri} Reference	y_{mi} (Map)	Error in y Dimension = Reference – Map $(y_{ri} - y_{mi}) = e_{yi}$	Absolute Error $\|e_y\|$	Error in y Dimension Squared $(y_{ri} - y_{mi})^2 = e_{yi}^2$	(Absolute y_i–$\text{RMSE}_y)^2$	Sum of Squared Errors $e_{xi}^2 + e_{yi}^2$	Sum of Absolute Error
232	6509030.6018	6509030.5422	0.0596	0.0596	0.0036	0.0760	1793670.4405	1793670.5122	–0.0717	0.0717	0.0051	0.2014	0.0087	0.1313
125	6436524.8263	6436525.0783	–0.2520	0.2520	0.0635	0.0069	1782491.7176	1782490.8878	0.8298	0.8298	0.6886	0.0957	0.7521	1.0818
126	6464717.9797	6464718.1392	–0.1595	0.1595	0.0254	0.0309	1778968.4568	1778968.1097	0.3472	0.3472	0.1205	0.0300	0.1459	0.5066
128	6536017.6536	6536017.6690	–0.0154	0.0154	0.0002	0.1023	1791341.5236	1791341.5819	–0.0583	0.0583	0.0034	0.2136	0.0036	0.0737
207	6523447.4186	6523447.4146	0.0040	0.0040	0.0000	0.1098	1781813.7690	1781813.5640	0.2050	0.2050	0.0420	0.0995	0.0420	0.2090
208	6458661.4231	6458661.4229	0.0002	0.0002	0.0000	0.1123	1763512.1326	1763512.0318	0.1008	0.1008	0.0102	0.1761	0.0102	0.1010
210	6432704.3136	6432704.3469	–0.0333	0.0333	0.0011	0.0912	1797681.4156	1797681.5817	–0.1661	0.1661	0.0276	0.1256	0.0287	0.1994
214	6524150.2574	6524150.1541	0.1033	0.1033	0.0107	0.0538	1766691.1359	1766691.0337	0.1022	0.1022	0.0105	0.1749	0.0211	0.2055
221	6490159.7535	6490159.5953	0.1582	0.1582	0.0250	0.0314	1774521.0769	1774520.9524	0.1245	0.1245	0.0155	0.1568	0.0405	0.2827
223	6464915.1344	6464915.4456	–0.3112	0.3112	0.0968	0.0006	1795190.2232	1795190.4426	–0.2194	0.2194	0.0481	0.0906	0.1450	0.5306
224	6446211.1711	6446211.4834	–0.3123	0.3123	0.0975	0.0005	1776288.7842	1776289.2253	–0.4411	0.4411	0.1946	0.0063	0.2921	0.7534
226	6513283.3804	6513283.4917	–0.1113	0.1113	0.0124	0.0502	1771237.4139	1771237.6203	–0.2064	0.2064	0.0426	0.0986	0.0550	0.3177
		Sum		**4.7780**	**3.3724**	**3.5409**				**9.2277**	**8.1266**	**6.6477**	**11.4990**	**14.0057**

TABLE 3.4
Horizontal Accuracy Example Equations and Statistics

Definitions	*X* Dimension Equations	*X* Dimension Values
Estimated root-mean-square of the population of errors	$\mathrm{RMSE}_x = \sqrt{\sum_i^n (e_{xi})^2/n}$	0.3353
Estimated absolute arithmetic mean of the population of errors	$\lvert\bar{e}_x\rvert = \sum_1^n \lvert\bar{e}_{xi}\rvert/n$	0.1593
Estimated variance of the population of errors	$S_x^2 = \sum_i^n (\lvert e_{xi}\rvert - \mathrm{RMSE}_x)^2/(n-1)$	0.1221
Estimated standard deviation of the population of errors	$S_x = \sqrt{\sum_i^n (\lvert e_{xi}\rvert - \mathrm{RMSE}_x)^2/(n-1)}$	0.3494
Estimated standard deviation of the population of RMSEs	$S_{\mathrm{RMSE}_x} = \sqrt{S_x^2/n}$	0.0638
Greenwalt and Schultz CMAS standard normal (*Z)* interval of the population of errors at 90% probability	$1.645 * S_x$	0.5748
Greenwalt and Schultz standard normal (*Z)* interval of the population of errors at 95% probability	$1.96 * S_x$	0.6849
NSSDA statistic	$1.96 * \mathrm{RMSE}_x$	0.6572
Confidence interval on the estimate of RMSE_x at 95% probability	$\mathrm{RMSE}_x \pm 1.96 * S_{\mathrm{RMSE}}$	0.3353 ± 0.1250 which results in a range from 0.2102 to 0.4603

Definitions	*Y* Dimension Equations	*Y* Dimension Values
Estimated root-mean-square of the population of errors	$\mathrm{RMSE}_y = \sqrt{\sum_i^n (e_{yi})^2/n}$	0.5205
Estimated absolute arithmetic mean of the population of errors	$\lvert\bar{e}_y\rvert = \sum_1^n \lvert e_{yi}\rvert/n$	0.3076
Estimated variance of the population of errors	$S_y^2 = \sum_1^n (\lvert e_{yi}\rvert - \mathrm{RMSE}_y)^2/(n-1)$	0.2292
Estimated standard deviation of the population of errors	$S_y = \sqrt{\sum_1^n (\lvert e_{yi}\rvert - \mathrm{RMSE}_y)^2/(n-1)}$	0.4788
Estimated standard deviation of the population of RMSEs	$S_{\mathrm{RMSE}_y} = \sqrt{S_y^2/n}$	0.0874
Greenwalt and Schultz CMAS standard normal (*Z*) interval of the population of errors at 90% probability	$1.645 * S_y$	0.7876
Greenwalt and Schultz standard normal (*Z*) interval of the population of errors at 95% probabilty	$1.96 * S_y$	0.9384
NSSDA statistic	$1.96 * \mathrm{RMSE}_y$	1.0201
Confidence interval on the estimate of RMSE_y at 95% probability	$\mathrm{RMSE}_y \pm 1.96 * S_{\mathrm{RMSE}}$	0.5202 ± .01713 which results in a range from 0.3491 to 0.6918

(*Continued*)

TABLE 3.4 (CONTINUED)
Horizontal Accuracy Example Equations and Statistics

Definitions	Circular Equations	Circular Values
Estimated root-mean-square of the populations of errors	$\text{RMSE}_h = \sqrt{\sum_i^n (e_{hi})^2/n}$	0.6191
Estimated absolute arithmetic mean of the population of errors	$\lvert\bar{e}_h\rvert = \sum_1^n \lvert e_{hi}\rvert / n$	0.4669
Estimated standard deviation of the population of errors	$S_h = (S_x + S_y)/2$	0.4141
Estimated standard deviation of the population of RMSEs	$S_{\text{RMSE}_h} = S_h/\sqrt{n}$	0.0756
Greenwalt and Schultz CMAS standard normal (Z) interval of the population of errors at 90% probability	$2.1460 * S_h$	0.8887
Greenwalt and Schultz standard normal (*Z*) interval of the population of errors at 95% probability	$2.4477 * S_h$	1.0136
Test for circularity	S_{min}/S_{max}	0.7298
$\text{NSSDA}_{\text{circular}}$ statistic	$1.7308 * \text{RMSE}_h$	1.0716
$\text{NSSDA}_{\text{elliptical}}$ statistic	$2.4477 * .5 * (\text{RMSE}_x + \text{RMSE}_y)$	1.0473
Confidence interval on the estimate of RMSE_h at 95% probability	$\text{RMSE}_h \pm 1.96 * S_{\text{RMSE}}$	0.6191 ± 0.1482 which results in a range from 0.4709 to 0.7673

accuracy assessment requires the collection of a separate and independent set of test sample points that were not used as control points in the registration process.

SUMMARY

A major tenet of most positional accuracy assessment standards is to report accuracy at a specified "confidence level" (FGDC, 1998; MPLMIC, 1999; NDEP, 2004). However, none of the existing standards provide equations for producing a confidence interval on the estimate of error. Table 3.5 provides a comparison of the commonly used positional accuracy standards. Each standard has both advantages and disadvantages:

1. NMAS is simple to implement, but it is based on map units instead of ground units, making it unusable for digital data. It also omits any guidance for estimating the range of errors at a given probability or for estimating a confidence interval around RMSE.
2. The Greenwalt and Schultz standard requires the assumption that the errors are normally distributed. The report states that the assumption of normality is "valid because positional error components generally follow a normal distribution pattern when sufficient data is available." However,

TABLE 3.5
Comparison of Commonly Used Positional Accuracy Standards to Each Other and to the Suggested New Standard

Positional Accuracy Standard	Uses a Maximum Distance of Error Allowed as the Standard	Provides Equations for Estimating Error Population Statistics	Requires the Assumption that the Errors Are Normally Distributed	Uses RMSE and 95% Percentile	Requires Stratification of the Landscape into Ground Cover Classes	Units
NMAS	Yes	No	Not required†	No, but uses 90th percentile	No	Map units
Greenwalt and Schultz	No	Yes	Yes	No	No	Unstated
ASPRS, 1989	Yes	Yes	Not required	No	No	Ground units
NSSDA	No	Yes	Yes	No	No	Ground units
FEMA, 2003	No	Yes	Yes	No	Yes	Ground units
ASPRS, 2004	No	Yes	Yes for RMSE, but not for 95th percentile	Yes	Yes	Ground units
NDEP, 2004	No	Yes	Yes for RMSE, but not for 95th percentile	Yes	Yes	Ground units
Combined standard: Z_i*S and a confidence interval on RMSE	No	Yes	Yes for calculating the probable range of errors, but no for calculating the confidence interval around the estimate of the mean error	No	Yes if desired	Ground units

† The assumption of normality is not required because probabilities of error distributions are not considered or calculated for the standard.

many practitioners are uncomfortable with this assumption and believe that positional errors are usually biased. Additionally, the Greenwalt and Schultz equations are often misinterpreted to calculate a level of confidence in the estimate of RMSE. However, a confidence level is a measure of the *reliability* of an estimation of a population parameter and is calculated using the standard error, not the standard deviation. The expression used in Greenwalt and Schultz is

$$Z_i\, S$$

and the range of errors around either vertical or horizontal *RMSE* at a specified probability is

$$\text{RMSE} \pm Z_i\, S$$

where RMSE and S were defined earlier and Zi is the Z statistic for the specified probability. The Greenwalt and Schultz equation calculates the distance on either side of the RMSE beyond which errors will not occur at a specified probability, and does not indicate the confidence of the estimate of RMSE. Contrary to statements in the NSSDA, the equations in Greenwalt and Schultz do not, nor do they pretend to, provide a confidence level for the estimate of the RMSE.

3. As an improvement to NMAS, the ASPRS standards use ground units, but do not provide for any guidance on estimating either the range of errors at given probabilities or a confidence level on the estimate of RMSE.
4. NSSDA provides excellent guidance on positional accuracy sample design and collection methods. It also attempts to provide a means for estimating the range of errors at specified probabilities (and not, as it states, the "confidence level" of the RMSE estimate). However, the NSSDA equation is incorrect because it applies the RMSE variable in its equations where the estimate of the standard deviation (*S*) should be used instead.

An alternative clarifying standard would be to require calculation of both

- the interval of errors around RMSE that captures 95% of the map errors using Greenwalt and Schultz's equations and assumptions:
  $\text{RMSE} \pm Z_i\, S$, and
- a confidence interval around the estimate of RMSE at 95% probability
  $\text{RMSE} \pm Z_i S_{RMSE}$ for large sample sizes, and
  $\text{RMSE} \pm t_i S_{\text{RMSE}}$ for small sample sizes.

This standard has several advantages:

1. It relies on widely accepted statistical theory and equations to characterize geospatial positional error.
2. It corrects for the equation mistakes in NSSDA.
3. It clarifies the difference between estimating the range of errors at a certain probability versus calculating a confidence interval on the estimate of RMSE.
4. Use of the confidence interval does not require the assumption that the errors be normally distributed, because the population of possible RMSE values is normally distributed even if the population of the errors is not, and
5. The equations incorporate all of the critical concepts of NSSDA accuracy standards, including:

- reporting the estimated accuracy of a spatial data layer rather than specifying a standard to be met,
- use of estimated RMSE to estimate positional accuracy, and
- the ability to express a confidence level in the estimate of RMSE.

APPENDIX 3.1

Determining the Required Sample Size

If we have a prior estimate of the mean and standard deviation of our population of errors, we can determine how many samples we have to take to provide a specified confidence interval around our estimate of the mean error.

If *d* is the interval on either side of the mean that we want to estimate, then the confidence interval is

$$\bar{X} \pm d$$

and

$$d = tS_{\text{RMSE}}$$

Because $S_{\text{RMSE}} = S/\sqrt{n}$, we can solve for *n* because

$$n = (t^2 S^2)/d^2$$

For example, let us assume that we want our confidence interval on our estimate of the mean vertical error in Table 3.1 to be no more than ±20% of the mean at the 95% confidence level. Using the value of RMSE of 0.320 and the variance (S^2) of 0.059, we can calculate how many samples we would need to take as

$$n = \frac{(1.96)^2(0.059)}{(20\%(0.320))^2}$$

$$= \frac{3.842\,(0.059)}{.0041}$$

or about 55 samples.

4 Thematic Accuracy

The major focus of this book is thematic accuracy assessment. Chapter 3 presented a summary of positional accuracy assessment and the standard measure for reporting it, RMSE. This chapter introduces the most widely accepted measure for representing thematic accuracy, the error matrix. The chapter also documents the evolution of thematic accuracy assessment, beginning with a discussion of early non-site-specific assessments. Next, site-specific assessment techniques employing the error matrix are presented, followed by the mathematical representation of the error matrix.

NON-SITE-SPECIFIC ASSESSMENTS

In a non-site-specific accuracy assessment, only total areas for each category mapped are computed, without regard to the location of these areas. In other words, a comparison between the number of acres or hectares of each category on the map generated from remotely sensed data and the reference data is performed. In this way, the errors of omission and commission tend to compensate for one another and the totals compare favorably. However, nothing is known about any specific location on the map or how it agrees or disagrees with the reference data.

A simple example quickly demonstrates the shortcomings of the non-site-specific approach. Figure 4.1 shows the distribution of the forest category on both a reference image and two different classifications generated from remotely sensed data. Classification #1 was generated using one type of classification algorithm (e.g., supervised, unsupervised, or nonparametric, etc.), while classification #2 employed a different algorithm. In this example, only the forest category is being compared. The reference data shows a total of 2,435 acres of forest, while classification #1 shows 2,322 acres and classification #2 shows 2,635 acres. In a non-site-specific assessment, you would conclude that classification #1 is better for the forest category because the total number of forest acres for classification #1 more closely agrees with the number of acres of forest on the reference image (2,435 acres – 2,322 acres = 113 acres difference for classification #1, while classification #2 differs by 200 acres). However, a visual comparison (see Figure 4.2) between the forest polygons on classification #1 and the reference data demonstrates little locational correspondence. Classification #2, despite being judged inferior by the non-site-specific assessment, appears to locationally agree much better with the reference data forest polygons (see Figure 4.2). Therefore, the use of non-site-specific accuracy assessment can be quite misleading. In the example shown here, the non-site-specific assessment actually recommends the use of the inferior classification algorithm.

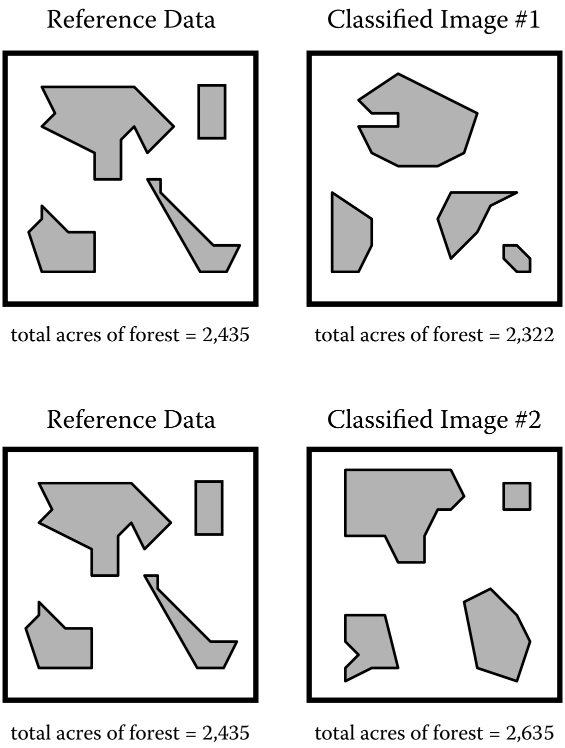

Reference Data
Classified Image #1
total acres of forest = 2,435
total acres of forest = 2,322
Reference Data
Classified Image #2
total acres of forest = 2,435
total acres of forest = 2,635

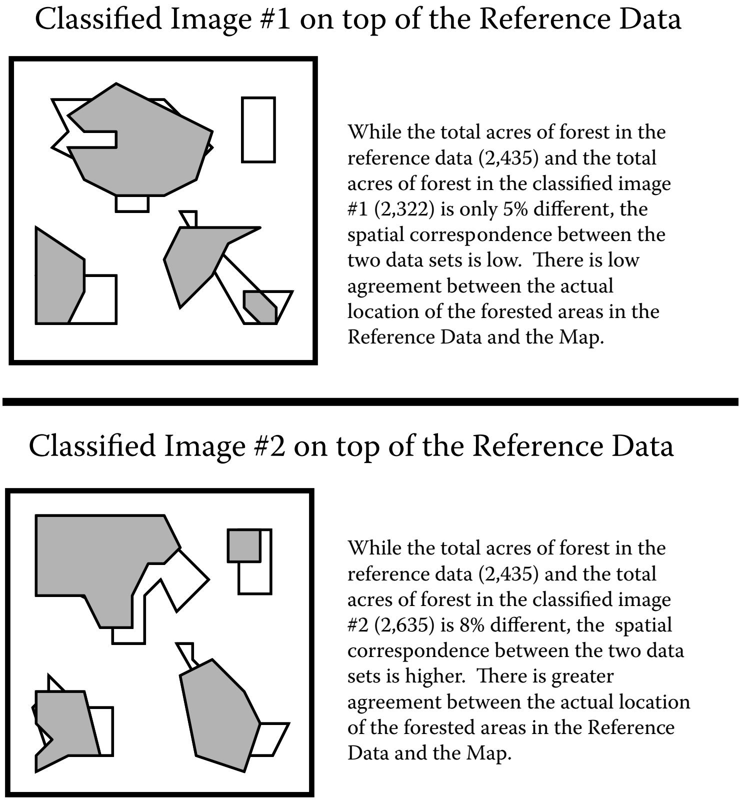
FIGURE 4.1 Example of non-site-specific accuracy assessment.

SITE-SPECIFIC ASSESSMENTS

Given the obvious limitations of non-site-specific accuracy assessment, there was a need to know how the map generated from the remotely sensed data compared to the reference data on a locational basis. Therefore, site-specific assessments were instituted. Initially, a single value representing the accuracy of the entire classification (i.e., overall accuracy) was presented. This computation was performed by comparing a sample of locations on the map with the same locations on the reference data and keeping track of the number of times there was agreement.

An overall accuracy level of 85% was adopted as representing the cutoff between acceptable and unacceptable results. This standard was first described in Anderson et. al (1976) and seems to be almost universally accepted despite there being nothing magical or even especially significant about the 85% correct accuracy level. Obviously, the accuracy of a map depends on a great many factors including the amount of effort, the level of detail (i.e., classification scheme), and the variability of the categories to be mapped. In some applications an overall accuracy of 85% is more than sufficient and in other cases it would not be accurate enough.

While having a single number to measure overall thematic map accuracy was an improvement over the non-site-specific assessment method, it was soon realized



Classified Image #1 on top of the Reference Data

While the total acres of forest in the reference data (2,435) and the total acres of forest in the classified image #1 (2,322) is only 5% different, the spatial correspondence between the two data sets is low. There is low agreement between the actual location of the forested areas in the Reference Data and the Map.

Classified Image #2 on top of the Reference Data

While the total acres of forest in the reference data (2,435) and the total acres of forest in the classified image #2 (2,635) is 8% different, the spatial correspondence between the two data sets is higher. There is greater agreement between the actual location of the forested areas in the Reference Data and the Map.

FIGURE 4.2 Spatial correspondence for the non-site-specific accuracy assessment example.

that this single number was not enough. The need to evaluate individual categories within the classification scheme was recognized, and thus began the use of the error matrix to represent map accuracy.

The Error Matrix

As previously introduced, an error matrix is a square array of numbers set out in rows and columns that expresses the number of sample units assigned to a particular category in one classification relative to the number of sample units assigned to a particular category in another classification (Table 4.1). In most cases, one of the classifications is considered to be correct (i.e., the reference data) and may be generated from aerial photography, airborne video, ground observation, or ground measurement. The columns usually represent this reference data, while the rows indicate the classification generated from the remotely sensed data (i.e., the map). It should be noted that the reference data has often been referred to as the “ground truth” data. Now, while it is true that the reference data are assumed to be more correct than the map it is being used to assess, it is by no means true that these data are perfect

TABLE 4.1
Example Error Matrix (Same as that Presented in Figure 2.6)

		Reference Data				Row
		D	C	AG	SB	Total
Classified Data	D	65	4	22	24	115
	C	6	81	5	8	100
	AG	0	11	85	19	115
	SB	4	7	3	90	104
Column Total		75	103	115	141	434

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

OVERALL ACCURACY =
(65 + 81 + 85 + 90)/434 =
321/434 = 74%

PRODUCER'S ACCURACY	USER'S ACCURACY
D = 65/75 = 87%	D = 65/115 = 57%
C = 81/103 = 79%	C = 81/100 = 81%
AG = 85/115 = 74%	AG = 85/115 = 74%
SB = 90/141 = 64%	SB = 90/104 = 87%

or represent "the truth." Therefore, the term "ground truth" is inappropriate and, in some cases, very misleading. Throughout this book, the authors will use the term *reference data* to identify the data being used to compare to the map generated from remotely sensed data (i.e., the map).

An error matrix is a very effective way to represent map accuracy in that the individual accuracies of each category are plainly described along with both the errors of inclusion (commission errors) and errors of exclusion (omission errors) present in the classification. A commission error is simply defined as including an area in a category when it does not belong to that category. An omission error is excluding an area from the category to which it belongs. Each and every error is an omission from the correct category and a commission to a wrong category.

For example, in the error matrix in Table 4.1, there are four areas that were classified as deciduous when the reference data shows that they were actually conifer. Therefore, four areas were omitted from the correct coniferous category and committed to the incorrect deciduous category. In addition to clearly showing errors of omission and commission, the error matrix can be used to compute other accuracy measures such as overall accuracy, producer's accuracy, and user's accuracy (Story and Congalton, 1986). Overall accuracy is simply the sum of the major diagonal (i.e., the correctly classified sample units) divided by the total number of sample units in the entire error matrix. This value is the most commonly reported accuracy assessment statistic and is probably most familiar to the reader. However, just presenting the overall accuracy is not enough. It is important to present the entire matrix so that other accuracy measures can be computed as needed and confusion between map classes is clearly presented and understood.

Producer's and user's accuracies are ways of representing individual category accuracies instead of just the overall classification accuracy, and were introduced by Story and Congalton (1986). Before error matrices became the standard accuracy reporting mechanism, it was common to report the overall accuracy and either only the producer's or user's accuracy. Sometimes, only the higher of the two accuracies (between the producer's and user's accuracies) was selected to be reported, resulting in misleading information about the map accuracy. A quick example will demonstrate the need to publish the entire matrix so that all three accuracy measures can be computed.

Studying the error matrix shown in Table 4.1 reveals an overall map accuracy of 74%. However, suppose we are most interested in the ability to classify hardwood forests so we calculate a "producer's accuracy" for this category. This calculation is performed by dividing the total number of correct sample units in the deciduous category (i.e., 65) by the total number of deciduous sample units as indicated by the reference data (i.e., 75 or the column total). This division results in a "producer's accuracy" of 87%, which is quite good. If we stopped here, one might conclude that although this classification appears to be average overall, it is more than adequate for the deciduous category. Drawing such a conclusion could be a very serious mistake. A quick calculation of the "user's accuracy" computed by dividing the total number of correct sample units in the deciduous category (i.e., 65) by the total number of sample units classified as deciduous (i.e., 115 or the row total) reveals a value of 57%. In other words, although 87% of the deciduous areas have been correctly identified as deciduous, only 57% of the areas called deciduous on the map are actually deciduous on the ground. The high producer's accuracy occurs because too much of the map is labeled deciduous. A more careful look at the error matrix reveals that there is significant confusion in discriminating deciduous from barren and shrub. Therefore, although the producer of this map can claim that 87% of the time an area that was deciduous on the ground was identified as such on the map, a user of this map will find that only 57% of the time that the map says an area is deciduous will it actually be deciduous on the ground.

Mathematical Representation of the Error Matrix

This subsection presents the error matrix in mathematical terms necessary to perform the analysis techniques described in the Chapter 7. The error matrix was presented previously in descriptive terms, including an example (Table 4.1) that should help the reader make this transition to equations and mathematical notation easier to understand.

Assume that n samples are distributed into k^2 cells, here each sample is assigned to one of k categories in the map (usually the rows), and independently to one of the same k categories in the reference data set (usually the columns). Let n_{ij} denote the number of samples classified into category i $(i = 1, 2, \ldots, k)$ in the map and category j $(j = 1, 2, \ldots, k)$ in the reference data set (Table 4.2).

TABLE 4.2
Mathematical Example of an Error Matrix

		j = Columns			Row
		(Reference)			Total
		1	2	k	n_{i+}
i = Rows (Classification)	1	n_{11}	n_{12}	n_{1k}	n_{1+}
	2	n_{21}	n_{22}	n_{2k}	n_{2+}
	k	n_{k1}	n_{k2}	n_{kk}	n_{k+}
Column Total n_{+j}		n_{+1}	n_{+2}	n_{+k}	n

Let

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

be the number of samples classified into category *i* in the remotely sensed classification, and

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

be the number of samples classified into category *j* in the reference data set.

Overall accuracy between remotely sensed classification and the reference data can then be computed as follows:

$$\text{overall accuracy} = \frac{\sum_{i=1}^{k} n_{ii}}{n}.$$

Producer's accuracy can be computed by

$$\text{producer's accuracy } j = \frac{n_{jj}}{n_{+j}}$$

and the user's accuracy can be computed by

$$\text{user's accuracy}_i = \frac{n_{ii}}{n_{i+}}$$

Finally, let p_{ij} denote the proportion of samples in the *i*, *j*th cell, corresponding to n_{ij}. In other words, $p_{ij} = n_{ij}/n$.

Then let p_{i+} and p_{+j} be defined by

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

and

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

This mathematical representation of the error matrix takes a little practice to get used to. Actually, understanding an error matrix the very first time can take a little effort. However, given the importance of the error matrix in thematic accuracy assessment and the need for the mathematical representation for some of the analysis techniques, readers are encouraged to spend a little time here until they feel comfortable. Many examples will be provided throughout the book, as well as some case studies, to aid every reader in becoming an error matrix expert.

5 Sample Design Considerations

Now that we understand thematic accuracy is typically represented using an error matrix, it is important to know how to correctly generate and populate the matrix. Assessing the thematic accuracy of maps or other spatial data requires sampling because it is not economically feasible to visit every place on the ground. Sampling design requires knowledge of the distribution of thematic classes across the landscape, determination of the types and number of samples to be taken, and choice of a sampling scheme for selecting the samples. Design of an effective and efficient sample to collect valid reference and map accuracy data is one of the most challenging and important components of any accuracy assessment, because the design will determine both the cost and the statistical rigor of the assessment.

Accuracy assessment assumes that the information displayed in the error matrix is a true characterization of the map being assessed. Thus, an improperly designed sample will produce misleading accuracy results. Several considerations are critical to designing an accuracy assessment sample that is truly representative of the map:

1. What are the thematic map classes to be assessed and how are they distributed across the landscape?
2. What is the appropriate sample unit?
3. How many samples should be taken?
4. How should the samples be chosen?

While seemingly straightforward, each of these steps has many potential pitfalls. Failure to consider even one of them can lead to serious shortcomings in the assessment process. This chapter considers each one of these factors.

WHAT ARE THE THEMATIC MAP CLASSES TO BE ASSESSED?

How we sample the map for accuracy will partially be driven by how the thematic classes of the map are distributed across the landscape. This distribution will, in turn, be a function of how we have chosen to categorize the features of the earth being mapped; referred to as the *classification scheme.* Once we know the classification scheme, we can learn more about how the map classes are distributed. Important considerations are the discrete nature of map information, and the spatial interrelationship or autocorrelation of that information. Assumptions made about the distribution of map categories will affect both how we select accuracy assessment samples and the outcome of the analysis.

The Classification Scheme

Maps categorize the earth's surface. For example, road maps tell us the type of road, its name, and location. Land cover maps typically enumerate the types, mix, and density of vegetation covering the earth (e.g., trees, shrubs, and grass). Land use maps characterize how land is utilized by humans (e.g., urban, agriculture, and forest management).

Thematic map categories are specified by the project's classification scheme. Classification schemes are a means of organizing spatial information in an orderly and logical way (e.g., Cowardin et al., 1979). Classification schemes are fundamental to any mapping project because they create order out of chaos and reduce the total number of items considered to some reasonable number. The classification scheme makes it possible for the map producer to characterize landscape features and for the user to readily recognize them. Without a classification scheme, no mapping is truly possible. The detail of the scheme is driven by (1) the anticipated uses of the map information, and (2) the features of the earth that can be discerned with the remotely sensed data (e.g., aerial or satellite imagery) being used to create the map. If a rigorous classification scheme is not developed before mapping begins, then any subsequent accuracy assessment of the map will be meaningless because it will be impossible to definitively label the accuracy assessment samples.

A classification scheme has two critical components: (1) a set of *labels* (e.g., urban residential, deciduous forest, palustrine emergent wetland, etc.); and (2) a set of *rules* or definitions such as a dichotomous key for assigning labels (e.g., a "deciduous forest must have at least 75% crown closure in deciduous trees"). Without a clear set of rules, the assignment of labels to classes can be arbitrary and lack consistency. For example, everyone has their own idea about what constitutes a forest and yet there are many definitions that could result in very different maps of forest distribution. Consider a situation in which one agency defines forest as an area where 10% of the ground area is covered by trees, and another agency uses a slightly different definition according to which forest exists only if 25% of the ground area is covered by trees. If analysts from each of these agencies were together in a specific plot of land, they could label the area differently based on their agency's definitions of a forest and both of the labels would be correct. Without class definitions expressed as quantifiable rules, there can be little agreement on what area on the ground or the image should be labeled.

The level of detail (i.e., number and complexity of the categories) in the scheme strongly influences the time and effort needed to make the map and to conduct the accuracy assessment. The more detailed the scheme, the more expensive the map and its assessment. Because the classification scheme is so important, no work should begin on a mapping project until the scheme has been thoroughly reviewed and as many problems as possible identified and solved.

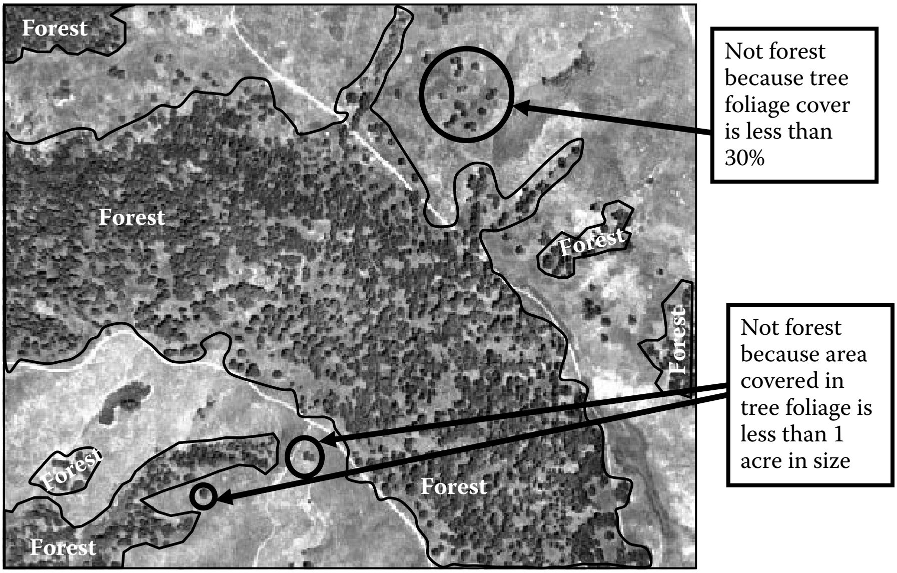
In addition to being composed of labels and a set of rules, a classification scheme should be (1) *mutually exclusive* and (2) *totally exhaustive*. Mutual exclusivity requires that each mapped area fall into one and only one category or class. For example, classification scheme rules would need to clearly distinguish between forest and water (seemingly simple), so that a mangrove swamp cannot receive both a forest and a water label. A totally exhaustive classification scheme results in every area on the mapped landscape receiving a map label; no area can be left unlabeled. One way to ensure that the scheme is totally exhaustive is to have a category labeled as other or unclassified.

If possible, it is also advantageous to use a classification scheme that is *hierarchical.* In hierarchical systems, specific categories within the classification scheme can be collapsed to form more general categories. This ability is especially important when it is discovered that certain map categories cannot be reliably mapped. For example, it may be impossible to separate interior live oak from canyon live oak in California's oak woodlands (these two oak types are almost indistinguishable on the ground). Therefore, these two categories may have to be collapsed to form a live oak category that can be reliably mapped.

Finally, the classification scheme must specify the minimum mapping unit (mmu) for each class being mapped. The mmu is the smallest area of the class to be delineated on the map. Figure 5.1 illustrates this concept. In this example, the rule for mapping a forest is specified as follows:

> An area of 1 acre or more where more than 30% of the ground, as seen from above the tree canopy, is covered by the foliage of hardwood or conifer trees.

The minimum mapping unit for forests is one acre. Areas covered with 30% tree foliage, but smaller than the 1 acre minimum mapping unit will not be labeled as forests. Additionally, areas larger than 1 acre but containing less than 30% tree foliage cover will also not be labeled as forests. Reference data must be collected at the



Not forest because tree foliage cover is less than 30%

Not forest because area covered in tree foliage is less than 1 acre in size

FIGURE 5.1 Example of the impact of a minimum mapping unit.

same minimum mapping unit as was applied to the map generated from the remotely sensed data. For example, it is not possible to assess the accuracy of a Landsat 30 m × 30 m pixel with a single 1/20 hectare ground inventory plot, nor is it possible to assess the accuracy of an AVHRR 1.1 km × 1.1 km pixel using a 30 m × 30 m pixel.

Figure 5.2 provides a dichotomous key for a simple, yet robust, classification scheme for a fire/fuel mapping project. Note how the scheme specifies a minimum mapping unit and is:

- Totally exhaustive—every piece of the landscape will be labeled,
- Mutually exclusive—no one piece of the landscape can receive more than one label, and
- Hierarchical—detailed fuel classes can be lumped into the more general groups of nonfuel, grass, shrub, timber slash, and timber litter.

It is critical that accuracy assessment reference data be collected and labeled using the same classification scheme as that used to generate the map. This may seem obvious until you are tempted to use an existing map to assess the accuracy of a new map. Rarely will any two maps be created using the same classification scheme. Any differences between the classification scheme of the map and the classification scheme of the reference data may result in discrepancies between map and reference accuracy assessment site labels. The result will be an assessment of classification scheme differences, and not of map accuracy.

Other Data Considerations

Continuous versus Noncontinuous Data

Most statistical analysis assumes that the population to be sampled is continuous and normally distributed, and that samples will be independent. Yet we know that classification systems, for all their power in organizing chaos, also take a continuous landscape and divide it into often arbitrarily discrete categories. For example, tree crown closure rarely exists in discrete classes. Yet when we make a map of crown closure, we impose discrete crown closure classes across the landscape. For example, we may create a crown closure map with 4 classes; class 1 from 0 to 10% crown closure, class 2 from 11 to 50% crown closure, class 3 from 51 to 75% crown closure, and class 4 from 76 to 100% crown closure. Given this boundary between two crown closure classes at 75%, one can expect to find confusion between a forest stand with a crown closure of 73% that belongs in class 3 and a stand of 77% that belongs in class 4 (see Chapter 9 for a discussion on fuzzy accuracy assessment). In addition, categories tend to be related spatially, resulting in autocorrelation (discussed next in this chapter). In most situations, some balance between what is statistically valid and what is practically obtainable is desired. Therefore, knowledge of these statistical considerations is a must.

Most students who have completed a beginning statistics course are familiar with sampling and analysis techniques for continuous, normally distributed data. It is these techniques such as analysis of variance (ANOVA) and linear regression that are most familiar to the reader.

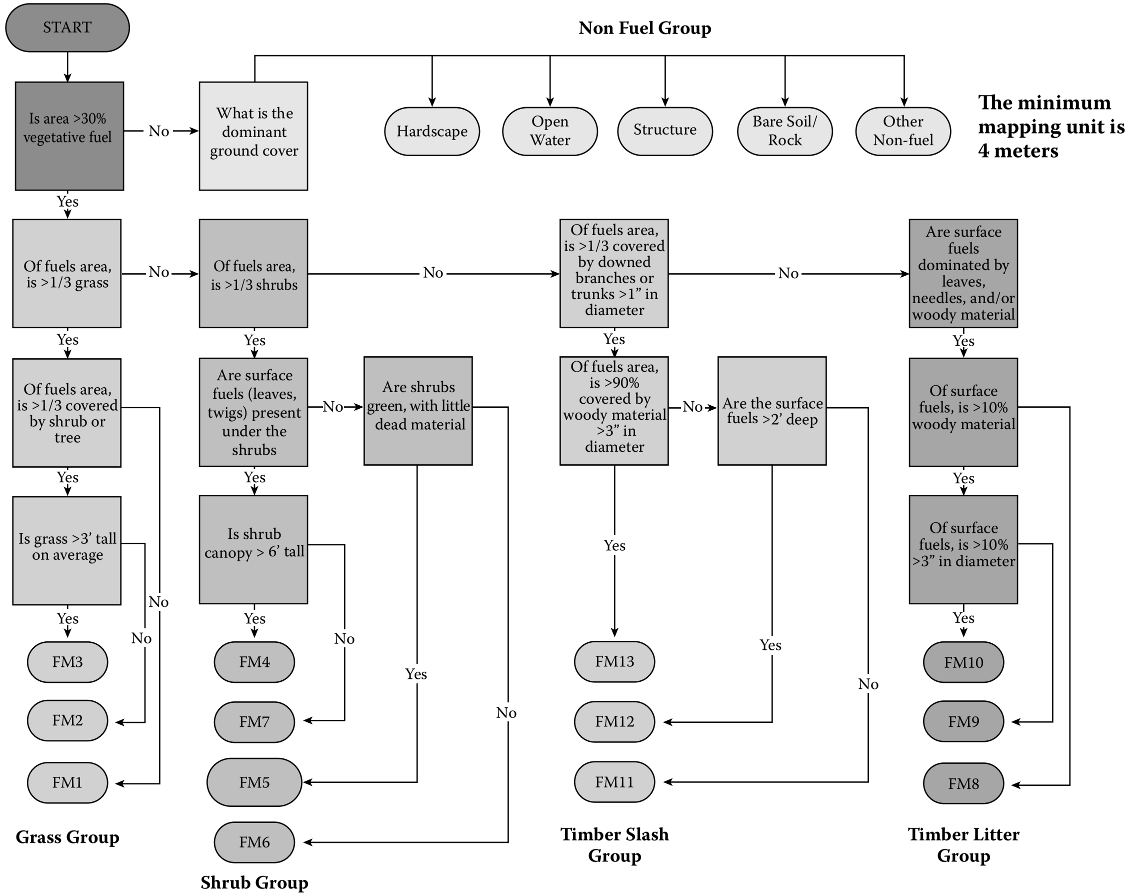

START
Non Fuel Group
Is area >30% vegetative fuel
No
Yes
What is the dominant ground cover
Hardscape
Open Water
Structure
Bare Soil/ Rock
Other Non-fuel
The minimum mapping unit is 4 meters
Of fuels area, is >1/3 grass
Of fuels area, is >1/3 shrubs
Of fuels area, is >1/3 covered by downed branches or trunks >1” in diameter
Are surface fuels dominated by leaves, needles, and/or woody material
Of fuels area, is >1/3 covered by shrub or tree
Are surface fuels (leaves, twigs) present under the shrubs
Are shrubs green, with little dead material
Of fuels area, is >90% covered by woody material >3” in diameter
Are the surface fuels >2’ deep
Of surface fuels, is >10% woody material
Is grass >3’ tall on average
Is shrub canopy > 6’ tall
Of surface fuels, is >10% >3” in diameter
FM3
FM2
FM1
FM4
FM7
FM5
FM6
FM13
FM12
FM11
FM10
FM9
FM8
Grass Group
Shrub Group
Timber Slash Group
Timber Litter Group

FIGURE 5.2 Example wildland fuel classification scheme.

However, thematic map information is discrete, not continuous, and frequently not normally distributed. Therefore, normal theoretical statistical techniques that assume a continuous normal distribution may be inappropriate for map accuracy assessment. It is important to consider how the data are distributed and what assumptions are being made before performing any statistical analysis. Sometimes there is little that can be done about the artificial delineations in the classification scheme; other times the scheme can be modified to better represent natural breaks. Care and thought must go into this process to achieve the best analysis possible.

Spatial Autocorrelation

Spatial autocorrelation occurs when the presence, absence, or degree of a certain characteristic affects the presence, absence, or degree of that same characteristic in neighboring units (Cliff and Ord, 1973), thereby violating the assumption of sample independence. This condition is particularly important in accuracy assessment if an error in a certain location can be found to positively or negatively influence errors in surrounding locations (Campbell, 1981). Clearly, if spatial autocorrelation exists, the sampling must ensure that the samples are separated by enough distance to minimize this effect, or else the sampling will not adequately represent the entire map.

The existence of spatial autocorrelation is clearly illustrated in work by Congalton (1988a) on Landsat MSS data from three areas of varying spatial diversity/complexity (i.e., an agriculture, a rangeland, and a forested site), which showed a positive influence over 1 mile away. Figure 5.3 presents the results of this analysis. Each image, called a difference image, is a comparison between the remotely sensed classification (i.e., the map) and the reference data. The black areas represent the error, those places where the map and the reference label disagree. The white areas represent the agreement.

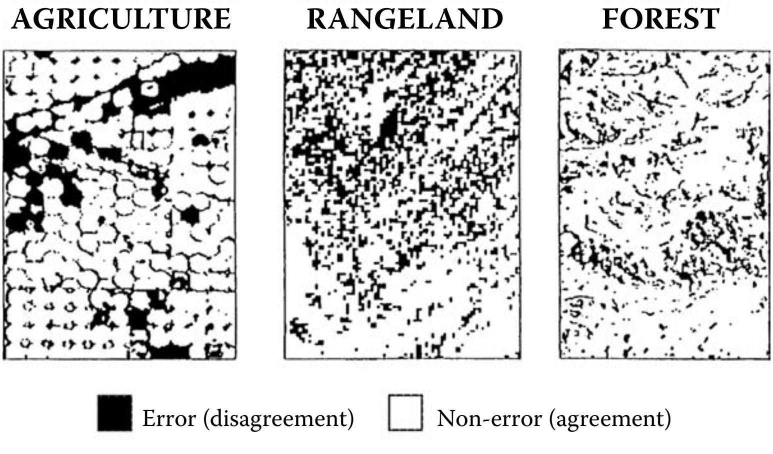


FIGURE 5.3 Difference images (7.5-minute quadrangles) showing the pattern of error for three ecosystems of varying complexity: agriculture, rangeland, and forest. (Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton, R. 1988. Using spatial autocorrelation analysis to explore errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*. 54(5): 587–592.)

The pattern of differences between the map and reference labels are readily explainable in an agricultural environment in which field sizes are large and typical misclassification would result in an error in labeling the entire field. In the agricultural difference image in Figure 5.3, the fields are circular fields employing center-pivot irrigation, and examples can be seen of misclassifying entire fields. For example, a field that is mapped as corn when it is actually wheat will result in an entire field (center pivot area in Figure 5.3) being mislabeled. Therefore, it is not surprising that the errors occur in large areas and that there is a positive autocorrelation over a large distance.

However, the results are more surprising for the rangeland and forested classes. Both classes are more spatially complex (i.e., have more fragmentation, edges, and mixtures of land cover) than the agriculture class, and therefore one would expect them to be less spatially autocorrelated. Primarily because of rangeland fencing, the rangeland class does have some of the fields similar to agriculture, but it also reflects some of the edge effects more common to the complex forest class.

The forest class is the most spatially complex, and most map error would be expected to occur along the edges or transition zones between forest types. Although viewing the forest difference image does tend to confirm these edge problems, the results of the analysis still indicate that there is strong positive autocorrelation between errors up to 30 pixels away. In other words, if an error occurs at a given location, it is more likely that another error will be found, even up to this rather large distance away (i.e., 30 MSS pixels or about 240 m), than a correct classification.

The existence of spatial autocorrelation can violate the assumption of sample independence which, in turn, can affect the sample size and especially the sampling scheme used in accuracy assessment. Spatial autocorrelation may indicate the existence of periodicity in the presence of a class across the landscape that could affect the results of any type of systematic sample if the systematic sample design repeats the same periodicity. For example, maple trees need ample water and, in arid landscapes, are usually located along streams. A systematic sampling scheme based on choosing samples near streams would repeat the periodicity of the maple forest class and would result in a biased choice of samples that would oversample maple forests and undersample other map classes.

In addition, autocorrelation may affect the size and number of samples used in cluster sampling because each sample unit may not be contributing new, independent information, but rather, redundant information. Therefore, it would not be effective to collect information in a large cluster sample since very quickly each new sample site in the cluster would be adding very little new information. However, cluster sampling is a very cost-effective method, especially in the field, when the cost of traveling from one sample to another can be very high. Even when the accuracy assessment samples are taken in the office from aerial imagery, cluster sampling can create savings in setup time for each image. Therefore, it is important to consider spatial autocorrelation and balance the impact of having spatially autocorrelated samples against the efficiencies of cluster sampling. This can be done by limiting the number of samples taken in the cluster to 2–4, making sure that each sample unit in the cluster is taken in a different thematic class, and spreading the samples as far apart as possible.

WHAT IS THE APPROPRIATE SAMPLE UNIT?

Sample units are the portions of the map that will be selected for accuracy assessment. There are four possible choices for the sampling unit: (1) a single pixel, (2) a cluster of pixels (often a 3 × 3 pixel square), (3) a polygon (or object), and (4) a cluster of polygons.

Single Pixel

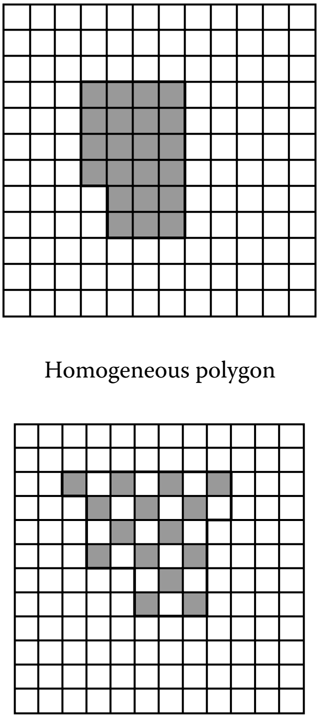
Historically, a large number of accuracy assessments have been conducted using a single pixel as the sampling unit. However, a single pixel is a very poor choice for the sampling unit for many reasons:

- First, a pixel is an arbitrary rectangular delineation of the landscape that may have little relation to the actual delineation of land cover or land use type. It can be a single land cover or vegetation category (i.e., a pure pixel) or more often than not, it can be a mixture of land cover or vegetation classes.
- Second, before the relatively new geocoding and terrain correction procedures were adopted, it was almost impossible to exactly align one pixel on a map to the exact same area in the reference data. Therefore, there was no way to guarantee that the location of the reference pixel was identical to the location of the map pixel. Even with terrain correction and georeferencing, it is still not possible to get an exact alignment of the boundaries of a pixel. Similarly, until global positioning system (GPS) came along, there was no practical way to ensure that ground-collected reference data was being collected for the exact map pixel being assessed. Even with GPS, this correspondence is not guaranteed to exactly match. Therefore, positional accuracy becomes a large issue, and the thematic accuracy of the map is affected because of positional error.
- Finally, few classification schemes specify a unit as small as a pixel as the minimum mapping unit. If the mmu is larger than a single pixel, then a single pixel is inappropriate as the sample unit.

Even with all the recent technological advances in GPS, terrain correction, and geocoding, accuracy assessment sample units will still have some positional inaccuracies. It is commonly accepted that a positional accuracy of one-half pixel is sufficient for sensors such as Landsat Thematic Mapper and SPOT Multispectral imagery. As sensors increase in spatial resolution, such as that collected from digital airborne cameras and high-resolution satellites, positional accuracy becomes more important and new standards need to be established. If an image with a pixel size of 10–30 m is registered to the ground to within half a pixel (i.e., 5–15 m) and a GPS unit is used to locate the unit on the ground to within 10–15 m, then it is impossible to use a single pixel as the sampling unit for assessing the thematic accuracy of the map. There would simply be no guarantee that the map and the reference data would be collected from the identical area. If the positional accuracy is not up to the standard or if GPS is not used to precisely locate the sample on the ground, then these factors become more important and can significantly affect the thematic accuracy assessment. This is all the more true for higher spatial resolution imagery, in which the pixels may be smaller.

Cluster of Pixels

Given the need to balance thematic accuracy with positional accuracy, a cluster of pixels, typically a 3 × 3 square for moderate resolution imagery, has recently been the most common choice for the sample unit. A cluster minimizes registration problems because it is easier to locate on the reference data or in the field. However, a cluster of pixels (especially a 3 × 3 window) may still be an arbitrary delineation of the landscape resulting in the sample unit encompassing more than one map category. To avoid this problem, many analysts require that only homogeneous clusters of pixels be sampled. However, such restrictions may result in a biased sample that avoids heterogeneous areas which are a function of a mix of pixels (e.g., a mixed hardwood-conifer stand of trees) as depicted in Figure 5.4.



Homogeneous polygon

Heterogeneous polygon

FIGURE 5.4 Comparison of accuracy assessment polygons comprising homogeneous versus heterogeneous pixels.

It is important to remember that the sample unit dictates the level of detail for the accuracy assessment. If the assessment is performed on a 3 × 3 cluster of pixels, then nothing can be said about an individual pixel, nor can anything be said about polygons (i.e., management areas, forest stands, agricultural fields, etc.). Additionally, each sample unit must be considered a single sample. If, for example, a 3 × 3 cluster of pixels is used as the sample unit, then it must be counted as one sample, and not as nine samples. There are numerous examples in the literature of authors mistakenly counting each pixel in a cluster as a separate accuracy assessment unit. Also, the presence of spatial autocorrelation in most thematic maps dictates that samples should be spaced adequately apart from one another.

Extending the concept of a cluster of pixels to higher-resolution imagery requires knowledge about the positional accuracy of the imagery. As previously stated, common registration (positional) accuracies for Landsat Thematic Mapper (30 m pixels) and SPOT (10 m pixels) satellite imagery are about half a pixel. Therefore, selecting a homogeneous cluster of 3 × 3 pixels ensures that the center of the sample will definitely fall within the 3 × 3 cluster. Higher spatial resolution imagery such as that from Ikonos or Digital Globe have pixel sizes of 4 m to below 1 m. However, because of the off-nadir acquisition and other issues, the positional accuracy of these data are often in the range of 10–20 m and can even be much larger. Therefore, a 3 × 3 pixel cluster as the sampling unit would not be appropriate in this case. If the registration accuracy was 10 m and the pixel size was 4 m, then the cluster would need to be at least 5 × 5 pixels to account for this positional error. It is imperative that the positional accuracy be considered in the selection of the sample unit cluster size or else the thematic assessment will be flawed.

Polygons

Most large-scale thematic maps delineate the landscape into polygons of homogeneous map classes. Polygons are delineated on edges of classes, where more "between" than "within" class polygon variation exists. While the pixels inside the polygons may vary dramatically (as in a sparse stand of trees), the class label across the pixels is constant. Usually the polygon map is created either through manual interpretation or through the use of image segmentation and object-oriented classification algorithms. If the map to be assessed is a polygon map, then the accuracy assessment sample units should also be polygons. The resulting accuracy values inform the map's user and producer about the level of detail in which they are interested: the polygons. More and more mapping projects using remotely sensed data are generating polygon rather than pixel products as a result of developments in image segmentation and object-based image analysis. As a result, the polygon is replacing the cluster of pixels as the sample unit of choice.

However, using polygons as sample units can cause confusion if the accuracy assessment polygons are collected during the initial training data/calibration fieldwork, which occurs before the map polygons are created. The result can often be manually delineated accuracy assessment polygons with dramatically different delineations than the final map polygons, as illustrated in Figure 5.5. When this occurs, some

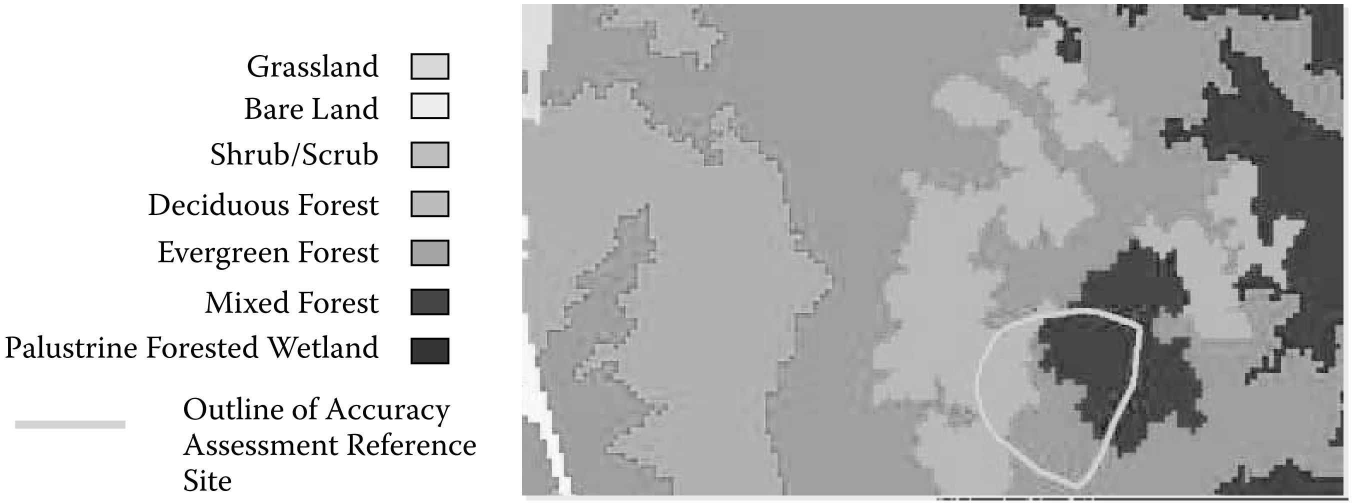


FIGURE 5.5 ***(Color version follows page 112)*** Mixed forest accuracy assessment reference polygon over the map polygons of evergreen, mixed, and deciduous forest. Determining the map label of the accuracy assessment polygon, when the polygon intersects with multiple map classes, can be problematic.

way of creating the map label for the accuracy assessment polygon must be developed. The simplest approach is to use the majority class of the polygon to create the map label. However, this may not work well in heterogeneous conditions, in which the label is more a function of the mix of the ground cover (e.g., patchy seagrass or mixed hardwood conifer forests) rather than the majority of the ground cover.

Another approach is to run the segmentation algorithms and finalize the delineation of the polygons prior to the initial field trip. The resulting polygons will probably vary only slightly from the final polygons and therefore can be reliably used as accuracy-sampling units.

Clusters of Polygons

Sampling clusters of polygons (or a grouping of polygons together), rather than single polygons, can reduce accuracy assessment costs dramatically because travel time and/or setup time is decreased. Unlike clusters of pixels, each polygon within a cluster of polygons can represent a single sampling unit because polygons are by definition separate map class types that have more between than within variation. However, care must be taken to provide some separation between polygons and to limit the number in the cluster. In other words, the impact of spatial autocorrelation still must be considered.

HOW MANY SAMPLES SHOULD BE TAKEN?

Accuracy assessment requires that an adequate number of samples per map class be gathered so that the assessment is a statistically valid representation of the accuracy of the map. However, the collection of reference data at each sample unit is very expensive, requiring that sample size be kept to a minimum to be affordable.

Of all the considerations discussed in this chapter, the most has probably been written about sample size. Many researchers, notably Hord and Brooner (1976), van Genderen and Lock (1977), Hay (1979), Ginevan (1979), Rosenfield et al. (1982), and Congalton (1988b), have published equations and guidelines for choosing the appropriate sample size.

The majority of work performed by early researchers used an equation based on the binomial distribution or the normal approximation to the binomial distribution to compute the required sample size. These techniques are statistically sound for calculating the sample size needed to compute the overall accuracy of a classification or even the overall accuracy of a single category. The equations are based on the proportion of correctly classified sample units and on some allowable error. However, these techniques were not designed to choose a sample size for generating an error matrix.

In the case of creating an error matrix, it is not simply a question of correct or incorrect (the binomial case). Instead, it is a matter of which error or which categories are being confused. Given an error matrix with n land cover categories, for a given category there is one correct answer and $(n - 1)$ incorrect answers. Sufficient samples must be acquired to be able to adequately represent this confusion (i.e., build a statistically valid error matrix). Therefore, the use of the binomial distribution for determining the sample size for an error matrix is not appropriate. Instead, the use of the multinomial distribution is recommended (Tortora, 1978).

The appropriate sample size can and should be computed for each project using the multinomial distribution. However, in our experience, a general guideline or good "rule of thumb" suggests planning to collect a minimum of 50 samples for each map class for maps of less than 1 million acres in size and fewer than 12 classes (Congalton, 1988b). Larger area maps or more complex maps should receive 75 to 100 accuracy assessment sites per class. These guidelines were empirically derived over many projects, and the use of the multinomial equation has confirmed that they are a good balance between statistical validity and practicality.

Because of the large number of potential samples (i.e., pixels, clusters of pixels, polygons, and clusters of polygons) in a remotely sensed image, traditional thinking about sampling in which a 2%, or even 5%, sample is not uncommon often does not apply. To illustrate this point, even as small a sample as 0.5% of a single Landsat Thematic Mapper scene is over 300,000 pixels. As we have previously concluded, accuracy assessment should not be performed on a per-pixel basis, but the same relative argument holds true for the other sample units. Therefore, practical considerations are often a key component of the sample size selection process. For example, the number of samples for each category may be adjusted on the basis of the relative importance of that category within the objectives of the mapping project or by the inherent variability within each of the categories. Sometimes, because of budget constraints or other factors, it is better to concentrate the sampling on the categories of interest and increase their number of samples while reducing the number of samples taken in the less important categories. Also, it may be useful to take fewer samples in categories that show little variability, such as water or forest plantations, and increase the sampling in the categories that are more variable, such as uneven-aged forests or riparian areas. However, in most instances, some minimum number of samples (e.g., 50 samples as per the guidelines or the result of the multinomial equation calculation) should be taken in each land cover category contained in the matrix. Perhaps most importantly, the entire accuracy assessment process should be documented so that others can know exactly what procedures were followed.

Finally, it may be tempting to design a sample that selects many samples in categories which are most accurate and a few in the confused categories. This strategy would guarantee a high accuracy value, but would not be representative of the map accuracy. Care should be taken to ensure that the sampling effort is carefully planned and implemented. It should also be noted that exactly how the sample is selected can affect the analysis performed on the sampled data. Again, the object here is to balance the statistical recommendations in order to get an adequate sample to generate an appropriate error matrix within the time, cost, and practical limitations associated with any viable mapping project. However this balance is achieved, it is critical to document the exact process so that future users of the map can know how the assessment was conducted.

Binomial Distribution

As mentioned earlier, the binomial distribution or the normal approximation to the binomial distribution is appropriate for computing the sample size for determining overall accuracy or the accuracy of an individual category. Later in this book, the binomial distribution will be used to assess a change/no change map (see Chapter 11). It is appropriate for the two-case situation in which only right and wrong are important. Choosing the appropriate sample size from the binomial or normal approximation is dependent on (1) the level of acceptable error one is willing to tolerate and (2) the desired level of confidence that the actual accuracy is within some minimum range. Numerous publications present look-up tables of the required sample size for a given acceptable error and desired level of confidence (e.g., Cochran, 1977 and Ginevan, 1979).

For example, suppose it is decided that a map is unacceptable if the overall accuracy is 90% or less. Also, let us say that we are willing to accept a 1 in 20 chance that we will make a mistake based on our sample and accept a map that actually has an accuracy of less than 90%. Finally, let us decide that we will accept the same risk, a 1 in 20 chance, of rejecting a map that is actually correct. The appropriate look-up table would then indicate that we must take 298 samples of which only 21 can be misclassified. If more than 21 samples were misclassified, we would conclude that the map is not acceptable.

Multinomial Distribution

As discussed earlier in this chapter, the multinomial distribution provides the appropriate equations for determining the sample size required to generate an error matrix. The procedure for generating the appropriate sample size from the multinomial distribution is summarized here and was originally presented by Tortora (1978).

Consider a population of units divided into k mutually exclusive and exhaustive categories. Let Π_i, $i = 1, \ldots, k$, be the proportion of the population in the ith category, and let n_i, $i = 1, \ldots, k$, be the frequency observed in the ith category in a simple random sample of size n from the population.

For a specified value of α, we wish to obtain a set of intervals S_i, $i = 1, \ldots, k$, such that

$$\Pr\left\{\bigcap_{i=1}^{k}(\Pi_i \in S_i)\right\} \geq 1-\alpha;$$

that is, we require the probability that every interval S_i contains Π_i to be at least $1-\alpha$. Goodman (1965) determined the approximate large-sample confidence interval bounds (when $n \rightarrow \infty$) as

$$\Pi_i^- \leq \Pi_i \geq \Pi_i^+,$$

where

$$\Pi_i^- = \Pi_i - \left[B\Pi_i(1-\Pi_i)/n\right]^{1/2} \tag{5.1}$$

$$\Pi_i^+ = \Pi_i + \left[B\Pi_i(1-\Pi_i)/n\right]^{1/2} \tag{5.2}$$

and B is the upper $(\alpha/k) \times$ 100th percentile of the χ^2 distribution with 1 degree of freedom. These equations are based on Goodman's (1965) procedure for simultaneous confidence interval estimation.

Examining these equations (Equations 5.1 and 5.2), we see that $[\Pi_i(1-\Pi_i)/n]^{1/2}$ is the standard deviation for the ith cell of the multinomial population. Also, it is important to realize that each marginal probability mass function is binomially distributed. If N is the total population size, then using the finite population correction (fpc) factor and the variance for each Π_i (from Cochran, 1977), the approximate confidence bounds are:

$$\Pi_i^- = \Pi_i - \left[B(N-n)\Pi_i(1-\Pi_i)/(N-1)n\right]^{1/2} \tag{5.3}$$

$$\Pi_i^+ = \Pi_i + \left[B(N-n)\Pi_i(1-\Pi_i)/(N-1)n\right]^{1/2} \tag{5.4}$$

Note as $N \to \infty$, Equations 5.3 and 5.4 converge to Equations 5.1 and 5.2, respectively.

Next, in order to determine the required sample size, the precision for each parameter in the multinomial population must be specified. If the absolute precision for each cell is set to b_i, then Equations 5.1 and 5.2 become

$$\Pi_i - b_i = \Pi_i - \left[B\Pi_i(1-\Pi_i)/n\right]^{1/2} \tag{5.5}$$

$$\Pi_i + b_i = \Pi_i + \left[B\Pi_i(1-\Pi_i)/n\right]^{1/2} \tag{5.6}$$

respectively. Similar results are obtained when the fpc is included. Equations 5.5 and 5.6 can be rearranged to solve for b_i (the absolute precision of the sample)

$$b_i = \left[B\Pi_i(1-\Pi_i)/n\right]^{1/2} \tag{5.7}$$

Then by squaring Equation 5.7 and solving for n, the result is:

$$n = B\Pi_i(1-\Pi_i)/b_i^2 \tag{5.8}$$

or, using the fpc,

$$n = BN\Pi_i(1-\Pi_i)/\left[b_i^2(N-1) + B\Pi_i(1-\Pi_i)\right] \tag{5.9}$$

Therefore, one should make k calculations to determine the sample size, one for each pair (b_i, Π_i), $i = 1, \ldots, k$, and select the largest n as the desired sample size. As functions of Π_i and b_i, Equations 5.8 and 5.9 show that n increases as $\Pi_i \to 1/2$ or $b_i \to 0$.

In rare cases, a relative precision b_i' could be specified for each cell in the error matrix and not just each category. Here $b_i = b_i' \; \Pi_i$. Substituting this into Equation 5.8 gives

$$n = B(1 - \Pi_i)/\Pi_i \, b_i'^2 \tag{5.10}$$

A similar sample size calculation including the fpc can be computed as before.

Here again, one should make k calculations, one for each pair (b_i', Π_i), $i = 1, \ldots, k$. The largest n computed is selected as the desired sample size. As $\Pi_i \rightarrow 1/2$ or $b_i' \rightarrow 0$ the sample size increases according to Equation 5.10. If $b_i' = b'$ for all i, then the largest sample size is $n = B(1 - \Pi)/\Pi b'^2$, where $\Pi = \min(\Pi_1, \ldots, \Pi_k)$.

In the majority of cases for assessing the accuracy of remotely sensed data, an absolute precision is set for the entire classification and not each category or each cell. Therefore, $b_i = b$ and the only sample size calculation required is for the Π_i closest to 1/2. If there is no prior knowledge about the values of the Π_i' s, a "worst-case" calculation of sample size can be made assuming some Π_i= 1/2 and $b_i = b$ for i = 1, …, k. In this worst-case scenario, the sample size required to generate a valid error matrix can be obtained from this simple equation as follows:

$$n = B/4b^2.$$

This approach can be made much clearer with a numerical example. First, let us look at an example using the full equation (Equation 5.8) and then at the corresponding sample size using the worst-case or conservative sample size equation. Assume that there are eight categories in our classification scheme (k = 8), that the desired confidence level is 95%, the desired precision is 5%, and that this particular class makes up 30% of the map area (Π_i = 30%). The value for B must be determined from a *chi*-square table with 1 degree of freedom and 1 – α/k. In this case, the appropriate value for B is $\chi^2_{(1,0.99375)} = 7.568$. Therefore, the calculation of the sample size is as follows:

$$n = B\Pi_i(1 - \Pi_i)/b_i^2$$

$$n = 7.568(0.30)(1\text{–}0.30)/(0.05)^2$$
$$n = 1.58928/0.0025$$
$$n = 636$$

A total of 636 samples should be taken to adequately fill an error matrix or approximately 80 samples per class given that there were 8 classes in this map.

If the simplified worst-case scenario equation is used, then the class proportion is assumed to be 50% and the calculation is as follows:

$$n = B/4b^2$$
$$n = 7.568/4(0.05)^2$$
$$n = 7.568/0.01 = 757$$

In this worst-case scenario, approximately 95 samples per class, or 757 total samples would be required.

If the confidence interval is relaxed from 95 to 85%, the required sample sizes decrease. In the earlier example, the new appropriate value for B would be $\chi^2_{(1,0.98125)} = 5.695$ and the total samples required would be 478 and 570 for the complete equation and the worst case scenario, respectively.

HOW SHOULD THE SAMPLES BE CHOSEN?

In addition to the considerations already discussed, the choice and distribution of samples, or sampling scheme, is an important part of any accuracy assessment. Selection of the proper scheme is critical to generating an error matrix that is representative of the entire map. First, to arrive at valid conclusions about a map's accuracy, the sample must be selected without bias. Failure to meet this important criterion affects the validity of any further analysis performed because the resulting error matrix may over- or underestimate the true accuracy. Second, further data analysis will depend on which sampling scheme is selected. Different sampling schemes assume different sampling models, and consequently, different variance equations to compute the required accuracy methods. Finally, the sampling scheme will determine the distribution of samples across the landscape, which will significantly affect accuracy assessment costs.

Sampling Schemes

Many researchers have expressed opinions about the proper sampling scheme to use (e.g., Hord and Brooner, 1976; Rhode, 1978; Ginevan, 1979; Fitzpatrick-Lins, 1981; and Stehman, 1992). These opinions vary greatly among researchers and include everything from simple random sampling to a scheme called stratified, systematic, unaligned sampling.

There are five common sampling schemes that have been applied for collecting reference data: (1) simple random sampling, (2) systematic sampling, (3) stratified random sampling, (4) cluster sampling, and (5) stratified, systematic, unaligned sampling. In a simple random sample, each sample unit in the study area has an equal chance of being selected. In most cases, a random number generator is used to pick random x, y coordinates to identify samples to be collected. The main advantage of simple random sampling is the good statistical properties that result from the random selection of samples (i.e., it results in the unbiased selection of samples).

Systematic sampling is a method in which the sample units are selected at some specified and regular interval over the study area. In most cases, the first sample is randomly selected and each successive sample is taken at some specified interval thereafter. The major advantage of systematic sampling is the ease in sampling somewhat uniformly over the entire study area.

Stratified random sampling is similar to simple random sampling; however, some prior knowledge about the study area is used to divide the area into groups or strata and then each stratum is randomly sampled. In the case of accuracy assessment, the map has been stratified into map classes. The major advantage of stratified random sampling is that all strata (i.e., map classes), no matter how small, will be included in the sample. This factor is especially important in making sure that sufficient samples are taken in rare but important map classes.

In addition to the sampling schemes already discussed, cluster sampling has also been frequently used in assessing the accuracy of maps from remotely sensed data, especially to collect information on many samples quickly. There are clear advantages to collecting a number of sample units in close proximity to one another. However, cluster sampling must be used intelligently and with great care. Simply taking a large number of sample units (whether they be pixels, a cluster of pixels, or polygons) together is not a valid method of collecting data, because each sample unit is not independent of the other and adds very little additional information. Congalton (1988b), looking at single pixels, recommended that no clusters larger than 10 pixels and certainly not larger than 25 pixels be used because each pixel beyond these cluster sizes did not add further information.

Finally, stratified, systematic, unaligned sampling attempts to combine the advantages of randomness and stratification with the ease of a systematic sample, without falling into the pitfalls of periodicity common to systematic sampling. This method is a combined approach that introduces more randomness than just a random start within each stratum.

Sampling Scheme Considerations

Congalton (1988b) performed sampling simulations on three spatially diverse areas (see Figure 5.5) using all five of these sampling schemes and concluded that in all cases, simple random and stratified random sampling provided satisfactory results.

Simple random sampling allows reference data to be collected simultaneously for both training and assessment. However, it is not always appropriate, because it tends to undersample rarely occurring, but possibly very important, map categories unless the sample size is significantly increased. For this reason, stratified random sampling, in which a minimum number of samples are selected from each stratum (i.e., map category), is often recommended. However, stratified random sampling can be impractical because stratified random samples can only be selected after the map has been completed (i.e., when the location of the strata is known). This limits the accuracy assessment data to being collected late in the project instead of in conjunction with the training data collection, thereby increasing the costs of the project. In addition, in some projects the time between the project beginning and the accuracy assessment may be so long as to cause temporal problems in collecting ground reference data. In other words, the ground may change (e.g., the crop may have been harvested) between the time the project is started and the accuracy assessment is begun.

The concept of randomness is a central issue when performing almost any statistical analysis because a random sample is one in which each member of the population has an equal and independent chance of being selected. Therefore, a random sample ensures that the samples will be chosen without bias. If in-office manual interpretation is used to label reference samples, then random sampling is feasible because access to the samples will not be a problem. However, a subset of the sample should be visited on the ground to verify the accuracy of the interpretation.

Despite the nice statistical properties of random sampling, access in the field to random sample units can often be problematic because many of the samples will be difficult to locate. Locked gates, fences, travel distances, and rugged terrain all combine to make random field sampling extremely costly and difficult. In forested and other wildland environments, randomly selected samples may be totally inaccessible except by helicopter. The cost of getting to each of the randomly located samples can be more than the cost of the rest of the entire mapping effort.

Obviously, one cannot spend the majority of project resources collecting accuracy assessment reference data. Instead, some balance must be struck. Often, some combination of random and systematic sampling provides the best balance between statistical validity and practical application. Such a system may employ systematic or simple random sampling to collect some assessment data early in a project, and stratified random sampling within strata after the classification is completed to ensure that enough samples were collected for each category and to minimize any periodicity in the data. However, results of Congalton (1988a) showed that periodicity in the errors, as measured by the autocorrelation analysis, could make the use of systematic sampling risky for accuracy assessment.

An example of a combined approach could include a systematic sample tied to existing aerial photography with sample selection based on the center of every *n*th photo. Sample choices based on flight lines should not be highly correlated with a factor determining land cover unless the flight lines were aligned with a landscape feature. Choice of the number of samples per photo and the sampling interval between photos would depend on the size of the area to map and the number of samples to collect. This systematic sample would ensure that the entire mapped area gets covered.

However, rarely occurring map classes will probably be undersampled. It may be necessary to combine this approach with stratified random sample when the map is completed to augment the underrepresented map categories. It may be practical to limit the stratified random field sample selection within some realistic distance of the roads. However, care must be taken because roads tend to occur on flatter areas and in valleys along streams, which will bias sample selection to land cover likely to exist there; so steps must be taken to mitigate these factors so that the most representative sample can be achieved. This type of combined approach minimizes the resources used and obtains the maximum information possible. Still, the statistical complexities of such a combination cannot be neglected. Again, a balance is desirable.

Finally, some analytic techniques assume that certain sampling schemes were used to obtain the data. For example, use of the Kappa analysis for comparing error matrices (see Chapter 7 for details of this analysis technique) assumes a multinomial sampling model. Only simple random sampling completely satisfies this assumption. If another sampling scheme or combination of sampling schemes is used, then it may be necessary to compute the appropriate variance equations for performing the Kappa analysis or other similar technique. The effects (i.e., bias) of using another of the sampling schemes discussed here and not computing the appropriate variances are unknown.

An interesting project would be to test the effect on the Kappa analysis of using a sampling scheme other than simple random sampling. If the effect is found to be small, then the scheme may be appropriate to use, subject to the conditions discussed earlier. If the effect is found to be large, then that sampling scheme should not be used to perform the Kappa analysis. If that scheme is to be used, then the appropriate correction to the variance equation must be applied. Stehman (1992) has done such an analysis for two sampling schemes (simple random sampling and systematic sampling). His analysis shows that the effect on the Kappa analysis of using systematic sampling is negligible. This result adds further credence to the idea of using a combined systematic initial sample followed by a random sample to fill in the gaps.

Table 5.1 presents a summary of the pros and cons of the different possible accuracy assessment sampling schemes.

TABLE 5.1
A Summary of the Pros and Cons of Various Accuracy Assessment Sampling Schemes

Sampling Scheme	Pros	Cons
Random	Unbiased sample selection. Excellent statistical properties.	Expensive, especially for fieldwork. Does not ensure that enough samples will be taken in each class. Does not ensure good distribution of samples across the landscape.
Stratified Random	Unbiased sample selection. Ensures adequate sample in each class because a minimum number of samples is selected from each stratum (class).	Requires prior knowledge about the distribution of map classes so that strata can be developed. Expensive, especially for fieldwork. Often difficult to find enough samples in rare map classes. Does not ensure good distribution of samples across the landscape.
Systematic	Easy to implement. Less expensive than random sampling. Ensures good distribution of samples across the landscape.	Can be biased if sampling pattern is correlated with a landscape pattern (periodicity). Weaker statistically, as each sample unit does not have equal probability of selection.
Cluster	Least expensive as samples are close to one another, reducing travel time in the field and/or set up time in the office.	Can be impacted by spatial autocorrelation, which results in the samples not being independent. If the samples are not independent from one another, then they are not distinct samples, and more independent samples must be taken.

FINAL CONSIDERATIONS

Because of the many assumptions required for statistical analysis, a few researchers have concluded that some sampling designs can be used for descriptive techniques and others for analytical techniques. However, this conclusion seems quite impractical. Accuracy assessment is expensive, and no one is going to collect data for only descriptive use. Eventually, someone will use that matrix for some analytical technique. It is best to pay close attention to both the practical limitations and the statistical requirements when performing any accuracy assessment.

6 Reference Data Collection

Collection of the reference data for use in an accuracy assessment is a key component of any assessment. Failure to collect appropriate reference data produces erroneous results, dooming the assessment. The collection of accuracy assessment data requires completing the following three steps while considering both the reference data being collected and the map being assessed:

- First, the accuracy assessment sample sites must be accurately located both on the reference source and on the map. This can be a relatively simple task in an urban area, or a far more difficult one in a wildland area where few recognizable landmarks exist. While the arrival of GPS has greatly increased our ability to locate accuracy sites, it is still possible to misidentify the location of a site.
- Next, the sample unit must be delineated. Sample units must represent exactly the same area on both the reference data and the map. Usually they are delineated once, either on the reference source data or on the map, and then transferred to the other. However, if the source of the reference data is not accurately coregistered to the map being assessed, slivers of unaligned assessment sites may be created that can greatly confuse the assessment.
- Finally, the reference and map labels must be assigned to each sample unit based on the map classification scheme. The reference labels may be collected from a variety of sources, and may be captured either through observation or measurement.

Serious oversights and problems can arise at each step of data collection. To adequately assess the accuracy of the remotely sensed classification, each step must be implemented correctly on each and every sample. If the reference labels are inaccurate, then the entire assessment becomes meaningless. Four basic considerations drive all reference data collection:

1. What should the source be for the reference data samples? Can existing maps or existing field data be used as the reference data? Should the information be collected from remotely sensed data or are new field visits required?
2. What type of information should be collected for each sample? Should measurements be taken or are observations adequate?

3. When should the reference data be collected? Should it be collected during initial field investigations when the map is being made, or should it only be collected after the map is completed? What are the implications of using old data for accuracy assessment?
4. How can we ensure that the reference data are collected correctly, objectively, and consistently?

There are many methods for collecting reference data, some of which depend on making observations (qualitative assessments) and some which require detailed, quantitative measurements. Given the varied reliability, difficulty, and expense of collecting reference data, it is critical to know which of these data collection techniques are valid and which are not for any given project.

We all understand that maps are rarely 100% correct. Each remote sensing project requires trade-offs between the remotely sensed data used to create the map and the scale and level of accuracy required by the project. We accept some level of map error as a trade-off for the cost savings inherent in using remotely sensed data to create the map. However, accuracy assessment reference labels must be correct if they are to constitute a fair assessment of the map. Thus, reference labels must be derived using source data or methods that are assumed to be more reliable than those used to make the map.

WHAT SHOULD BE THE SOURCE OF THE REFERENCE DATA?

The first decision in data collection requires determining what source data will be used for the determination of reference labels. The type of source data required will depend on the complexity of the map classification scheme and the budget. It is best to keep in mind this general rule: the simpler the classification scheme, the simpler and less expensive the reference data collection.

Sometimes, previously existing maps or ground data can be used as the reference data. More often, the reference source data are newly collected information that is at least one level more accurate than the remotely sensed data and methods used to make the map. Thus, aerial photography is often used to assess the accuracy of maps made from moderate-resolution satellite imagery (e.g., SPOT and Landsat TM), ground visits are often used to assess the accuracy of maps created from high-resolution airborne imagery, and manual image interpretation is often used to assess the accuracy of automated classification methods.

Using Existing versus Newly Collected Data

When a new map is produced, the usual first reaction may be to compare the map to some existing source of information about the mapped area. Using previously collected ground information or existing maps for accuracy assessment is tempting because of the cost savings resulting from avoiding new data collection. While this can be a valuable qualitative tool, existing data are rarely acceptable for accuracy assessment because:

1. The classification schemes employed to create existing maps usually differ from the one being used to create the new map. Comparisons between the two maps can result in the error matrix expressing merely differences between the reference data and the map data classification schemes, rather than map error. Developing a crosswalk that specifically translates the map classification scheme to the new map classification scheme can sometimes solve this problem. However, this method rarely produces a perfect crosswalk and, therefore, some error is unavoidable.
2. Existing data are older than those being used to create the new map. Changes on the landscape (e.g., fire, urban development, etc.) will not be reflected in the existing data. Therefore, differences in the error matrix caused by these changes will incorrectly be assumed to be caused by map error.
3. Errors in existing maps are rarely known (it is unlikely that an accuracy assessment has been performed on the existing map). Often, the differences caused by existing map errors are then blamed on the new map, thereby wrongly lowering the new map's accuracy.
4. Existing field inventory data usually were collected for a purpose other than accuracy assessment. Often the size of the inventory plot is too small (e.g., a 1 m ecology site cannot be used to assess a map with a 4 m minimum mapping unit) or the measurements made on the plots cannot be transformed into measurements that are useful for the accuracy assessment.

If existing information is the only available source of reference data, then consideration should be given to not performing quantitative accuracy assessment. Instead, a qualitative comparison of the new map and existing map or field data should be performed, and the differences between the two should be identified and scrutinized. If a quantitative assessment is performed with existing data, it is vital to document the issues with the reference data so as to allow the potential user of the map to understand the limitations of such an assessment.

Photos versus Ground

If new data are to be collected for reference samples, then a choice must be made between using ground visits and using aerial imagery, video, or reconnaissance as the source of the reference data. The accuracy assessment professional must assess the reliability of each data type for obtaining an accurate reference site label.

Simple classification schemes with a few general classes can often be reliably assessed from air reconnaissance or interpretation of aerial imagery or video. As the level of detail in the map classification scheme increases, so does the complexity of the reference data collection. Eventually, even the very largest-scale photography cannot provide valid reference data. Instead, the data must be collected on the ground.

In some situations, the use of image interpretation or videography for generating reference data may not be appropriate. For example, aerial photo interpretation is often used as reference data for assessing a land cover map generated from satellite imagery such as Landsat Thematic Mapper. The photo interpretation is assumed to be correct because the photos have greater spatial resolution than the satellite imagery and because photo interpretation has become a time-honored skill that is accepted as accurate. Unfortunately, errors do occur in photo interpretation and air recognizance depending on the skill of the photo interpreter and the level of detail required by the classification system. Inappropriately using photo interpretation as reference data could severely impact the conclusions regarding the accuracy of the satellite-based land cover map. In other words, one may conclude that the satellite-based map is of poor accuracy when actually it is the photo interpretation that is in error.

In such situations, actual ground visitation may be the only reliable method of data collection. At the very least, a subset of data should be collected on the ground and compared with the airborne data to verify the reliability of the reference labels interpreted from airborne imagery. Even if the majority of reference labels will come from image interpretation or videography, it is critical that a subsample of these areas be visited in the field to verify the reliability of the interpretation. Much work is yet to be done to determine the proper level of effort and collection techniques necessary to provide this vital information. When the labels developed from image interpretation and the ground begin to disagree regularly, it may be time to switch to ground-based reference data collection. However, the collection of ground reference data is extremely expensive and, therefore, the collection effort must be sufficient to meet the needs of the accuracy assessment while being efficient enough to stay within the budget.

For example, Biging et al. (1991) compared photo interpretation to ground measurements for characterizing forest structure. These characteristics included forest species, tree size class, and crown closure, which were photo-interpreted by a number of expert photo interpreters. The ground data used for comparison were a series of measurements made in a sufficient number of ground plots to characterize each forest polygon (i.e., stand). The results showed that the overall accuracy of photo interpretation of species ranged in accuracy between 75 and 85%. The accuracy of size class was around 75%, and the accuracy of crown closure was less than 40%. This study reinforces the need to be careful if assuming that the results of the photo interpretation are sufficient for use as reference data in an accuracy assessment.

HOW SHOULD THE REFERENCE DATA BE COLLECTED?

The next decision in reference data collection involves deciding how information will be collected from the source data to obtain a reliable label for each reference site. Reference data must be labeled using the same classification scheme that was used to make the map. In many instances, simple observations/interpretations are sufficient for labeling a reference sample. In other cases, observation is not adequate, and actual measurements in the field are required.

The purpose of collecting reference data for a sample site is to derive the "correct" reference label for the site for comparison to the map label. Often, the reference label can be obtained by merely observing the site from an airplane, car, or aerial imagery. For example, in most cases, a golf course can be accurately identified through observation from quite a distance away.

Whether or not accuracy assessment reference data should be obtained from observations or measurements will be determined by the complexity of the landscape, the detail of the classification scheme, the required precision of the accuracy assessment, and the project budget. Reference data for simple classification schemes that distinguish homogeneous land cover or use types from one another (e.g., water versus agriculture) usually can be obtained from observations and/or estimations either on the ground or from larger-scale remotely sensed data. For example, distinguishing conifer forest from an agricultural field from a golf course can be determined from observation alone. Collecting reference data may be as simple as looking at aerial imagery or observing sites on the ground.

However, complex classification schemes may require some measurements to determine the precise (i.e., nonvarying) reference site labels. For example, a more complex forest classification scheme may involve collecting reference data for tree size class (related to the diameter of the trunk). Tree size class is important both as a determinant of endangered species habitat and as a measurement of wood product merchantability. Size class can be ocularly estimated on aerial imagery and on the ground. However, different individuals may produce different estimations, introducing variability into the observation. Not only will this variation exist between individuals, but also within the same individual. The same observer may see things differently depending on whether it is Monday or Friday; or whether it is sunny or raining; or especially depending on how much coffee he or she has consumed. To avoid variability in human estimation, size class can be measured, but a great many trees will need to be measured to precisely estimate the size class for each sample unit. In such instances, the accuracy assessment professional must decide whether to require measurement (which can be time consuming and expensive) or to accept the variation inherent in human estimations.

Whether or not measurements are required depends on the level of precision required by the map users and on the project budget. For example, information on spotted owl habitat requirements indicates that the owls prefer older, multistoried stands that include large trees. "Large" in this context is relative, and precise measurements of trees will probably not be needed as long as the map accurately distinguishes between stands of single-storied small trees and multistoried large trees. In contrast, many wood product mills can only accept trees within a specific size class. Trees that are one inch smaller or larger than the prescribed range cannot be accepted by the machinery in the mill. In this case, measurements will probably be required.

Observer variability is especially evident in estimates of vegetation cover that cannot be precisely measured from aerial imagery. In addition, ground verification of aerial estimates of vegetative cover is problematic as estimates of cover from the ground (i.e., below tree canopies) are fundamentally different from estimates made from above the canopy. Spurr (1960) asserts that forest crown closure is overestimated from aerial imagery and underestimated from the ground. Therefore, using ground estimates as reference data for aerial cover estimates can be like comparing apples and oranges.

The trade-offs inherent between observation and measurement are exemplified in a pilot study conducted to determine the level of effort needed to collect appropriate ground reference data for use in forest inventory. The objective of this study was to determine if visual calls made by trained experts walking into forest polygons are sufficient or whether actual ground measurements need to be made. There are obviously many factors influencing the accuracy of ground data collection, including the complexity of the vegetation itself. A variety of vegetation complexities were represented in this study. The results are enlightening to those remote sensing specialists who routinely collect forest ground data only by visual observation. The pilot study was part of a larger project aimed at developing the use of digital remotely sensed data for commercial forest inventory (Biging and Congalton, 1989).

Commercial forest inventory involves much more than creating a land cover map derived from digital remotely sensed data. Often the map is used only to stratify the landscape; a field inventory is conducted on the ground to determine tree volume statistics for each type of stand of trees. A complete inventory requires that the forest type, the size class, and the crown closure of a forested area be known in order to determine the volume of the timber in that area. If a single species dominates, the forest type is commonly named by that species (Eyre, 1980). However, if a combination of species is present, then a mixed label is used (e.g., the mixed conifer type). The size of the tree is measured by the diameter of the tree at 4.5 ft above the ground (i.e., diameter at breast height, DBH) and then is divided into size classes such as poles, small saw timber, and large saw timber. This measure is obviously important because large-diameter trees contain more volume of high-quality wood (i.e., valuable timber) than small-diameter trees. Crown closure, as measured by the amount of ground area the tree crowns occupy (canopy closure), is also an important indicator of tree size and numbers. Therefore, in this pilot study, it was necessary to collect ground reference data not only on tree species/type, but on crown closure and size class as well.

Ground reference data were collected using two approaches. In the first approach, a field crew of four entered a forest stand (i.e., polygon), observed the vegetation, and came to a consensus regarding a visual call of dominant species/type, size class of the dominant species, crown closure of the dominant size class, and crown closure of all tree species combined. Dominance was defined as the species or type comprising the majority of forest volume. In the second approach, measurements were conducted on a fixed-radius plot to record the species, DBH, and height of each tree falling within the plot. A minimum of two plots (1/10th or 1/20th acre) were measured for each forested polygon. Because of the difficulty of making all the required measurements (precise location and crown width for each tree in the plot) to estimate crown closure on the plot, an approach using transects was developed to determine crown closure. A minimum of four 100-ft-long transects randomly located within the polygon were used to collect crown closure information. The percentage of crown closure was determined by the presence or absence of tree crown at 1-ft intervals along the transects. All the measurements were input into a computer program that categorized the results as dominant species/type, the size class of the dominant species/type, the crown closure of the dominant size class, and the crown closure of all tree species for each forested area. The results of the two approaches were compared by using an error matrix.

Table 6.1 shows the results of field measurement versus visual call as expressed in an error matrix for the dominant species. This table indicates that species can be fairly well determined from a visual call because there is strong agreement between the field measurements and the visual call. Of course, this conclusion requires one to assume that the field measurements are a better measure of ground reference data, a reasonable assumption in this case. Therefore, ground reference data collection of species information can be maximized using visual calls, and field measurements appear to be unnecessary.

TABLE 6.1
Error Matrix for the Field Measurement versus Visual Call for Dominant Species

		Field Measurement							Row
		TF	MC	LP	DF	PP	PD	OAK	Total
Visual Call	TF	14	0	0	0	0	0	0	14
	MC	0	10	0	0	0	2	0	12
	LP	0	0	1	0	0	0	0	1
	DF	0	1	0	8	0	0	0	9
	PP	1	1	0	0	0	0	0	2
	PD	0	0	0	1	0	0	0	1
	OAK	0	0	0	0	0	0	0	0
	Column Total	15	12	1	9	0	2	0	39

Species

TF = true fir
MC = mixed conifer
LP = lodgepole pine
DF = Douglas fir
PP = Ponderosa pine
PD = PP and DF
OAK = oaks

OVERALL ACCURACY
= 33/39 = 85%

PRODUCER'S ACCURACY			USER'S ACCURACY		
TF	= 14/15	= 93%	TF	= 14/14	= 100%
MC	= 10/12	= 83%	MC	= 10/12	= 83%
LP	= 1/1	= 100%	LP	= 1/1	= 100%
DF	= 8/9	= 89%	DF	= 8/9	= 89%
PP	= 0/0	= —	PP	= 0/2	= 0%
PD	= 0/2	= 0%	PD	= 0/1	= 0%
OAK	= 0/0	= —	OAK	= 0/0	= —

Source: Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton R. and G. Biging. 1992. A pilot study evaluating ground reference data collection efforts for use in forest inventory. *Photogrammetric Engineering and Remote Sensing*. 58(12): 1669–1671.

TABLE 6.2
Error Matrix for the Field Measurement versus Visual Call for Dominant Size Class

Visual Call	Field Measurement 1	2	3	4	Row Total
1	1	0	0	0	1
2	1	3	1	0	5
3	0	0	17	5	22
4	0	0	1	11	12
Column Total	2	3	19	16	40

Size Classes

1 = 0–5″ dbh
2 = 5–12″ dbh
3 = 12–24″ dbh
4 = >24″ dbh

OVERALL ACCURACY
= 32/40 = 80%

PRODUCER'S ACCURACY		USER'S ACCURACY	
1 = 1/2	= 50%	1 = 1/1	= 100%
2 = 3/3	= 100%	2 = 3/5	= 60%
3 = 17/19	= 89%	3 = 17/22	= 77%
4 = 11/16	= 69%	4 = 11/12	= 92%

Source: Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton R. and G. Biging. 1992. A pilot study evaluating ground reference data collection efforts for use in forest inventory. *Photogrammetric Engineering and Remote Sensing*. 58(12): 1669–1671.

Table 6.2 presents the results of comparing the two ground reference data collection approaches for the dominant size class. As in species, the overall agreement is relatively high, with most of the confusion occurring between the larger classes. The greatest inaccuracies result from visually classifying the dominant size class (i.e., the one with the most volume) as size class three (12–24 in. DBH) when, in fact, size class four (>24 in. DBH) trees contained the most volume. This visual classification error is easy to understand. Tree volume is directly related to the square of DBH. There are numerous cases when a small number of large trees contribute the majority of the volume in the stand, while there may be many more medium-sized trees present. The dichotomy between prevalence of medium-sized trees, but dominance in volume by a small number of large trees can be difficult to assess visually. It is likely that researchers and practitioners would confuse these classes in cases where the size class with the majority of volume was not readily evident. In cases like this, simply improving one's ability to visually estimate diameter would not improve one's ability to classify size class. The ability to weigh numbers and sizes to estimate volume requires considerable experience and would certainly require making plot and tree measurements to gain and retain this ability.

Tables 6.3 and 6.4 show the results of comparing the two collection approaches for crown closure. Table 6.3 presents the crown closure of the dominant size class results, while Table 6.4 shows the results of overall crown closure. In both matrices,

TABLE 6.3
Error Matrix for the Field Measurement versus Visual Call for Density (Crown Closure) of the Dominant Species

		Field Measurement				Row
		O	L	M	D	Total
Visual Call	O	10	8	3	0	21
	L	2	8	1	1	12
	M	0	3	1	1	5
	D	0	1	0	0	1
Column Total		12	20	5	2	39

Density Classes

O = Open (0–10% crown closure)
L = Low (11–25% crown closure)
M = Medium (26–75% crown closure)
D = Dense (> 75% crown closure)

OVERALL ACCURACY
= 19/39 = 49%

PRODUCER'S ACCURACY	USER'S ACCURACY
O = 10/12 = 83%	O = 10/21 = 48%
L = 8/20 = 40%	L = 8/12 = 67%
M = 1/5 = 20%	M = 1/5 = 20%
D = 0/2 = 0%	D = 0/1 = 0%

Source: Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton R. and G. Biging. 1992. A pilot study evaluating ground reference data collection efforts for use in forest inventory. *Photogrammetric Engineering and Remote Sensing*. 58(12): 1669–1671.

there is very low agreement (46–49%) between the observed estimate and the field measurements. Therefore, it appears that field measurements may be necessary to obtain precise measures of crown closure and that visual calls, although less expensive and quicker, may vary at an unacceptable level.

In conclusion, it must be emphasized that this is only a small pilot study. Further work needs to be conducted in this area to evaluate ground reference data collection methods and to include the validation of aerial methods (i.e., image interpretation and videography). The results demonstrate that making visual calls of species are relatively easy and accurate, except where many species occur simultaneously. Size class is more difficult to assess than species, because of the implicit need to estimate the size class with the majority of volume. Crown closure is by far the toughest to determine. It is most dependent on where one is standing when the call is made. Field measurements, such as the transects used in this study, provide an alternative means of determining crown closure. This study has shown that at least some ground data must be collected using measurements, and it has suggested that a multilevel effort may result in the most efficient and practical method for collection of ground reference data.

Table 6.5 presents the pros and cons of the different sources of reference data.

TABLE 6.4
Error Matrix for the Field Measurement versus Visual Call for Overall Density (Crown Closure)

		Field Measurement				Row
		O	L	M	D	Total
Visual Call	O	0	1	1	0	2
	L	1	3	7	0	11
	M	0	0	8	10	18
	D	0	0	0	6	6
	Column Total	1	4	16	16	37

Density Classes

O = Open
L = Low
M = Medium
D = Dense

OVERALL ACCURACY
= 17/37 = 46%

PRODUCER'S ACCURACY	USER'S ACCURACY
O = 0/1 = 0%	O = 0/2 = 0%
L = 3/4 = 75%	L = 3/11 = 27%
M = 8/16 = 50%	M = 8/18 = 44%
D = 6/16 = 38%	D = 6/6 = 100%

Source: Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton R. and G. Biging. 1992. A pilot study evaluating ground reference data collection efforts for use in forest inventory. *Photogrammetric Engineering and Remote Sensing*. 58(12): 1669–1671.

WHEN SHOULD THE REFERENCE DATA BE COLLECTED?

The world's landscape is constantly changing. If change occurs between the date of capture of the remotely sensed data used to create a map and the date of the reference data collection, accuracy assessment reference sample labels may be affected. When a crop is harvested, a wetland drained, or a field developed into a shopping mall, the error matrix may show a difference between the map and the reference label that is not caused by map error, but rather by landscape change.

As previously noted, aerial or high-resolution satellite imagery is often used as reference source data for accuracy assessment of forest type maps created from Landsat TM or SPOT satellite data. Because aerial imagery is relatively expensive to obtain, existing imagery that may be 5–10 years old may be used. If an area has changed because of fire, disease, harvesting, or growth, the resulting reference labels in the changed areas will be incorrect. Harvests and fire are clearly visible on most satellite imagery, making it possible to detect the changes by looking at the imagery.[†]

[†] Using satellite imagery to correct the reference information collected from the photos seems a little convoluted since the photos are supposedly being used to assess the accuracy of a map produced from the imagery.

TABLE 6.5
Comparison of Sources of Reference Data

Source of Reference Data	Pros	Cons
Existing maps/data	Least expensive and quickest.	Can be out of date if change has occurred on the landscape. Must ensure that the minimum mapping unit and classification scheme used to label the existing data are identical to the scheme used to label the map.
New office interpreted data from remote sensing	Less expensive and time consuming than field collected data. Provides the same perspective as the remotely sensed data used to make the map (i.e., view from above).	Less accurate for vegetation species identification than field-collected data. Can be out of date if change has occurred on the landscape since the capture of the remotely sensed reference data.
New field collected data	More accurate for vegetation species identification.	Most expensive. Does not offer the same perspective as captured by the remotely sensed data (i.e., view from below versus view from above). Often difficult to establish because of terrain or access restrictions.

However, stand growth and partial defoliation from disease or pests are not as readily observable on the imagery, making the use of older photos especially problematic in the northwest and southeast portions of the U.S., where trees can grow through several size classes in a 10-year period.

In general, accuracy assessment reference data should be collected as close as possible to the date of the collection of the remotely sensed data used to make the map. However, trade-offs may need to be made between the timeliness of the data collection and the need to use the resulting map to stratify the accuracy assessment sample. In most, if not all, remote sensing mapping projects, it is necessary to go to the field to become familiar with what causes variation in the classes to be mapped, to calibrate the eye of the image analyst, and to collect information for training the classifier (i.e., supervised classification or object-oriented classification) or to aid in labeling the clusters (i.e., unsupervised classification). If the reference data for accuracy assessment can be collected independently, but simultaneously, during this trip,

TABLE 6.6
Error Matrix Showing Number of Samples in Each Crop Type

		REFERENCE DATA															
		A	**C**	**SG**	**CN**	**L**	**M**	**BG**	**CS**	**T**	**SU**	**O**	**CR**	**F**	**D**	**S**	**Total**
MAP DATA	**A**	157		8				3						3			171
	C		1			1	1										3
	SG	3		163		6						12	2	1			187
	CN																0
	L			4		3					1		1				9
	M						5							1			6
	BG	1						10									11
	CS								69								69
	T																0
	SU																0
	O			1		3						7					11
	CR												2				2
	F													224			224
	D														11		11
	S																0
	Total	161	1	176	0	13	6	13	69	0	1	19	5	229	11	0	704

LEGEND		**Producer's Accuracy**	**User's Accuracy**
A = Alfalfa	A	98%	92%
C = Cotton	C	100%	33%
SG = Small Grains	SG	93%	87%
CN = Corn	CN	—	—
L = Lettuce	L	23%	33%
M = Melons	M	83%	83%
BG = Bermuda Grass	BG	77%	91%
CS = Citrus	CS	100%	100%
T = Tomatoes	T	—	—
SU = Sudan Grass	SU	0%	—
O = Other Veg.	O	37%	64%
CR = Crucifers	CR	40%	100%
F = Fallow	F	98%	100%
D = Dates	D	100%	100%
S = Safflowers	S	—	—

then a second trip to the field is eliminated, saving costs and ensuring that reference data collection is occurring close to the time the remotely sensed data are captured.

However, if accuracy assessment reference data are collected at the beginning of the project before the map is generated, then it is not possible to stratify the samples by map class since the map has yet to be created. It is also not possible to have a proportional to area allocation of the samples since the total area of each map class is still unknown.

An example helps illustrate these points. The USDI Bureau of Reclamation maps the crops of the lower Colorado River region four times a year using Landsat TM data. Farmland in this region is so productive and valuable that growers plant three to four crops per year and will plow under a crop to plant a new one in response to the futures market. With so much crop change, ground data collection and accuracy assessment must occur at the same time the imagery is collected. The Bureau sends a ground data collection crew to the field for two weeks surrounding the date of image acquisition. A random number generator is used to determine the fields to be visited and the same fields are visited during each field effort, regardless of the crops being grown. Therefore, the accuracy assessment sample is random, but not stratified by crop type. As Table 6.6 illustrates, some crops are oversampled and others are undersampled each time. The Bureau believes it is more important to ensure correct crop identification than it is to ensure that enough samples are collected in rarely occurring crop types.

Table 6.7 compares and contrasts the trade-offs required when deciding when to collect reference data.

TABLE 6.7
Pros and Cons of the Timing of Reference Data Collection

When Should the Reference Data Be Collected?	Pros	Cons
When the remotely sensed data are collected	Eliminates any chance of landscape change between the date of the acquisition and the date of the reference data. Cost effective as information needed to make and assess the map are collected at the same time.	Because the map has not been made, there is no way to ensure that enough samples will be taken in each map class.
After the map has been made	Because the map has been made, it is poosible to ensure that enough samples for each map class are collected.	Can be more expensive. Introduces the possibility of landscape change occurring between the date of the map and the date of the reference data collection.

ENSURING OBJECTIVITY AND CONSISTENCY

For accuracy assessment to be useful, map users must have faith that the assessment is a realistic representation of the map's accuracy. They must believe that the assessment is objective and the results are repeatable. Maintaining the following three conditions will ensure objectivity and consistency:

1. Accuracy reference data must always be kept independent of any training data.
2. Data must be collected consistently from sample site to sample site.
3. Quality control procedures must be developed and implemented for all steps of data collection.

DATA INDEPENDENCE

It was not uncommon for early accuracy assessments to use the same information to assess the accuracy of a map as was used to create the map. This unacceptable procedure obviously violates all assumptions of independence and biases the assessment in the favor of the map. Independence of the reference data can be assured in one of two ways. First, the reference and training data collection can be performed at a completely different time and/or by different people. Collecting information at different times is expensive and can introduce landscape change problems as discussed earlier. Using different people can also be expensive, as more personnel need to be completely trained in the details of the project.

The second method of ensuring independence involves collecting reference and training data simultaneously, and then using a random number generator to select and remove the accuracy assessment sites from the training data set. The accuracy assessment sites are not reviewed again until it is time to perform the assessment. In both cases, accuracy assessment reference data must be kept absolutely independent (i.e., separate) of any training/labeling data, and it must not be accessible during manual map editing.

DATA COLLECTION CONSISTENCY

Data collection consistency can be ensured through personnel training and the development of objective data collection procedures. Training should occur simultaneously for all personnel at the initiation of data collection. One to three days of intensive training is often necessary and must include reference collection on numerous example sites that represent the broad array of variation between and within map classes. Trainers must ensure that reference data collection personnel are (1) applying the classification scheme correctly and (2) accurately identifying characteristics of the landscape that are inherent in the classification scheme. For example, if a classification scheme depends on the identification of plant species, then all reference data personnel must be able to accurately identify species on the reference source data. The classification scheme must use the same minimum mapping unit as was applied to create the map.

In addition to personnel training, objective data collection procedures are key to consistent data collection. The more measurement (as opposed to estimation) involved in reference data collection, the more consistent and objective the collection. However, measurement increases the cost of accuracy assessment, so most assessments rely heavily on ocular estimation. If ocular estimates are to be used, then the variance inherent in estimation must be accepted as an unavoidable part of the assessment, and some method of assessing it must be included in the assessment. Several of these methods are discussed in Chapter 9.

An important mechanism for imposing objectivity is the use of a reference data collection form to force all data collection personnel through the same collection process. The complexity of the reference data collection form will depend on the level of the complexity of the classification scheme. The form should lead the collector through a rule-based process to a definitive reference label from the classification scheme. Forms also provide a means of performing a quality assessment/quality control check on the collection process. Figure 6.1 is an example data collection form for a relatively simple classification scheme. An important portion of this form is the dichotomous key that leads data collection personnel to the land cover class label solely on the basis of the classification scheme rules.

Reference data collection forms, regardless of their complexity, have some common components. These include (1) the name of the collector and the date of the collection, (2) locational information about the site, (3) some type of table or logical progression that represents what the collector is seeing, (4) a place to fill in the actual reference label from the classification scheme, and (5) a place to describe any anomalies, any variability, or interesting findings at the site.

These days it may be more common to have the form on your laptop computer, data logger, or PDA, rather than as a piece of paper. Regardless of how the form is represented, it is vital to make use of some form to ensure objectivity.

Quality Control

Quality control is necessary at every step of data collection. Each error in data collection can translate into an incorrect indication of map accuracy. Data collection errors result in both over- and underestimations of map accuracy.

The following text discusses some of the most common quality control problems in each step of accuracy assessment data collection. Because accuracy assessment requires collecting information from both the reference source data and the map, each step involves two possible occasions of error: during collection from the map and during collection from the reference source data.

1. *Location of the accuracy assessment sample site.* It is not uncommon for accuracy assessment personnel to collect information at the wrong location because inadequate procedures were used to locate the site on either the map or the reference data. As discussed in Chapter 3, any errors in the position of either the location of the accuracy assessment sample site on the reference data or on the map will result in a thematic error. Positional accuracy cannot be ignored when conducting a thematic accuracy assessment.

Site Name: ________________________________ Site ID #: ________

Quad Name: ________________________ Satellite Scene ID #: ________

Latitude: __________ Longitude: __________ Elevation Range: ________

Method for Determining Position: ________________________________

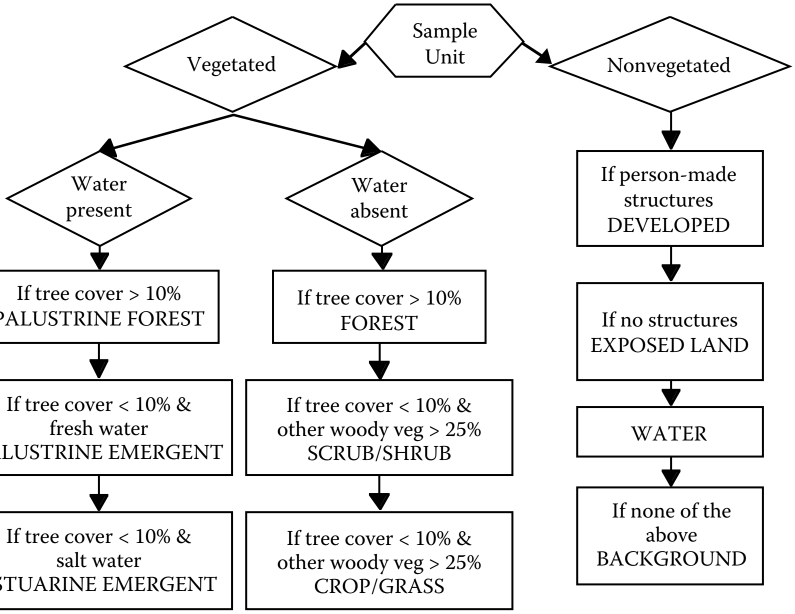
Comments on Position: ____

Comments on Weather: ____

Crew Names: ________________________ Date: ______________

CATEGORIES

Background	**Forest**	**Palustrine Forest**
Developed	**Scrub/Shrub**	**Estuarine Emergent**
Crop/Grass	**Exposed Land**	**Palustrine Emergent**
	Water	


Sample Unit
Vegetated
Nonvegetated
Water present
Water absent
If tree cover > 10% PALUSTRINE FOREST
If tree cover < 10% & fresh water PALUSTRINE EMERGENT
If tree cover < 10% & salt water ESTUARINE EMERGENT
If tree cover > 10% FOREST
If tree cover < 10% & other woody veg > 25% SCRUB/SHRUB
If tree cover < 10% & other woody veg > 25% CROP/GRASS
If person-made structures DEVELOPED
If no structures EXPOSED LAND
WATER
If none of the above BACKGROUND

Actual category as determined from flowchart: ________________________________

Comments on anomalies, variability, or interesting finding: ______________________

FIGURE 6.1 Example of a reference data collection form for a simple classification scheme.

A common method for locating accuracy assessment sites on reference aerial imagery is to view the site on the map and then "eyeball' the location onto the photos based on similar patterns of land cover and terrain in both the map and the reference data. In this situation, it is critical to provide the reference personnel with as many tools and as much information as possible to help them locate the site. GPS equipment has become critical to assuring location during fieldwork. Helpful information includes digitized flight line maps and other ancillary data such as stream, road, or ownership coverages. A laptop computer linked to a GPS and loaded with GIS software; the imagery, and ancillary data can reduce field time and increase reference location accuracy immeasurably.

Field location is always problematic, especially in wildlands (e.g., tundra, open water, wilderness areas, etc.) with few recognizable landscape characteristics. GPS is extremely helpful and should always be used to ensure the correct location of field sample sites.

2. *Sample unit delineation.* Both the reference site and the map accuracy assessment sites must represent exactly the same location. Thus, not only must the sites be properly located, they must also be delineated precisely and correctly transferred to a planimetric base. For example, if an existing map is used as the reference source, and the map was not registered correctly, then all accuracy assessment reference sites will not register with the new map being assessed, and a misalignment will occur when the reference site and the map site are compared. This was not an uncommon situation when aerial photography was used to create the existing map, and the transfer from the photo to the map was performed ocularly without the use of photogrammetric equipment.

   Another common error in accuracy assessment occurs when the reference and the map sites are in the same general location, but are of different sizes or shapes. For example, if map polygons constitute the sampling unit, and the reference data are aerial photographs, the selected polygons will often fall across two or more photos as depicted in Figure 6.2.

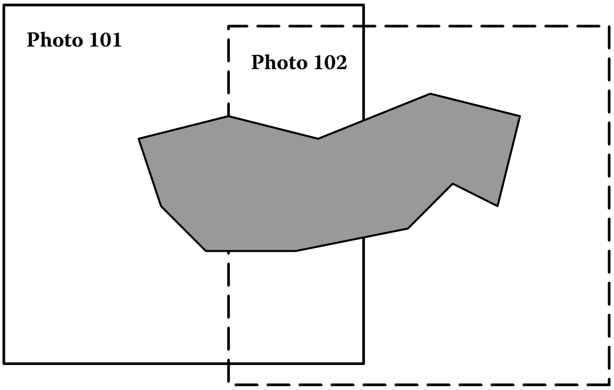


FIGURE 6.2 The shaded accuracy assessment site polygon falls across two aerial photographs.

In this case, the analyst must either collect reference data across all the photos (which can be time consuming), or the sample site must be reduced in size (though not below the mmu of the classification scheme) on both the map and the reference data so that it fits on one photo.

Also, given the need and ability to automate this entire process today, it is possible to use map coordinates to represent both the reference data samples and map locations and to automate the process of creating the error matrix such that the analyst is completely removed from the process. It happens entirely inside the computer, and only the final error matrix is revealed. While it is very appealing for this process to be automated, it is also very dangerous if left completely unverified because no visual comparison of the sample unit points is performed and none of the potential problems discovered. Historically, many of these issues would have been discovered during the manual comparison of reference sample sites to map sites during the error matrix generation. Therefore, even if using a fully automated method, it is recommended that some sample of these points be visually examined to check for errors.

3. *Data collection and data entry* are the most common sources of quality control problems in accuracy assessment. Data collection errors occur when measurements are done incorrectly, variables of the classification scheme are misidentified (e.g., species), or the classification scheme is misapplied. In addition, weak classification schemes will also create ambiguity in data collection. Unfortunately, the first indication of a weak classification system often occurs during accuracy assessment, when the map is already completed, and refinement of the classification scheme is not possible.

   Data collection errors are usually monitored by selecting a subsample of the accuracy assessment sites and having reference data on them collected simultaneously by two different personnel. Usually, the most experienced personnel are assigned to the subsample. When differences are detected, the sources of the differences are immediately identified. If data collection errors are the source of the differences, then personnel are either retrained or removed from reference data collection.

   When aerial imagery is used as the reference source data, it is critical that a ground assessment of the imagery interpretation be conducted. In addition, reference samples chosen from an existing map must also be assessed for accuracy.

   Data entry errors can be reduced by using digital data entry forms that restrict each field of the form to an allowable set of characters. Data can also be entered twice and the two data sets compared to identify differences and errors. Data entry errors also can occur when the site is digitized. Quality control must include a same-scale comparison of the digitized site to the source map.

Finally, although no reference data set may be completely accurate, it is important that the reference data have high accuracy or else the assessment is not a fair characterization of map accuracy. Therefore, it is critical that reference data collection be carefully considered in any accuracy assessment. Much work is yet to be done to determine the proper level of effort and collection techniques necessary to provide this vital information.

7 Basic Analysis Techniques

This chapter presents the basic analysis techniques needed to perform an accuracy assessment. Once an error matrix has been properly generated, any or all of the following analysis techniques can be performed. These techniques clearly demonstrate why the error matrix is such a powerful tool and should be included in any published accuracy assessment. Without having the error matrix as a starting point, none of these analysis techniques would be possible.

KAPPA

The Kappa analysis is a discrete multivariate technique used in accuracy assessment to statistically determine if one error matrix is significantly different from another (Bishop et al., 1975). The result of performing a Kappa analysis is a KHAT statistic (actually $\hat{K}$, an estimate of Kappa), which is another measure of agreement or accuracy (Cohen, 1960). This measure of agreement is based on the difference between the actual agreement in the error matrix (i.e., the agreement between the remotely sensed classification and the reference data as indicated by the major diagonal) and the chance agreement that is indicated by the row and column totals (i.e., marginals). In this way, the KHAT statistic is similar to the more familiar *chi*-square analysis.

Although this analysis technique has been in the sociology and psychology literature for many years, the method was not introduced to the remote sensing community until 1981 (Congalton, 1981) and not published in a remote sensing journal before Congalton et al. (1983). Since then numerous papers have been published recommending this technique. Consequently, the Kappa analysis has become a standard component of most every accuracy assessment (Congalton et al., 1983; Rosenfield and Fitzpatrick-Lins, 1986; Hudson and Ramm, 1987; Congalton, 1991) and is considered a required component of most image analysis software packages that include accuracy assessment procedures.

The following equations are used for computing the KHAT statistic and its variance. Let

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

be the actual agreement, and

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

with p_{i+} and p_{+j} as previously defined, the "chance agreement."

Assuming a *multinomial sampling model,* the maximum likelihood estimate of Kappa is given by

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

For computational purposes

$$\hat{K} = \frac{n \sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i+} n_{+i}}{n^2 - \sum_{i=1}^{k} n_{i+} n_{+i}}$$

with n_{ii}, n_{i+}, and n_{+i} as previously defined.

The approximate large sample variance of Kappa is computed using the Delta method as follows:

$$\hat{\text{var}}(\hat{K}) = \frac{1}{n}\left\{\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4}\right\}$$

where

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{k} n_{ii},$$

$$\theta_2 = \frac{1}{n^2}\sum_{i=1}^{k} n_{i+} n_{+i},$$

$$\theta_3 = \frac{1}{n^2}\sum_{i=1}^{k} n_{ii}(n_{i+} + n_{+i}),$$

and

$$\theta_4 = \frac{1}{n^3}\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}(n_{j+} + n_{+i})^2.$$

A KHAT value is computed for each error matrix and is a measure of how well the remotely sensed classification agrees with the reference data. Confidence intervals around the KHAT value can be computed using the approximate large sample variance and the fact that the KHAT statistic is asymptotically normally distributed. This fact also provides a means for testing the significance of the KHAT statistic for a single error matrix to determine if the agreement between the remotely sensed classification and the reference data is significantly greater than 0 (i.e., better than a random classification).

It is always satisfying to perform this test on a single matrix and confirm that your classification is meaningful and significantly better than a random classification. If it is not, you know that something has gone terribly wrong during the classification process.

Finally, there is a test to determine if two independent KHAT values, and therefore two error matrices, are significantly different. With this test, it is possible to statistically compare two analysts, the same analyst over time, two algorithms, two types of imagery, or even two dates of imagery and see which produces the higher accuracy. Both the single error matrix and paired error matrix tests of significance rely on the standard normal deviate as follows:

Let $\hat{K}_1$ and $\hat{K}_2$ denote the estimates of the Kappa statistic for error matrix #1 and #2, respectively. Also, let $\hat{\text{var}}(\hat{K}_1)$ and $\hat{\text{var}}(\hat{K}_2)$ be the corresponding estimates of the variance as computed from the appropriate equations. The test statistic for testing the significance of a single error matrix is expressed by:

$$Z = \frac{\hat{K}_1}{\sqrt{\hat{\text{var}}(\hat{K}_1)}}.$$

Z is standardized and normally distributed (i.e., standard normal deviate). Given the null hypothesis $H_O{:}K_1 = 0$, and the alternative $H_1 : K_1 \neq 0$, H_0 is rejected if $Z \geq Z_{\alpha/2}$, where $\alpha/2$ is the confidence level of the two-tailed Z test and the degrees of freedom are assumed to be ∞ (infinity).

The test statistic for testing if two independent error matrices are significantly different is expressed by:

$$Z = \frac{\left|\hat{K}_1 - \hat{K}_2\right|}{\sqrt{\hat{\text{var}}(\hat{K}_1) + \hat{\text{var}}(\hat{K}_2)}}.$$

Z is standardized and normally distributed. Given the null hypothesis $H_O{:}(K_1 - K_2) = 0$, and the alternative $H_1{:}(K_1 - K_2) \neq 0$, H_0 is rejected if $Z \geq Z_{\alpha/2}$.

It is prudent at this point to provide an actual example so that the equations and theory can come alive for the reader. The error matrix presented as an example in Table 7.1 was generated from Landsat Thematic Mapper data using an unsupervised classification approach by analyst #1. A second error matrix was generated using precisely the same imagery and same classification approach; however, the clusters were labeled by analyst #2 (Table 7.2). It is important to note that analyst #2 was not as ambitious as analyst #1, and did not collect as much accuracy assessment data.

Table 7.3 presents the results of the Kappa analysis on the individual error matrices. The KHAT values are a measure of agreement or accuracy. The values can range from +1 to −1. However, since there should be a positive correlation between

TABLE 7.1
Error Matrix Produced Using Landsat Thematic Mapper Imagery and an Unsupervised Classification Approach by Analyst #1

		Reference Data				Row
		D	C	AG	SB	Total
Classified Data	D	65	4	22	24	115
	C	6	81	5	8	100
	AG	0	11	85	19	115
	SB	4	7	3	90	104
	Column Total	75	103	115	141	434

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

OVERALL ACCURACY =
(65+81+85+90)/434 =
321/434 = 74%

PRODUCER'S ACCURACY	USER'S ACCURACY
D = 65/75 = 87%	D = 65/115 = 57%
C = 81/103 = 79%	C = 81/100 = 81%
AG = 85/115 = 74%	AG = 85/115 = 74%
SB = 90/141 = 64%	SB = 90/104 = 87%

the remotely sensed classification and the reference data, positive KHAT values are expected. Landis and Koch (1977) characterized the possible ranges for KHAT into three groupings: a value greater than 0.80 (i.e., >80%) represents strong agreement; a value between 0.40 and 0.80 (i.e., 40–80%) represents moderate agreement; and a value below 0.40 (i.e., <40%) represents poor agreement.

Table 7.3 also presents the variance of the KHAT statistic and the *Z* statistic used to determine if the classification is significantly better than a random result. At the 95% confidence level, the critical value would be 1.96. Therefore, if the absolute value of the test *Z* statistic is greater than 1.96, the result is significant and you would conclude that the classification is better than random. The *Z* statistic values for the two error matrices in Table 7.3 are both 20 or more, and so both classifications are significantly better than random.

Table 7.4 presents the results of the Kappa analysis that compares the error matrices, two at a time, to determine if they are significantly different. This test is based on the standard normal deviate and the fact that although remotely sensed data are discrete, the KHAT statistic is asymptotically normally distributed. The results of this pairwise test for significance between two error matrices reveals that these

TABLE 7.2
Error Matrix Using the Same Imagery and Classification Algorithm as in Table 7.1 Except that the Work Was Performed by a Different Analyst

		Reference Data				Row
		D	C	AG	SB	Total
Classified Data	D	45	4	12	24	85
	C	6	91	5	8	110
	AG	0	8	55	9	72
	SB	4	7	3	55	69
	Column Total	55	110	75	96	336

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

OVERALL ACCURACY =
(45+91+55+55)/336 =
246/336 = 73%

PRODUCER'S ACCURACY

D = 45/55 = 82%
C = 91/110 = 83%
AG= 55/75 = 73%
SB = 55/96 = 57%

USER'S ACCURACY

D = 45/85 = 53%
C = 91/110= 83%
AG = 55/72 = 76%
SB = 55/69 = 80%

TABLE 7.3
Individual Error Matrix Kappa Analysis Results

Error Matrix	KHAT	Variance	Z Statistic
Analyst #1	0.65	0.0007778	23.4
Analyst #2	0.64	0.0010233	20.0

TABLE 7.4
Kappa Analysis Results for the Pairwise Comparison of the Error Matrices

Pairwise Comparison	Z Statistic
Analyst #1 vs. Analyst #2	0.3087

two matrices are not significantly different. This is not surprising since the overall accuracies were 74 and 73% and the KHAT values were 0.65 and 0.64, respectively. Therefore, it could be concluded that these two analysts may work together because they produce approximately equal classifications. If two different techniques or algorithms were being tested and if they were shown to be not significantly different, then it would be best to use the cheaper, quicker, or more efficient approach.

MARGFIT

In addition to the Kappa analysis, a second technique called Margfit can be applied to "normalize" or standardize the error matrices for comparison purposes. Margfit uses an iterative proportional fitting procedure that forces each row and column (i.e., marginal) in the matrix to sum to a predetermined value; hence the name Margfit (marginal fitting). If the predetermined value is 1, then each cell value is a proportion of 1 and can easily be multiplied by 100 to represent percentages or accuracies. The predetermined value could also be set to 100 to obtain percentages directly, or to any other value the analyst chooses.

In this normalization process, differences in sample sizes used to generate the matrices are eliminated and, therefore, individual cell values within the matrix are directly comparable. In addition, because, as part of the iterative process, the rows and columns are totaled (i.e., marginals), the resulting normalized matrix is more indicative of the off-diagonal cell values (i.e., the errors of omission and commission). In other words, all the values in the matrix are iteratively balanced by row and column, thereby incorporating information from that row and column into each individual cell value. This process then changes the cell values along the major diagonal of the matrix (correct classifications) and, therefore, a normalized overall accuracy can be computed for each matrix by summing the major diagonal and dividing by the total of the entire matrix.

Consequently, one could argue that the normalized accuracy is a better representation of accuracy than is the overall accuracy computed from the original matrix, because it contains information about the off-diagonal cell values. Table 7.5 presents the normalized matrix generated from the original error matrix presented in Table 7.1 (an unsupervised classification of Landsat TM data by analyst #1) using the Margfit procedure. Table 7.6 presents the normalized matrix generated from the original error matrix presented in Table 7.3, which used the same imagery and classifier, but was performed by analyst #2.

In addition to computing a normalized accuracy, the normalized matrix can also be used to directly compare cell values between matrices. For example, we may be interested in comparing the accuracy each analyst obtained for the conifer category. From the original matrices we can see that analyst #1 classified 81 sample units correctly, while analyst #2 classified 91 correctly. Neither of these numbers means much because they are not directly comparable due to the differences in the number of samples used to generate the error matrix by each analyst. Instead, these numbers would need to be converted into percentages or user's and producer's accuracies so that a comparison could be made.

TABLE 7.5
Normalized Error Matrix from Analyst #1

		Reference Data				
		D	C	AG	SB	
Classified Data	D	0.7537	0.0261	0.1300	0.0909	
	C	0.1226	0.7735	0.0521	0.0517	
	AG	0.0090	0.1042	0.7731	0.1133	
	SB	0.1147	0.0962	0.0448	0.7440	
						3.0443

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

NORMALIZED ACCURACY =
0.7537+0.7735+0.7731+0.7440 =
3.0443/4.0 = 76%

TABLE 7.6
Normalized Error Matrix from Analyst #2

		Reference Data				
		D	C	AG	SB	
Classified Data	D	0.7181	0.0312	0.1025	0.1488	
	C	0.1230	0.7607	0.0541	0.0619	
	AG	0.0136	0.1017	0.7848	0.0995	
	SB	0.1453	0.1064	0.0587	0.6898	
						2.9534

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

NORMALIZED ACCURACY =
0.7181+0.7607+0.7848+0.6898 =
2.9534/4.0 = 74%

TABLE 7.7
Comparison of the Accuracy Values for an Individual Category

Error Matrix	Original Cell Value	Producer's Accuracy	User's Accuracy	Normalized Value
Analyst #1	81	79%	81%	77%
Analyst #2	91	83%	83%	76%

Here, another problem arises: do we divide the total correct by the row total (user's accuracy) or by the column total (producer's accuracy)? We could calculate both and compare the results or we could use the cell value in the normalized matrix. Because of the iterative proportional fitting routine, each cell value in the matrix has been balanced by the other values in its corresponding row and column. This balancing has the effect of incorporating producer's and user's accuracies together. Also, since each row and column adds to one, an individual cell value can quickly be converted to a percentage by multiplying by 100. Therefore, the normalization process provides a convenient way of comparing individual cell values between error matrices regardless of the number of samples used to derive the matrix (Table 7.7).

Table 7.8 provides a comparison of the overall accuracy, the normalized accuracy, and the KHAT statistic for the two analysts. In this particular example, there is agreement among all three measures of accuracy about the relative ranking of the results. However, it is possible for these rankings to disagree simply because each measure incorporates various levels of information from the error matrix into its computations. Overall accuracy only incorporates the major diagonal and excludes the omission and commission errors. As already described, normalized accuracy directly includes the off-diagonal elements (omission and commission errors) because of the iterative proportional fitting procedure. As shown in the KHAT equation, KHAT accuracy indirectly incorporates the off-diagonal elements as a product of the row and column marginals. Therefore, depending on the amount of error included in the matrix, these three measures may not agree.

It is not possible to give clear-cut rules as to when each measure should be used. Each accuracy measure incorporates different information about the error matrix and therefore must be thought of as different computations attempting to explain the error. Our experience has shown that if the error matrix tends to have a great many

TABLE 7.8
Summary of the Three Accuracy Measures for Analyst #1 and #2

Error Matrix	Overall Accuracy	KHAT	Normalized Accuracy
Analyst #1	74%	65%	76%
Analyst #2	73%	64%	74%

off-diagonal cell values with zeros in them, then the normalized results tend to disagree with the overall and Kappa results.

Many zeros occur in a matrix when an insufficient sample has been taken or when the classification is exceptionally good. Because of the iterative proportional fitting routine, these zeros tend to take on positive values in the normalization process, showing that some error could be expected. The normalization process then tends to reduce the accuracy because of these positive values in the off-diagonal cells. If a large number of off-diagonal cells do not contain zeros, then the results of the three measures tend to agree. There are also times when the Kappa measure will disagree with the other two measures. Because of the ease of computing all three measures and because each measure reflects different information contained within the error matrix, we recommend an analysis such as the one performed here to glean as much information from the error matrix as possible.

CONDITIONAL KAPPA

In addition to computing the Kappa coefficient for an entire error matrix, it may be useful to look at the agreement for an individual category within the matrix. Individual category agreement can be tested using the conditional Kappa coefficient. The maximum likelihood estimate of the Kappa coefficient for conditional agreement for the ith category is given by

$$\hat{K}_i = \frac{nn_{ii} - n_{i+}n_{+i}}{nn_{i+} - n_{i+}n_{+i}},$$

where n_{i+} and n_{+i} are as previously defined and the approximate large sample variance for the ith category is estimated by

$$\hat{\text{var}}(\hat{K}_i) = \frac{n(n_{i+} - n_{ii})}{[n_{i+}(n - n_{+i})]^3}[(n_{i+} - n_{ii})(n_{i+}n_{+i} - nn_{ii}) + nn_{ii}(n - n_{i+} - n_{+i} + n_{ii})].$$

The same comparison tests available for the Kappa coefficient apply to this conditional Kappa for an individual category.

WEIGHTED KAPPA

The Kappa analysis is appropriate when all the errors in the matrix can be considered of equal importance. However, it is easy to imagine a classification scheme in which errors may vary in their importance. In fact, this latter situation is really the more realistic approach. For example, it may be far worse to classify a forested area as water than to classify it as shrub. In this case, the ability to weight the Kappa analysis would be very powerful (Cohen, 1968). The following section describes the procedure to conduct a weighted Kappa analysis.

Let w_{ij} be the weight assigned to the i, jth cell in the matrix. This means that the proportion p_{ij} in the i, jth cell is to be weighted by w_{ij}. The weights should be restricted to the interval $0 \leq w_{ij} \leq 1$ for $i \neq j$, and the weights representing the maximum agreement are equal to 1; that is, $w_{ij} = 1$ (Fleiss et al., 1969).

Therefore, let

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

be the weighted agreement, and

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

where p_{ij}, p_{i+}, and p_{+j} are, as previously defined, the weighted "chance agreement."

Then the weighted Kappa is defined by

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

To compute the large sample variance of the weighted Kappa, define the weighted average of the weights in the *i*th category of the remotely sensed classification by

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

where p_{+j} is as previously defined and the weighted average of the weights in the *j*th category of the reference data set by

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

where p_{i+} is as previously defined.

The variance may be estimated by

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

The same tests of significant difference described previously for the Kappa analysis apply to the weighted Kappa. An individual weighted Kappa value can be evaluated to see if the classification is significantly better than random. Two independent weighted Kappas can also be tested to see if they are significantly different.

Although the weighted Kappa has been in the literature since the 1960s and was even suggested to the remote sensing community by Rosenfield and Fitzpatrick-Lins (1986), it has not received widespread attention. The reason for this lack of use is undoubtedly the need to select appropriate weights. Manipulating the weighting scheme can significantly change the results. Therefore, comparisons between different projects using different weighting schemes would be very difficult. The subjectivity of choosing the weights is always hard to justify. Using the unweighted Kappa analysis avoids these problems.

COMPENSATION FOR CHANCE AGREEMENT

Some researchers and scientists have objected to the use of the Kappa coefficient for assessing the accuracy of remotely sensed classifications because the degree of chance agreement may be overestimated (Foody, 1992). Remember from the equation for computing the Kappa coefficient,

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

that p_O is the observed proportion of agreement (i.e., the actual agreement) and p_C is the proportion of agreement that is expected to occur by chance (i.e., the chance agreement). However, in addition to the chance agreement, p_C also includes some actual agreement (Brennan and Prediger, 1981) or agreement for cause (Aickin, 1990). Therefore, since the chance agreement term does not consist solely of chance agreement, the Kappa coefficient may underestimate the classification agreement.

This problem is known to occur when the marginals are free (not fixed *a priori*), which is most often the case with remotely sensed classifications. Foody (1992) presents a number of possible solutions to this problem, including two Kappa-like coefficients that compensate for chance agreement in different ways. Others have suggested additional measures. However, given the very powerful properties of the Kappa coefficient, including the ability to test for significant differences between two independent coefficients, it must still be considered a vital accuracy assessment measure.

CONFIDENCE LIMITS

Confidence intervals are extremely common and are an expected component of any statistical estimate. However, computing confidence intervals for values in an error matrix are significantly more complex than simply computing a confidence interval for a traditional statistical analysis. The following example illustrates the calculations derived from the error matrix (Card, 1982). This example is designed assuming simple random sampling. If another sampling scheme is used, the variance equations change slightly.

TABLE 7.9
Error Matrix Showing Map Marginal Proportions

		True (i) Reference Data				Row	Map Marginal
		D	C	AG	SB	Total	Proportions, π_j
Map (j) Classified Data	D	65	4	22	24	115	0.3
	C	6	81	5	8	100	0.4
	AG	0	11	85	19	115	0.1
	SB	4	7	3	90	104	0.2
	Column Total	75	103	115	141	434	OVERALL ACCURACY = (65+81+85+90)/434 = 321/434 = 74%

The same error matrix as in Table 7.1 will be used to compute the confidence intervals. However, the map marginal proportions, π_j, computed as the proportion of the map falling into each map category, are also required (Table 7.9). The map marginal proportions are not derived from the error matrix but are simply the proportion of the total map area falling into each category. These proportions can quickly be obtained by dividing the area of each category by the total map area.

Given this matrix, the first step is to compute the individual cell probabilities using the following equation:

$$\hat{p}_{ij} = \pi_j \, n_{ij}/n_{.j}$$

The individual cell probabilities are simply the map marginal proportion multiplied by the individual cell value all divided by the row marginal. The results of these computations are shown in Table 7.10.

The true marginal proportions, $\hat{p}_i$, can then be computed using the equation:

$$\hat{p}_i = \sum_{j=1}^{r} \pi_j \, n_{ij}/n_{.j}$$

The true marginal proportions can also be computed simply by summing the individual cell probabilities in each column. For example, $\hat{p}_1 = 0.170 + 0.024 + 0.000 + 0.008 = 0.202$, $\hat{p}_2 = 0.357$, $\hat{p}_3 = 0.157$, and $\hat{p}_4 = 0.285$.

The third step is to compute the probability correct given the true class *i;* in other words, the producer's accuracy. It should be noted that the values here differ somewhat from those computed in the error matrix discussion because these values have been corrected for bias by incorporating the true marginal proportions as shown in

TABLE 7.10
Error Matrix of Individual Cell Probabilities, $\hat{p}_{ij}$

		True (i) Reference Data			
		D	C	AG	SB
Map (j) Classified Data	D	0.170	0.101	0.057	0.063
	C	0.024	0.324	0.020	0.032
	AG	0.000	0.010	0.074	0.017
	SB	0.008	0.013	0.006	0.0173

the following equation:

$$\hat{\theta}_{ii} = (\pi_i/\hat{p}_i)(n_{ii}/n_{.i}) \quad \text{or} \quad \hat{p}_{ii}/\hat{p}_i$$

As expected, the producer's accuracy is computed taking the diagonal cell value from the cell probability matrix (Table 7.10) and dividing by the true marginal proportion. For example, θ_{11} = 0.170/0.202 = 0.841, or 84%; θ_{22} = 0.908; θ_{33} = 0.471; and θ_{44} = 0.607.

The next step is to compute the probability correct given map class *j;* in other words, the user's accuracy. This computation is made exactly as described in the error matrix discussion by taking the diagonal cell value and dividing by the row (*j*) marginal. The equation for this calculation is as follows:

$$\hat{l}_{jj} = n_{jj}/n_{.j}$$

Therefore, $\hat{l}_{11}$ = 65/115 = 0.565, or 57%; $\hat{l}_{22}$ = 0.810; $\hat{l}_{33}$ = 0.739; and $\hat{l}_{44}$ = 0.865.

Step five is to compute the overall probability correct by summing the major diagonal of the cell probabilities or using the equation:

$$\hat{P}_c = \sum_{j=1}^{r} \pi_j \, n_{jj}/n_{.j}$$

Therefore, in this example, $\hat{P}_c$ = 0.170 + 0.324 + 0.074 + 0.173 = 0.741, or 74%.

We have now made essentially the same calculations as described in the error matrix discussion except that we have corrected for bias by using the true marginal proportions. The next step is to compute the variances for those terms (overall, producer's, and user's accuracies) that we wish to calculate confidence intervals.

Variance for overall accuracy, $\hat{P}_c$

$$V(\hat{P}_c) = \sum_{i=1}^{r} p_{ii}(\pi_i - p_{ii})/(\pi_i n)$$

Therefore, in this example, $\hat{P}_c$ = [0.170(0.3 – 0.170)/(0.3)(434)
+ 0.324(0.4 – 0.324)/(0.4)(434)
+ 0.074(0.1 – 0.074)/(0.1)(434)
+ 0.173(0.2 – 0.173)/(0.2)(434)]
= 0.00040

Confidence interval for overall accuracy, $\hat{P}_c$

$$\hat{P}_c = 2[V(\hat{P}_c)]^{1/2}$$

Therefore, in this example, the confidence interval for $\hat{P}_c$ = 0.741 ± $2(0.0004)^{1/2}$
= 0.741 ± 2(0.02)
= 0.741 ± 0.04
= (0.701, 0.781) or 70% to 78%

Variance for producer's accuracy, $\hat{\theta}_{ii}$

$$V(\hat{\theta}_{ii}) = p_{ii} p_i^{-4}\left[p_{ii}\sum_{j\neq 1}^{r} p_{ij}(\pi_j - p_{ij})/\pi_j n + (\pi_i - p_{ii})(p_i - p_{ii})^2/\pi_i n \right]$$

Therefore, in this example, $V(\hat{\theta}_{11})$ = 0.170 $(0.202)^{-4}$ {0.170[0.024(0.4 – 0.024)
/(0.4)(434) + 0.008(0.2 – 0.008)/(0.2)(434)]
+ (0.3 – 0.170)$(0.202 – 0.170)^2$ /(0.3)(434)}
= 0.00132

Confidence interval for producer's accuracy, $\hat{\theta}_{ii}$

$$\hat{\theta}_{ii} \pm 2[V(\hat{\theta}_{ii})]^{1/2}$$

Therefore, in this example, the confidence interval for $\hat{\theta}_{11}$ = 0.841 ± $2(0.00132)^{1/2}$
= 0.841 ± 2(0.036)
= 0.841 ± 0.072
= (0.768, 0.914) or 77% to 91%

Variance for user's accuracy, $\hat{l}_{ii}$

$$V(\hat{l}_{ii}) = p_{ii}(\pi_i - p_{ii})/\pi_i^2 n$$

Therefore, in this example, $V(\hat{l}_{11})$ $= 0.170(0.3 - 0.170)/(0.3)^2(434)$
$= 0.00057$

Confidence interval for

$$\hat{l}_{ii} \pm 2[V(\hat{l}_{ii})]^{1/2}$$

Therefore, in this example, the confidence interval for $\hat{l}_{11}$ $= 0.565 \pm 2(0.00057)^{1/2}$
$= 0.565 \pm 2(0.024)$
$= 0.741 \pm 0.048$
$= (0.517, 0.613)$ or 52% to 61%

It must be remembered that these confidence intervals are computed from asymptotic variances. If the normality assumption is valid, then these are 95% confidence intervals. If not, then by Chebyshev's inequality, they are at least 75% confidence intervals.

AREA ESTIMATION/CORRECTION

In addition to all the uses of an error matrix already presented, it can also be used to update the areal estimates of the map categories. The map derived from the remotely sensed data is a complete enumeration of the ground. However, the error matrix is an indicator of where misclassification occurred between what the map said is on the ground and what is actually on the ground. Therefore, it is possible to use the information from the error matrix to revise the estimates of total area for each map category. It is not possible to update the map itself or to revise a specific location on the map, but it is possible to revise total area estimates. Updating in this way may be especially important for small, rare categories whose estimates of total area could vary greatly depending on even small misclassification errors.

Czaplewski and Catts (1990) and Czaplewski (1992) have reviewed the use of the error matrix to update the areal estimates of map categories. They propose an informal method, both numerically and graphically, to determine the magnitude of bias introduced in the areal estimates by the misclassification. They also review two methods of statistically calibrating the misclassification bias. The first method is called the classical estimator and was proposed to the statistical community by Grassia and Sundberg (1982) and used in a remotely sensed application by Prisley and Smith (1987) and Hay (1988). The classical estimator uses the probabilities from the omission errors for calibration.

The second method is the inverse estimator, and it uses the probabilities from the commission errors to calibrate the areal estimates. Tenenbein (1972) introduced this technique in the statistical literature, and Chrisman (1982) and Card (1982) have used it for remote sensing applications. The confidence calculations derived in the previous section are from Card's (1982) work using the inverse estimator for calibration. More recently, Woodcock (1996) proposed a modification of the Card approach incorporating fuzzy set theory into the calibration process.

Despite all this work, not many users have picked up on these calibration techniques or the need to perform the calibration. From a practical standpoint, overall total areas are not that important. We have already discussed this in terms of non-site-specific accuracy assessment. However, as more and more work is done with looking at change, and especially changes of small, rare categories, the use of these calibration techniques may gain in importance.

8 Analysis of Differences in the Error Matrix

After testing the error matrix for statistical significance, the next step in the accuracy assessment analysis involves exploring why some of the map labels do not match the reference labels. While much attention is commonly placed on overall accuracy percentages (the sum of the major diagonal divided by the total), by far the more interesting analysis concerns discovering why some of the accuracy assessment samples did not fall on the diagonal of the error matrix. To both effectively use the map and to make better maps in the future, we need to know what causes these off-diagonal samples or differences in the matrix to occur.

All off-diagonal samples or differences in the error matrix will be the result of one of four possible sources:

1. Errors in the reference data,
2. Sensitivity of the classification scheme to observer variability,
3. Inappropriateness of the remote sensing data employed for mapping a specific land cover class, and
4. Mapping error.

This chapter reviews each one of these sources and discusses the impacts of each one to the accuracy assessment results.

ERRORS IN THE REFERENCE DATA

A major and required assumption of the error matrix is that the label from the reference data represents the "correct" label of the site and that all differences between the map and the reference label are due to classification and/or delineation error. Although this assumption is necessary, the reference data will never be perfect. As previously discussed, the term "ground truth" should be avoided for exactly this reason. Throughout this book, the authors prefer the use of reference data or reference label to refer to the sample data set being compared to the map that is being assessed. Unfortunately, error matrices can be inadequate indicators of map error because they are often confused by errors in the reference data (Congalton and Green, 1993). Errors in the reference data can be a function of the following:

- *Registration differences.* Registration differences between the reference data and the remotely sensed map classification caused by delineation and/or digitizing errors. For example, if GPS is not used in the field during accuracy assessment, it is possible for field personnel to collect data in the wrong area. Other registration errors can occur when an accuracy assessment site is incorrectly delineated or digitized, or when an existing map used for reference data is not precisely registered to the map being assessed.

- *Data entry errors.* Data entry errors are common in any database project and can be controlled only through rigorous quality control. Developing digital data entry forms that will only allow a certain set of characters for specific fields can catch errors during data entry. One of the best—but expensive—methods for catching data entry errors is to enter all data twice and then compare the two data sets. Differences usually indicate an error.
- *Classification scheme errors.* Every accuracy assessment map and reference site must have a label derived from the classification scheme used to create the map. Classification scheme errors occur when personnel misapply the classification scheme to the map or reference data, a common occurrence with complex classification schemes. If the reference data are in a database, then such errors can be avoided, or at least highlighted, by programming the classification scheme rules and using the program to label accuracy assessment sites. Classification scheme errors also occur when the classification scheme used to label the reference site is different from the one used to create the map, a common occurrence when existing data or maps are used as reference data.
- *Change.* Changes in land cover between the date of the remotely sensed imagery collection and the date of the reference data. As discussed in Chapter 6, land cover change can have a profound effect on accuracy assessment results. Tidal differences, crop or tree harvesting, urban development, fire, and pests all can cause the landscape to change in the time period between capturing the remotely sensed data and the accuracy assessment reference data collection.
- *Mistakes in labeling reference data.* Labeling mistakes usually occur because inexperienced personnel are used to collect reference data. Even with experienced personnel, the more detailed the classification scheme, the more likely it is that an error in labeling the reference data will occur. For example, some conifer and hardwood species are difficult to distinguish on the ground, much less from aerial photography. Young crops of broccoli, brussels sprouts, and cauliflower are easily confused. Thus, an accuracy assessment must also be performed on the reference data. If manual photo interpretation is used to assess a map created through semi-automated methods, then a sample of the photo-interpreted sites must be visited on the ground. If field data are used, then some sample of the sites must be visited by two different personnel and their answers compared. If the answers mostly agree, then the collection is satisfactory. If the answers mostly disagree, then there is a problem with the reference data collection method.

TABLE 8.1
Analysis of Map and Reference Label Differences

Map versus Ref. Difference	Number of Sites Different	Map Error	Reference Label Error	Date Change	Class. Scheme Difference	Variation in Estimation
Barren vs Water	19	0	6	8	0	5
Hardwood vs Water	6	0	0	0	0	6
Herb vs Forested	50	6	17	4	0	23
Wetland vs All Other Types	50	0	0	0	50	0
Total	125	6	23	12	50	34

Table 8.1 summarizes the reference data errors discovered during the quality control process for an actual accuracy assessment. Only 6 out of 125 of the differences between the map and reference labels were caused by actual errors in the map. Over two thirds of the differences (85 sites) were caused by mistakes in the reference data. The most significant error occurred from using different classification schemes (50 sites). In this project, National Wetlands Inventory (NWI) maps were used exclusively to map wetlands (i.e., wetlands were defined in the classification scheme to be those areas identified by NWI data as wetlands). However, when the accuracy assessment was performed, the photo interpreters collecting the reference data used a different definition of wetlands and disagreed with all the NWI labels. The remaining differences were caused by landscape change, reference label error, and observer variation, which is discussed in the next section of this chapter.

SENSITIVITY OF THE CLASSIFICATION SCHEME TO OBSERVER VARIABILITY

Classification scheme rules often impose discrete boundaries on continuous conditions in nature such as vegetation cover, soil type, or land use. In situations where breaks in the classification scheme represent artificial distinctions along a continuum, observer variability is often difficult to control and, although unavoidable, can have profound effects on accuracy assessment results (Congalton, 1991; Congalton and Green, 1993). Analysis of the error matrix must include exploring how many of the matrix differences are the result of observers being unable to precisely distinguish between classes when the accuracy assessment site is on the margin between two or more classes in the classification scheme.

Plato's parable of the shadows in the cave is useful for thinking about observer variability. In the parable, Plato describes prisoners who cannot move:

> "Above and behind them a fire is blazing in the distance, and between the fire and the prisoners there is a … screen which marionette players have in front of them over which they show puppets … (the prisoners) see only their own shadows, or the shadows of one another which the fire throws on the opposite wall of the cave … To them … the truth would be literally nothing but the shadows of the images." (Plato, *The Republic,* Book VII, 515-B, from Benjamin Jowett's translation as published in Vintage Classics, Random House, New York).

Like Plato's prisoners in the cave, we all perceive the world within the context of our experience. The difference between reality and perceptions of reality is often as fuzzy as Plato's shadows. Our observations and perceptions vary day to day and depend on our training, experience, or mood.

The analysis in Table 8.1 shows the impact that variation in interpretation can have on accuracy assessment. In the project, two photo interpreters were asked to label the same accuracy assessment reference sites. Almost 30% (34 of 125) of the differences between the map and reference label were caused by variation in interpretation.

Consider, for example, the assessment of a map of tree crown closure with classification scheme rules defining classes as:

Unvegetated	0–10%
Sparse	11–30%
Light	31–50%
Medium	51–70%
Heavy	71–100%

An accuracy assessment reference site from photo interpretation estimated at 45% tree crown cover would be labeled "Light." However, since it is recognized that crown closure can only be interpreted on aerial photos to ±10% (Spurr, 1960), it is also feasible that the proper label could be "Medium." Either the label of Light or the label of Medium is within the variability of the reference data collection. The map user would be much more concerned with a difference caused by a map label of Unvegetated compared to a reference label of Heavy tree crown cover. Differences on class margins are both inevitable and far less significant to the map user than other types of differences.

Classification schemes that employ estimates of percentage of vegetative cover are particularly susceptible to this type of confusion in the error matrix. Appendix 8.1 of this chapter presents a set of very complex classification scheme rules for a mapping project of Wrangell-St. Elias National Park in Alaska. The classification scheme is highly sensitive to estimates of percent vegetative cover. Sensitivity analysis on 140 accuracy assessment sites revealed that nearly 33% of the sites received new class labels when estimates of vegetative cover were varied by as little as 5%.

Several researchers have noted the impact of the variation in human interpretation on map results and accuracy assessment (Gong and Chen, 1992; Lowell, 1992;

Congalton and Biging, 1992; Congalton and Green, 1993). Woodcock and Gopal (1992) state, "The problem that makes accuracy assessment difficult is that there is ambiguity regarding the appropriate map label for some locations. The situation of one category being exactly right and all other categories being equally and exactly wrong often does not exist." Lowell (1992) calls for "a new model of space which shows transition zones for boundaries, and polygon attributes as indefinite." As Congalton and Biging (1992) conclude in their study of the validation of photo-interpreted stand-type maps, "The differences in how interpreters delineated stand boundaries was most surprising. We were expecting some shifts in position, but nothing to the extent that we witnessed. This result again demonstrates just how variable forests are and the subjectiveness of photo interpretation."

While it is difficult to control observer variation, it is possible to measure the variation, and to use the measurements to compensate for differences between reference and map data that are caused not by map error but by variation in interpretation. One option is to measure each reference site precisely to reduce observer variance in reference site labels. This method can be prohibitively expensive, usually requiring extensive field sampling. The second option incorporates fuzzy logic into the reference data to compensate for nonerror differences between reference and map data, and is discussed in Chapter 9.

INAPPROPRIATENESS OF THE REMOTE SENSING DATA EMPLOYED TO MAKE THE MAP

Early satellite remote sensing projects were primarily concerned with testing the viability of various remote sensing data for mapping certain types of land cover. Researchers tested the hypotheses of whether or not the imagery could be used to detect land use, or crop types, or forest types. Many accuracy assessment techniques were developed primarily to test these hypotheses.

Recently, accuracy assessment has focused more on learning about the reliability of a map for land management or policy analysis. However, some of the differences in the error matrix will be because the map producer was attempting to use remote sensing data or methods that were incapable of distinguishing certain land cover or vegetation class types. Understanding what differences are caused by the technology is useful to the map producer when the next map is being made.

In the Wrangell-St. Elias example cited earlier, Landsat TM data was employed as the primary remotely sensed data, with 1:60,000 aerial photography as ancillary data. The classification scheme included distinguishing between pure and mixed stands of black and white spruce. Accuracy assessment analysis consistently showed success at differentiating pure stands of black versus white spruce. However, consistently differentiating these species in mixed or occasional hybrid stands was found to be unreliable. This phenomenon is not surprising considering the difficulty often associated with differentiating these species in mixed and hybrid stands on the ground. In summary, moderate resolution multispectral remotely sensed data, at the scales employed, cannot be used to reliably and consistently differentiate between mixed classes of these two tree species.

To make the map more reliable, the map user can collapse the classification system across classes. In this example, the nonpure spruce classes of Closed, Open, and Woodland were collapsed into an "Unspecified Interior Spruce" class. In the difference matrix, "Unspecified Interior Spruce," map labels were considered to be mapped correctly if they corresponded to a pure or mixed white spruce or black spruce reference site demonstrating the same density class of Closed, Open, or Woodland. For example, a map label of "Open Unspecified Interior Spruce" was considered to be correctly mapped if its corresponding reference label for the site was "Open Black Spruce," "Open White Spruce," or "Open Black/White Spruce" mix. While less information is displayed on the map, the remaining information is more reliable.

MAPPING ERROR

The final cause of differences in the error matrix is the result of mapping error (i.e., the actual real errors). Often, these are difficult to distinguish from an inappropriate use of remote sensing data. Usually, they are errors that are particularly obvious and unacceptable. For example, it is not uncommon for an inexperienced remote sensing professional to produce a map of land cover from satellite data that misclassifies northeast-facing forests on steep slopes as water. Because water and shadowed wooded slopes both absorb most energy, this type of error is explainable but unacceptable, and must be avoided. Many map users will be appalled at this type of error and are not particularly interested in having the electromagnetic spectrum explained to them as an excuse. However, careful editing and comparison with aerial photography, checking that all water exists in areas without slope, and comparison to existing maps of waterways and lakes will reduce the possibility of this type of map error.

Understanding the causes of true error can point the map producer to additional methods of improving the accuracy of the map. Perhaps other bands or band combinations will improve accuracy. Incorporation of ancillary data such as slope, aspect, or elevation may be useful. In the Wrangell-St. Elias example, confusion existed between the Dwarf Shrub classes and the Graminoid class. The confusion was addressed through the use of unsupervised classifications and parkwide models utilizing digital elevation data, field-based data, and aerial photography. First, an unsupervised classification with 20 classes was run for only those areas of the imagery classified as Dwarf Shrub in the map. A digital elevation coverage was utilized to stratify the study area for subsequent relabeling of unsupervised classes previously mapped as Dwarf Shrub but actually representing areas of Graminoid cover on the ground. From the unsupervised classification, two spectral classes were found to consistently represent Graminoid cover throughout the study area while another spectral class was found to represent Graminoid cover in areas below 3500 ft elevation. These spectral classes were subsequently recoded to the Graminoid class.

SUMMARY

Analysis of the causes of differences in the error matrix can be one of the most important and interesting steps in the creation of a map from remotely sensed data. In the past, too much emphasis was placed on the overall accuracy of the map, without investigating the conditions that give rise to that accuracy. By understanding what caused the reference and map data to differ, we can use the map more reliably, and produce both better maps and better accuracy assessments in the future.

APPENDIX 8.1

Wrangell-St. Elias National Park and Preserve: Land Cover Mapping Classification Key

```
If tree total ≥ 10% (Forested)
        If Conifer ≥ 75% of tree total
                If (Pigl + Pima) ≥ 67% of conifer total
                        If (Pigl/(Pigl+Pima)) ≥ 75%             PIGL
                        If (Pima/(Pigl+Pima)) ≥ 75%             PIMA
                        Else                                    Unspecified Spruce
        If Broadleaf ≥ 75% of tree total                        Broadleaf
        Else (mixed conifer/broadleaf)                          Spruce/Broadleaf

Else if shrub total ≥ 25% (Shrub)
        If tall shrub total ≥ 25%                               Tall Shrub
        If low shrub total ≥ 25%                                Low Shrub
        If dwarf shrub total ≥ 25%                              Dwarf Shrub
Else (tall, low, or dwarf are not individually > 25%)
        If tall shrub total ≥ 67% of shrub total                Tall Shrub
        If low shrub total ≥ 67% or shrub total                 Low Shrub
        If dwarf shrub total ≥ 67% of shrub total               Dwarf Shrub
        Else "pick the largest percent of ":
                    tall shrub                                  Tall Shrub
                    low shrub                                   Low Shrub
                    dwarf shrub                                 Dwarf Shrub
                    (ties go to the "tallest")

Else if herbaceous ≥ 15% (Herbaceous)
      If graminoid ≥ 50% or (graminoid/herb total) ≥ 50%        Graminoid
      Else if forb ≥ 50% or (forb/herb total) ≥ 50%             Forb
      Else if moss ≥ 50% or (moss/herb total) ≥ 50%             Moss/Lichen
      Else if lichen ≥ 50% or (lichen/herb total) ≥ 50%)        Moss/Lichen
      Else "pick the largest percent of ":
                    graminoid
                    forb
                    moss
                    lichen
                    (preference for ties go in the order listed)

Else if total vegetation ≥ 10% and < 30%                        Sparse Vegetation

Else (nonvegetated)
```

Water
Barren
Glacier/Snow
Clouds/Cloud Shadow

- **Forested** (>10% tree cover)
  - **Conifer** (>75% conifer)
    - **Closed** (60–100%)
      - **Pigl**
      - **Pima**
      - **Pigl/Pima**
      - **Pisi**
      - **Tshe**
      - **Tsme**
      - **Pisi/Tsme**
      - **Pisi/Tshe**
      - **Tshe/Tsme**
      - **Spruce**
      - **Mixed Conifer**
    - **Open** (25–59%)
      - **Pigl**
      - **Pima**
      - **Pigl/Pima**
      - **Pisi**
      - **Tshe**
      - **Tsme**
      - **Pisi/Tsme**
      - **Pisi/Tshe**
      - **Tshe/Tsme**
      - **Spruce**
      - **Mixed Conifer**
    - **Woodland** (10–24%)
      - **Pigl**
      - **Pima**
      - **Pigl/Pima**
      - **Pisi**
      - **Tshe**
      - **Tsme**
      - **Pisi/Tsme**
      - **Pisi/Tshe**
      - **Tshe/Tsme**
      - **Spruce**
      - **Mixed Conifer**
  - **Broadleaf** (>75% broadleaf)
    - **Closed** (60–100%)
      - **Closed Broadleaf**
    - **Open** (10–59%)
      - **Open Broadleaf**

Mixed

Closed (60–100%)

Pigl/Pima-Broadleaf

Pisi-Broadleaf

Tshe-Broadleaf

Conifer-Broadleaf

Open (10–59%)

Pigl/Pima-Broadleaf

Pisi-Broadleaf

Tshe-Broadleaf

Conifer-Broadleaf

Shrub (>25% shrub)

Tall (tall shrub > 25% or dominant)

Closed (>75%)

Open (25–74%)

Low (low shrub > 25% or dominant)

Closed (>75%)

Open (25–74%)

Dwarf (dwarf shrub > 25% or dominant)

Herbaceous (herbaceous > 15%)

Graminoid

Forb

Moss

Lichen

Sparse vegetation

Sparse vegetation

Nonvegetated

Water

Barren

Glacier/Snow

Clouds/Cloud Shadow

9 Fuzzy Accuracy Assessment

As our use of remotely sensed data and maps has grown in complexity, so have the classification schemes associated with these efforts. The classification scheme then becomes an even more important factor influencing the accuracy of the entire project. A review of the accuracy assessment literature points out some of the limitations of using only the traditional error matrix approach to accuracy assessment of a map with a complex classification scheme. Congalton and Green (1993) recommend the error matrix as a jumping-off point for identifying sources of confusion (i.e., differences between the map created from remotely sensed data and the reference data) and not simply the "error." For example, variation in human interpretation can have a significant impact on what is considered correct. If photo interpretation is used as the source of the reference data and that interpretation is flawed, then the results of the accuracy assessment could be very misleading. This is true even for observations made in the field. As classification schemes become more complex, more variation in human interpretation is introduced. In addition, in situations where the breaks (i.e., divisions between classes) in the classification system represent artificial distinctions along a continuum, variation in human interpretation is often very difficult to control and, while unavoidable, can have profound effects on accuracy assessment results (Congalton, 1991; Congalton and Green, 1993). Several researchers have noted the impact of the variation in human interpretation on map results and accuracy assessment (Gong and Chen, 1992; Lowell, 1992; McGwire, 1992; Congalton and Biging, 1992).

Gopal and Woodcock (1994) proposed the use of fuzzy sets to "allow for explicit recognition of the possibility that ambiguity might exist regarding the appropriate map label for some locations on the map. The situation of one category being exactly right and all other categories being equally and exactly wrong often does not exist." In this fuzzy set approach, it is recognized that instead of a simple system of correct (agreement) and incorrect (disagreement), there can be a variety of responses such as "absolutely right," "good answer," "acceptable," "understandable but wrong," and "absolutely wrong."

Fuzzy set theory or fuzzy logic is a form of set theory. Although initially introduced in the 1920s, fuzzy logic gained its name and its algebra in the 1960s and 1970s from Zadeh (1965), who developed fuzzy set theory as a way to characterize the ability of the human brain to deal with vague relationships. The key concept is that membership in a class is a matter of degree. Fuzzy logic recognizes that, on the margins of classes that divide a continuum, an item may belong to both classes. As Woodcock and Gopal (1992) state, "The assumption underlying fuzzy set theory is that the transition from membership to non-membership is seldom a step function." Therefore, while a 100% hardwood stand can be labeled hardwood, and a 100%

conifer stand may be labeled conifer, a 49% hardwood and 51% conifer stand may be acceptable if labeled either conifer or hardwood.

Lowell (1992) calls for "a new model of space which shows transition zones for boundaries, and polygon attributes as indefinite." As Congalton and Biging (1992) conclude in their study of the validation of photo interpreted stand type maps, "The differences in how interpreters delineated stand boundaries was most surprising. We were expecting some shifts in position, but nothing to the extent that we witnessed. This result again demonstrates just how variable forests are and is a strong indicator of human variation in photo interpretation."

There are a number of techniques that have been proposed to incorporate ambiguity or fuzziness into the accuracy assessment process. Three methods are presented in this chapter including (1) expanding the major diagonal of the error matrix, (2) measuring map class variability, and (3) using a fuzzy error matrix approach.

EXPANDING THE MAJOR DIAGONAL OF THE ERROR MATRIX

The simplest and most straightforward method for incorporating fuzziness into the accuracy assessment process is to expand the major diagonal of the error matrix. Remember that the major diagonal of the error matrix represents agreement between the reference data and the map, and is represented by a single cell in the matrix for each map class. By acknowledging some fuzziness in the classification, the class boundaries may be expanded to accept as correct plus or minus one class of the actual class. In other words, the major diagonal is no longer just a single cell for each map class, but rather wider. This method works well if the classification scheme is continuous, such as elevation or tree size class or forest crown closure. If the classification scheme is discrete, such as in a vegetation or land cover mapping project, then this method probably cannot be used.

Table 9.1 presents the traditional error matrix for a classification of forest crown closure (a continuous classification scheme divided into 6 discrete classes). Only exact matches are considered correct; they are tallied along the major diagonal. The overall accuracy of this classification is 40%. Table 9.2 presents the same error matrix with the major diagonal expanded to include plus or minus one crown closure class. In other words, for crown closure class 3 both crown closure classes 2 and 4 are also accepted as correct. This revised major diagonal then results in a tremendous increase in overall accuracy to 75%.

The advantage of using this method of accounting for fuzzy class boundaries is obvious; the accuracy of the classification can increase dramatically. The disadvantage is that if the reason for accepting plus or minus one class cannot be adequately justified or does not meet the map user's requirements, then it may be thought that you are cheating to try to get higher accuracies. Therefore, although this method is very simple to apply, it should only be used when agreement exists that it is a reasonable course of action. The other techniques described next may be more difficult to apply, but are easier to justify.

TABLE 9.1
Error Matrix Showing the Ground Reference Data versus the Image Classification for Forest Crown Closure

		Ground Reference						Row
		1	2	3	4	5	6	Total
Image Classification	1	2	9	1	2	1	1	16
	2	2	8	3	6	1	1	21
	3	0	3	3	4	9	1	20
	4	0	0	2	8	7	10	27
	5	0	1	2	1	6	16	26
	6	0	0	0	0	3	31	34
Column Total		4	21	11	21	27	60	144

Crown Closure Categories		
Class 1 =	0%	CC
Class 2 =	1–10%	CC
Class 3 =	11–30%	CC
Class 4 =	31–50%	CC
Class 5 =	51–70%	CC
Class 6 =	71–100%	CC

OVERALL ACCURACY = 58/144 = 40%

PRODUCER'S ACCURACY	USER'S ACCURACY
Class 1 = 2/4 = 50%	Class 1 = 2/16 = 13%
Class 2 = 8/21 = 38%	Class 2 = 8/21 = 38%
Class 3 = 3/11 = 27%	Class 3 = 3/20 = 15%
Class 4 = 8/21 = 38%	Class 4 = 8/27 = 30%
Class 5 = 6/27 = 22%	Class 5 = 6/26 = 23%
Class 6 = 31/60 = 52%	Class 6 = 31/34 = 91%

MEASURING MAP CLASS VARIABILITY

The second method for incorporating fuzziness into the accuracy assessment process is not as simple as expanding the major diagonal of the error matrix. While it is difficult to control variation in human interpretation, it is possible to measure the variation, and to use these measurements to compensate for differences between reference and map data that are caused not by map error but by variation in interpretation. There are two options available to control the variation in human interpretation to reduce the impact of this variation on map accuracy. One is to measure each reference site with great precision to minimize the variance in the reference site labels. This method can be prohibitively expensive, usually requiring extensive field sampling and detailed measurements. The second option is to measure the variance and use these measurements to compensate for nonerror differences between reference and map data. Measuring the variance requires having multiple analysts assess each reference site. This assessment

TABLE 9.2
Error Matrix Showing the Ground Reference Data versus the Image Classification for Forest Crown Closure within Plus or Minus One Tolerance Class

		Ground Reference						Row
		1	2	3	4	5	6	Total
Image Classification	1	2	9	1	2	1	1	16
	2	2	8	3	6	1	1	21
	3	0	3	3	4	9	1	20
	4	0	0	2	8	7	10	27
	5	0	1	2	1	6	16	26
	6	0	0	0	0	3	31	34
Column Total		4	20	11	21	27	60	144

Crown Closure Categories

Class 1 =	0%	CC
Class 2 =	1–10%	CC
Class 3 =	11–30%	CC
Class 4 =	31–50%	CC
Class 5 =	51–70%	CC
Class 6 =	71–100%	CC

OVERALL ACCURACY = 108/144 = 75%

PRODUCER'S ACCURACY	USER'S ACCURACY
Class 1 = 4/4 = 100%	Class 1 = 11/16 = 69%
Class 2 = 20/21 = 95%	Class 2 = 13/21 = 62%
Class 3 = 8/11 = 73%	Class 3 = 10/20 = 50%
Class 4 = 13/21 = 62%	Class 4 = 17/27 = 63%
Class 5 = 16/27 = 59%	Class 5 = 23/26 = 88%
Class 6 = 47/60 = 78%	Class 6 = 34/34 = 100%

could be done by field visitation or using photo interpretation, and requires an objective and repeatable method of capturing the impacts of human variation. The collection of reference data for accuracy assessment is an expensive component of any mapping project; multiple visits to every reference site to capture variation may be prohibitively expensive. Therefore, while theoretically possible, measuring map class variability is not a viable component of most remotely sensed data mapping projects.

THE FUZZY ERROR MATRIX APPROACH

The previous approaches of expanding the major diagonal to incorporate fuzziness in the accuracy assessment process may be hard to justify, and the effort needed to measure the variability may be cost-prohibitive. Therefore, another method is required to incorporate fuzziness into the map accuracy assessment process. As mentioned earlier, the difficult task in using fuzzy logic is the development of the specific rules for its application. Fuzzy systems often rely on experts for development of these rules. Hill (1993) developed an arbitrary but practical fuzzy set rule that determined “sliding class widths” for assessing the accuracy of maps produced for The California Department of Forestry and Fire Protection of the Klamath Province in Northwestern California. Woodcock and Gopal (1992) relied on experts in their application of fuzzy sets to assess the accuracy of maps generated for Region 5 of the U.S. Forest Service. While both of their methods incorporated fuzziness into the accuracy assessment process, neither used an error matrix approach. Instead, a number of other metrics to represent map accuracy and agreement were computed.

The Fuzzy Error Matrix

Given the wide acceptance of the error matrix as the standard for reporting the accuracy of thematic maps, it would be far better to employ some approach that combines both the error matrix and some measure of fuzziness. Such a technique, called the *fuzzy error matrix approach*, was introduced by Green and Congalton (2004) and is described here. The use of the fuzzy error matrix is a very powerful tool in the accuracy assessment process because the fuzzy error matrix allows the analyst to compensate for situations in which the classification scheme breaks represent artificial distinctions along a continuum of landcover and/or where observer variability is often difficult to control. While one of the assumptions of the traditional or deterministic error matrix used in the rest of this book is that a reference data sample site can have only one label, this is not the case with the fuzzy error matrix approach.

Let us continue with the example used so far in this chapter. Table 9.3 presents a fuzzy error matrix generated from a set of fuzzy rules applied to the same classification that was used to generate the deterministic (i.e., nonfuzzy) error matrix that was presented in Table 9.1. In this case, the classification was defined using the following fuzzy rules:

- Class 1 was defined as always 0% crown closure. If the reference data indicated a value of 0%, then only a map classification of 0% was accepted.
- Class 2 was defined as acceptable if the reference data was within 5% of that of the map classification. In other words, if the reference data indicates that a sample has 15% crown closure and the map classification put it in Class 2, the answer would not be absolutely correct, but would be considered acceptable.
- Classes 3–6 were defined as acceptable if the reference data were within 10% of that of the map classification. In other words, a sample classified as Class 4 on the image, but found to be 55% crown closure on the reference data would be considered acceptable.

As a result of these fuzzy rules, off-diagonal elements in the matrix contain two separate values. The first value in the off-diagonal represents those labels that, although not absolutely correct, are considered acceptable labels within the fuzzy rules. The second value indicates those labels that are still unacceptable (i.e., wrong). The major diagonal still only tallies those labels considered to be absolutely correct. Therefore, in order to compute the accuracies (overall, producer’s, and user’s), the values along

TABLE 9.3
Error Matrix Showing the Ground Reference Data versus the Image Classification for Forest Crown Closure Using the Fuzzy Logic Rules

		Ground Reference						Row
		1	2	3	4	5	6	Total
Image Classification	1	2	6,3	1	2	1	1	16
	2	0,2	8	2,1	6	1	1	21
	3	0	2,1	3	4,0	9	1	20
	4	0	0	0,2	8	5,2	10	27
	5	0	1	2	1,0	6	12,4	26
	6	0	0	0	0	2,1	31	34
Column Total		4	21	11	21	27	60	144

Crown Closure Categories		
Class 1 =	0%	CC
Class 2 =	1–10%	CC
Class 3 =	11–30%	CC
Class 4 =	31–50%	CC
Class 5 =	51–70%	CC
Class 6 =	71–100%	CC

OVERALL ACCURACY = 92/144 = 64%

PRODUCER'S ACCURACY	USER'S ACCURACY
Class 1 = 2/4 = 50%	Class 1 = 8/16 = 50%
Class 2 = 16/21 = 76%	Class 2 = 10/21 = 48%
Class 3 = 5/11 = 45%	Class 3 = 9/20 = 45%
Class 4 = 13/21 = 62%	Class 4 = 13/27 = 48%
Class 5 = 13/27 = 48%	Class 5 = 19/26 = 73%
Class 6 = 43/60 = 72%	Class 6 = 33/34 = 97%

the major diagonal (i.e., absolutely correct) and those deemed acceptable (i.e., those in the first value) in the off-diagonal elements are combined. In Table 9.3, this combination of absolutely correct and acceptable answers results in an overall accuracy of 64%. This overall accuracy is significantly higher than the original error matrix (Table 9.1), but not as high as Table 9.2.

It is much easier to justify the fuzzy rules used in generating Table 9.3 than it is to simply extend the major diagonal to plus or minus one whole class as was done in Table 9.2. For crown closure, it is recognized that mapping typically varies by plus or minus 10% (Spurr, 1948). Therefore, it is reasonable to define as acceptable a range within 10% for classes 3–6. Class 1 and class 2 take an even more conservative approach and are therefore even easier to justify.

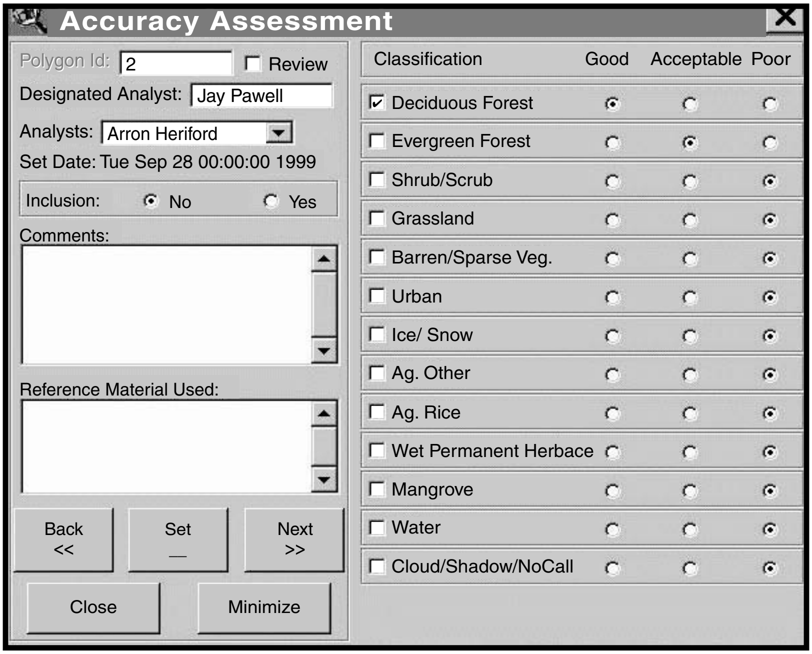
In addition to this fuzzy set theory working for continuous variables such as crown closure, it also applies to more categorical data. All that is required is a set of fuzzy rules to explain or capture the variation. For example, in the hardwood range area of California, many land cover types differ only by which hardwood species is dominant. In many cases, the same species are present and the specific land cover type is determined by which species is most abundant. Also, in some of these situations, the species look very much alike on aerial photography and on the ground. It is clear that there is a great deal of acceptable and unavoidable variation in mapping the hardwood range.

In a worldwide mapping effort funded by the National Geospatial-Intelligence Agency (NGA) using Landsat imagery, no ground visitation was possible for collecting the reference data. In some areas of the world, the imagery used for the reference data was of such low resolution as to make interpretation of the individual classes very difficult. The use of this fuzzy error matrix approach was the only viable solution in this case (Green and Congalton, 2004). A traditional, deterministic accuracy assessment conducted with such highly variable reference data would have unfairly represented the accuracy of this mapping effort.

Therefore, in many situations, the use of these fuzzy rules, which allow for the incorporation of acceptable reference labels in addition to the absolutely correct reference labels into the construction of the error matrix, makes a great deal of sense. Using fuzzy rules in an error matrix approach combines all the established descriptive and analytical power of the error matrix while incorporating variation into the assessment.

Implementation of the Fuzzy Error Matrix

Implementation of the fuzzy error matrix approach is greatly simplified with the use of a special reference data collection form (Figure 9.1). Each reference site can be evaluated for the likelihood of being identified as each of the possible map classes



Polygon Id: 2 □ Review

Designated Analyst: Jay Pawell

Analysts: Arron Heriford

Set Date: Tue Sep 28 00:00:00 1999

Inclusion: ● No ○ Yes

Comments:

Reference Material Used:

Classification	Good	Acceptable	Poor
☑ Deciduous Forest	●	○	○
□ Evergreen Forest	○	●	○
□ Shrub/Scrub	○	○	●
□ Grassland	○	○	●
□ Barren/Sparse Veg.	○	○	●
□ Urban	○	○	●
□ Ice/ Snow	○	○	●
□ Ag. Other	○	○	●
□ Ag. Rice	○	○	●
□ Wet Permanent Herbace	○	○	●
□ Mangrove	○	○	●
□ Water	○	○	●
□ Cloud/Shadow/NoCall	○	○	●

FIGURE 9.1 Form for labeling accuracy assessment reference sites.

given the fuzzy rules for that project. First, the analyst determines the most appropriate ("good") label for the site and enters this label in the appropriate box under the "classification" column on the form. This label determines which row of the matrix the site will be tallied in and is also the value used for calculation of the deterministic error matrix. After assigning the most appropriate label for the site, the remaining possible map labels are evaluated as either "acceptable" or "poor" candidates for the site's label, again as indicated by the fuzzy rules. For example, a site might fall near the classification scheme margin between deciduous forest and evergreen forest because of the exact mix of species and/or the difficulty interpreting the exact mixture on the reference data imagery. In this instance, the analyst might rate deciduous forest as the most appropriate label (i.e., "good"), but also rate evergreen forest as "acceptable" (see Figure 9.1). In this case, no other map classes would be acceptable; all the others would be rated as "poor."

Using this fuzzy error matrix approach allows for the analyst to compensate for interpreter variability and difficulty in determining only a single label for each reference data sample site. While this method can be used for any assessment, it works best when (1) there are issues in collecting good reference data because of limitations in the reference data collection methods, (2) when interpreter variability cannot be controlled, or (3) when the ecosystem being mapped is highly heterogeneous. If there is little variation or fuzziness in the classification scheme or if detailed measurements can be taken to minimize the variation, then there may be little need for this approach. However, in most projects creating maps from remotely sensed imagery, the use of the fuzzy error matrix approach can significantly help to incorporate variation inherent in the project.

Another Fuzzy Error Matrix Example

Table 9.4 shows a fuzzy error matrix for a categorical classification scheme (i.e., a land cover mapping project). Again, the power of this approach lies in the ability to compute the same descriptive metrics as in the traditional deterministic error matrix. Computation of the overall, producer's and user's accuracy statistics for the fuzzy error matrix follows the same methodology as the traditional deterministic error matrix with the following additions. Nondiagonal cells in the matrix contain two tallies, which can be used to distinguish class labels that are uncertain or that fall on class margins, from class labels that are most probably in error. The first number represents those sites in which the map label matched an "acceptable" reference label in the fuzzy assessment (Table 9.4). Therefore, even though the label was not considered the most appropriate, it was considered acceptable given the fuzziness of the classification system and/or the minimal quality of some of the reference data. These sites are considered a "match" for estimating fuzzy accuracy. The second number in the cell represents those sites where the map label was considered poor (i.e., an error).

The fuzzy assessment overall accuracy is estimated as the percentage of sites where the "good" and "acceptable" reference labels matched the map label. Producer's and user's accuracies are computed in the traditional way, but again instead of just using

TABLE 9.4
Example of Fuzzy Error Matrix Showing Both Deterministic and Fuzzy Accuracy Assessment

		REFERENCE DATA								User's Accuracies			
		Decid. Forest	EG Forest	Scrub/ Shrub	Grass	Barren/ Sparse	Urban	Agric.	Water	Deterministic Totals	Percent Deterministic	Fuzzy Totals	Percent Fuzzy
	Deciduous Forest	**48**	24,7	0,1	0,3	0,0	0,1	0,11	0,18	48/113	42.5%	72/113	63.7%
	Evergreen Forest	4,0	**17**	0,1	0,0	0,0	0,0	0,1	0,3	17/26	65.4%	21/26	80.8%
	Shrub/Scrub	2,0	0,1	**15**	8,1	0,0	0,0	2,2	0,0	15/31	48.4%	27/31	87.1%
	Grassland	0,1	0,0	5,1	**14**	0,0	0,0	3,0	0,0	14/24	58.3%	22/24	91.7%
M A P	**Barren/Sparse Veg.**	0,0	0,0	0,2	0,0	**0**	0,0	0,1	0,0	0/3	0.0%	0/3	0.0%
	Urban	0,0	0,0	0,0	0,0	0,0	**20**	2,0	0,0	20/22	90.9%	22/22	100.0%
	Agriculture	0,1	0,1	7,15	18,6	0,0	2,0	**29**	1,2	29/82	35.4%	57/82	69.5%
	Water	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**8**	8/8	100.0%	8/8	100.0%
	Producer's Accuracies												
	Deterministic Totals	48/56	17/50	15/47	14/50	NA	20/24	29/51	8/33	**Overall Accuracies**			
	Percent Deterministic	85.7%	34.0%	31.9%	28.0%	NA	83.3%	56.9%	24.2%	**Deterministic**		**Fuzzy**	
	Fuzzy Totals	54/56	41/50	27/47	40/50	NA	22/24	36/51	10/33	151/311	48.6%	230/311	74.0%
	Percent Fuzzy	96.4%	82.0%	57.4%	80.0%	NA	91.7%	70.6%	30.3%				

the value on the major diagonal (“good”), the value in the first off-diagonal position (“acceptable”) is also included (Table 9.4).

SUMMARY

While three methods are presented in this chapter for dealing with variation or fuzziness in the accuracy assessment process, the fuzzy error matrix approach is by far the most useful and operational. The elegance of this approach is that it combines all of the power of the error matrix, including computing overall, producer’s, and user’s accuracies, with the ability to incorporate the variation inherent in many classification schemes or resulting from the reference data collection process. Given that the matrix contains the information to compute both the traditional deterministic accuracy measures and fuzzy accuracy measures, there is strong impetus to use this approach. It is highly recommended that this approach be considered whenever map class variation or variation in the reference data collection process is a significant issue in the mapping project.

10 Case Study *Accuracy Assessment for the NOAA Next-Generation C-CAP Pilot Project*

This chapter details an actual case study of thematic accuracy assessment design, data collection, and analysis. The first section reviews the goals of a case study mapping project and briefly summarizes the classification methods used. Next, the questions raised in Chapter 1 are answered for this specific case study. The chapter concludes with a review of lessons learned during the design and implementation of the accuracy assessment.

OVERVIEW OF THE CASE STUDY

The National Oceanic and Atmospheric Administration (NOAA) currently relies on Landsat TM and ETM+ moderate resolution imagery for the creation of its Coastal Change Analysis Program (C-CAP) land use and change products. Recognizing the power of higher-resolution imagery, in September of 2004, NOAA contracted with several organizations to develop methodologies for successfully introducing high spatial resolution imagery and land cover products into NOAA's current C-CAP effort. This case study reviews the accuracy assessment of one of those efforts.

The project area is located in the vicinity of Panama City, Florida, as displayed in Figure 10.1. The area is characterized by little elevation variation and is highly diverse in both land cover and land use. Spanning parts of Bay and Washington Counties, the project area also crosses six Florida physiographic groups including Crystal Lake Karst, Delta Plain, Coastal Strip, Hosford Delta, and a sliver of Fountain Delta and Betts Delta (see Figure 10.2). Land use types are intermixed throughout the area, as are uplands and wetlands.

The mapping methods used in the project were fairly straightforward. To understand how the land use/cover classes vary on the ground, a training data/calibration field trip was conducted and field samples of all classes to be mapped were collected. To understand how the variation in land cover/use classes was correlated with variation in the imagery and ancillary data, a Classification and Regression Tree (CART) analysis was performed on the nonaccuracy assessment sample data

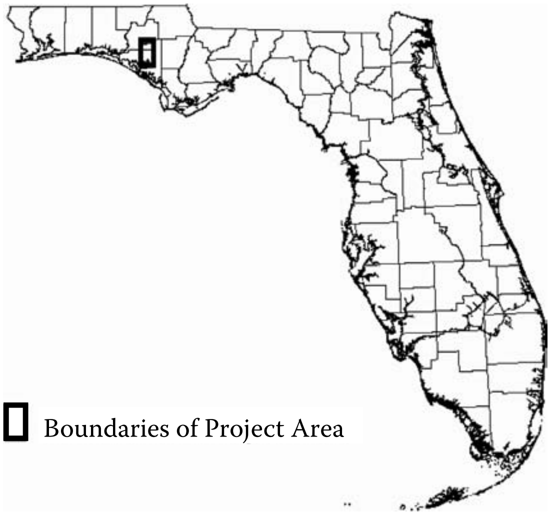


FIGURE 10.1 Location of the pilot project area in Florida.

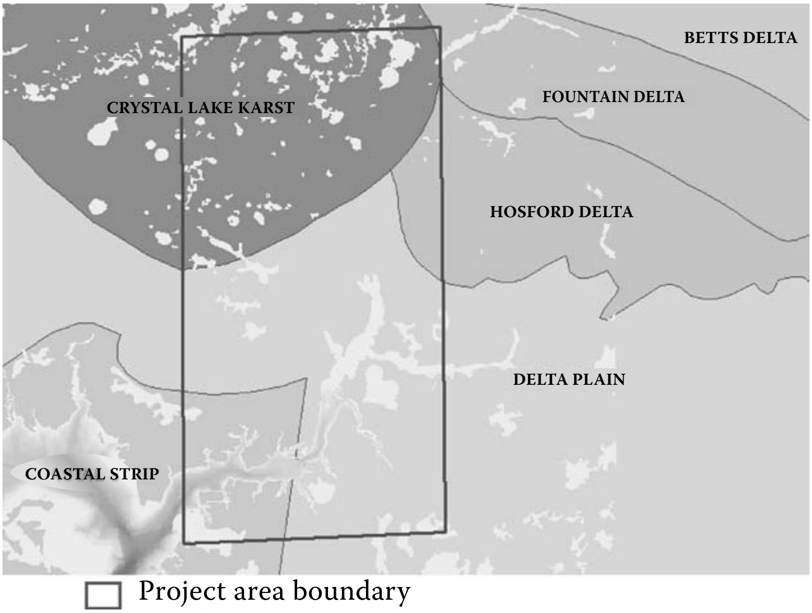


FIGURE 10.2 (***Color version follows page 112***) Project area boundary over Florida physiographic groups.

from the imagery and ancillary data layers. To link variation in land cover/land use with variation in the imagery and ancillary data, Visual Learning Systems' Feature Analyst software was used to classify DigitalGlobe QuickBird imagery and ancillary data into 26 classes of land use and land cover.

Following development and review of the draft map, a validation trip to the project area was conducted. Field visits focused on known areas of confusion, and specific areas noted by NOAA. Upon return from the field, additional Feature Analyst classifications were conducted on subareas, and extensive editing was performed. Figure 10.3 presents a portion of the imagery and the final map of the project area.

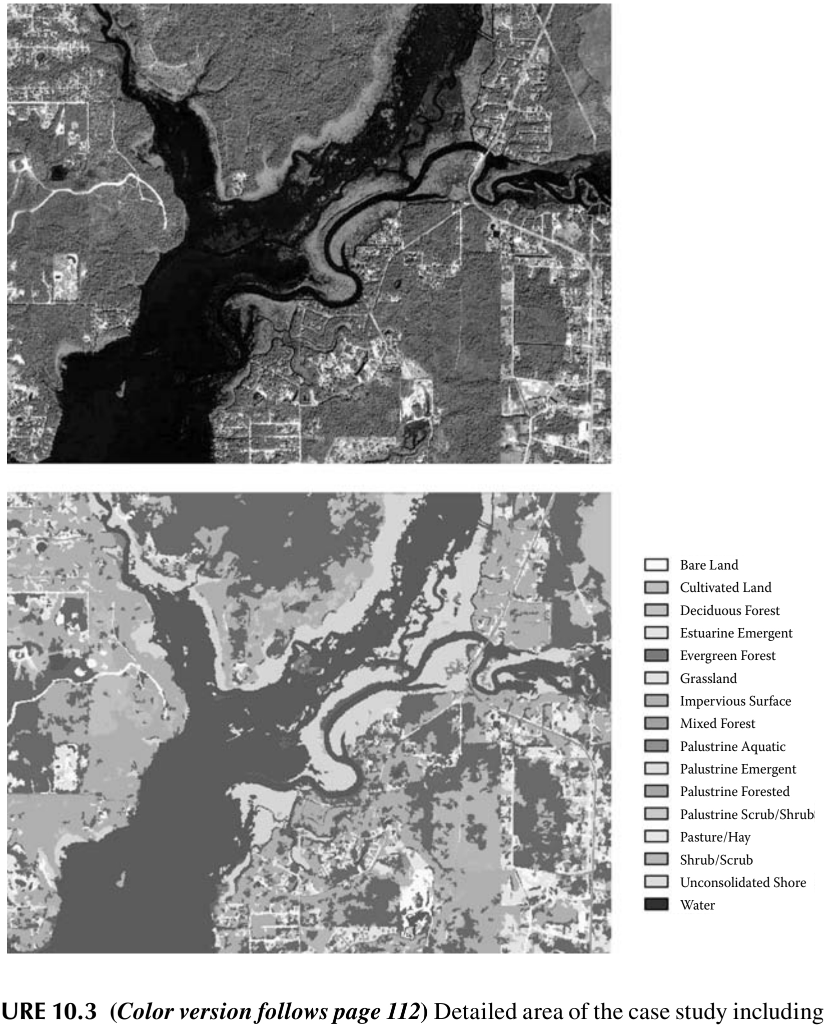


FIGURE 10.3 (***Color version follows page 112***) Detailed area of the case study including the QuickBird multispectral imagery and the final map.

TABLE 10.1
Land Cover/Land Use Classes and Subclasses Mapped for the Project

C-CAP Class	Subclass
Impervious	None
Cultivated Land	None
Pasture/Hay	None
Grassland	None
Deciduous Forest	None
Evergreen Forest	None
Mixed Forest	None
Scrub/Shrub	None
Water	None
Palustrine Forested Wetland	Deciduous
	Evergreen
Palustrine Scrub/Shrub Wetland	Deciduous
	Evergreen
Palustrine Emergent Wetland	Persistent (*Typha*/*Cladium*)
	Persistent (Sedges)
Estuarine Scrub/Shrub Wetland	Deciduous
	Evergreen
Estuarine Emergent Wetland	Persistent — High Marsh (*Juncus*)
	Persistent — High Marsh (*Salicornia*)
Unconsolidated Shore	None
Bare Land	Dirt Roads
	Other Bare Land
Palustrine Aquatic Bed	Floating Vascular
	Rooted Vascular
Estuarine Aquatic Band	Algal
	Rooted Vascular

DESIGN OF THE ACCURACY ASSESSMENT

What Are the Thematic Classes to Be Assessed?

NOAA chose the revised C-CAP Coastal Land Cover Classification[†] for this project and requested the development of a more detailed classification scheme that could be collapsed up into the C-CAP classes. Table 10.1 lists the land cover/use class labels used. Appendix 10.1 presents the totally exhaustive, mutually exclusive, and hierarchical classification scheme rules used in the project.

[†] http://www.csc.noaa.gov/crs/lca/tech_cls.html

The minimum mapping units for the project were

- 1/20th of an acre for impervious areas,
- 1/10th of an acre for areas that had been classified with moderate-resolution imagery as high, medium, low, or open-space development, and
- 1/8th of an acre for all other areas.

What Is the Appropriate Sampling Unit?

Sample units were polygons of land cover/land use class manually delineated on the Quickbird imagery. Polygons were chosen as the appropriate sample unit because the final map was a polygon coverage.

How Many Samples Should Be Taken?

The calibration field trip resulted in data collection for over 152 field samples. Upon return to the office, an additional 1500+ sites were collected through manual interpretation of the imagery for a total of 1720 project sample sites.

The accuracy assessment was planned with a goal of selecting 50 accuracy assessment sites per class from the total pool of 1720 project samples. However, as Table 10.2 illustrates, less than 50 samples were selected in some classes while more than 50 were collected in others. Reasons for the discrepancies are as follows:

- Because there is not much Palustrine Aquatic Bed or Unconsolidated Shore in the project area, fewer sites were selected for accuracy assessment of these classes, so that enough sites could be retained for making the map.
- Prior to the validation trip, image analysts started to believe that some of the deciduous sites were mislabeled. The confusion occurred with the mistaken identification of live oak trees as deciduous rather than evergreen hardwood trees. During the validation trip, several of these sites were visited, and it was confirmed that this mistake had been made. As a result, the image analyst reinterpreted all of the deciduous accuracy assessment sites[†] and relabeled 16 of them to either evergreen, mixed upland forest, or palustrine forested wetland.

How Should the Samples Be Chosen?

Accuracy assessment samples were selected from the total pool of project samples using a stratified random number generator in a statistical software package (S-PLUS), which randomly selected samples by relevant C-CAP moderate-resolution class (i.e., no tundra sites).

[†] At no time did the analyst have knowledge of which sites were in agreement or disagreement with the map site label.

TABLE 10.2
Accuracy Assessment Sites by C-CAP Class and Subclass

	Total Sites by	
Land Cover Land Use Class/Subclass	**Subclass**	**Class**
Bare Land		50
Impervious		50
Cultivated Land		50
Grassland		50
Scrub/Shrub		50
Deciduous Forest		34
Mixed Forest		58
Evergreen Forest		54
Pasture/Hay		50
Palustrine Forest		54
Palustrine Deciduous Forested Wetland	30	
Palustrine Mixed Forested Wetland	10	
Palustrine Evergreen Forested Wetland	14	
Palustrine Aquatic Bed		19
Palustrine Floating Vascular Aquatic Bed	18	
Palustrine Rooted Vascular Aquatic Bed	1	
Palustrine Emergent Wetland		50
Palustrine Emergent Sedge Wetland	23	
Palustrine Emergent Typha Wetland	27	
Palustrine Evergreen Scrub/Shrub Wetland		50
Estuarine Emergent Juncus Wetland		50
Unconsolidated Shore		12
Water		50
Total		**731**

DATA COLLECTION

What Should Be the Source of the Reference Data?

Reference labels were determined through field and office manual interpretation of the QuickBird multispectral 2.4 m Digital Globe imagery collected on November 9, 2004. While at first it may seem unusual to use the same imagery to make the map and also manually interpret for the reference data, the high spatial resolution of this imagery makes this feasible and reasonable. Historically, manual interpretation of medium-resolution imagery, such as Landsat TM imagery, was not typically used to create reference data. The quality of the manual interpretation could not be considered of higher accuracy as required of the reference data set. However, high-resolution satellite imagery and digital camera imagery are of such high spatial quality that manual interpretation provides for excellent reference data collection. While digital image analysis of the entire image to produce a map is more cost-effective and less time consuming than producing the map from manual interpretation, the use of manually interpreted reference labels is certainly reasonable and efficient.

How Should the Reference Data Be Collected?

The reference data were collected through the manual interpretation of the QuickBird 2.4 m resolution imagery either in the field or office. Project samples (the combination of accuracy and training samples) were chosen by field and office personnel at their discretion and were governed by the following criteria:

- *Informational homogeneity*—The site must represent one and only one land use per land cover class.
- *Spectral homogeneity*—The site should have less spectral variation within the polygons than between other polygons.
- *Minimum size*—Sites should be larger than the minimum mapping unit.
- *Projectwide distribution*—For a given class, analysts attempted to distribute the sites evenly across that type's distribution in the project area.

In addition, an unsupervised classification was run on the imagery to capture important spectral variation classes and an ongoing list was kept of the sites in each unsupervised class to ensure that all of the spectral variation in the imagery was captured.

Project samples were manually delineated on the imagery in ArcGIS because the map polygons had not been completed prior to sample selection. Manually delineating the samples, rather than choosing map polygons, creates the possibility that the sample polygons will cross multiple map polygons. This problem did occur in this project, as illustrated in Figure 10.4 by the multiple map polygons crossed by the turquoise accuracy assessment polygon. To ensure consistency and lack of bias, map

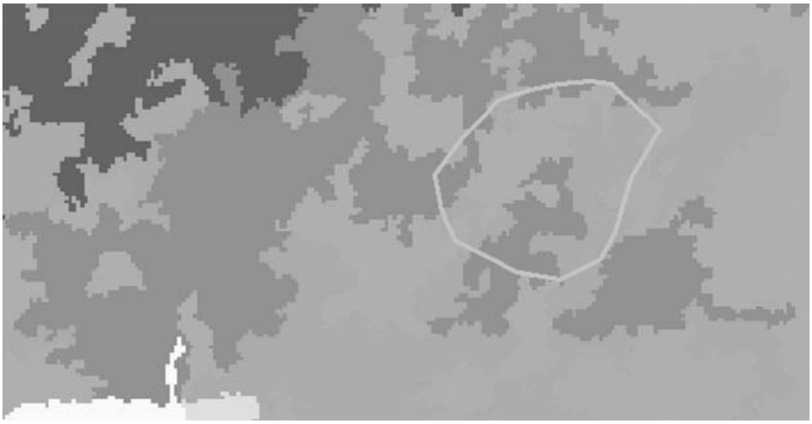


FIGURE 10.4 (***Color version follows page 112***) Illustration of a heterogeneous map accuracy assessment site.

labels for each sample polygon were determined by calculating the majority map class in each sample.

Polygon reference labels were derived using the classification scheme and all information available about the site. Manual determination of the label occurred following:

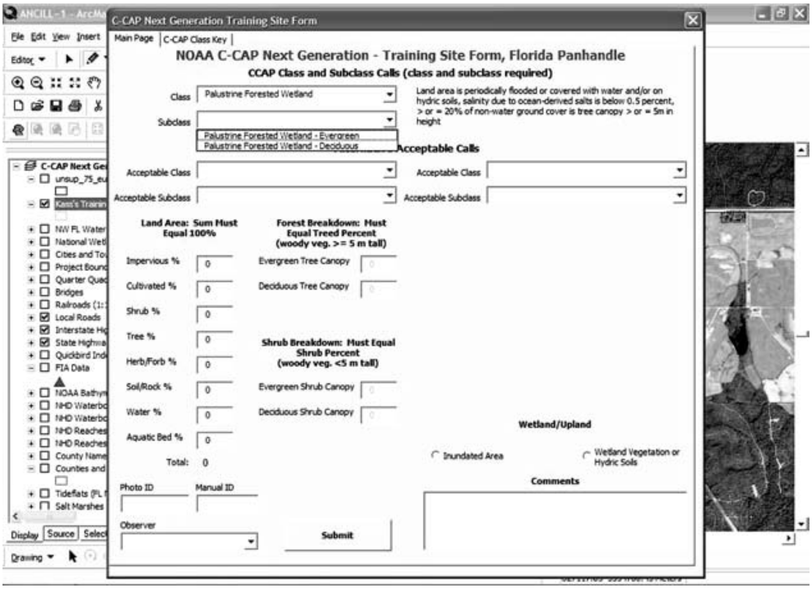
- Review of ancillary data concerning the site,
- A walkthrough of the site if the analyst was in the field,
- Review of field notes concerning the site if the analyst was in the office, and
- Review of the QuickBird 0.6 m resolution (colorized) imagery.

When Should the Reference Data Be Collected?

The reference data were collected either during the calibration trip or immediately thereafter in the office. Collecting the samples prior to creation of the map is cost-effective and allows for interim accuracy assessment as the map is being created. However, it does not ensure that an adequate number of samples per map class will be collected, which was the situation in this case study as illustrated in Table 10.3, which compares the number of accuracy assessment samples per map to reference class.

TABLE 10.3
Numbers of Map and Reference Samples by Class

	Number of Reference Sites	Number of Map Sites
Bare Land	50	56
Impervious	50	48
Cultivated Land	50	49
Grassland	50	71
Scrub/Shrub	50	51
Deciduous Forest	34	28
Mixed Forest	58	53
Evergreen Forest	54	65
Pasture/Hay	50	39
Palustrine Forested Wetland	54	59
Palustrine Aquatic Bed	19	18
Palustrine Emergent Wetland	50	40
Palustrine Scrub/Shrub	50	48
Estuarine Emergent Wetland	50	50
Unconsolidated Shore	12	6
Water	50	55



C-CAP Next Generation Training Site Form

Main Page | C-CAP Class Key

NOAA C-CAP Next Generation - Training Site Form, Florida Panhandle

CCAP Class and Subclass Calls (class and subclass required)

Class: Palustrine Forested Wetland

Land area is periodically flooded or covered with water and/or on hydric soils, salinity due to ocean-derived salts is below 0.5 percent, > or = 20% of non-water ground cover is tree canopy > or = 5m in height

Subclass:
- Palustrine Forested Wetland - Evergreen
- Palustrine Forested Wetland - Deciduous

Acceptable Calls

Acceptable Class: | Acceptable Class:

Acceptable Subclass: | Acceptable Subclass:

Land Area: Sum Must Equal 100%		Forest Breakdown: Must Equal Treed Percent (woody veg. >= 5 m tall)	
Impervious %	0	Evergreen Tree Canopy	0
Cultivated %	0	Deciduous Tree Canopy	0
Shrub %	0		
Tree %	0	Shrub Breakdown: Must Equal Shrub Percent (woody veg. <5 m tall)	
Herb/Forb %	0		
Soil/Rock %	0	Evergreen Shrub Canopy	0
Water %	0	Deciduous Shrub Canopy	0
Aquatic Bed %	0		
Total:	0		

Wetland/Upland

○ Inundated Area ○ Wetland Vegetation or Hydric Soils

Photo ID

Manual ID

Observer

Comments

Submit

FIGURE 10.5 (*Color version follows page 112*) Digital field form used in the project.

How Do I Ensure Consistency and Objectivity in My Data Collection?

Consistency and objectivity were ensured by:

1. Simultaneous training of image analysts with NOAA personnel in the field for identification of vegetative cover species, recognition of ecological relationships, delineation of polygon samples, and use of the field form (Figure 10.5).
2. Implementation of a digital field form linked to GPS (Figure 10.5). The form's functionality included pull-down menus and automated error checking, and also included the classification scheme rules for easy reference.
3. After all samples had been selected, each sample was reviewed one by one to ensure that the information collected for each site was complete and correct.

ANALYSIS

What Are the Different Analysis Techniques for Continuous versus Discontinuous Map Data?

The project mapped 26 discrete classes of land use/land cover. Because the classes are discontinuous, the only accuracy assessment analysis technique applicable is the error matrix.

What Is an Error Matrix and How Should It Be Used?

Table 10.4 displays the C-CAP class error matrix, which compares the final map and reference labels of the accuracy assessment samples.

What Are the Statistical Properties Associated with the Error Matrix and What Analysis Techniques Are Applicable?

Kappa analysis was performed on the error matrix, and the results are displayed in Table 10.4.

What Is Fuzzy Accuracy and How Can You Conduct a Fuzzy Accuracy Assessment?

As discussed in Chapter 9, one of the assumptions of the traditional or deterministic error matrix is that an accuracy assessment sample site can have only one reference label. However, classification scheme rules often impose discrete boundaries on continuous conditions in nature. In situations where classification scheme breaks represent artificial distinctions along a continuum of land cover, observer variability is often difficult to control, and although unavoidable, can have profound effects on results. While it is difficult to control observer variation, it is possible to use fuzzy logic to compensate for differences between reference and map data that are caused not by map error, but by variation in interpretation (Gopal and Woodcock, 1994). In this project, both deterministic and fuzzy error matrices were compiled and analyzed.

Table 10.5 displays both the deterministic and fuzzy error matrix for the C-CAP class map.

The overall deterministic accuracy is 83% and overall fuzzy accuracy is 91%. Table 10.6 summarizes the user's and producer's accuracies for the C-CAP class map, sorted first by producer's and then by user's accuracies.

The following summarizes the major findings of the accuracy assessment:

- The most accurate classes with combined user's and producer's deterministic and fuzzy accuracies above 80% at the class level are estuarine emergent wetland, water, cultivated land, impervious, bare land, palustrine forested wetland, palustrine aquatic bed, and palustrine scrub/shrub wetland.
- At the class level, all fuzzy user's accuracies exceed 80% except deciduous forest (75%). All producer's fuzzy accuracies exceed 80% except palustrine emergent wetland (72%) and unconsolidated shore (75%).
- Confusion exists between mixed forest and evergreen or deciduous forest. Some of the confusion between mixed forest and evergreen or deciduous forest occurs because the high resolution of the imagery allows for the classification of small polygons of homogeneous evergreen and deciduous trees. Taken together the clumps represent a mixed forest.
- Spatial autocorrelation exists with 4 of the 6 mixed forest reference sites, which are confused with deciduous forest in the error matrix. The 4 sites are all contained within one very large deciduous forest map polygon.

TABLE 10.4
Case Study Error Matrix

REFERENCE DATA

User's Accuracies

		Bare		Cult.	Grass	Scrub/	Decid.	Mixed	EG	Past/	PF	PA	PE	PS/S	EE	UnCon		Totals	Percent
		Land	Imperv	Land	Land	Shrub	Forest	Forest	Forest	Hay	Wet	Bed	Wet	Wet	Wet	Shore	Water		
	Bare Land	47	0	0	3	0	0	0	0	0	0	0	6	0	0	0	0	47/56	84.0%
	Impervious	0	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48/48	100.0%
	Cultivated Land	0	0	49	0	0	0	0	0	0	0	0	0	0	0	0	0	49/49	100.0%
	Grassland	2	2	1	39	8	0	0	0	19	0	0	0	0	0	0	0	39/71	55.0%
	Scrub/Shrub	0	0	0	3	38	1	0	1	0	0	0	3	5	0	0	0	38/51	75.0%
	Deciduous Forest	0	0	0	0	0	19	8	0	0	1	0	0	0	0	0	0	19/28	68.0%
M	Mixed Forest	0	0	0	0	0	13	37	1	0	2	0	0	0	0	0	0	37/53	70.0%
A	Evergreen Forest	0	0	0	0	2	0	11	50	0	0	0	2	0	0	0	0	50/65	77.0%
P	Pasure/Hay	0	0	0	3	0	0	0	0	31	0	0	0	0,0	0	0	0	31/34	91.0%
	Palustrine Forest Wetland	0	0	0	0	0	1	2	2	0	48	1	1	2	0	2	0	48/59	81.0%
	PalustrineAquatic Bed	0	0	0	0	0	0	0	0	0	0	16	1	0	0	1	0	16/18	89.0%
	Palustrine Emergent Wetland	0	0	0	2	0	0	0	0	0	1	1	35	0	0	1	0	35/40	88.0%
	Palustrine Scrub/Shrub Wetland	1	0	0	0	2	0	0	0	0	1	1	2	41	0	0	0	41/48	85.0%
	Estuarine Emergent Wetland	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	50/50	100.0%
	Unconsolidated Shore	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	6/6	100.0%
	Water	0	0	0	0	0	0	0	0	0	1	0	0	2	0	2	50	50/55	91.0%
	Producer's Accuracies																		
	Totals	47/50	48/50	49/50	39/50	38/50	19/34	37/58	50/54	31/50	48/54	16/19	35/50	41/50	50/50	6/12	50/50		
	Percent	94.0%	96.0%	98.0%	78.0%	78.0%	56.0%	64.0%	93.0%	62.0%	89.0%	84.0%	70.0%	82.0%	100.0%	50.0%	100.0%		

Overall Accuracy = 604/731 = 83.0 %

KAPPA = 0.814

TABLE 10.5
Case Study Error Matrix Showing Both Deterministic and Fuzzy Accuracies

		REFERENCE DATA																*User's Accuracies*			
		Bare Land	Imperv	Cult. Land	Grass Land	Scrub/ Shrub	Decid. Forest	Mixed Forest	EG Forest	Past/ Hay	PF Wet	PA Bed	PE Wet	PS/S Wet	EE Wet	UnCon Shore	Water	Determin. Totals	Percent Determin.	*Fuzzy Totals*	*Percent Fuzzy*
	Bare Land	47	0,0	0,0	1,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,5	0,0	0,0	0,0	0,0	47/56	84.0%	*49/56*	*88.0%*
	Impervious	0,0	48	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	48/48	100.0%	*48/48*	*100.0%*
	Cultivated Land	0,0	0,0	49	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	49/49	100.0%	*49/49*	*100.0%*
	Grassland	1,1	0,2	0,1	39	1,7	0,0	0,0	0,0	19,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	39/71	55.0%	*60/71*	*85.0%*
	Scrub/Shrub	0,0	0,0	0,0	2,1	38	0,1	0,0	1,0	0,0	0,0	0,0	0,3	3,2	0,0	0,0	0,0	38/51	75.0%	*44/51*	*86.0%*
	Deciduous Forest	0,0	0,0	0,0	0,0	0,0	19	2,6	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	19/28	68.0%	*21/28*	*75.0%*
M	Mixed Forest	0,0	0,0	0,0	0,0	0,0	9,4	37	0,1	0,0	1,1	0,0	0,0	0,0	0,0	0,0	0,0	37/53	70.0%	*47/53*	*89.0%*
A	Evergreen Forest	0,0	0,0	0,0	0,0	1,1	0,0	7,4	50	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0	50/65	77.0%	*58/65*	*89.0%*
P	Pasure/Hay	0,0	0,0	0,0	3,0	0,0	0,0	0,0	0,0	31	0,0	0,0	0,0	0,0	0,0	0,0	0,0	31/34	91.0%	*34/34*	*100.0%*
	Palustrine Forest Wetland	0,0	0,0	0,0	0,0	0,0	1,0	2,0	0,2	0,0	48	0,1	0,1	0,2	0,0	1,1	0,0	48/59	81.0%	*52/59*	*88.0%*
	Palustrine Aquatic Bed	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	16	0,1	0,0	0,0	0,1	0,0	16/18	89.0%	*16/18*	*89.0%*
	Palustrine Emergent Wetland	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,1	0,1	35	0,0	0,0	0,1	0,0	35/40	88.0%	*35/40*	*88.0%*
	Palustrine Scrub/Shrub Wetland	0,1	0,0	0,0	0,0	1,1	0,0	0,0	0,0	0,0	0,1	0,1	0,2	41	0,0	0,0	0,0	41/48	85.0%	*42/48*	*88.0%*
	Estuarine Emergent Wetland	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	50	0,0	0,0	50/50	100.0%	*50/50*	*100.0%*
	Unconsolidated Shore	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	6	0,0	6/6	100.0%	*6/6*	*100.0%*
	Water	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,0	0,0	2,0	0,0	2,0	50	50/55	91.0%	*55/55*	*100.0%*
	Producer's Accuracies																				
	Deterministic Totals	47/50	48/50	49/50	39/50	38/50	19/34	37/58	50/54	31/50	48/54	16/19	35/50	41/50	50/50	6/12	50/50				
	Percent Deterministic	94.0%	96.0%	98.0%	78.0%	76.0%	56.0%	64.0%	93.0%	62.0%	89.0%	84.0%	70.0%	82.0%	100.0%	50.0%	100.0%				
	Fuzzy Totals	48/50	48/50	49/50	45/50	41/50	29/34	48/58	51/54	50/50	50/54	16/19	36/50	46/50	50/50	9/12	50/50				
	Percent Fuzzy	96.0%	96.0%	98.0%	90.0%	82.0%	85.0%	83.0%	94.0%	100.0%	93.0%	84.0%	72.0%	92.0%	100.0%	75.0%	100.0%				

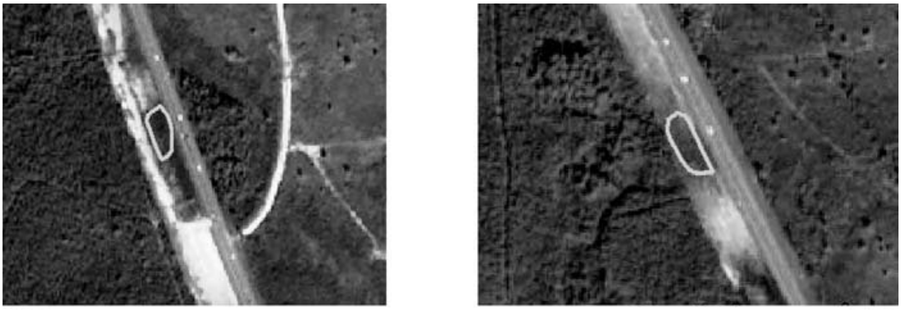
Overall Accuracies			
Deterministic		Fuzzy	
604/731	83.0%	666/731	91.0%

TABLE 10.6
Final Producer's and User's Accuracies by C-CAP Class

	Deterministic Producer's Accuracy (%)	Fuzzy Producer's Accuracy (%)	Deterministic User's Accuracy (%)	Fuzzy User's Accuracy (%)
Estuarine Emergent Wetland	100	100	100	100
Water	100	100	91	100
Cultivated Land	98	98	100	100
Impervious	96	96	100	100
Bare Land	94	96	84	88
Evergreen Forest	93	94	77	89
Palustrine Forested Wetland	89	93	81	88
Palustrine Aquatic Bed	84	84	89	89
Palustrine Scrub/Shrub Wetland	82	92	85	88
Grassland	78	90	55	85
Scrub/Shrub	76	82	75	86
Palustrine Emergent Wetland	70	72	88	88
Mixed Forest	64	83	70	89
Pasture/Hay	62	100	91	100
Deciduous Forest	56	85	68	75
Unconsolidated Shore	50	75	100	100

Use of the random number generator to choose the accuracy assessment sites should have (but did not) preclude this type of problem. Relying on map polygons as sample units, rather than manually delineating samples, would have negated this type of spatial autocorrelation.

- Nineteen pasture hay sites have an acceptable alternative reference label of grassland. Thus, pasture/hay has low deterministic class values, but high fuzzy class values. Of the 19 sites, 18 sites were office interpreted, where it is almost impossible to distinguish between pasture/hay and grassland. Thus, the sites were given fuzzy labels. These 18 sites contribute almost one third of the eight-point difference between the overall deterministic (83%) and fuzzy (91%) accuracies.
- Four palustrine emergent sedge wetland reference sites were confused with bare land map sites. Three of the sites are along newly constructed roads and were probably bare land in the imagery, but populated with vegetation in the months between the capture of the imagery and the calibration trip. Figure 10.6 shows two of these sites.
- Seven scrub/shrub reference sites are confused with grassland. All seven sites are regenerating pine forests with an overstory of turkey oak (which did not have its leaves at the time of the imagery) and an understory of grass, shrubs, and pine seedlings.



Site 1088

Site 1491

FIGURE 10.6 (***Color version follows page 112***) Examples of sites with palustrine emergent scrub/shrub reference labels confused with bare land map labels.

LESSONS LEARNED

As in most projects, lessons are learned continually as the project progresses. Specific lessons learned during the accuracy assessment portion of the case study include:

1. To eliminate spatial autocorrelation in accuracy assessment sites, no more than one accuracy assessment site should be allowed to fall within one map polygon.
2. During the initial training data/calibration trip, the location of accuracy assessment sites should not be delineated on the hardcopy calibration imagery. The sites should only be delineated digitally on a laptop, leaving the hardcopy calibration images for notes only, so that the calibration images can be used in map editing.
3. Determining the map label of a "mixed" (e.g., mixed deciduous/evergreen) class accuracy assessment site is problematic with classification of high-resolution imagery which is often capable of individually distinguishing the components of a "mixed" area. Using the simple majority of the site to create the map label (as was done in this project) is relatively easy but may produce an incorrect label, especially in a site composed of close-to-equal proportions of the components of a mixed site. For example, an accuracy assessment site could intersect several map polygons and be composed of the following components:

- 30% mixed
- 32% evergreen
- 28% deciduous

If a majority rule is used, the map label would be evergreen, but clearly the polygon is actually a mixed forest. Two solutions to this problem are possible:

- If possible, sample polygons should be chosen from the actual map polygons. This is easy to do, if sample selection is to follow map finalization. In situations where sample selection occurs prior to map finalization, it is often possible to create unlabeled polygons early in the mapping project that can be used as the population from which samples are chosen. Reference labels for the polygons can be determined during the calibration trip. Map labels are determined when the mapping portion of the project is complete.

- If it is not possible to create polygons early, and sample selection must be carried out before the map is final, then more complex rules than a simple majority should be considered for labeling the map samples. Using the actual rules from the classification scheme is the best alternative. For example, the classification scheme for the case study project labeled a nonwetland forested polygon as evergreen or deciduous only if the percentage cover of the polygon was 75% evergreen or deciduous, respectively. Under that rule, our example polygon considered earlier would have been labeled "mixed forest" rather than "evergreen."

APPENDIX 10.1

Decision Rules for the Classification Scheme

If land area is > or = to 80% impervious surface over 1/20th acre or more, then **Impervious (1)**

*If land area is designated by moderate resolution map as High, Medium, Low, or Open Space Developed, then minimum mapping unit is 1/10*th *of an acre.*

*Else minimum mapping unit is 1/8*th *of an acre.*

Developed and Undeveloped areas will be determined by the Moderate Resolution map for the project area.

Else if land area is > or = to 75% open water, then **Water (2)**

Else if land area is periodically flooded and/or covered with water, or if image signature is "wet," then **Wetland (3)**

> If salinity due to ocean-derived salts is below 0.5%, then **Palustrine Wetland (3.1)**
>
> > If > or = 20% of nonwater ground cover is tree canopy > or = 5 m in height, then **Palustrine Forested Wetland (3.1.1)**
> >
> > > If > or = 75% of nonwater cover is deciduous tree, then **Palustrine Deciduous Forested Wetland (3.1.1.1)**
> > >
> > > Else if > or = 75% of nonwater cover is evergreen tree, then **Palustrine Evergreen Forested Wetland (3.1.1.2)**
> > >
> > > Else **Palustrine Mixed Forested Wetland (3.1.1.3)**
>
> Else if > or = 20% of nonwater ground cover is woody < 5 m in height, then **Palustrine Scrub/Shrub Wetland (3.1.2)**
>
> > > If a majority (> or = 51%) of shrub cover is deciduous, then **Palustrine Deciduous, Shrub Wetland (3.1.2.1)**
>
> Else **Palustrine Evergreen Shrub Wetland (3.1.2.2)**
>
> Else if > 50% plants growing and forming a continuous surface principally on or at the water surface, then **Palustrine Aquatic Bed (3.1.3)**

If a majority of vegetative cover is floating vascular, then **Palustrine Floating Vascular Aquatic Bed (3.1.3.1)**

Else if a majority of vegetative cover is rooted vascular, then **Palustrine Rooted Vascular Aquatic Bed (3.1.3.2)**

Else **Palustrine Emergent Wetland (3.1.4)**

If a majority of cover is persistent *Typha* spp. *or Cladium* spp*,* then **Palustrine Typha/Cladium Persistent Wetland (3.1.4.1)**

Else if a majority of vegetative cover is persistent *Scirpus* spp*,* then **Palustrine Scirpus Persistent Wetland (3.1.4.2)**

Else if a majority of vegetative cover is persistent Sedges, then **Palustrine Sedge Persistent Wetland (3.1.4.3)**

Else if a majority of vegetative cover is persistent *Phragmites* spp*,* then **Palustrine Phragmites Persistent Wetland (3.1.4.4)**

Else **Palustrine Emergent Mixed Wetland (3.1.4.5)**

Else if (salinity due to ocean-derived salts is equal to or > 0.5%), then **Estuarine Wetland (3.2)**

If > or = 20% of nonwater ground cover is tree canopy that = or > 5 m in height, then **Estuarine Forested Wetland (3.2.1)**

Else if > or = 20% of nonwater ground cover is woody < 5 m in height, then **Estuarine Scrub/Shrub Wetland (3.2.2)**

If a majority (> or = 51%) of shrub cover is deciduous, then **Estuarine Deciduous Shrub Wetland (3.2.2.1)**

Else **Estuarine Evergreen Shrub Wetland (3.2.2.2)**

Else if > 50% plants growing and forming a continuous surface principally on or at the water surface, then **Estuarine Aquatic Bed (3.2.3)**

If a majority of vegetative cover is rooted vascular, then **Estuarine Rooted Vascular Aquatic Bed (3.2.3.1)**

Else if a majority of vegetative cover is algal, then **Estuarine Algal Aquatic Bed (3.2.3.2)**

Else **Estuarine Emergent Wetland (3.2.4)**

If majority of vegetative cover is low marsh *Spartina* spp*,* then **Estuarine Emergent Spartina Wetland (3.2.4.1)**

Else if majority of vegetative cover is low marsh *Juncus* spp, then **Estuarine Emergent Juncus Wetland (3.2.4.2)**

Else if majority of vegetative cover is high marsh *Salicornia* spp*,* then **Estuarine Emergent Salicornia Wetland (3.2.4.3)**

Else if majority of vegetative cover is nonpersistent, then **Estuarine Emergent Non-persistent Wetland (3.2.4.4)**

Else if land area is characterized by herbaceous vegetation that has been planted or is intensely managed for the production of food, feed, or fiber, then **Cultivated Land (4)**

Else if land area is characterized by grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seeds or hay crops, then **Pasture/Hay (5)**

Else if land area is > 50% tundra vegetation, then **Tundra (6)**

Else if land area is > 75% snow/ice throughout the year, then **Snow/Ice (7)**

Else if land area is > 85% covered with bare rock, gravel, sand, silt, clay, or other earthen materials, then **Undeveloped Bare Land (8)**

If characterized by intertidal, or intermittently flooded areas (mud flats), then **Unconsolidated Shore (8.1)**

Else **Bare Land (8.2)**

Else if tree canopy (woody vegetation) > 20% of land area and tree canopy (woody vegetation) > or = 5 m tall, then **Forest (9)**

If tree canopy (woody vegetation) > or = 75% deciduous, then **Deciduous Forest (9.1)**

Else if tree canopy > or = 75% evergreen, then **Evergreen Forest (9.2)**

Else **Mixed Forest (9.3)**

Else if tree canopy (woody vegetation) > 20% of land area and tree canopy ≤ 5 m tall, then **Shrub/Scrub (10)**

Else **Grassland (11)**

11 Advanced Topics

This chapter begins with a discussion of change detection accuracy assessment. The complexities of conducting such an assessment are presented along with the formulation of the change detection error matrix. A key issue in any change detection accuracy assessment is the realization that change is a rare event and sampling must occur to specifically deal with this issue. While it is possible to create a change detection error matrix, it requires a tremendous amount of work. A compromise two-step method is proposed and demonstrated that may provide a more practical approach to assessing the accuracy of change. The chapter concludes with a short discussion of multilayer accuracy assessment.

CHANGE DETECTION

An increasingly popular application of remotely sensed data is for use in change detection. Change detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times (Singh, 1989). Four aspects of change detection are important: (1) detecting that changes have occurred, (2) identifying the nature of the change, (3) measuring the areal extent of the change, and (4) assessing the spatial pattern of the change (Brothers and Fish, 1978; Malila, 1985; Singh, 1986). Techniques to perform change detection with digital imagery have become numerous because of increasing versatility in manipulating digital data, better image analysis software, and increasing computing power. Change detection accuracy assessment is an important component of any change analysis project.

Assessing the accuracy of a single date or one point in time (OPIT) thematic map generated from remotely sensed data as presented in this book is a complex but attainable endeavor. In addition to the complexities associated with a single-date accuracy assessment of remotely sensed data, change detection presents even more difficult and challenging issues to consider. The very nature of change detection makes quantitative analysis of the accuracy difficult. For example, how does one obtain reference data for images that were taken in the past? How does one sample enough areas that will change in the future to have a statistically valid assessment? Which change detection technique will work best for a given change in the environment? Positional accuracy also plays a big role in change detection. It is critical to determine if an increase in size or shape of an area has actually occurred or if the apparent change is simply due to a positional error. Figure 11.1 is a modification of the sources of error figure for a single-date assessment presented early on in this book (Chapter 2, Figure 2.5) and shows how complicated the error sources get when performing a change detection. Most of the studies on change detection conducted up to now do not present quantitative results of their work, which makes it difficult to determine which method should be applied to a future project.

The following section presents the topics to be considered when preparing to perform a change detection accuracy assessment. There are three critical components

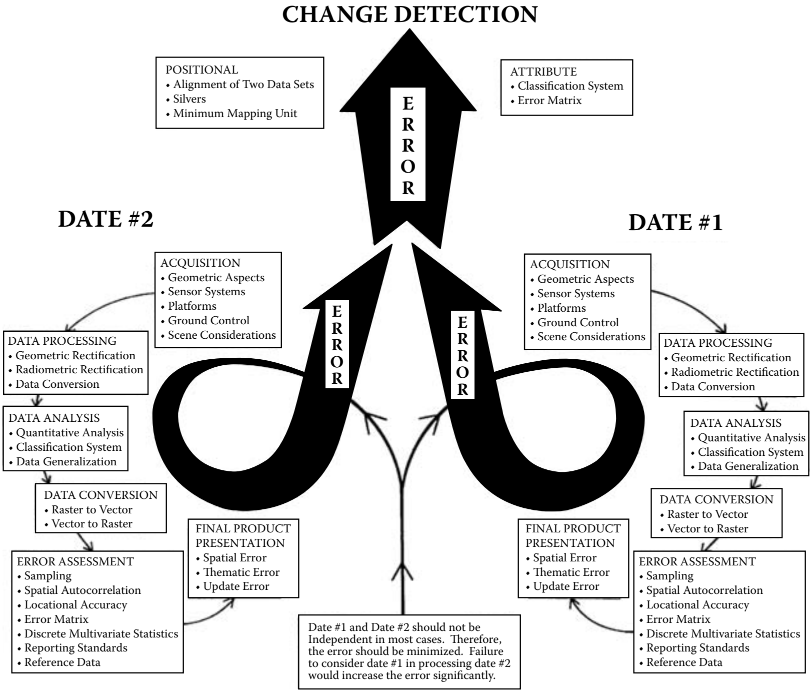


FIGURE 11.1 Sources of error in a change detection analysis from remotely sensed data. (Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton R.G. 1996. Accuracy assessment: A critical component of land cover mapping, in *Gap Analysis: A Landscape Approach to Biodiversity Planning.* A peer-reviewed proceedings of the ASPRS/GAP Symposium. Charlotte, NC, pp. 119–131.)

that must be considered in any change detection accuracy assessment. These are (1) reference data, (2) sampling, and (3) the change detection error matrix.

Reference Data

A collection of valid reference data is central to any accuracy assessment, whether it is a single-date assessment or for evaluating a change detection. Let us imagine conducting a change detection project in 2008 by comparing a vegetation/land cover map generated from 1998 Landsat Thematic Mapper (TM) imagery (call this time 1) with another map generated from 2008 TM imagery (call this time 2). Let us further suppose that the classification scheme used for both maps is the same because we created both maps. Reference data for evaluating the 2008 map could be collected on the ground in 2008 or even 2009 and still be considered valid. However, how can reference data for assessing the 1998 map and, therefore, the change detection, be obtained?

There are a few possible answers. The most probable answer is that there are no reference data available and really no way to assess the change. Second, there might be some aerial photography of the area that was acquired around the same time as the 1998 TM imagery. Of course, scale is an issue here. If the photos are of such small scale that sufficient detail cannot be accurately interpreted from them given the classification scheme used in the mapping project, then the photos cannot be used to provide reference data. Even if the scale is sufficient, photo interpretation is subject to error and the reference data may be flawed. Third, there may be some ground inventory of the area in question that can be used as reference data. This third possibility is extremely slim. Therefore, the lack of valid reference data is often a limiting factor when attempting to conduct a change detection accuracy assessment.

Sampling

There is one overriding issue that must be considered when sampling for change detection accuracy assessment that is beyond the sampling issues already presented in this book for a single-date assessment. Failure to consider this issue dooms the assessment to a wasted effort. It must be remembered that change is a rare event. Under normal circumstances, it would be unusual for more than 10% of a given area to change in a 5–10 year period. More likely, the change would be closer to 5%. In extreme cases, high change rates like 20% are possible. Of course, in certain catastrophic situations, change may be even higher.

Now consider sampling to find the change areas. Using a random sampling approach, even in a map with high change rates (20%), on average only 2 out of 10 samples will find any change. In the more usual case, it could take up to 20 samples before an area that has changed is found. Given the time and effort to collect samples for accuracy assessment, this sampling in the nonchange areas must be avoided. Stratification of the area to prioritize sampling in the change areas should be employed. However, exactly how to delineate these strata is not always obvious. If all the change areas were known, then no new map of change would be needed.

Fortunately, for many applications, logic or experience dictates the likely places for change to occur. For example, urban change occurs in areas around existing urban centers. It is extremely rare to find a new city built in the middle of nowhere. Sampling for urban change in a buffer zone around an urban center increases your chances of finding it when compared to a randomly placed sample. In this scenario, taking some portion of your sample in high-priority areas makes sense. Macleod and Congalton (1998) conducted a change detection accuracy assessment for monitoring eel grass change in Great Bay, New Hampshire. Because change is such a rare event, it was necessary to proportionally allocate more sampling effort to areas where change was more likely to happen. In this example, for mapping eel grass, we know that it is very unlikely that eel grass will grow in the channel (i.e., the deep water areas). Sampling should be limited in the channel. On the other hand, the eel grass is more likely to expand around existing eel grass beds and in shallow areas where no eel grass currently grows. The sampling effort should be increased in these areas. Therefore, we modified our sampling efforts in the following ways: (1) Only 10% of our sampling effort occurred in the deep water areas, (2) 40% of our sampling effort was dedicated to a buffer area within one sample grid (i.e., pixel) of existing eel grass, and (3) 50% of our sampling effort was dedicated to shallow areas where new eel grass seedlings could occur. In this way, the sampling was designed to find the change areas (Congalton and Brennan, 1998).

There are many other factors to consider when sampling for change detection accuracy assessment. However, failure to note that change is a rare event influences all these other factors and must be considered first.

Change Detection Error Matrix

In order to apply established accuracy assessment techniques to change detection, the standard single-date classification error matrix needs to be adapted to a change detection error matrix as proposed by Congalton and Macleod (1994) and Macleod and Congalton (1998). This new matrix has the same characteristics of the single-date classification error matrix, but also assesses errors in changes between two time periods (between time 1 and time 2) and not simply a single classification.

Table 11.1 reviews a single-date error matrix and the associated descriptive statistics: overall, producer's, and user's accuracies that have already been presented in this book. This single-date error matrix is for three vegetation/land cover categories (F = Forest, U = Urban, and W = Water). The matrix is of dimension 3 × 3. The y-axis of the error matrix represents the three vegetation/land cover categories as derived from the remotely sensed classification (i.e., the map) and the x-axis shows the three categories identified in the reference data.

The major diagonal of this matrix is highlighted and indicates correct classification. In other words, when the classification indicates the category was F and the reference data agrees that it is F, then the [F, F] cell in the matrix is tallied. The same logic follows for the other categories: U and W. Off-diagonal elements in the matrix indicate the different types of confusion (called omission and commission errors)

TABLE 11.1
An Example of a Single-Date (One Point in Time) Error Matrix Showing Overall, User's, and Producer's Accuracies

		Reference Data			Row
		F	**U**	**W**	**Total**
Classified Data	**F**	40	9	8	57
	U	1	15	5	21
	W	1	1	20	22
	Column Total	42	25	33	100

Land Cover Categories

F = Forest
U = Urban
W = Water

OVERALL ACCURACY
= 40 + 15 + 20
= 75/100 = 75%

PRODUCER'S ACCURACY

F = 40/42 = 95%
U = 15/25 = 60%
W = 20/33 = 61%

USER'S ACCURACY

F = 40/57 = 70%
U = 15/21 = 71%
W = 20/22 = 91%

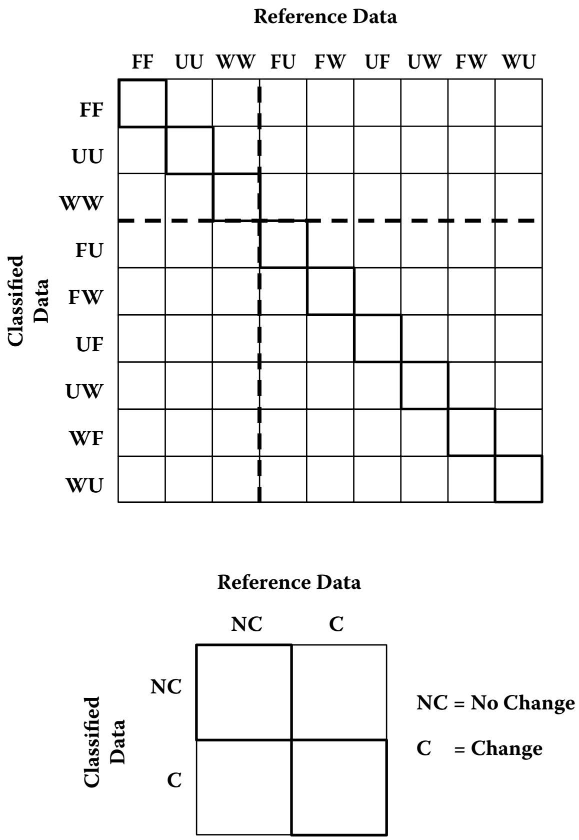


FIGURE 11.2 A change error matrix for the same three map categories (Forest, Urban, Water) as the single-date matrix and the collapsed no change/change matrix. (Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton R.G. 1996. Accuracy assessment: A critical component of land cover mapping, in *Gap Analysis: A Landscape Approach to Biodiversity Planning.* A peer-reviewed proceedings of the ASPRS/GAP Symposium. Charlotte, NC, pp. 119–131.)

that exist in the classification. Omission error occurs when an area is omitted from the correct category. Commission error occurs when an area is placed in the wrong category. This information is helpful in guiding the user to where the major problems exist in the classification.

The top part of Figure 11.2 shows a change detection error matrix generated for the same three vegetation/land cover categories (F, U, and W). Note, however, that the matrix is no longer of dimension 3 × 3, but rather 9 × 9. This is because we are no longer looking at a single classification, but rather a change between two different maps generated at different times. Remember, in a typical error matrix, there is one row and column for each map category. However, in assessing change detection, the error matrix is: the size of the number of categories squared. Therefore, the question of interest is: What category was this area at time 1 and what is it at time 2? The answer has 9 possible outcomes for each axis of the matrix (F at time 1 and F at time 2, U at time 1 and U at time 2, W at time 1 and W at time 2, F at time 1 and U at time 2, F at time 1 and W at time 2, U at time 1 and F at time 2, U at time 1 and W at time 2, W at time 1 and F at time 2, or W at time 1 and U at time 2), all of which are indicated along the rows and columns of the error matrix. It is then important to note what the remotely sensed data said about the change and compare it to what the reference data indicates. This comparison uses precisely the same logic as for the single classification error matrix; it is just complicated by the two time periods (i.e., the change). Again, the major diagonal indicates correct classification, while the off-diagonal elements indicate the errors or confusion. The descriptive statistics (i.e., overall, user's, and producer's accuracies) can also be computed.

It is important to note that the change detection error matrix can also be simplified or collapsed into a 2 × 2 no change/change error matrix (bottom of Figure 11.2). The no change/change error matrix can be generated by summing the appropriate cells in the four sections of the complete change detection error matrix partitioned by the dotted lines. For example, to get the number of areas for which both the classification and reference data determined that no change had occurred between the two dates, you would simply sum all nine cells in the upper left box (the areas that did not change in either the classification or reference data). To summarize or collapse the cells in which change occurred in both the classification and reference data, you would sum the 36 cells in the lower right box. The other two cells in the no change/change matrix would be determined in a similar manner. From this no change/change error matrix, the analysts can easily determine if a low accuracy was due to a poor change detection technique, misclassification, or both.

It should be obvious to the reader that performing a change detection accuracy assessment is a very complex undertaking. By simply scaling the single-date assessment methodology, the size of the error matrix increases, as does the number of samples required for the assessment. In the example error matrix for a single-date, three-class map (Chapter 10, Table 10.1), 150 samples (3 classes × 50 samples/class) are required. By adding a second time period, the number of samples grows to 450 (9 change classes × 50 samples/class). If a single-date mapping project had 10 classes, the required sample size would be 5000 (10 × 10 × 50 samples per class) samples. Since not all changes are logical within a given period of time (e.g., one would not expect water to become forest in 5 years), that number would likely be smaller, but the number of required samples is still much greater than for a single-date accuracy assessment and probably not feasible under most time and budget conditions.

Therefore, while it may not be possible to perform a complete change detection accuracy assessment and generate a change detection error matrix for every change detection project, it is still relevant to try to answer the following two questions: (1) How accurate have the areas that have changed between time 1 and time 2 been mapped? and (2) How well was the change captured? To answer these questions, the change detection accuracy assessment process can be divided into two steps instead of using a single assessment and the change detection error matrix approach.

Two-Step Approach to Change Detection Accuracy Assessment

If it is not possible to use the change detection error matrix approach to perform your change detection accuracy assessment, then you may wish to use this two-step approach. This method does not allow you to obtain the accuracy of all the change classes (e.g., the map was forest in time 1 and is now residential in time 2; from forest to residential), but it does provide for assessing the accuracy of the areas that changed in time 2 and to assess how well the overall changes were captured.

The first step in this process is to assess the accuracy of just the areas that changed between the two time periods in question. In other words, conduct a single-date accuracy assessment only on the areas that changed between time 1 and time 2. The sampling procedure is similar to that of a traditional single-date accuracy assessment with the requisite number of samples per land cover class selected using a chosen sampling strategy from the map area. However, in this case, only areas classified as change (i.e., the map class is different in time 2 than it was in time 1) are used to select the samples. The accuracy assessment only needs to be conducted for the change areas for time 2 because the rest of the map has the same accuracy as the map did in time 1 for all the areas that did not change.

The second step in this process is simply a change/no change validation. This step is similar to collapsing the change detection error matrix to the change/no change (2 × 2) matrix presented at the bottom of Figure 11.2. The difference here is that instead of having to sample to fill in the entire change detection error matrix, the sampling is performed to only assess the change/no change. Treating the map as a binary or two-class scheme (change/no change) requires a simpler sampling technique than the multinomial situation of a complete change detection error matrix. Since we are working with a two-case situation where we only wish to know whether the classification is change or no change, we can use the binomial distribution to calculate the sample size. Ginevan (1979) introduced this sampling method to the remote sensing community and concluded that:

- The method should have a low probability of accepting a map of low accuracy.
- It should have a high probability of accepting a map of high accuracy.
- It should require a minimum number of samples.

Computing the sample size for the binomial approach requires the use of a look-up table that presents the required sample size for a given minimum error and a desired level of confidence. For example, a map with a chosen accuracy of 90% (10% error) and using a 95% confidence level (at 95%, we run the risk of a 1 in 20 chance that we reject a map that is actually correct), the minimum number of samples required for the assessment is 298. Given this sample size, the map is rejected as inaccurate if more than 21 samples are misclassified.

Therefore, this two-step approach is quite effective. While not producing a complete change detection error matrix or assessing the accuracy of each change (to–from) class, it does provide a means of assessing the accuracy of the labeling (thematic accuracy) of the areas that changed between the two time frames. In addition, an assessment of whether or not the change is accurately captured can be generated using the binomial change/no change approach. These two steps are considerably easier and require significantly less time, money, and resources than using the change detection error matrix approach. However, if the required resources are available, the change detection error matrix provides the most information about the change analysis and is the recommended approach to use.

Case Study

This case study details the change detection accuracy assessment for the Kentucky Landscape Census (KLC), National Land Cover Dataset (NLCD) update from 2001 to 2005. Appendix 11.1 presents a list of the land cover classes and a brief description of each. It was not possible in this project, due to limited time and resources, to collect enough data to generate a change detection error matrix. Instead, the goal for this change analysis was to assess the accuracy of the change classification (the areas that were changed between 2001 and 2005) and to determine how well change, in general, was captured between the two dates.

To accomplish this task, the accuracy assessment was completed in two steps. First, the 2005 change areas were assessed as a single-date land cover map. Validation samples were collected by interpreting high-resolution imagery collected within a year of the 2005 Landsat imagery (NLCD classification). Generating an error matrix with at least 30 samples per class, the overall map accuracy was computed as well as the omission and commission error rates for each individual thematic class within the map. Second, a change mask was assessed as a binary change/no change map. Samples were collected using a stratified random selection approach within Kentucky. To limit the selection area to areas of likely change, various strata layers were created to prioritize the selection of the samples. By conducting the assessment in these two separate steps, the following questions were answered: How accurate is the 2005 change map and how well was land cover change captured?

Step 1: Accuracy of the Change Areas

The first step in the change detection accuracy assessment was to assess the accuracy of the areas that changed as a separate, single-date map. The sampling procedure is similar to that of any traditional accuracy assessment with between 30 and 50 samples per land cover class randomly selected from the mapping area. However, in this case, only areas classified as change between 2001 and 2005 were used to draw samples, and only the 2005 classification was assessed. The reference data for this time period was 1 m color imagery from the National Agricultural Imagery Program (NAIP). Complete coverage for all of Kentucky was available from NAIP.

The results of the accuracy assessment of the 2005 change areas are presented in error matrix form in Table 11.2. Inspection of the error matrix shows that not all classes were assessed for accuracy and included in the error matrix. While all classes in the USGS classification scheme (Appendix 11.1) were classified, the bulk of the change occurred in only some of the land cover classes. Changes to cover types such as wetland features or forest regrowth classes did not occur in sufficient amounts and, therefore, too few samples were available with which to assess the accuracy of these classes.

TABLE 11.2
Error Matrix Shows the Accuracy of the 2005 Change Areas

		REFERENCE								User's Accuracies			
	LABELS	Water	Developed Open Space	Developed Low Intensity	Developed Medium Intensity	Developed High Intensity	Bare Land	Shrub	Grassland	Deterministic Totals	Deterministic Accuracies	Fuzzy Totals	Fuzzy Accuracies
	Water	15	0,0	0,0	0,0	0,0	0,4	0,0	0,0	15/19	78.9%	15/19	78.9%
	Developed Open Space	0,0	9	21,1	4,0	0,0	2,5	0,0	1,2	9/45	20.0%	37/45	82.2%
	Developed Low Intensity	0,0	4,0	24	3,0	1,1	1,3	0,0	0,1	24/38	63.2%	33/38	86.8%
MAP	Developed Medium Intensity	0,0	0,0	4,0	5	1,0	0,3	0,0	0,0	5/13	38.5%	10/13	76.9%
	Developed High Intensity	0,0	0,0	2,0	0,0	11	0,0	0,0	0,0	11/13	84.6%	13/13	100.0%
	Bare Land	0,0	0,0	1,2	0,1	0,0	49	0,0	1,0	49/54	90.7%	51/54	94.4%
	Shrub	0,0	0,1	0,1	0,0	0,0	0,6	31,0	19,12	31/61	50.8%	50/61	82.0%
	Grassland	0,0	1,0	0,0	0,0	0,0	9,18	0,2	40,0	40/70	57.1%	50/70	71.4%
	Producer's Accuracies												
	Deterministic Totals	15/15	9/15	24/56	5/13	11/14	49/100	31/33	40/76				
	Deterministic Accuracies	100.0%	60.0%	42.9%	38.5%	78.6%	49.0%	91.7%	52.6%	***Overall Accuracies***			
	Fuzzy Totals	15/15	14/15	52/56	12/13	13/14	61/100	31/33	61/76	**Deterministic**		**Fuzzy**	
	Fuzzy Accuracies	100.0%	93.3%	92.9%	92.3%	92.9%	61.0%	91.7%	80.3%	184/313	58.8%	250/313	79.9%

The overall deterministic accuracy for the 2005 change areas is 58.8%, and the fuzzy accuracy is 79.9%. The 21.1% difference between deterministic and fuzzy accuracies can be attributed to two similar effects. Much of the increases in fuzzy accuracy are related to class confusion within the developed classes and a separate but similar confusion between grassland and shrub land. The four developed classes are defined by the percentage of impervious surface in each class:

- Developed, Open: 0–25%
- Developed, Low Intensity: 26–50%
- Developed, Medium Intensity: 51–75%
- Developed, High Intensity: 76–100%

While this division results in well-defined class boundaries, there is a degree of uncertainty associated with the percent impervious map that translates to the final classification. As the pixels in the percent impervious map are assigned through a statistical regression analysis technique, a degree of error is associated with each estimated value, generally ±10%. This results in pixels within less than 10% of the class boundaries potentially being in two developed categories. For example, a pixel with a value of 55% would be categorized as "Developed, Medium Intensity"; however, by factoring in the degree of uncertainty with the estimate, it could also be categorized as "Developed, Low Intensity." For the purposes of this accuracy assessment, a developed accuracy point was given a fuzzy interpretation if, by factoring in the degree of uncertainty, it satisfied the categorization criteria for more than one developed class. The predominance of the developed categories in the final map and their inherent uncertainty contribute to the variance between the deterministic and fuzzy estimates.

Similar confusion between grass and shrub is the second major contributor to the difference between the deterministic and fuzzy accuracies. These classes are rarely found naturally in Kentucky. Instead, the two classes more often represent a transition or succession of vegetation growth after a disturbance related to forestry or mining. Determining the amount of shrub to grass vegetation based on the hard class breaks leads to fuzziness in some accuracy calls.

Step 2: Change/No Change Assessment

Treating the map as a binary scheme (change/no change) requires a simpler sampling technique than generating a complete change detection error matrix with all the "from" and "to" classes. We can use a binomial distribution to calculate the sample size (Ginevan, 1979). A simple look-up table can be used to determine the required sample size for a given minimum error and a desired level of confidence. For a map accuracy of 90% and using a 95% confidence level (at 95%, we run the risk of a 1 in 20 chance that we reject a map that is actually correct), the minimum number of samples required is 298, with the map being rejected as not meeting the accuracy standard if more than 21 are misclassified.

In order to compensate for the rarity of change within the landscape, an approach was designed that employed five strata layers to increase the sampling to areas of likely change. The first stratum is called the change mask and incorporates all the

TABLE 11.3
Sampling Breakdown Based on Strata Layer

Strata Layer	Percentage of Total Samples	Number of Samples
Change mask	30	88
Distance from change mask	25	75
Spectral magnitude	25	75
Probability of change	10	30
Remaining unsampled area	10	30

areas indicated to have changed by the image analysis change methodology used in this project. Thirty percent of the sampling was performed within the change mask. The second area sampled was a buffer surrounding the change mask. It is expected that change will occur near change, so it follows that sampling should occur around the change areas. Twenty-five percent of the samples were taken in this buffer area around the change mask. A third stratum used for another 25% of the sampling included those areas indicated by spectral analysis of the two images as changed. Fourth, 10% of the samples were allocated to those map classes that had the highest amounts of change. In other words, sampling was increased for those map categories for which significant change occurred between 2001 and 2005. Finally, the last 10% of the sampling was allocated to the rest of the map. Table 11.3 presents a summary of the sampling allocation by strata along with the number of samples taken in each stratum.

The overall accuracy for the change/no change assessment was 96% (Table 11.4). Seven samples were labeled "change" on the map, but were not "change" on the

TABLE 11.4
Final Change/No Change Matrix

		REFERENCE		*Producer's Accuracies*	
		Change	No Change	**Totals**	**Accuracies**
M A P	Change	75	7	75/82	92%
	No Change	6	210	210/216	97%
User's Accuracies					
Totals		75/81	210/217	*Overall Accuracy*	
Accuracies		93%	97%	285/298	96.0%

reference data whereas 6 samples were labeled “no change” on the map but actually did change. Only 13 total errors were found. Given the binomial sampling selected with a desired map accuracy of 90% and a 95% confidence level, 21 errors were permitted. Therefore, this map was accurate at the 90% level, and the error matrix shows the true accuracy to be 96%.

Determining the accuracy of the KLC change map was a critical component of this project. The process demonstrated by this case study was designed to assess the accuracy of the change areas on the 2005 map and evaluate how well change was captured between 2001 and 2005. It was not possible, in this project, to conduct a full change detection accuracy assessment and generate a change detection error matrix. This two-step approach is an effective compromise when the available time and resources do not permit a full assessment. These results show that change was captured with a success rate of 96%.

While the deterministic accuracy assessment is low at 58.8%, the fuzzy assessment of the classification shows a favorable overall classification accuracy of 79.9%.

MULTILAYER ASSESSMENTS

Everything that has been presented in the book up to this point, with the exception of the last section on change detection, has dealt with the accuracy of a single map layer. However, it is important to at least mention multilayer assessments. Figure 11.3

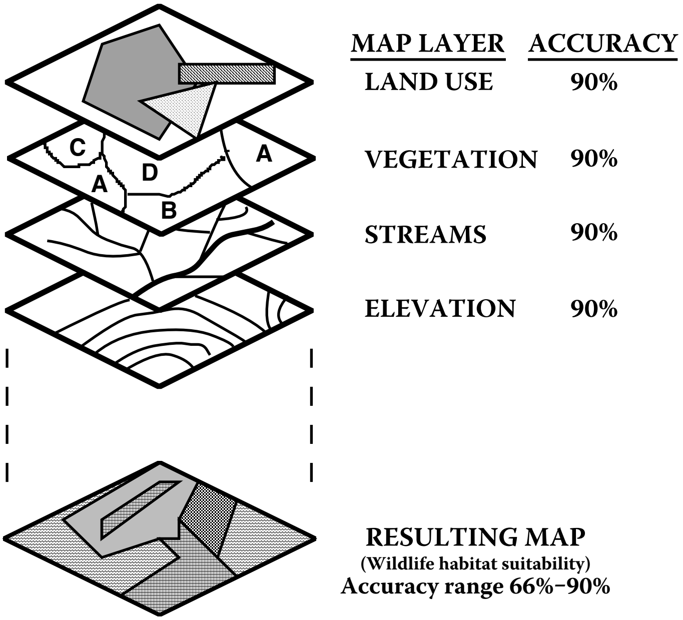


FIGURE 11.3 The range of accuracies for a decision made from combining multiple layers of spatial data.

demonstrates a scenario in which four different map layers are combined to produce a map of wildlife habitat suitability. In this scenario, accuracy assessments have been performed on each of the map layers; each layer is 90% accurate. The question is, how accurate is the wildlife suitability map?

If the four map layers are independent (i.e., the errors in each map are not correlated), then probability tells us that the accuracy would be computed by multiplying the accuracies of the layers together. Therefore, the accuracy of the final map is 90% × 90% × 90% × 90% = 66%. However, if the four map layers are not independent but completely correlated with one another (i.e., the errors are in the exact same place in all four layers), then the accuracy of the final map is 90%. In reality, neither of these cases is very likely. There is usually some correlation between the map layers. For instance, vegetation is certainly related to proximity to a stream and also to elevation. Therefore, the actual accuracy of the final map could only be determined by performing another accuracy assessment on this layer. We do know that this accuracy will be between 66 and 90%, and will probably be closer to 90% than to 66%.

One final observation should be mentioned here. It is quite an eye-opener that using four map layers, all with very high accuracies, could result in a final map of only 66%. On the other hand, we have been using these types of maps for a long time without any knowledge of their accuracy. Certainly, this knowledge can only help us to improve our ability to effectively use spatial data.

APPENDIX 11.1

Class Descriptions of the 2005 NLCD Land Cover

Task	Duration
Open Water	All areas of open water, generally with less than 25% cover of vegetation or soil.
Developed, Open Space	Includes areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes.
Developed, Low Intensity	Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20–49% of total cover. These areas most commonly include single-family housing units.
Developed, Medium Intensity	Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50–79% of the total cover. These areas most commonly include single-family housing units.
Developed, High Intensity	Includes highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses, and commercial/industrial. Impervious surfaces account for 80 to 100% of the total cover.

Bare Land	Barren areas of bedrock, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits, and other accumulations of earthen material. Generally, vegetation accounts for less than 15% of total cover.
Deciduous Forest	Areas dominated by trees generally more than 5 m tall, and greater than 20% of total vegetation cover. More than 75% of the tree species shed foliage simultaneously in response to seasonal change.
Evergreen Forest	Areas dominated by trees generally more than 5 m tall, and greater than 20% of total vegetation cover. More than 75% of the tree species maintain their leaves all year. Canopy is never without green foliage.
Mixed Forest	Areas dominated by trees generally more than 5 m tall, and greater than 20% of total vegetation cover. Neither deciduous nor evergreen species are greater than 75% of total tree cover.
Scrub Shrub	Areas dominated by shrubs less than 5 m tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in an early successional stage, or trees stunted from environmental conditions.
Grassland Herbaceous	Areas dominated by Grammanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.
Pasture Hay	Areas of grasses, legumes, or grass–legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20% of total vegetation.
Cultivated Crop	Areas used for the production of annual crops, such as corn, soybeans, vegetables, tobacco, and cotton, and also perennial woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20% of total vegetation. This class also includes all land being actively tilled.
Woody Wetland	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.
Emergent Herbaceous Wetland	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.

Bibliography

Ager, T. 2004. An Analysis of Metric Accuracy Definitions and Methods of Computation. Unpublished memo prepared for the National Geospatial-Intelligence Agency. InnoVision. March 2004.

Aickin, M. 1990. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*. Vol. 46, pp. 293–302.

American Society of Photogrammtery. 1960. *Manual of Photographic Interpretation*. ASP, Washington, DC.

Anderson, J. R., E. E. Hardy, J. T. Roach, and R. E. Witner. 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data. USGS Professional Paper 964. 28 pp.

Aronoff, S. 1982. Classification accuracy: A user approach. *Photogrammetric Engineering and Remote Sensing*. Vol. 48, No. 8, pp. 1299–1307.

Aronoff, S. 1985. The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering and Remote Sensing.* Vol. 51, No. 1, pp. 99–111.

ASPRS. 1989. ASPRS interim accuracy standards for large-scale maps. *Photogrammetric Engineering and Remote Sensing.* Vol. 54, No. 7, pp. 1038–1041.

ASPRS. (American Society for Photogrammetry and Remote Sensing Specifications and Standards Committee). 1990. ASPRS Accuracy Standards for Large-Scale Maps. *Photogrammetric Engineering and Remote Sensing*. Vol. 56, No. 7, pp. 1068–1070.

ASPRS. 2004. *ASPRS Guidelines, Vertical Accuracy Reporting for Lidar Data.* American Society for Photogrammetry and Remote Sensing. May 24, 2004.

ASPRS and ASCE. 1994. *Glossary of the Mapping Sciences.* ASPRS, Bethesda Maryland and ASCE, New York.

Biging, G. and R. Congalton. 1989. Advances in forest inventory using advanced digital imagery. Proceedings of Global Natural Resource Monitoring and Assessments: Preparing for the 21st Century. Venice, Italy. September 1989. Vol. 3, pp. 1241–1249.

Biging, G., R. Congalton, and E. Murphy. 1991. A comparison of photointerpretation and ground measurements of forest structure. Proceedings of the Fifty-Sixth Annual Meeting of the American Society of Photogrammetry and Remote Sensing, Baltimore, MD. Vol. 3, pp. 6–15.

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

Bolstad, P. 2005. *GIS Fundamentals*. 2nd edition. Eider Press, White Bear Lake, MN. 543 pp.

Brennan, R. and D. Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*. Vol. 41, pp. 687–699.

Brothers, G. L. and E. B. Fish. 1978. Image enhancement for vegetation pattern change analysis. *Photogrammetric Engineering and Remote Sensing.* Vol. 44, No. 5, pp. 607–616.

Campbell, J. B. 1981. Spatial autocorrelation effects upon the accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing*. Vol. 47, No. 3, pp. 355–363.

Card, D. H. 1982. Using known map categorical marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*. Vol. 48, No. 3, pp. 431–439.

Chrisman, N. 1982. Beyond accuracy assessment: Correction of misclassification. Proceedings of the 5th International Symposium on Computer-assisted Cartography. Crystal City, VA. pp. 123–132.

Cliff, A. D. and J. K. Ord. 1973. *Spatial Autocorrelation.* Pion Limited. London, England. 178 pp.

Cochran, W. G. 1977. *Sampling Techniques.* John Wiley & Sons, New York. 428 pp.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* Vol. 20, No. 1, pp. 37–40.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin.* Vol. 70, No. 4, pp. 213–220.

Colwell, R. N. 1955. The PI picture in 1955. *Photogrammetric Engineering.* Vol. 21, No. 5, pp. 720–724.

Congalton, R. G. 1981. The use of discrete multivariate analysis for the assessment of Landsat classification accuracy. MS Thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA. 111 pp.

Congalton, R. G. and R. A. Mead. 1983. A quantitative method to test for consistency and correctness in photo-interpretation. *Photogrammetric Engineering and Remote Sensing.* Vol. 49. No. 1, pp. 69–74.

Congalton, R. G., R. G. Oderwald, and R. A. Mead. 1983. Assessing Landsat classification accuracy using discrete multivariate statistical techniques. *Photogrammetric Engineering and Remote Sensing.* Vol. 49, No. 12, pp. 1671–1678.

Congalton, R. G. 1984. A comparison of five sampling schemes used in assessing the accuracy of land cover/land use maps derived from remotely sensed data. Ph.D. dissertation. Virginia Polytechnic Institute and State University. Blacksburg, VA. 147 pp.

Congalton, R. and R. Mead. 1986. A review of three discrete multivariate analysis techniques used in assessing the accuracy of remotely sensed data from error matrices. *IEEE Transactions of Geoscience and Remote Sensing*. Vol. GE-24, No 1, pp. 169–174.

Congalton, R. G. 1988a. Using spatial autocorrelation analysis to explore errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*. Vol. 54, No. 5, pp. 587–592.

Congalton, R. G. 1988b. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*. Vol. 54, No. 5, pp. 593–600.

Congalton, R. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment.* Vol. 37, pp. 35–46.

Congalton, R. and G. Biging. 1992. A pilot study evaluating ground reference data collection efforts for use in forest inventory. *Photogrammetric Engineering and Remote Sensing.* Vol. 58, No. 12, pp. 1669–1671.

Congalton, R. and K. Green. 1993. A practical look at the sources of confusion in error matrix generation. *Photogrammetric Engineering and Remote Sensing.* Vol. 59, No. 5, pp. 641–644.

Congalton, R. G. and R. D. Macleod. 1994. Change detection accuracy assessment on the NOAA Chesapeake Bay pilot study. Proceedings of the International Symposium of Spatial Accuracy of Natural Resource Data Bases, Williamsburg, VA. pp. 78–87.

Congalton, R. and M. Brennan. 1998. Change detection accuracy assessment: Pitfalls and considerations. Proceedings of the Sixty Fourth Annual Meeting of the American Society of Photogrammetry and Remote Sensing, Tampa, Florida. pp. 919–932 (CD-ROM).

Cowardin, L. M., V. Carter, F. Golet, and E. LaRoe. 1979. A Classification of Wetlands and Deepwater Habitats of the United States. Office of Biological Services. U.S. Fish and Wildlife Service. U.S. Department of Interior, Washington, DC. 103 pp.

Czaplewski, R. 1992. Misclassification bias in aerial estimates. *Photogrammetric Engineering and Remote Sensing.* Vol. 58, No. 2, pp. 189–192.

Czaplewski, R. and G. Catts. 1990. Calibrating area estimates for classification error using confusion matrices. Proceedings of the 56th Annual Meeting of the American Society for Photogrammetry and Remote Sensing. Denver, CO. Vol. 4, pp. 431–440.

DMA (Defense Mapping Agency). 1991. Error Theory as Applied to Mapping, Charting, and Geodesy. Defense Mapping Agency Technical Report 8400.1. Fairfax, VA. 71 pages plus appendices.

Eyre, F. H. 1980. Forest Cover Types of the United States and Canada. Society of American Foresters, Washington, DC. 148 pp.

FEMA (Federal Emergency Management Agency). 2003. *Guidelines and Specifications for Flood Hazard Mapping Partners*.

Ferris State University. 2007. http://www.ferris.edu/faculty/burtchr/sure340/notes/History.pdf

FGDC (Federal Geographic Data Committee). 1998. Subcommittee for Base Cartographic Data. *Geospatial Positioning Accuracy Standards. Part 3: National Standard for Spatial Data Accuracy*. FGDC-STD-007.3-1998: Washington, DC. Federal Geographic Data Committee, 24 pp.

Fitzpatrick-Lins, K. 1981. Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogrammetric Engineering and Remote Sensing*. Vol. 47, No. 3, pp. 343–351.

Fleiss, J., J. Cohen, and B. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin.* Vol. 72, No. 5, pp. 323–327.

Foody, G. 1992. On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*. Vol. 58, No. 10, pp. 1459–1460.

Freese, F. 1960. Testing accuracy. *Forest Science*. Vol. 6, No. 2, pp. 139–145.

Ginevan, M. E. 1979. Testing land-use map accuracy: another look. *Photogrammetric Engineering and Remote Sensing*. Vol. 45, No. 10, pp. 1371–1377.

Gong, P. and J. Chen. 1992. Boundary uncertainties in digitized maps: Some possible determination methods. IN: Proceedings of GIS/LIS '92. Annual Conference and Exposition. San Jose, CA. pp. 274–281.

Goodman, L. 1965. On simultaneous confidence intervals for multinomial proportions. *Technometrics*. Vol. 7, pp. 247–254.

Gopal, S. and C. Woodcock. 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*. Vol. 60, No. 2, pp. 181–188.

Grassia, A. and R. Sundberg. 1982. Statistical precision in the calibration and use of sorting machines and other classifiers. *Technometrics*. Vol. 24, pp. 117–121.

Green, K. and R. Congalton. 2004. An error matrix approach to fuzzy accuracy assessment: the NIMA Geocover project. A peer-reviewed chapter in Lunetta, R. S. and J. G. Lyon (Eds.), *Remote Sensing and GIS Accuracy Assessment*. CRC Press, Boca Raton, FL. 304 pp.

Greenwalt, C. and M. Schultz. 1962 and 1968. *Principles of Error Theory and Cartographic Applications*. United States Air Force. Aeronautical Chart and Information Center. ACIC Technical Report Number 96. St. Louis, Missouri. 60 pages plus appendices. This report is cited in the ASPRS standards as ACIC, 1962.

Hay, A. M. 1979. Sampling designs to test land-use map accuracy. *Photogrammetric Engineering and Remote Sensing*. Vol. 45, No. 4, pp. 529–533.

Hay, A. M. 1988. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*. Vol. 9, pp. 1395–1398.

Hill, T. B. 1993. Taking the " " out of "ground truth": Objective accuracy assessment. In Proceedings of the 12th Pecora Conference. Sioux Falls, SD. pp. 389–396.

Hopkirk, P. 1992. *The Great Game. The Struggle for Empire in Central Asia.* Kodansha International. 565 pp.

Hord, R. M. and W. Brooner. 1976. Land-use map accuracy criteria. *Photogrammetric Engineering and Remote Sensing*. Vol. 42, No. 5, pp. 671–677.

Hudson, W. and C. Ramm. 1987. Correct formulation of the kappa coefficient of agreement. *Photogrammetric Engineering and Remote Sensing*. Vol. 53, No.4, pp. 421–422.

Katz, A. H. 1952. Photogrammetry needs statistics. *Photogrammetric Engineering and Remote Sensing.* Vol. 18, No. 3, pp. 536–542.

Landis, J. and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*. Vol. 33. pp. 159–174.

Lewis, Meriwether, William Clark, Nicholas Biddle, Paul Allen. 1814 *Map of Lewis and Clark's Track across the Western Portion of North America*. Bradford and Inskeep, Philadelphia.

Lopez, A., F. Javier, A. Gordo, and A. David. 2005. Sample Size and Confidence When Applying the NSSDA. XXII International Cartographic Conference (ICC2005). Hosted by The International Cartographic Association. Coruna, Spain. July 11–16, 2005.

Lowell, K. 1992. On the incorporation of uncertainty into spatial data systems. In Proceedings of GIS/LIS '92. Annual Conference and Exposition. San Jose, CA. pp. 484–493.

Lunetta, R., R. Congalton, L. Fenstermaker, J. Jensen, K. McGwire, and L. Tinney. 1991. Remote sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering and Remote Sensing.* Vol. 57, No. 6, pp. 677–687.

Macleod, R. and R. Congalton. 1998. A quantitative comparison of change detection algorithms for monitoring eelgrass from remotely sensed data. *Photogrammetric Engineering and Remote Sensing.* Vol. 64, No. 3, pp. 207–216.

Malila, W. 1985. Comparison of the information contents of Landsat TM and MSS data. *Photogrammetric Engineering and Remote Sensing*. Vol. 51, No. 9, pp. 1449–1457.

Maune, D. Ed. 2007. *Digital Elevation Model Technologies and Applications: The DEM Users Manual,* 2nd edition. American Society of Photogrammetry and Remote Sensing. Bethesda, MD. 655 pp.

McGlone, J. C. 2004. Ed. *Manual of Photogrammetry*. American Society for Photogrammetry and Remote Sensing. Bethesda, MD. 1151 pp.

McGuire, K. 1992. Analyst variability in labeling unsupervised classifications. *Photogrammetric Engineering and Remote Sensing*. Vol. 58, No. 12, pp. 1705–1709.

Mikhail, E. M. and G. Gracie. 1981. *Analysis and Adjustment of Survey Measurements.* Van Nostrand Reinhold, 340 pp.

MPLMIC. 1999. *Positional Accuracy Handbook. Using the National Standard for Spatial Data Accuracy to Measure and Report Geographic Data Quality.* Minnesota Planning land Management Information Center. St. Paul, Minnesota. 29 pp.

NDEP. 2004. *Guidelines for Digital Elevation Data*. Version 1.0. National Digital Elevation Program. May 10, 2004.

Prisley, S. and J. Smith. 1987. Using classification error matrices to improve the accuracy of weighted land-cover models. *Photogrammetric Engineering and Remote Sensing*. Vol. 53, No. 9, pp. 1259–1263.

Rhode, W. G. 1978. Digital image analysis techniques for natural resource inventories. National Computer Conference Proceedings. pp. 43–106.

Rosenfield, G. and K. Fitzpatrick-Lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*. Vol. 52, No. 2, pp. 223–227.

Rosenfield, G. H., K. Fitzpatrick-Lins, and H. Ling. 1982. Sampling for thematic map accuracy testing. *Photogrammetric Engineering and Remote Sensing*. Vol. 48, No. 1, pp. 131–137.

Sammi, J. C. 1950. The application of statistics to photogrammetry. *Photogrammetric Engineering.* Vol. 16, No. 5, pp. 681–685.

Singh, A. 1986. Change detection in the tropical rain forest environment of Northeastern India using Landsat. In *Remote Sensing and Tropical Land Management*. Edited by Eden, M. J. and Parry, J. T. London: John Wiley & Sons. pp. 237–254.

Singh, A. 1989. Digital change detection techniques using remotely sensed data. *International Journal of Remote Sensing,* Vol. 10, No 6, pp. 989–1003.

Spurr, S. 1948. *Aerial Photographs in Forestry*. Ronald Press, New York. 340 pp.

Spurr, S. 1960. *Photogrammetry and Photo-Interpretation with a Section on Applications to Forestry.* Ronald Press, New York. 472 pp.

Stehman, S. 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*. Vol. 58, No. 9, pp. 1343–1350.

Stewart, G. 1960. *Ordeal by Hunger. The Classic Story of the Donner Party.* Pocket Book edition. 320 pp.

Story, M. and Congalton, R. 1986. Accuracy assessment: A user's perspective. *Photogrammetric Engineering and Remote Sensing*. Vol. 52, No. 3, pp. 397–399.

Tenenbein, A. 1972. A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics*. Vol. 14, pp. 187–202.

Tortora, R. 1978. A note on sample size estimation for multinomial populations. *The American Statistician*. Vol. 32, No. 3. pp. 100–102.

U.S. Bureau of the Budget. 1941. 1947 *U.S. National Map Accuracy Standards.* Washington, DC.

van Genderen, J. L. and B. F. Lock. 1977. Testing land use map accuracy. *Photogrammetric Engineering and Remote Sensing*. Vol. 43, No. 9. pp. 1135–1137.

van Genderen, J. L., B. F. Lock, and P. A. Vass. 1978. Remote sensing: statistical testing of thematic map accuracy. Proceedings of the Twelfth International Symposium on Remote Sensing of Environment. ERIM. pp. 3–14.

Woodcock, C. 1996. On roles and goals for map accuracy assessment: A remote sensing perspective. Proc: Second International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, USDA Forest Service Rocky Mountain Forest and Range Experiment Station, Gen. Tech. Rep. RM-GTR-277. Fort Collins, CO. pp. 535–540.

Woodcock, C. and S. Gopal. 1992. Accuracy assessment of the Stanislaus Forest vegetation map using fuzzy sets. In Remote Sensing and Natural Resource Management. Proceedings of the 4th Forest Service Remote Sensing Conference, Orlando, FL. pp. 378–394.

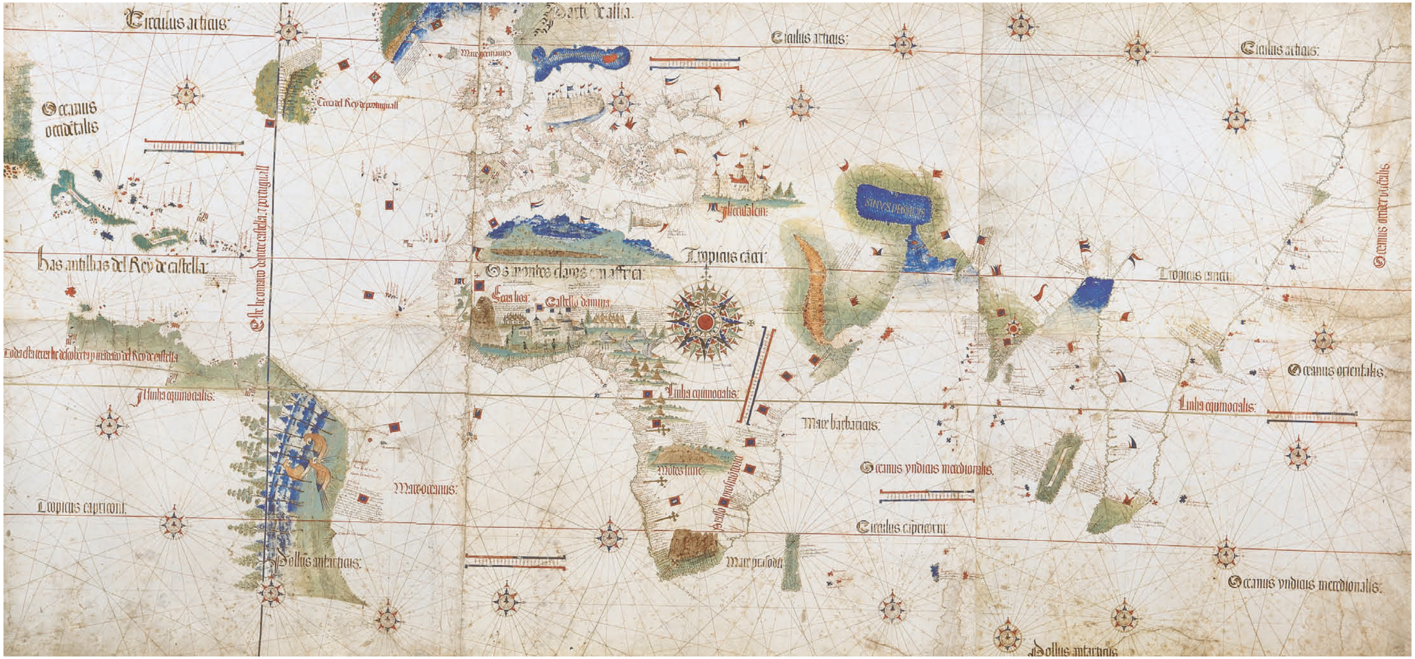
Young, H. E. 1955. The need for quantitative evaluation of the photo interpretation system. *Photogrammetric Engineering.* Vol. 21, No. 5. pp. 712–714.

Young, H. E. and E. G. Stoeckler. 1956. Quantitative evaluation of photo interpretation mapping. *Photogrammetric Engineering.* Vol. 22, No. 1. pp. 137–143.

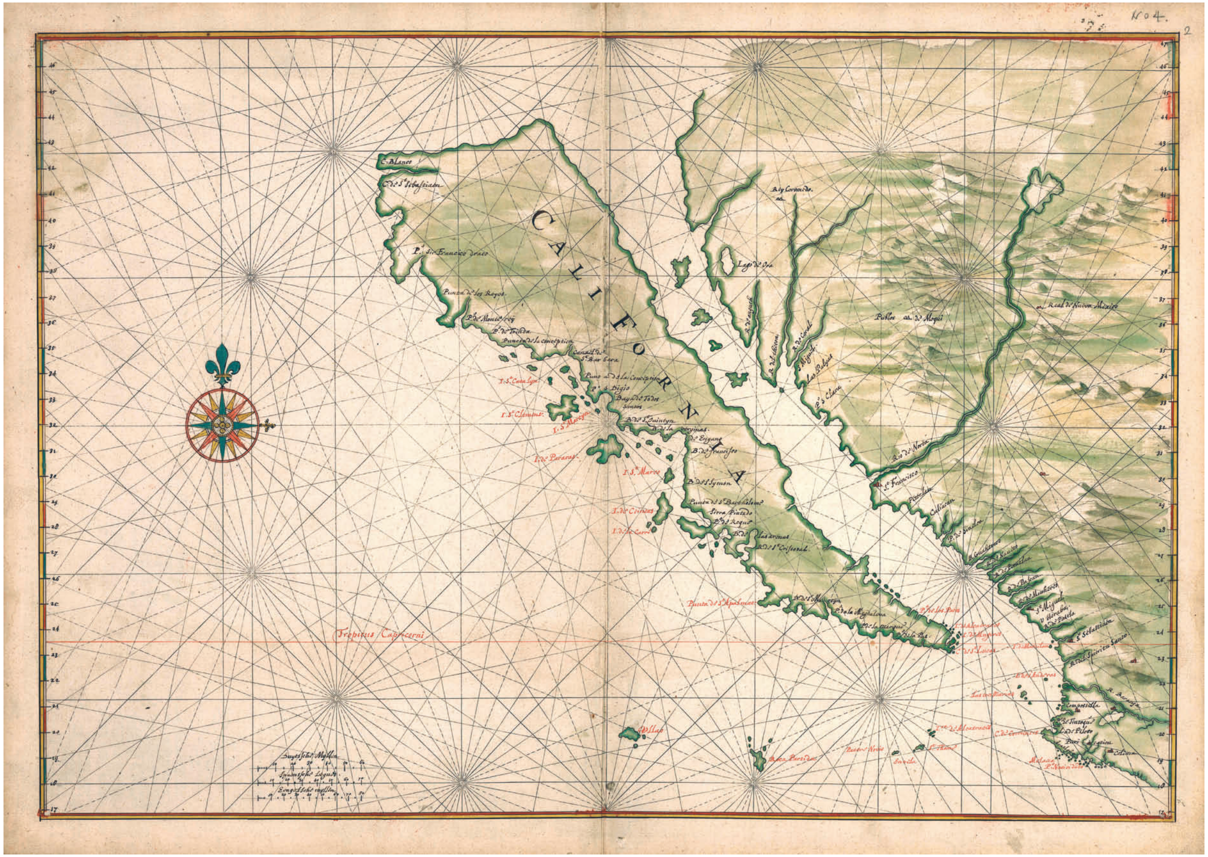
Zadeh, L. A. 1965. Fuzzy sets. *Information and Control.* Vol. 8, pp. 338–353.

Zar, J. 1974. *Biostatistical Analysis.* Prentice-Hall. 620 pp.

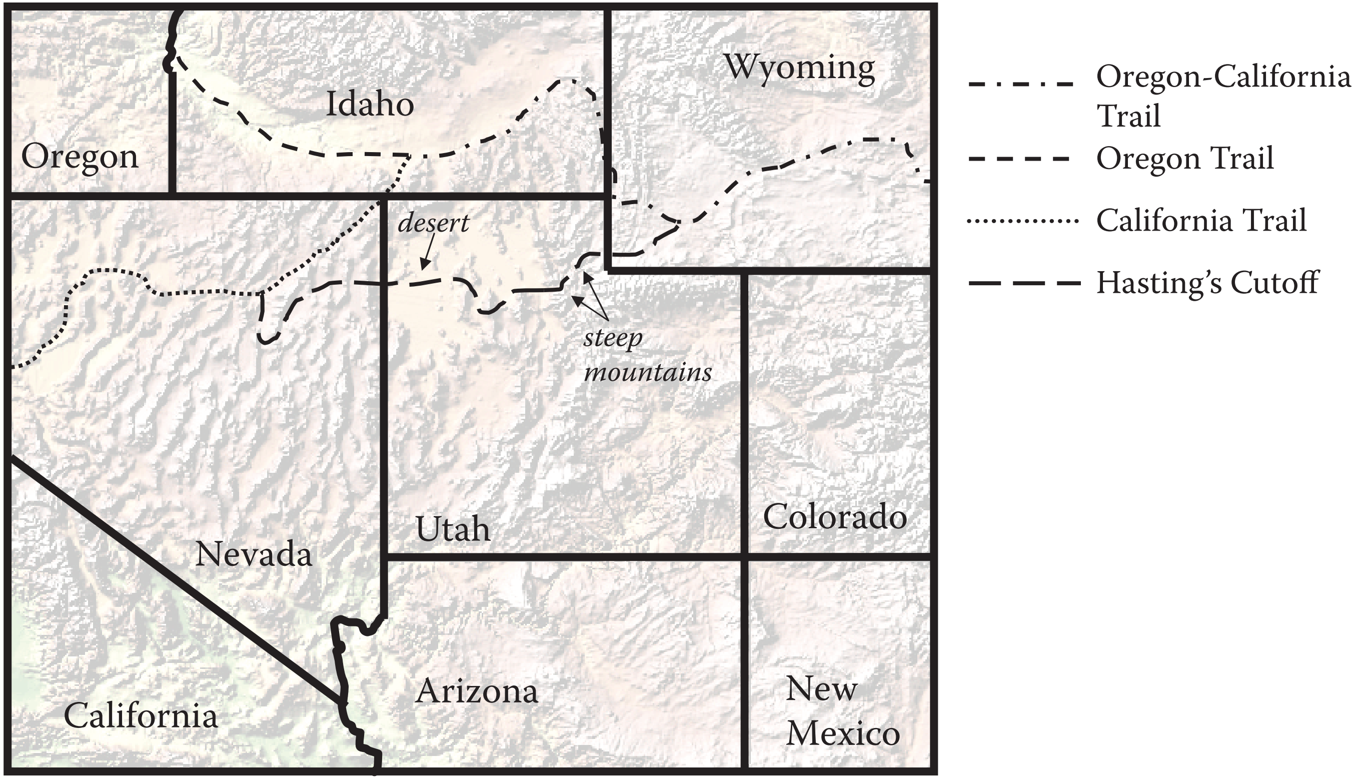
Index



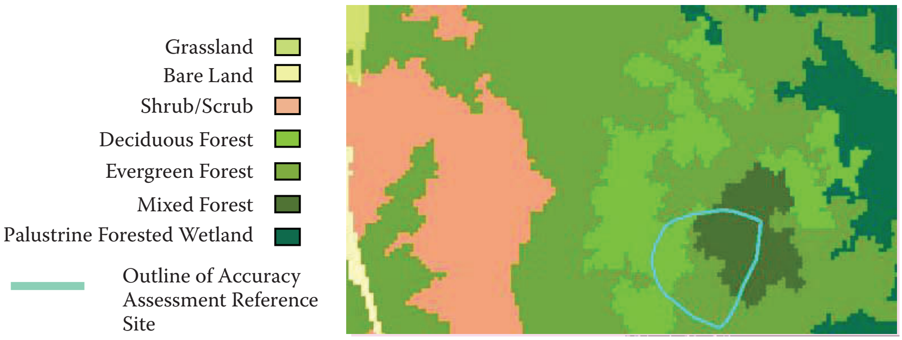
COLOR FIGURE 2.1 The Cantino World Map, which is a map of the known coastlines of the world, created by sixteenth century navigators.



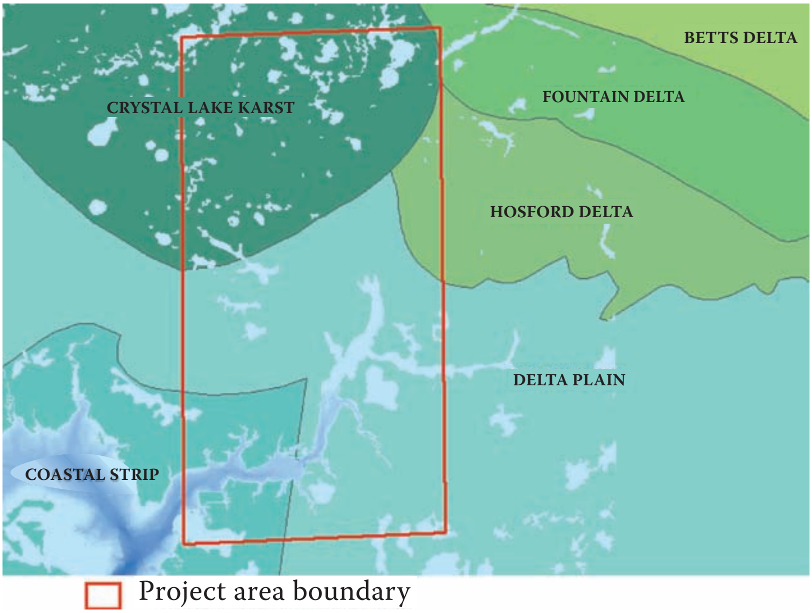
COLOR FIGURE 2.3 A seventeenth century map of California.



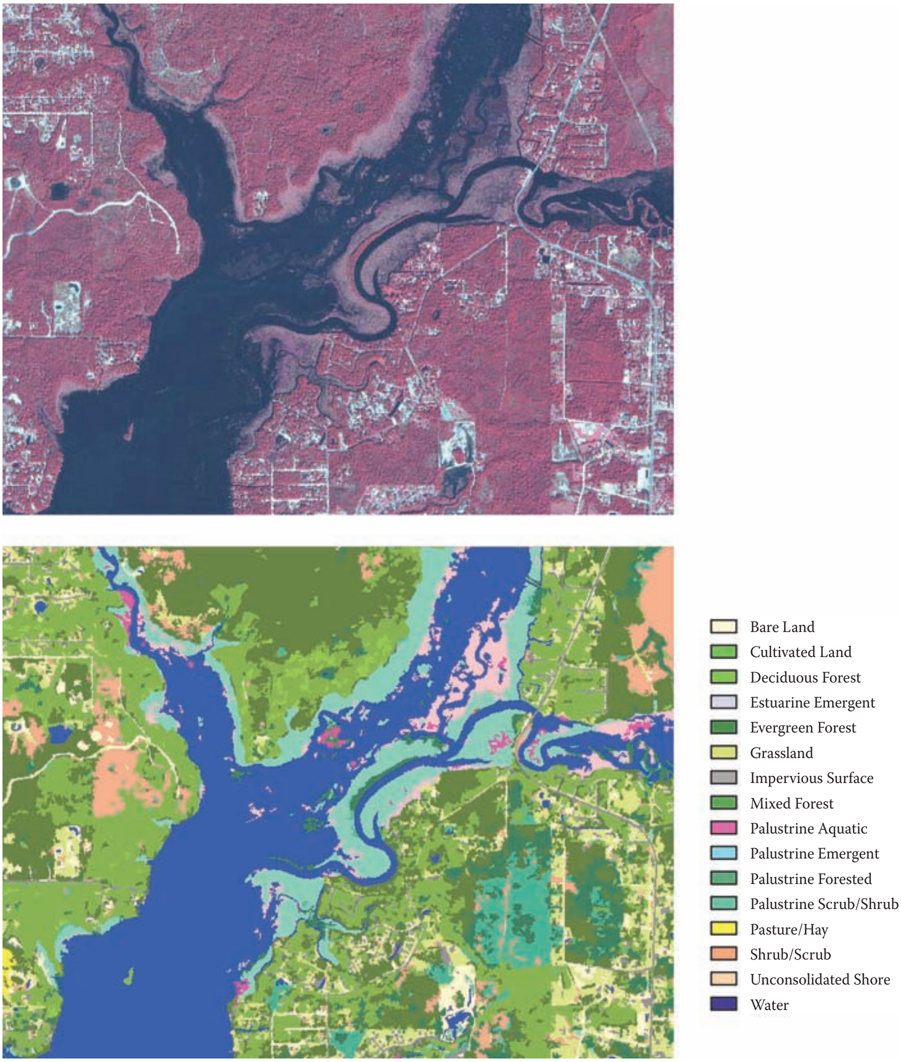
COLOR FIGURE 2.4 Hasting's Cutoff versus the safer California and Oregon trails used in 1846 by emigrants.



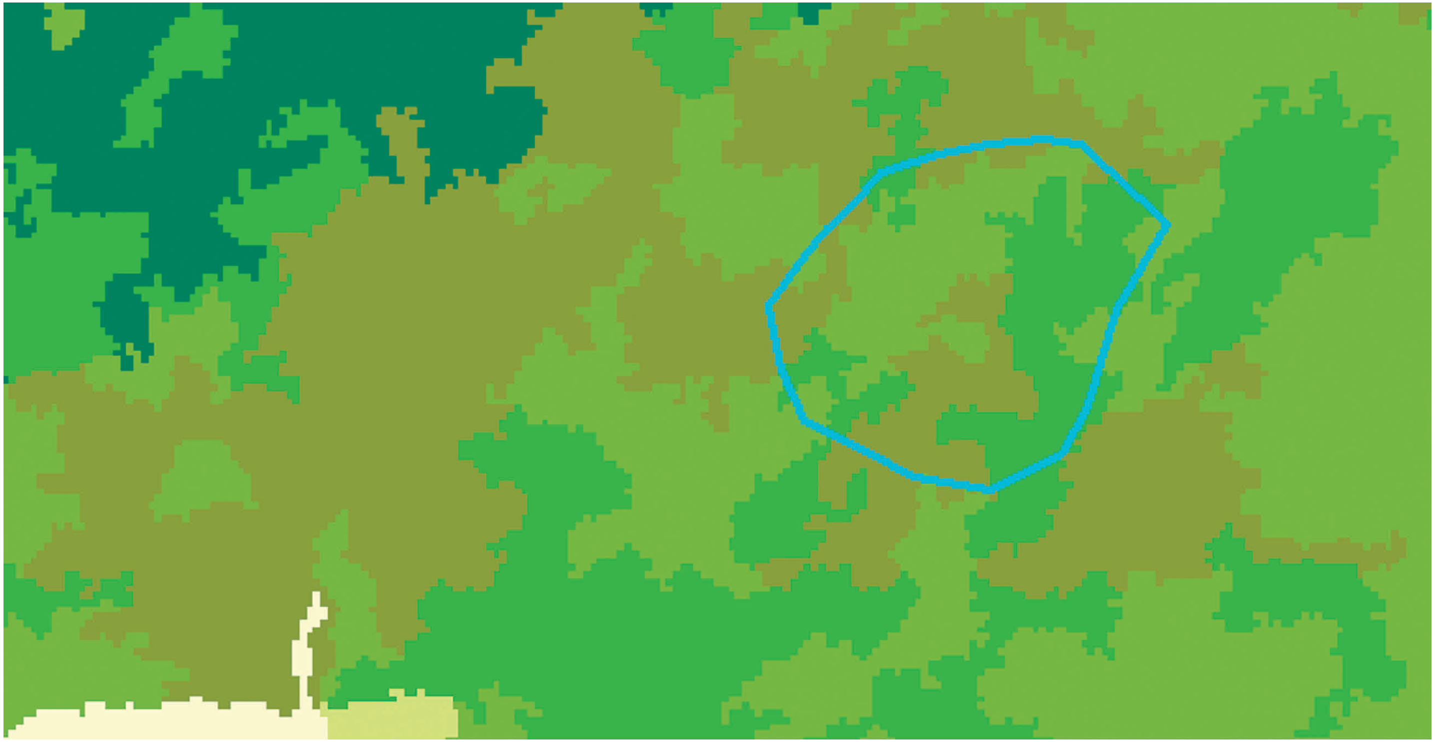
COLOR FIGURE 5.5 Mixed forest accuracy assessment reference polygon (in turquoise) over the map polygons of evergreen, mixed, and deciduous forest. Determining the map label of the accuracy assessment polygon, when the polygon intersects with multiple map classes, can be problematic.



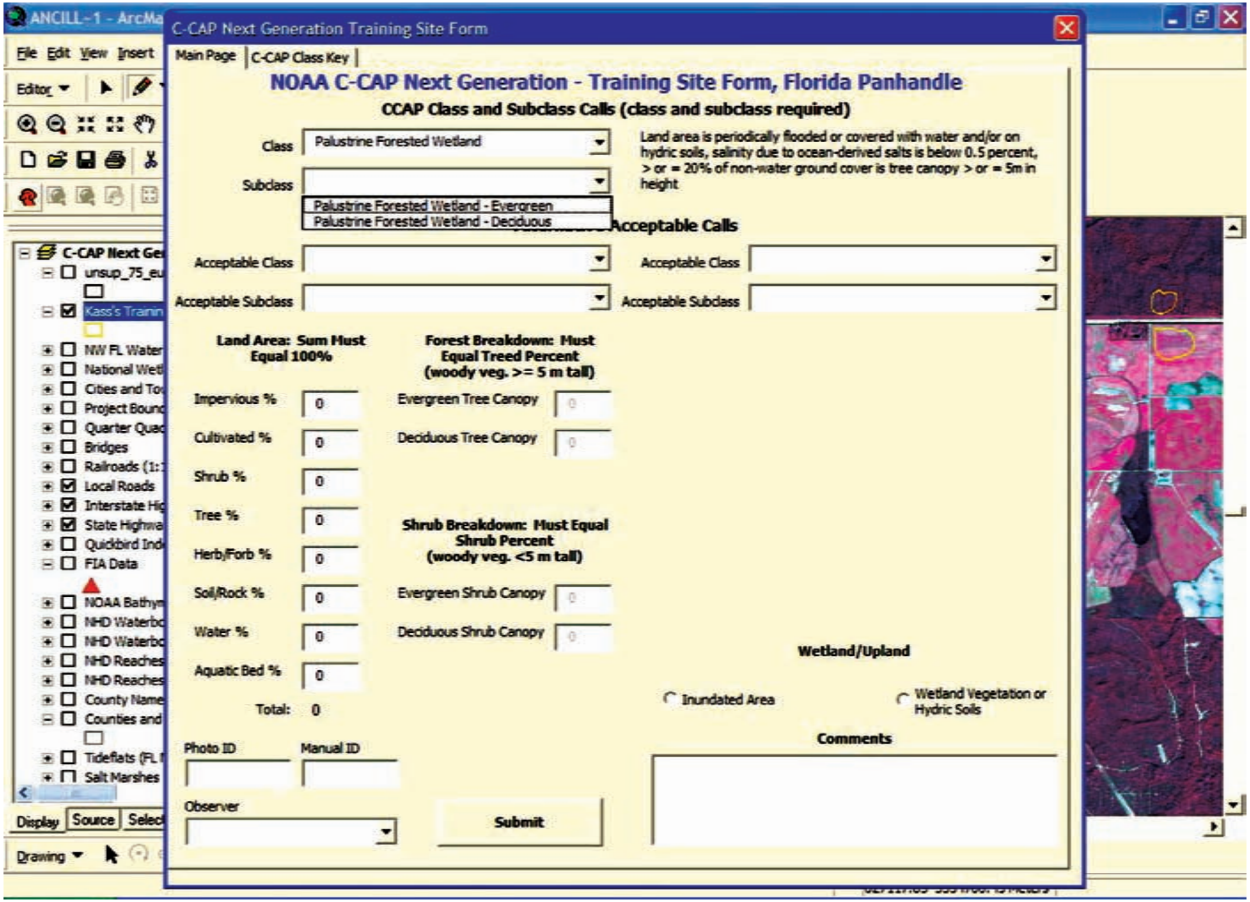
COLOR FIGURE 10.2 Project area boundary over Florida physiographic groups.



COLOR FIGURE 10.3 Detailed area of the case study including the quickbird multispectral imagery and the final map.



COLOR FIGURE 10.4 Illustration of a heterogeneous map accuracy assessment site.



C-CAP Next Generation Training Site Form

Main Page | C-CAP Class Key

NOAA C-CAP Next Generation - Training Site Form, Florida Panhandle

CCAP Class and Subclass Calls (class and subclass required)

Class: Palustrine Forested Wetland ▾

Land area is periodically flooded or covered with water and/or on hydric soils, salinity due to ocean-derived salts is below 0.5 percent, > or = 20% of non-water ground cover is tree canopy > or = 5m in height

Subclass: ▾

Palustrine Forested Wetland - Evergreen
Palustrine Forested Wetland - Deciduous

Acceptable Calls

Acceptable Class	▾	Acceptable Class	▾
Acceptable Subclass	▾	Acceptable Subclass	▾

Land Area: Sum Must Equal 100%

Impervious %	0
Cultivated %	0
Shrub %	0
Tree %	0
Herb/Forb %	0
Soil/Rock %	0
Water %	0
Aquatic Bed %	0
Total:	0

Forest Breakdown: Must Equal Treed Percent (woody veg. >= 5 m tall)

Evergreen Tree Canopy	0
Deciduous Tree Canopy	0

Shrub Breakdown: Must Equal Shrub Percent (woody veg. <5 m tall)

Evergreen Shrub Canopy	0
Deciduous Shrub Canopy	0

Wetland/Upland

○ Inundated Area ○ Wetland Vegetation or Hydric Soils

Comments

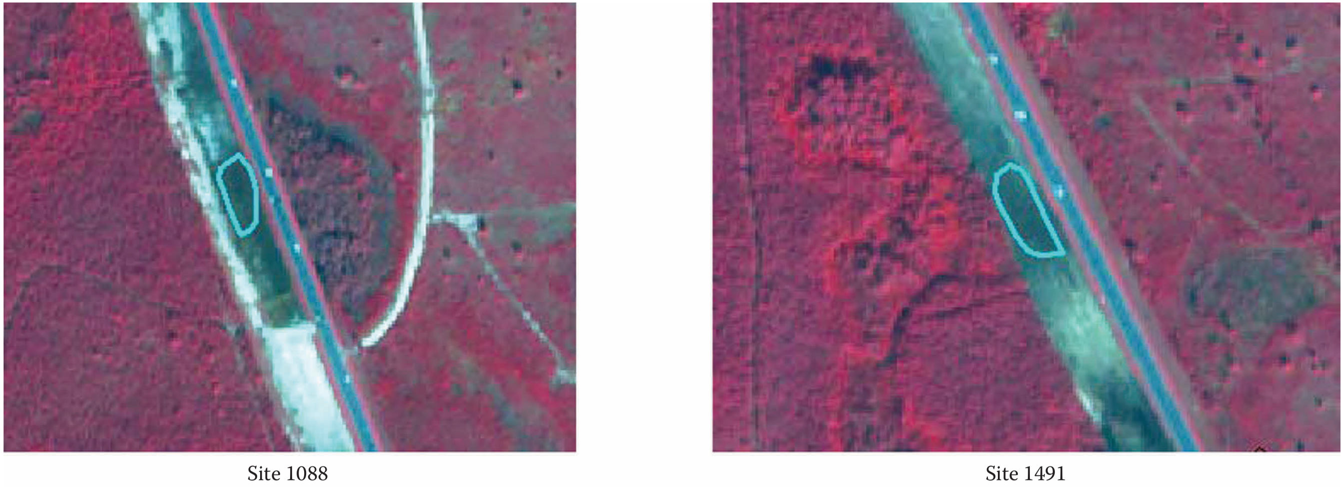
Photo ID

Manual ID

Observer ▾

Submit

COLOR FIGURE 10.5 Digital field form used in the project.



Site 1088

Site 1491

COLOR FIGURE 10.6 Examples of sites with palustrine emergent scrub/shrub reference labels confused with bare land map labels.

ENVIRONMENTAL ENGINEERING

Second Edition

Assessing the Accuracy of Remotely Sensed Data

Principles and Practices

Design and Implement a Successful Accuracy Assessment

Accuracy assessment of maps derived from remotely sensed data has continued to grow since the first edition of this groundbreaking book. As a result, the much-anticipated new edition is significantly expanded and enhanced to reflect growth in the field. It features three new chapters, including fuzzy accuracy assessment, positional accuracy, and a case study related to mapping land cover and land use in the Florida panhandle.

The authors provide a complete presentation of how to assess the positional accuracy of a map along with a discussion of the impact of positional accuracy on thematic accuracy. They also include a more thorough discussion of the special sampling issues that must be considered to assess change effectively. Among other features, this informative reference:

- Reviews the current state of remotely sensed image accuracy assessment
- Includes updates, revisions, a new case study, and new chapters
- Devotes a full chapter to fuzzy accuracy assessment
- Adopts a new chapter order presenting positional accuracy and thematic accuracy before sample design considerations

Complete with an 8-page color insert, this second edition continues to provide a complete guide to designing and conducting a state-of-the-art accuracy assessment.

55127

ISBN: 978-1-4200-5512-2

90000

9 781420 055122

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK