Workshop on New Research Directions for the National Geospatial-Intelligence Agency

White Papers

May 17-19, 2010, Workshop
National Academies Keck Center
Washington, D.C.
NOTICE

To set the stage for the May 17-19 workshop, the committee commissioned white papers from leading experts in NGA’s traditional core areas:

- remote sensing and imagery science
- photogrammetry and geomatics
- geodesy and geophysics
- cartographic science
- geographic information systems and geospatial analysis

The white papers, reproduced in this volume, may not always reflect the views of the committee or vice versa. Please contact the white paper authors for permission to use their material.
# Table of Contents

Agenda........................................................................................................................................vii

Calibration of the Aerial Photogrammetric System..............................................................1  
Dean Merchant, Ohio State University

Remote Sensing: An Overview...............................................................................................9  
Russell Congalton, University of New Hampshire

Cartographic Research in the United States: Current Trends and Future Directions.........25  
Robert McMaster, University of Minnesota

A Brief History of Satellite Geodesy – October 4, 1957 to Present.................................41  
Lewis Lapine, South Carolina Geodetic Survey

Geospatial Analysis..............................................................................................................57  
Luc Anselin, Arizona State University
Workshop on New Research Directions for the National Geospatial-Intelligence Agency
Keck Center
500 Fifth Street, NW, Washington, D.C.
May 17-19, 2010

Agenda

Monday, May 17, Room 201

8:15  Plenary session I: Overview

Goals and scope of the workshop
Overview of the first day  
Keith Clarke, UC Santa Barbara

8:30  National Geospatial-Intelligence Agency

• Overview of the agency and its challenges
• Motivation for the workshop and how the results will be used

Greg Smith, NGA

9:00  Plenary Session II: Future directions in NGA’s core areas (20 minute talks)

Photogrammetry: “Spatial Information Extraction from Imagery: Recent Trends in Geomatics”
Clive Fraser, U. Melbourne

Remote sensing: “Advanced Sensors and Information Extraction: Synergies for Optical Sensing”
Melba Crawford, Purdue

10:00 Instructions to the working groups and break

• Other ideas not raised in the presentations
• Key challenges and opportunities for NGA in each core area

K. Clarke

10:30 Working groups on photogrammetry and remote sensing

Working group 1, Room 201
Carolyn Merry (chair) and Douglas Lemon (vice chair)

Working group 2, Room 202
Luc Anselin (chair) and Melba Crawford (vice chair)

Working group 3, Room 204
Annette Krygiel (chair) and Russell Congalton (vice chair)

Working group 4, Room 206
Scott Sandgate (chair) and Clive Fraser (vice chair)

Working group 5, Room 213
James Thomas (chair) and Ayman Habib (vice chair)

12:00  Working lunch

1:00  Plenary session III: Future directions in NGA’s core areas (20 minute talks)
Cartography: “Cartographic Research in the United States: Current Trends and Future Directions”  
Robert McMaster, U. Minnesota

Dru Smith, NOAA NGS

GIS and geospatial analysis: “GIS as Geospatial Inspiration”  
May Yuan, U. Oklahoma

2:30 Instructions to the working groups  
Keith Clarke

- Other ideas not raised in the presentations
- Key challenges and opportunities for NGA in each core area

Working groups on cartography, geodesy, and GIS and geospatial analysis

Working group 6, Room 201  
Carolyn Merry (chair) and Michael Peterson (vice chair)

Working group 7, Room 202  
Luc Anselin (chair) and Chris Rizos (vice chair)

Working group 8, Room 204  
Annette Krygiel (chair) and Alan MacEachren (vice chair)

Working group 9, Room 206  
Scott Sandgathe (chair) and Alfred Leick (vice chair)

Working group 10, Room 213  
James Thomas (chair) and Alan Murray (vice chair)

4:00 Working groups prepare reports

4:30 Plenary session IV: Working group reports (5 minutes each)

Photogrammetry and remote sensing  
Douglas Lemon

Working group 1  
Melba Crawford

Working group 2  
Russell Congalton

Working group 3  
Clive Fraser

Working group 4  
Ayman Habib

Working group 5  

Cartography, geodesy, and GIS and geospatial analysis  
Michael Peterson

Working group 6  
Chris Rizos

Working group 7  
Alan MacEachren

Working group 8  
Alfred Leick

Working group 9  
Alan Murray

Working group 10  

Discussion  
All
5:30 Workshop adjourns for the day
5:30 Reception
6:00 Workshop dinner

Tuesday, May 18, Room 109

8:30 Overview of plans for the day

Plenary session V: Cross cutting themes (20 minutes each)

Forecasting: “Technosocial Predictive Analysis: Bridging the Gap between Human Judgment and Machine Reasoning” Antonio Sanfillipo, PNNL

Participatory sensing: “Participatory Urban Data Collection: Planning and Optimization” Cyrus Shahabi, U. Southern California

Visual Analytics: “Proactive and Predictive Visual Analytics” David Ebert, Purdue

10:00 Instructions to the working groups and break

- Other ideas not raised in the presentations
- How do advances in the cross-cutting themes shape the 5 core areas?

10:30 Working groups on forecasting, participatory sensing, and visual analytics

Working group 1, Room 109
Carolyn Merry (chair) and Michael Zyda (vice chair)

Working group 2, Room 202
Luc Anselin (chair) and William Ribarsky (vice chair)

Working group 3, Room 208
Mani Srivastava (chair) and Amitabh Varshney (vice chair)

Working group 4, Room 213
Scott Sandgathe (chair) and Mike Jackson (vice chair)

Working group 5, Room 600
James Thomas (chair) and Michael Zink (vice chair)

12:00 Working lunch

1:00 Plenary session VI: Cross cutting themes (20 minutes each)

Beyond fusion: “Data and Visual Analytics for Information Fusion” Haesun Park, Georgia Tech
Human terrain: “Geospatially Enabled Network Analysis”  
*Kathleen Carley, Carnegie Mellon*

2:00 Instructions to the working groups  
*K. Clarke*

- Other ideas not raised in the presentations
- How do advances in the cross-cutting themes shape the 5 core areas?

Working groups on beyond fusion and human terrain

Working group 1, *Room 109*  
Carolyn Merry (chair) and Huan Liu (vice chair)

Working group 2, *Room 202*  
Luc Anselin (chair) and May Yuan (vice chair)

Working group 3, *Room 208*  
Mani Srivastava (chair) and James Llinas (vice chair)

Working group 4, *Room 213*  
Scott Sandgathe (chair) and Mahendra Mallick (vice chair)

Working group 5, *Room 600*  
James Thomas (chair) and Joseph Young (vice chair)

3:30 Working groups prepare reports

4:00 Plenary session VII: Working group reports (*5 minutes each*)

Forecasting, participatory sensing, and visual analytics

- Working group 1: *Michael Zyda*
- Working group 2: *William Ribarsky*
- Working group 3: *Amitabh Varshney*
- Working group 4: *Mike Jackson*
- Working group 5: *Michael Zink*

Beyond fusion and human terrain

- Working group 6: *Huan Liu*
- Working group 7: *May Yuan*
- Working group 8: *James Llinas*
- Working group 9: *Mahendra Mallick*
- Working group 10: *Joseph Young*

Discussion: *All*

5:00 Workshop adjourns for the day
Wednesday, May 19, Room 109

8:30 Overview of plans for the day  

Instructions to the working groups  
- Synthesize the results of the 10 breakout reports  
- Identify implications of implementing the results for the scientific infrastructure

8:45 Working groups on new research directions for the NGA

- Working group 1, Room 109  
  Carolyn Merry (chair) and Alan MacEachren (vice chair)

- Working group 2, Room 202  
  Luc Anselin (chair) and Ayman Habib (vice chair)

- Working group 3, Room 204  
  Mani Srivastava (chair) and Joseph Young (vice chair)

- Working group 4, Room 205  
  Annette Krygiel (chair) and Chris Rizos (vice chair)

- Working group 5, Room 213  
  James Thomas (chair) and Mike Jackson (vice chair)

10:30 Working groups prepare reports

11:00 Plenary session VIII: Working group reports (10 minutes each)

- Working group 1  
  Alan MacEachren

- Working group 2  
  Ayman Habib

- Working group 3  
  Joseph Young

- Working group 4  
  Chris Rizos

- Working group 5  
  Mike Jackson

Discussion  
All

12:00 Wrap up and next steps  
Keith Clarke

12:15 Workshop adjourns
Calibration of the Aerial Photogrammetric System

Dean C. Merchant
Geodetic Science Department
Ohio State University

ABSTRACT: A brief history of aerial photogrammetry is presented, which leads to recent applications of modern digital cameras with associated sensors. The concept of measurement system calibration is emphasized since it appears to have potential for greatly improving the geospatial accuracy and reliability of NGA sensors. Although the modern digital mapping camera is emphasized, calibration has a useful role in development of most geospatial sensors. An example of recent results of calibration is provided for an airborne digital camera, the most recent Zeiss airborne digital camera, the RMK-D II. The quality of the result is good with the fit of image residuals at about 2 microns rmse. Based on the quality of these results, recommendations for calibration research are made in the context of the NGA mission.

Forward

I have been asked to write concerning the areas of photogrammetry and geomatics as they pertain to the geo-spatial interests of the NGA. Because of my area of interest, I will concentrate on photogrammetry in terms of current capability, potential areas of technical improvements, and future use. Improvements will be framed in the context of NGA’s interests, particularly in terms of timeliness and geo-spatial accuracy.

To begin, I will comment briefly on history of photogrammetry, particularly as it leads to applications of interest to the NGA. This will be followed with comments on the state of the art and the central topic of calibration of measurement systems, particularly of the type used by NGA. All comments are made without benefit of classified information.

BRIEF HISTORY OF PHOTOGRAMMETRY

Most of the early development of photogrammetry was done in Europe during the early part of the 20th century. The theory was well-developed by Otto Von Gruber (1942) and others. The bulk of this early photogrammetric equipment was developed and marketed by Zeiss of Germany. Of particular interest were their photo theodolites and aerial cameras. Zeiss was also an early developer of the stereo comparator and of map compilation equipment. Other early European developments occurred in France, England, Italy, Austria and Holland.

In the United States, credit for early developments in the theory of photogrammetry must be given to Prof. Earl Church at Syracuse University. His published work begins in the 1930s with his theory aimed at practical applications. It is said that one would see Prof. Church, walking about the campus followed by several graduate students manually computing photogrammetric solutions by what was termed the “Post Card Method”. It provided for solutions of such things as photogrammetric resection and orientation.
During this early period, Fairchild Camera became the primary producer of aerial mapping cameras. In fact, Fairchild claims to have produced 90% of all aerial cameras, including reconnaissance cameras, used by the Allies during WWII. Cameras were also made by Park Aerial Surveys and Abrams Aerial Surveys. Little map compilation equipment was developed in the US during this early period. Users relied on, for example, the Bausch and Lomb (B&L) copy of the Zeiss Multiplex, a stereoscopic projection type plotting instrument.

Early and mid-history of the developments in photogrammetry are well-documented in the sequence of Manuals of Photogrammetry published by the American Society of Photogrammetry (ASP & ASPRS, 1980).

The term “analytical photogrammetry” evolved as digital methods of computation became available. Much of the credit for advancements both in theory and application of analytical methods in the U.S. belong first to Dr. Hellmut Schmid, then working for the US Army and subsequently the US Coast and Geodetic Survey. While at the Army Aberdeen Proving Ground, Dr. Schmid guided Dwane Brown in the theory of analytical photogrammetry.

These interests led Brown to work for the US Air Force at the Air Force Missal Test Center (AFMTC) at Patrick AFB. Here, he was instrumental in developing the theory of spatial intersection for tracking test missiles by photo-theodolites. For improved spatial tracking accuracies, Brown adapted the 36 inch focal length lens from the B&L K-38 camera to a ground based camera for tracking. This approach depended on a star-field background to establish angular orientation for a series of widely spaced tracking cameras. With the known locations of these tracking cameras, an extremely accurate tracking by means of photogrammetric intersection was accomplished.

Brown began to develop his interest in camera calibration with early work conducted in a large hangar at Patrick AFB. By converging three or more cameras focusing on common targets placed on the hangar floor and using other exterior orientation steps, he was able to establish a procedure leading to both exterior and interior orientation for all participating cameras. Subsequently it was established that only first approximations of target locations were required for a solution of not only camera orientations but also target locations. This aspect of photogrammetry leads to some extended applications of calibration and target location that may be of interest to the NGA.

Later, then working for Duane Brown and Associates (DBA) and under contract to USAF, Brown demonstrated the internal accuracy of the Fairchild KC-6A camera as being one part in one million of the flying height. This camera, as a part of the USAF USQ-28 Geodetic Sub-System, was flown in the RC-135 aircraft at an altitude of 20,000 feet, augmented by an inertial system, and tracked by ground based ballistic cameras. This experiment clearly established aerial analytical photogrammetry as a reliable geodetic survey tool.

With the advent of GPS and with the refinement of software for automatic image measurement, airborne analytical photogrammetry found its way into operational mapping procedures. Bundle block aerial triangulation, production of digital elevation models, orthophotography, and of large scale topographic mapping have become the basic services and products provided by aerial
photogrammetry. Modern digital storage devices, electronic (digital) cameras, and optical scanning devices (LIDAR) are opening opportunities for wider applications for geospatial image collection devices. Rapid response times and higher geospatial accuracies will be some of the directions taken by research to meet the needs of NGA. Central to these developments will be the notion of measurement system calibration.

CONCEPT OF MEASUREMENT SYSTEM CALIBRATION

Common to all NGA image collection systems is the concept of measurement system calibration. During the designing and testing of new measurement systems, it is common practice to simply sum the estimated error contributions of each component to generate an estimated error representing the system. A more sophisticated approach has been to “cascade” the errors producing a more theoretically correct result. In such cases of error estimation of actual operating measurement systems, the environment in which the systems operate is neglected. The preferred approach is to assess the system’s operational measurement performance based on comparison of measurements to a standard of higher accuracy. In the aerial case, this is done by comparing independent resected photos to the exposure stations coordinates produced by GPS. In this case, GPS represents the independent standard of higher accuracy. As an alternative, the results of an aerial, bundle block adjustment can be used to produce coordinates for comparison to ground targets as the independent standard. After sufficient comparisons of the “final product” to its higher accuracy equivalent, the system can be said to have been calibrated by an in situ method. Comparisons of geospatial measurement accuracies based on the laboratory calibration to the in situ methods will be discussed below.

Reference is made to a classic treatment of the concept of measurement system calibration by Churchill Eisenhart (1962) then a member of the staff of the US Bureau of Standards. His complete approach to calibration is ideally suited to the needs of many of the NGA’s various geospatial data collectors. Accordingly, it will be discussed in some detail and serve as a basis for future R&D comments. To summarize Eisenhart’s approach to measurement system calibration, he states that two conditions must be met:

- Prepare measurement system specifications.
- Establish a state of statistical control.

Measurement system specifications fully describe all elements of the measurement system along with appropriate ranges. For the aerial case, elements include the camera, platform, software, and provisions for geodetic control to name a few. As an example of range, the altitude may be specified as between 100 and 1000 meters. Environment can also play a role. Temperature of the camera at collection time can greatly affect the focal length.

Establishment of a state of statistical control requires repeated exercise of the measurement system and comparison of its system produced coordinates to the standard of higher accuracy. In the case of the aerial system, this could be comparisons to system produced spatial coordinates to well-defined, independent, and accurately located ground targets. The comparisons are done randomly both in time and within the stated ranges of the system. After sufficient samples have been acquired to establish a level of consistency of results, the system can be said to be in a state
of statistical control. As long as results over time remain within acceptable limits of geospatial accuracy, the operational aerial photogrammetric system is said to be acceptable for application, and continues in a state of calibration.

**JUSTIFICATION FOR CALIBRATION**

The justification for NGA interest in measurement system calibration rests primarily in their need for geospatial coordinate accuracy. The USAF experiences with results of their USQ-28 project gave early confidence that airborne analytical photogrammetric methods can produce geospatial coordinates considered to be of geodetic quality. As indicated earlier, Brown was able to achieve accuracies approaching one part in one million of the camera flight altitude.

Early work of Goad (1989) indicated the potential for GPS to accurately determine the position of a moving vehicle. At White Sands Proving Ground he measured the track position of a moving sled equipped with GPS. Comparison of the track to the GPS positions produced mm discrepancies of 2 mm rmse.

It was not until GPS evolved to a satisfactory efficiency that it could replace the operationally expensive use of ground based, stellar oriented cameras to position the aerial camera at exposure time. The first well-documented application of use of GPS to position the camera over a controlled and targeted field for purposes of calibration was accomplished by Lapine (1990). Targeted ground control was established through the cooperation of the Ohio Department of Transportation (Aerial Engineering) and The Ohio State University, and located on the Ohio Transportation Research Center near Columbus. The aircraft was the NOAA Citation carrying the NGS Wild RC-10 15/23 camera at 15,000 feet above the ground. Airborne control was supplied by a Trimble (L1 only) receiver in the aircraft and at the base station about 10 miles from the range. Results of this early GPS controlled system calibration, allowed the NGS to collect suitable photography along the Eastern coast of Florida for coastal mapping purposes. Subsequently, results of the calibration proved useful for many airborne photogrammetric applications.

Since that early airborne calibration work, use of the airborne GPS controlled calibration method (*in situ*) has been promoted as an alternative to that provided for mapping cameras by USGS in their calibration laboratory located in Reston, Virginia. Some progress has been made in that regard but no *in situ* method of film based aerial camera calibration has been accepted as an alternative to their laboratory calibration. USGS is the responsible agency for calibrating cameras used for projects involving federal funds.

With the objective of promoting use of aerial calibration for mapping cameras, a series of investigations were conducted. These indicated that the design of the targeted range can be based on a simple crossroad pattern with targets radiating from the center at about 7.5 degree intervals. Based on synthetic data, one photo flying in the direction of a road, a second flying in the direction 45 degrees from the first and both taken over the intersection of roads, provided sufficient data to calibrate the camera. All functional correlations were suppressed below a practical level of significance.
For the purpose of comparing geospatial accuracies produced by laboratory and by *in situ* calibrations, photogrammetric resections were computed producing spatial coordinates of the exposure stations. A Zeiss LMK 15/23 camera was used for this test. Comparisons were based on photography taken at typical large scale operational altitude of 500 meters above ground. For both cases of resection, one using the laboratory calibration and the other the calibration produced from aerial photography (*in situ*), resected coordinate results were compared to the GPS positions, the independent standard of higher accuracy. Results of using a typical film based mapping camera, after comparing a number of sets of resected exposure stations, are presented in Table 1. These results are typical for film based mapping cameras compared over the last 20 years. Cause of such large elevation differences between laboratory and *in situ* calibrations may be lack of temperature corrections.

<table>
<thead>
<tr>
<th>COORDINATE</th>
<th>GPS – <em>IN SITU</em></th>
<th>FLT HT/DELTA</th>
<th>GPS - LAB</th>
<th>FLT HT/DELTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.05</td>
<td>10,000</td>
<td>-0.15</td>
<td>3,333</td>
</tr>
<tr>
<td>Y</td>
<td>0.06</td>
<td>8,333</td>
<td>0.02</td>
<td>25,000</td>
</tr>
<tr>
<td>Z</td>
<td>0.03</td>
<td>16,666</td>
<td>0.95</td>
<td>526 *</td>
</tr>
</tbody>
</table>

Figure 1. Relative Accuracies, GPS Compared to Calibration Methods@ Altitude 500 Meters [* Note: This proportional closure error exceeds the old, 3*rd* order survey closure error by a factor of 100]

Clearly, to improve geospatial accuracy, the concepts of calibration should follow an *in situ* approach and be in accord with Eisenhart.

**POTENTIAL NGA APPLICATIONS**

Viewing aerial analytical photogrammetry as having potential for providing high accuracy, geospatial coordinates by remote means; it then has application to NGA’s mission of providing accurate, reliable target locations on a timely basis. Use of unmanned aerial vehicles (UAV) for collection of intelligence is an important tool for such applications. It is the improvement in the current UAV intelligence collection technology, with emphasis on geospatial coordinates measurement accuracy that is the primary contribution made here to NGA’s research directions.

NGA will continue to be involved with the development of advanced geospatial data collection systems. Before calibration, these systems should demonstrate an adequate level of stability. Components of measurement systems may display high levels of stability; however, when operating as a system, stability in production of final measurements must also be demonstrated. To have a reasonable assurance of stability, the necessary number of measurements needs to be determined. After demonstration of adequate stability of a measurement system under development, it remains to calibrate under operational conditions. The crossroad method of target layout mentioned before, would be a likely choice, but is also simple enough to lend itself to calibrations application near aircraft deployment bases. It is likely that new systems will employ digital collection systems. This combination of crossroad range and digital aerial camera has been recently demonstrated. The range located in central Ohio, named the Buckeye Range, established by the Aerial Engineering Office of the Ohio Department of Transportation, is an example of a crossroad range. The range is patterned after the multi-collimator laboratory device.
located at the USGS facility in Reston, Virginia. The crossroads represent the crossing banks of optical collimators. Collimators are set at seven and one half degree intervals radiating from the nadir. The range targets are set accordingly. This spatial geometry has been used for many years by USGS to accomplish calibration of film-based mapping cameras thus verifying the geometry of the range.

The new digital camera used to demonstrate the calibration procedures is the Zeiss RMK-D using the green head. The camera (green head) has a nominal focal length of 46.8 mm with a digital array size of 41.5 x 46.8 mm. It was flown at 3900 feet above the Buckeye Range, and located in a Piper Aztec aircraft operated by Midwestern Aerial Photography. Airborne control was provided by a L1/L2 GPS receiver. Base station control was provided by a CORS station located 5 miles north of the range. The calibration, after adjustment, provided an rmse of image observational residuals of 1.9 microns; a result indicating a quality of fit equal to those experienced by laboratory calibrations.

**CONCLUDING REMARKS**

After reviewing the mission of NGA and current trends in geospatial data collection methods, some suggestions for further research and development can be made. Although the digital aerial metric camera has been the central data gathering device discussed here, other sensors such as multispectral sensors and LIDAR should be further analyzed for applications. These devices should be explored as more detailed mission requirements evolve.

In view of the NGA mission, the following research in photogrammetry should include:

- assess metric stability as new systems develop
- establish measurement system calibration procedures in accord with Eisenhart
- emphasize calibration concepts to UAV in field deployed applications

These recommendations are made with the objective of refining the metric, geospatial data collection systems of the NGA.

**REFERENCES**


Remote Sensing: An Overview

Russell G. Congalton
Department of Natural Resources and the Environment
University of New Hampshire

Forward

This paper presents a summary of current remote sensing technology. It begins with a historical review of the basics of this discipline. Current methodologies are then presented and discussed. The paper concludes with an examination of some future considerations. The material summarized in this paper, as well as more detail on all of these topics, can be found in the classic remote sensing texts by Campbell (2007), Jensen (2005) and Lillesand, Kiefer, and Chipman (2008).

1. Introduction

Remote sensing can be defined as learning something about an object without touching it. As human beings, we remotely sense objects with a number of our senses including our eyes, our noses, and our ears. In this paper we will define remote sensing only as it applies to the use of our eyes (i.e., sensing electromagnetic energy) to learn about an object of interest. Remote sensing has both a scientific component that includes measurements and equations as well as a more artistic or qualitative component that is gained from experience.

The field of remote sensing can be divided into two general categories: analog remote sensing and digital remote sensing. Analog remote sensing uses film to record the electromagnetic energy. Digital remote sensing uses some type of sensor to convert the electromagnetic energy into numbers that can be recorded as bits and bytes on a computer and then displayed on a monitor.

2. Remote Sensing

2.1 Analog Remote Sensing

The first photograph was taken around 1826 and the first aerial photograph was taken from a balloon above Paris in 1858 (Jensen 2007). However, the real advantages of analog remote sensing were not evident until World War I and became even more important in World War II. The military advantages of collecting information remotely are obvious and the methods for using aerial photography were rapidly developed. By the end of World War II there was a large group of highly trained personnel capable of doing amazing things with aerial photographs.

The field of analog remote sensing can be divided into two general categories: photo interpretation and photogrammetry. Photo interpretation is the qualitative or artistic component of analog remote sensing. Photogrammetry is the science, measurements, and the more quantitative component of analog remote sensing. Both components are important in the
understanding of analog remote sensing. However, this paper will concentrate more on the photo interpretation component, especially how it relates to the development of digital remote sensing. It is interesting to note that photogrammetry can stand alone as its own discipline and has developed a large base of knowledge and applications. Early uses of digital remote sensing tended to not be very dependent on many of the concepts of photogrammetry. Recently, however, with the increase in spatial resolution of many digital remote sensors, photogrammetry has again become a very relevant component of not only analog but also digital remote sensing.

Photo interpretation allows the analyst to deduce the significance of the objects that are of interest on the aerial photograph. In order to deduce the significance of a certain object, the interpreter uses the elements of photo interpretation. There are seven elements of photo interpretation: (1) size, (2) shape, (3) site, (4) shadow, (5) pattern, (6) tone/color, and (7) texture. Each element by itself is helpful in deducing the significance of an object. Some objects can easily be interpreted from a single element. For example, the Pentagon building is readily recognized just based on its shape. However, the interpreter/analyst gains increased confidence in their interpretation if a greater number of elements point to the same answer. This concept is called “the confluence of evidence”. For example, if the object of interest looks like pointy-shaped trees and is in a regular pattern and the shapes are small and they cast a triangular shadow and the photo is from the New England states, then the analyst can be very certain in concluding that this area is a Christmas tree farm.

Photo interpretation is accomplished by training the interpreter to recognize the various elements for an object of interest. The training is done by viewing the aerial photos, usually stereoscopically (i.e., in 3D), and then visiting the same areas on the ground. In this way, the interpreter begins to relate what an object looks like on the ground to what it looks like on the aerial photograph. The uses of the elements of photo interpretation aid the analyst in deducing the significance of the objects of interest.

Analog aerial photographs are limited to the wavelengths of the electromagnetic spectrum (i.e., light) that can be sensed with film. These wavelengths include the visible portion of the spectrum plus ultraviolet and near or photographic infrared. Since there are only 3 additive primary colors (Blue, Green, and Red) and the complementary 3 subtractive primary colors (Yellow, Magenta, and Cyan), film is limited to a maximum of 3 emulsion layers. This limitation is by definition of primary colors. A primary color can not be further broken down into any component parts. Therefore, the origins of any object on the film can be readily derived since it is composed of these subtractive primary colors (one for each emulsion layer). While historically black and white films (including ultraviolet, panchromatic, and black & white infrared) were commonly used, today the majority of aerial film is color film (either natural color or color infrared). Because of the 3 emulsion layer limitation, it is not possible to have the three wavelengths of light that our eyes see (Blue, Green, and Red) plus near infrared (NIR) on the same film. Natural color film senses the blue, green, and red light and creates photographs that emulate what our eyes see. Color infrared film (CIR) is not sensitive to blue light, but only green, red, and NIR light. These photographs look odd to the novice because human beings are not used to seeing in NIR. However, CIR aerial photographs are incredibly useful for many applications especially what has been termed “pre-visual stress detection” in vegetation. Healthy vegetation reflects highly in the NIR portion of the electromagnetic spectrum, causing healthy vegetation to appear
Remote Sensing: An Overview

as magenta on a CIR photo. Vegetation that is stressed because of drought or pollution or insect infestation, etc. has lower NIR reflectance that is readily visible in a CIR photograph. This stress is then “visible” on the CIR photo before it is visible to the human eye.

2.2 Digital Remote Sensing

While analog remote sensing has a long history and tradition, the use of digital remote sensing is relatively new and was built on many of the concepts and skills used in analog remote sensing. Digital remote sensing effectively began with the launch of the first Landsat satellite in 1972. Landsat 1 (initially called the Earth Resources Technology Satellite – ERTS) had two sensors aboard it. The return beam vidicon (RBV) was a television camera-like system that was expected to be the more useful of the two sensors. The other sensor was an experimental sensor called the Multispectral Scanner (MSS). This device was an optical-mechanical system with an oscillating mirror that scanned perpendicular to the direction of flight. It sensed in four wavelengths including the green, red, and two NIR portions of the electromagnetic spectrum. The RBV failed shortly after launch and the MSS became a tremendous success that set us on the path of digital remote sensing.

Since the launch of Landsat 1, there have been tremendous strides in the development of not only other multispectral scanner systems, but also hyperspectral and digital camera systems. However, regardless of the digital sensor there are a number of key factors to consider that are common to all. These factors include: (1) spectral resolution, (2) spatial resolution, (3) radiometric resolution, (4) temporal resolution, and (5) extent.

2.2.1 Spectral Resolution

Spectral resolution is typically defined as the number of portions of the electromagnetic spectrum that are sensed by the remote sensing device. These portions are referred to as “bands”. A second factor that is important in spectral resolution is the width of the bands. Traditionally, the band widths have been quite wide in multispectral imagery, often covering an entire color (e.g., the red or the blue portions) of the spectrum. If the remote sensing device captures only one band of imagery, it is called a panchromatic sensor and the resulting images will be black and white, regardless of the portion of the spectrum sensed. More recent hyperspectral imagery tends to have much narrower band widths with several to many bands within a single color of the spectrum.

Display of digital imagery for visual analysis has the same limitation as analog film (i.e., 3 primary colors). However, since digital imagery is displayed through a computer monitor, the additive primary colors (red, green, and blue) are used. Therefore, a computer display is called an RGB monitor. Again, a maximum of 3 bands of digital imagery can be displayed simultaneously. If panchromatic imagery is displayed then that single band is displayed through all three components of the computer monitor (R, G, and B) and the image will be in gray scale (black and white). However, if a multispectral image is displayed, then the analyst must select which bands to display through which components of the monitor. To display a natural color composite, the analyst would display the blue band through the blue component of the monitor, the green band through the green component, and the red band through the red component. This
image would appear just as our eyes would see it. However, many other composite images can be displayed. For example a Color Infrared Composite would display the green band through the blue component of the monitor, the red band through the green component, and the NIR band through the red component. This composite image looks just like a CIR aerial photograph. It is important to remember that only 3 bands can be displayed simultaneously and yet many different composite images are possible.

2.2.2 Spatial Resolution

Spatial resolution is defined by the pixel size of the imagery. A pixel or picture element is the smallest two-dimensional area sensed by the remote sensing device. An image is made up of a matrix of pixels. The digital remote sensing device records a spectral response for each wavelength of electromagnetic energy or “band” for each pixel. This response is called the brightness value (BV) or the digital number (DN). The range of brightness values depends on the radiometric resolution. If a pixel is recorded for a homogeneous area then the spectral response for that pixel will be purely that type. However, if the pixel is recorded for an area that has a mixture of types, then the spectral response will be an average of all that the pixel encompasses. Depending on the size of the pixels, many pixels may be mixtures.

2.2.3 Radiometric Resolution

The numeric range of the brightness values that records the spectral response for a pixel is determined by the radiometric resolution of the digital remote sensing device. These data are recorded as numbers in a computer as bits and bytes. A bit is simply a binary value of either 0 or 1 and represents the most elemental method of how a computer works. If an image is recorded in a single bit then each pixel is either black or white. No gray levels are possible. Adding bits adds range. If the radiometric resolution is 2 bits, then 4 values are possible (2 raised to the second power = 4). The possible values would be 0, 1, 2, and 3. Early Landsat imagery has 6 bit resolution (2 raised to the sixth power = 64) with a range of values from 0 – 63. Most imagery today has a radiometric resolution of 8 bits or 1 byte (range from 0 – 255). Some of the more recent digital remote sensing devices have 11 or even 12 bits.

It is important to note the difference between radiometric resolution and dynamic range. The radiometric resolution defines the potential range of values a digital remote sensing device can record. Dynamic range is the actual values within the radiometric resolution for a particular image. Given that the sensor must be designed to be able to record the darkest and brightest object on the Earth, it is most likely that the dynamic range of any given image will be less than the radiometric resolution.

2.2.4 Temporal Resolution

Temporal resolution is defined by how often a particular remote sensing device can image a particular area of interest. Sensors in airplanes and helicopters can acquire imagery of an area whenever it is needed. Sensors on satellites are in a given orbit and can only image a selected area on a set schedule. Landsat is a nadir sensor; it only images perpendicular to the Earth’s surface and therefore, can only sense the same place every 14 days. Other sensors are pointable
Remote Sensing: An Overview

and can acquire off-nadir imagery. This ability increases the revisit period (temporal resolution) of that sensor. However, there can be issues introduced in the image geometry and spectral response for images taken from too great an angle from vertical.

Temporal resolution often determines whether a particular remote sensing device can be used for a desired application. For example, using Landsat to monitor river flooding in the Mississippi River on a daily basis is not possible given that only one image of a given place is collected every 14 days.

2.2.5 Extent

Extent is the size of the area covered by a single scene (i.e., footprint). Typically, the larger the pixel size (spatial resolution), the larger the extent of the scene. Historically, satellite digital imagery covered large extents and had large to moderately large pixels. Recent, high spatial resolution digital sensors on both satellites and airplanes have much smaller extents.

3. Digital Image Analysis

Digital image analysis in digital remote sensing is analogous to photo interpretation in analog remote sensing. It is the process by which the selected imagery is converted/processed into information in the form of a thematic map. Digital image analysis is performed through a series of steps. These steps include: (1) image acquisition/selection, (2) pre-processing, (3) classification, (4) post-processing, and (5) accuracy assessment.

3.1 Image Acquisition/Selection

Selection or acquisition of the appropriate remotely sensed imagery is foremost determined by the application or objective of the analysis and the budget. Once these factors are known, the analyst should answer the questions presented previously. These questions include: what spectral, spatial, radiometric, temporal resolution and extent are required to accomplish the objectives of the study within the given budget? Once the answers to these questions are known, then the analyst can obtain the necessary imagery either from an archive of existing imagery or request acquisition of a new image from the appropriate image source.

3.2 Pre-processing

Pre-processing is defined as any technique performed on the image prior to the classification. There are many possible pre-processing techniques. However, the two most common techniques are geometric registration and radiometric/atmospheric correction.

Geometric registration is performed for three possible purposes: (1) to match one date of imagery to another, (2) to match the imagery to another spatial data layer, or (3) to match the imagery to the ground. The overwhelming advances in GIS (geographic information systems) have made the use of geometric registration an integral part of digital remote sensing. The process of geometric registration requires resampling of the imagery to fit the image to the desired source (i.e., other image, spatial data layer, or ground). There are a number of possible resampling algorithms.
including: nearest neighbor, bilinear interpolation, and cubic convolution. The most commonly used approach has traditionally been cubic convolution. However, this technique uses an average computed from the nearest 9 pixels and has the effect of smoothing (or reducing the variance) in the imagery. Since this variance can represent real information in the imagery, smoothing the image is not always recommended. The nearest neighbor technique does not average the image, but rather uses the nearest pixel as the label for the registered pixel. However, linear features can be split between lines of imagery using the nearest neighbor technique. Bilinear interpolation also averages, but only uses the nearest 4 pixels. This approach represents a compromise between the other approaches. If the goal of the remote sensing analysis is to get the most information from the imagery, then the use of the nearest neighbor technique should be employed.

Radiometric/atmospheric correction is done to remove the effects of spectral noise, haze, striping and other atmospheric effects so that multiple dates or multiple scenes of imagery can be analyzed together or so that the imagery can be compared to ground data. Historically, these corrections required collection of information about the atmospheric conditions during the time of the image acquisition. Algorithms/models used to correct for these effects relied on this ground information to accurately remove the atmospheric effects. More recently, algorithms/models have been developed that use information gained from certain wavelengths in the imagery to provide the required knowledge to accomplish the correction. These methods are preferred given the difficulty of collecting the required ground data during image acquisition especially for imagery that has already been acquired.

3.3 Classification

Classification of digital data has historically been limited to spectral information (tone/color). While these methods attempted to build on the interpretation methods developed in analog remote sensing, the use of the other elements of photo interpretation beyond just color/tone has been problematic. The digital imagery is simply a collection of numbers that represent the spectral response for a given pixel for a given wavelength of electromagnetic energy (i.e., band of imagery). In other words, the numbers represent the “color” of the objects. Other elements of photo interpretation such as texture and pattern were originally not part of the digital classification process.

In addition, digital image classification has traditionally been pixel-based. A pixel is an arbitrary sample of the ground and represents the average spectral response for all objects occurring within the pixel. If a pixel happens to fall on an area of a single land cover type (e.g., white pine), then the pixel is representative of the spectral response for white pine. However, if the pixel falls on an area that is a mixture of different cover types (white pine, shingle roof, sidewalk, and grass), then the “mixel” is actually the areaweighted average of these various spectral responses.

The earliest classification techniques tended to mimic photo interpretation and were called supervised classification techniques. These methods were followed by statistical clustering routines that were called unsupervised classification techniques. Both techniques were based completely on the spectral (color/tone) data in the imagery.
3.3.1 Supervised vs. Unsupervised Classification

Supervised classification is a process that mimics photo interpretation. The analyst “trains” the computer to recognize informational types such as land cover or vegetation in a similar way that the photo interpreter trains themselves to do the same thing. However, the interpreter uses the elements of photo interpretation while the computer is limited to creating statistics (means, minimums, maximums, variances, and co-variances) from the digital spectral responses (color/tone). The analyst is required to identify areas on the image of known informational type and create a training area (grouping of pixels) from which the computer creates a statistics file. Informational types that are rather homogeneous and distinct (e.g., water or sand) require only a few training areas. Complex, heterogeneous informational types (e.g., urban areas) require more training areas to adequately “train” the computer to identify these types. Supervised classification works especially well when there are only a few (4-8) informational types that are distinctly different from each other.

Unsupervised classification uses a statistical clustering algorithm to group the pixels in the imagery into spectral clusters. These clusters are spectrally unique, but may not be informationally unique. In other words, a single cluster may be a combination of a number of informational types (e.g., cluster 4 may be a combination of white pine and grass). The analyst decides how many clusters to use to group the imagery. Once the imagery has been grouped into the designated number of clusters, each cluster must be labeled with the corresponding informational type (i.e., land cover or vegetation type). Labeling the clusters can be difficult and frustrating as many of the clusters will be a combination of two or more informational types. Also, many clusters will be small and intimate knowledge of the area may be necessary in order to label them. In most cases, this knowledge is lacking which is why the analyst is trying to generate a thematic map of the area using remote sensing. The major advantage of the unsupervised classification technique is that all the spectral variation in the image is captured and used to group the imagery into clusters. The major disadvantage is that it is difficult to informationally label all the clusters to produce the thematic map.

3.3.2 Combined Approaches

Many remote sensing scientists have attempted to combine the supervised and unsupervised techniques together to take the maximum advantage of these two techniques while minimizing the disadvantages. Many of these examples can be found in the literature. A technique by Jensen (2005) selects larger clusters that result from an unsupervised classification and iteratively runs additional unsupervised classifications to break the clusters down into meaningful information classes. Another method by Chuvieco and Congalton (1988) combines the results of the supervised and unsupervised classification together using a statistical grouping routine. This method objectively labels the unsupervised clusters while also identifying issues of spectral uniqueness. Many other methods for combining these two basic classification techniques also exist.
3.3.3 Advanced Approaches

Using supervised or unsupervised classification approaches only work moderately well. Even the combined approaches only improve our ability to create accurate thematic maps a little more than using each technique separately. Therefore, a large amount of effort has been devoted to developing advanced classification approaches to improve our ability to create accurate thematic maps from digital remotely sensed imagery.

While there are many advanced approaches, this paper will only mention three: Classification and Regression Tree (CART) analysis, Artificial Neural Networks (ANN), and Support Vector Machines (SVM). CART analysis is a non-parametric algorithm that uses a set of training data to develop a hierarchical decision tree. This decision tree is created using a binary partitioning algorithm that selects the best variable to split the data into separate categories at each level of the hierarchy. Once the final tree is generated, it can then be used to label all the unknown pixels in the image. CART analysis has become widely used in the last few years both for pixel-based and object-based image classification. The method is extremely robust and provides significantly better map accuracies than have been achieved using the more basic approaches.

The ANN approach to image classification has been applied for the last ten years or more. The approach also relies on a training data set like CART does. The power of the ANN is that it has the ability to “learn”. An ANN consists of a set of nodes (called neurons) that are connected by links and organized by layers. There are at least three layers: the input layer, the output layer, and the intermediate layer(s). The training data are used to build the model that creates the weighting used in the intermediate layer(s) to label the unknown pixels. While this approach seems to have great potential, the method is subject to over-fitting the data and unnecessary complexity. Therefore, while much work is still being done on this technique, it has not experienced the same widespread adoption as CART.

SVM are derived from the field of statistical learning theory and have been used in the machine vision field for the last 10 years. More recently, these methods have been developed for use in creating thematic maps from remotely sensed imagery. SVM act by projecting the training data nonlinearly into feature space of a higher dimension than that of the input data. This projection is accomplished using a kernel function and results in a data set that now can be linearly separated. The ability to separate out the various informational classes in the imagery is a powerful advantage. Like ANN, SVM is subject to over-fitting. However, a technique that is part of the analysis works to minimize this issue. The use of SVM is relatively new and offers great potential for creating thematic maps from digital imagery.

3.3.4 Object-based Approaches

By far the greatest advance in classifying digital remotely sensed data in this century has been the widespread development and adoption of object-based image analysis (OBIA). Traditionally, all classifications were performed on a pixel basis. Given that a pixel is an arbitrary delineation of an area of the ground, any selected pixel may or may not be representative of the vegetation/land cover of that area. In the OBIA approach, unlabeled pixels are grouped into meaningful polygons that are then classified as polygons rather than individual pixels. This
method increases the number of attributes such as polygon shape, texture, perimeter to area ratio and many others that can be used to more accurately classify that grouping of pixels.

Polygons are created from pixels in OBIA using a method called segmentation. There are a number of current image analysis software packages that provide the means of performing object-based image analysis. In all these algorithms, the analyst must select a series of parameters that dictate how the segments or polygons are generated. Depending on the parameters selected, it is possible to create large polygons that may incorporate very general vegetation/land cover types or very small polygons that may divide even a specific cover type into multiple polygons.

The power of the segmentation process is two-fold. First, the imagery is now divided into polygons that can, in many ways, mimic the polygons that may have been drawn by an analyst that was manually interpreting this same image. In this way, some of the additional elements of manual interpretation mentioned earlier in this paper become relevant for digital image analysis. Secondly, as previously mentioned, the creation of polygons results in a powerful addition of attributes about the polygons that can be used by the classification algorithm to label the polygons. Both these factors significantly add to our ability to create accurate thematic maps.

3.4 Post-processing

Post-processing can be defined as those techniques applied to the imagery after it has been through the classification process. In other words, any techniques applied to the thematic map. It has been said that one analyst’s pre-processing is another analyst’s post-processing. It is true that many techniques that could be applied to the digital imagery as a pre-processing step may also be applied to the thematic map as a post-processing step. This statement is especially true of geometric registration. While currently most geometric correction is performed on the original imagery, such was not always the case. Historically, to avoid resampling the imagery and potentially removing important variation (information), the thematic map was geometrically registered to the ground instead of the original imagery. In addition, there are a vast variety of filtering processes that can be performed on the original imagery to enhance or smooth the imagery that can also be used on the thematic map to accomplish similar purposes.

3.5 Accuracy Assessment

Accuracy assessment is a vital step in any digital remote sensing project. The methods summarized here can be found in detail in Congalton and Green (2009). Historically, thematic maps generated from analog remotely sensed data through the use of photo interpretation were not assessed for accuracy. However, with the advent of digital remote sensing, quantitatively assessing the accuracy of thematic maps became a standard part of the mapping project.

Figure 1 outlines the process of map accuracy assessment. It is important to note that there are two components for assessing the accuracy of a map derived from remotely sensed imagery. The first component is positional accuracy and the second is thematic accuracy. The accuracy of the map is really a combination of these two accuracies and neither can be ignored in a valid assessment process. An error can occur if the map label is wrong or if the location is wrong.
Without assessing both positional and thematic accuracy it is impossible to determine the source of the errors.

Conducting an accuracy assessment is not as simple as following a series of steps. Rather there are a number of important considerations that must be balanced between statistical validity and what is practically attainable. The outline in Figure 1 documents these considerations.

The basic processes for assessing positional accuracy and thematic accuracy are very similar. However, positional accuracy is simpler and involves fewer issues to consider than thematic accuracy. Positional accuracy provides an assessment of the difference in distance between a sample of locations on the map and those same locations on a reference data set. Sampling is performed because it is not possible to visit every location due to time, cost, and efficiency. There must be an adequate number of samples and these samples must be appropriately distributed across the map. A number of statistics are commonly computed including the root mean square error (RMSE) and the National Standard for Spatial Data Accuracy (NSSDA). The NSSDA is the national standard currently used in the US for positional accuracy. However, there is an error in the calculation of this measure that results in it being too conservative. This issue is currently being investigated for possible correction by a number of scientists and government agencies.

Assessing the thematic accuracy of a map uses the same basic process as assessing positional accuracy. However, a greater number of issues must be considered (see the more complicated flow chart in Figure 1). Thematic accuracy assessment involves a number of initial considerations including taking into account the sources of error and the proper selection of a classification system. In addition, the collection of the reference data is more complicated including deciding the source of these data, how and when they will be collected, and insuring consistency in the collection process. Finally, sampling is a major component of the reference data collection. Sampling for thematic accuracy is much more complex than sampling for positional accuracy and requires selecting the appropriate sampling scheme along with the sample size and sample unit. Sample size for thematic accuracy assessment is considerably larger than positional accuracy assessment and requires more careful planning to be as efficient as possible. While many thematic accuracy assessments have been conducted using the pixel as the sampling unit, this sampling unit is not appropriate as it fails to account for any positional error in the image and reference data. Therefore, a grouping of pixels or a polygon is the far better choice for a sampling unit.

Once the reference data are appropriately collected, the method used to compute thematic accuracy uses a technique called an error matrix. An error matrix is a cross tabulation or contingency table generated from comparing the samples of the cover type on the thematic map and on the reference data. The error matrix can be generated using reference data accepted as either correct or incorrect (i.e., deterministic error matrix) or using the qualifiers of good, acceptable, and poor to produce a fuzzy error matrix. Once the error matrix is generated some basic descriptive statistics including overall, producer’s and user’s accuracies can be computed. In addition, there are a number of analysis techniques that can be performed from the error matrix. Most notable of these techniques is the Kappa analysis that allows the analyst to statistically test if one error matrix is significantly different than another.
Figure 1. A flow chart of the accuracy assessment process (Adapted from Congalton 2010).
It should be noted that Anderson et al. (1976) in a publication for the USGS stated that thematic maps generated from digital remotely sensed imagery should be at least 85 percent accurate. This value has been used as a guideline ever since. However, the overwhelmingly vast majority of thematic maps generated from remotely sensed data created from 1976 to the present have failed to meet this accuracy level.

4. Digital Image Types

4.1 MultiSpectral Imagery

The dominant digital image type for the last 40 years has been multispectral imagery from the launch of the first Landsat in 1972 through the launch of the latest GeoEye and DigitalGlobe sensors. Multispectral imagery contains multiple bands (more than 2 and less than 20) across a range of the electromagnetic spectrum. While there has been a marked increase in spatial resolution, especially of commercial imagery, during these 40 years it should be noted that there is a solid place that remains for mid-resolution imagery. The importance of continuing imagery with a spatial resolution of 20-30 meters and with a good spectral resolution that includes the visible, near-, and middle infrared portions of the electromagnetic spectrum cannot be understated. There is a special niche that this imagery fills that cannot be replaced by the higher spatial resolution imagery that costs significantly more to purchase. There will be increased uses of the higher spatial resolution data that continue to improve all the time, but this increase will not reduce the need for mid-resolution multispectral imagery.

4.2 Hyperspectral Imagery

Hyperspectral imagery is acquired using a sensor that collects many tens to even hundreds of bands of electromagnetic energy. This imagery is distinguished from multispectral imagery not only by the number of bands, but also by the width of each band. Multispectral imagery senses a limited number of rather broad wavelength ranges that are often not continuous along the electromagnetic spectrum. Hyperspectral imagery, on the other hand, senses many very narrow wavelength ranges (e.g., 10 microns in width) continuously along the electromagnetic spectrum.

Hyperspectral imagery has changed the way we perform digital image analysis. Given this imagery collected over narrow bandwidths across a large portion of the electromagnetic spectrum, it is possible to create spectral libraries of various information types and compare these for identification on the imagery. These libraries exist for a variety of rock and mineral types and have even been created for some simple land cover/vegetation classifications. These detailed spectral patterns also allow for the analysis of the chemical content of vegetation and other land cover. The uses of hyperspectral imagery for environmental studies, especially related to pollution and other hazards, have tremendous potential. Currently, significant research is occurring in this field. As the costs associated with this technology continue to decline, more and more uses of hyperspectral imagery will be developed.
4.3 Digital Camera Imagery

Digital cameras are revolutionizing the collection of digital imagery from airborne platforms. These cameras sense electromagnetic energy using either a charged-coupled device (CCD) or a Complimentary Metal Oxide Semiconductor (CMOS) computer chip and record the spectral reflectance from an object with greater sensitivity than the old analog film cameras. While more expensive than analog film cameras, new digital camera systems, especially large format systems, are rapidly replacing traditional analog systems. Large Federal Programs, such as the USDA National Agriculture Imagery Program (NAIP), are providing a tremendous source of digital imagery that the analyst can readily digitally analyze. Most digital camera imagery is collected as a natural color image (blue, green, and red) or as a color infrared image (green, red, and near infrared). Recently, more projects are acquiring all four wavelengths of imagery (blue, green, red, and near infrared). The spatial resolution of digital camera imagery is very high with 1 – 2 meter pixels being very common and some imagery having pixels as small as 15 cm.

4.4 Other Imagery

There are other sources of digital remotely sensed imagery that have not been presented in this paper. These sources include RADAR and LiDAR. Both these sources of imagery are important, but beyond the scope of this paper. RADAR imagery has been available for many years. However, only recently has the multispectral component of RADAR imagery become available (collecting multiple bands of imagery simultaneously and not just multiple polarizations) that significantly improves the ability to create thematic maps from this imagery. LiDAR has revolutionized the collection of elevation data and is an incredible source of information that can be used in creating thematic maps. In the last few years, these data have become commercially available and are being used as a vital part of many mapping projects.

5. Future Issues and Considerations

There are many issues and considerations related to the future of remote sensing. Four have been selected here for further discussion.

5.1 Software

The software for performing digital analysis is becoming more powerful and easier to use all the time. As computer processing continues to grow, the efficiencies by which the imagery can be analyzed also increases. However, there remains a major problem in that the software needed to perform a complete digital remote sensing analysis is not wholly contained within any single product. Rather, the tools needed are distributed among many different sources and a great deal of effort is required to put everything together. Currently there are digital image analysis packages, GIS software packages, statistics packages, and even software that only does object-based image analysis. It is cumbersome to work between all these various software packages. Great efficiencies could be gained if everything that the analyst needed was together inside a single package.
5.2 Data Exploration Approach

The major reason that remotely sensed imagery can be used to create thematic maps is that there is a very strong correlation between what is sensed on the imagery and what is actually occurring in the same area on the ground. The key to making the most use of this correlation is through a process of data exploration. While there are many techniques that are part of most image analysis projects that provide insight into this correlation (e.g., spectral pattern analysis, bi-spectral plots, Principal Components Analysis, band ratios, vegetation indices, etc.), most analysts have not embraced a data exploration approach to accomplish their work. The statistics community has embraced this philosophy and has developed an entire branch of statistics for “data mining”. Incorporation of a similar philosophy of data exploration to investigate all aspects of the correlation between the imagery and the ground could significantly improve our ability to create accurate thematic maps.

5.3 Landsat

Landsat is an invaluable source of remotely sensed imagery with a long history and an incredible archive. There is a tremendous need for the collection of imagery at this spatial and spectral resolution to continue well into the future. It is vital to our ability to look at global climate change, carbon sequestration, and other environmental issues that we have such imagery available. There is a need for a program to insure the future of Landsat and not simply a campaign each time a new Landsat is needed. Landsat imagery has become a public good and provides information that is not readily attainable anywhere else.

5.4 Remote Sensing and the General Public

Since the turn of the century, the general public has become increasingly aware of the use of remotely sensed imagery and geospatial technologies. Events such as the World Trade Center attack on September 11, 2001 and various natural disasters in New Orleans and elsewhere have made these technologies commonplace on the nightly news and the Internet. Applications such as Google Earth, Bing Maps, and others have given everyone the opportunity to quickly and easily use remotely sensed imagery anywhere in the world. This increased awareness and widespread understanding of these technologies is very positive. As more people become aware of the potential of these technologies more uses will be found for employing them for a myriad of applications.

6.0 Literature Cited


Cartographic Research in the United States: Current Trends and Future Directions

Robert McMaster
Department of Geography
University of Minnesota

This white paper will focus on three basic areas of cartographic research: scale and generalization, geographic visualization, and public participation mapping. It should be noted that the University Consortium for Geographic Information Science has developed and published a research agenda for GIS science, which consists of fourteen priorities, including (Usery and McMaster, 2004):

- Spatial Data Acquisition and Integration.
- Cognition of Geographic Information
- Scale
- Extensions to Geographic Representations
- Spatial Analysis and Modeling in a GIS Environment
- Uncertainty in Geographic Data and GIS-Based Analysis
- The Future of the Spatial Data Infrastructure
- Distributed and Mobile Computing
- GIS and Society: Interrelation, Integration, and Transformation
- Geographic Visualization
- Ontological Foundations for Geographic Information Science
- Remotely Acquired Data and Information in GIScience
- Geospatial Data Mining and Knowledge Discovery

1.0 SCALE AND GENERALIZATION
(Parts taken from Chapter 7, Thematic Cartography and Geographic Visualization, Vol. 3, Slocum, McMaster, Kessler, and Howard)

Geographic Scale

Scale is a fundamental concept in all of science and is of particular concern to geographers, cartographers, and others interested in geospatial data. Astronomers work at a spatial scale of light years, physicists work at the atomic spatial scale in mapping the Brownian motion of atoms, and geographers work at spatial scales from the human to the global. Within the fields of geography and cartography, the terms **geographic scale** and **cartographic scale** are often confused. Geographers and other social scientists use the term scale to mean the extent of the study area, such as a neighborhood, city, region, or state. Here, large scale indicates a large area—such as a state—whereas small scale represents a smaller entity—such as a neighborhood. Climatologists, for instance, talk about large-scale global circulation in relation to the entire Earth; in contrast, urban geographers talk about small-scale gentrification of a part of a city. Alternatively, cartographic scale is based on a strict mathematical principle: the representative fraction (RF). The RF, which expresses the relationship between map and Earth distances, has
become the standard measure for map scale in cartography. The basic format of the RF is quite simple, where RF is expressed as a ratio of map units to earth units (with the map units standardized to 1). For example, an RF of 1:25,000 indicates that one unit on the map is equivalent to 25,000 units on the surface of the Earth. The elegance of the RF is that the measure is unitless—with our example the 1:25,000 could represent inches, feet, or meters. Of course, in the same way that $\frac{1}{2}$ is a larger fraction than $\frac{1}{4}$, 1:25,000 is a larger scale than 1:50,000. Related to this concept, a scale of 1:25,000 depicts relatively little area but in much greater detail, whereas a scale of 1:250,000 depicts a larger area in less detail. Thus, it is the cartographic scale that determines the mapped space and level of geographic detail possible. At the extreme, architects work at very large scales, perhaps 1:100, where individual rooms and furniture can be depicted, whereas a standard globe might be constructed at a scale of 1:30,000,000, allowing for only the most basic of geographic detail to be provided. There are design issues that have to be considered when representing scales on maps, and a variety of methods for representing scale, including the RF, the verbal statement, and the graphical bar scale.

The term data resolution, which is related to scale, indicates the granularity of the data that is used in mapping. If mapping population characteristics of a city—an urban scale—the data can be acquired at a variety of resolutions, including census blocks, block groups, tracts, and even minor civil divisions (MCDs). Each level of resolution represents a different “grain” of the data. Likewise, when mapping biophysical data using remote sensing imagery, a variety of spatial resolutions are possible based on the sensor. Common grains are 79 meters (Landsat Multi-Spectral Scanner), 30 meters (Landsat Thematic Mapper), 20 meters (SPOT HRV multispectral), and 1 meter (Ikonos panchromatic). Low resolution refers to coarser grains (counties) and high resolution refers to finer grains (blocks). Cartographers must be careful to understand the relationship among geographic scale, cartographic scale, and data resolution, and how these influence the information content of the map.

Multiple-Scale Databases
Increasingly, cartographers and other geographic information scientists require the creation of multiscale/multiresolution databases from the same digital data set. This assumes that one can generate, from a master database, additional versions at a variety of scales. The need for such multiple-scale databases is a result of the requirements of the user. For instance, when mapping census data at the county level a user might wish to have significant detail in the boundaries. Alternatively, when using the same boundary files at the state level, less detail is needed. Because the generation of digital spatial data is extremely expensive and time-consuming, one master version of the database is often created and smaller scale versions are generated from this master scale. Further details are provided later.

Cartographic Generalization is the process of reducing the information content of maps due to scale change, map purpose, intended audience, and/or technical constraints. For instance, when reducing a 1:24,000 topographic map (large scale) to 1:250,000 (small scale), some of the geographical features must be either eliminated or modified because the amount of map space is significantly reduced. Of course, all maps are to some degree generalizations, as it is impossible to represent all features from the real world on a map, no matter what the scale.
The conceptual elements of generalization include reducing complexity, maintaining spatial accuracy, maintaining attribute accuracy, maintaining aesthetic quality, maintaining a logical hierarchy, and consistently applying the rules of generalization. Reducing complexity is perhaps the most significant goal of generalization. The question for the cartographer is relatively straightforward: How does one take a map at a scale of, perhaps, 1:24,000 and reduce it to 1:100,000? More important, the question is how the cartographer reduces the information content so that it is appropriate for the scale. Obviously, the complexity of detail that is provided at a scale of 1:24,000 cannot be represented at 1:100,000; some features must be eliminated and some detail must be modified. For centuries, through considerable experience, cartographers developed a sense of what constituted appropriate information content. The set of decisions required to generalize cartographic features based on their inherent complexity is difficult if not impossible to quantify, although as described next, several attempts have been made over the past two decades.

Clearly, there is a direct and strong relationship among scale, information content, and generalization. John Hudson explained the effect of scale by indicating what might be depicted on a map 5 by 7 inches:

- A house at a scale of 1:100
- A city block at a scale of 1:1,000
- An urban neighborhood at a scale of 1:10,000
- A small city at a scale of 1:100,000
- A large metropolitan area at a scale of 1:1,000,000
- Several states at a scale of 1:10,000,000
- Most of a hemisphere at a scale of 1:100,000,000
- The entire world with plenty of room to spare at a scale of 1:1,000,000,000

He explained that these examples, which range from largest (1:10^2) to smallest (1:10^9), span eight orders of magnitude and a logical geographical spectrum of scales. Geographers work at a variety of scales, from the very large—the neighborhood—to the very small—the world. Generalization is a key activity in changing the information content so that it is appropriate for these different scales. However, a rough guideline that cartographers use is that scale change should not exceed 10^x. Thus if you have a scale of 1:25,000, it should only be used for generalization up to 1:250,000. Beyond 1:250,000, the original data are “stretched” beyond their original fitness for use.

Two additional theoretical objectives important in generalization are maintaining the spatial and attribute accuracy of features. Spatial accuracy deals primarily with the geometric shifts that necessarily take place in generalization. For instance, in simplification coordinate pairs are deleted from the data set. By necessity, this shifts the geometric location of the features, creating “error.” The same problem occurs with feature displacement, where two features are pulled apart to prevent a graphical collision. A goal in the generalization process is to minimize this shifting and maintain as much spatial accuracy as possible. Attribute accuracy deals with the subject being mapped, such as population density or land use. For instance, classification, a key component of generalization, often degrades the original “accuracy” of the data through data aggregation.
When Generalization Is Required

In a digital cartographic environment, it is necessary to identify those specific conditions when generalization will be required. Although many such conditions can be identified, four of the fundamental ones include:

1. Congestion
2. Coalescence
3. Conflict
4. Complication

As explained by McMaster and Shea, *congestion* refers to the problem when, under scale reduction, too many objects are compressed into too small a space, resulting in overcrowding due to high feature density. Significant congestion results in decreased communication of the mapped message, for instance, when too many buildings are in close proximity. *Coalescence* refers to the condition in which features graphically collide due to scale change. In these situations, features actually touch. This condition thus requires the implementation of the displacement operation, as discussed shortly. *Conflict* results when, due to generalization, an inconsistency between or among features occurs. For instance, if generalization of a coastline eliminates a bay with a city located on it, either the city or the coastline must be moved to ensure that the urban area remains on the coast. Such spatial conflicts are difficult to both detect and correct. The condition of *complication* is dependent on the specific conditions that exist in a defined space. An example is a digital line that changes in complexity from one part to the next, such as a coastline that progresses from very smooth to very crenulated, like Maine’s coastline.

Despite the fact that many problems in generalization require the development and implementation of mathematical, statistical, or geometric measures, little work on generalization measurement has been reported. Two basic types of measures can be identified: procedural and quality assessment. *Procedural measures* are those needed to invoke and control the process of generalization. Such measures might include those to: (1) select a simplification algorithm, given a certain feature class; (2) modify a tolerance value along a feature as the complexity changes; (3) assess the density of a set of polygons being considered for agglomeration; (4) determine whether a feature should undergo a type change (e.g., area to point) due to scale modification; and (5) compute the curvature of a line segment to invoke a smoothing operation. *Quality assessment measures* evaluate both individual operations, such as the effect of simplification, and the overall quality of the generalization (e.g., poor, average, excellent).

A Framework for the Fundamental Operations

In the McMaster and Shea model discussed earlier, the third major component involves the fundamental operations or how to generalize. Most of the research in generalization assumes that the process can be broken down into a series of logical operations that can be classified according to the type of geometry of the feature. For instance, a simplification operation is designed for linear features, whereas an amalgamation operator works on areal features. Geographical features are normally represented in either a “vector” or “raster” format inside of a computer. The vector-representation uses x-y coordinate pairs to represent point features such as a house (a single x-y coordinate pair), a line feature such as a river (a string of connected x-y coordinate pairs), or an areal feature such as a park boundary (a string of x-y coordinate pairs in which the first pair matches the last pair). The raster approach uses a matrix of cells of a given
resolution (e.g., 30 meters) to represent features. Many standard GIS books describe these two data structures in more detail. Vector-based operators require more complicated strategies because they operate on strings of \(x\)-\(y\) coordinate pairs and require complex searching strategies.

To illustrate the practical application of generalization, we will consider how information is generalized when using the National Historical Geographic Information Systems (NHGIS) database. Much of this database was acquired from United States Geological Survey (USGS) Digital Line Graphs that were digitized at a scale of 1:150,000.

Consider the above illustration, which depicts three different representations of coastlines and county boundaries along the Florida Gulf Coast from Tampa Bay northward. The top right illustration shows the NHGIS raw data at a scale of 1:2,000,000. Here we can see excessive detail along the coast where lines coalesce and converge, a problem that is created from having far too much information (too many coordinate pairs) for the space. The NHGIS-produced generalization is depicted in the bottom right illustration. The left illustration enlarges a part of the coastline and compares the ungeneralized and generalized versions. The below illustration depicts further generalizations of the coastlines, including one from the Bureau of the Census.
County boundaries along the Florida gulf coast drawn at 1:2,000,000: a) base data from the Census TIGER files, with inland water extensions clipped, b) our generalization for a 1:2,000,000 target scale, and c) the Census cartographic boundary files.

Vector-Based Operations

Simplification is the most commonly used generalization operator. The concept is relatively straightforward, because it involves at its most basic level a “weeding” of unnecessary coordinate data. The goal is to retain as much of the geometry of the feature as possible, while eliminating the maximum number of coordinates. Below, we provide more detail on the simplification process.

Smoothing

Although often assumed to be identical to simplification, smoothing is a much different process. The smoothing operation shifts the position of points to improve the appearance of the feature. Smoothing algorithms relocate points in an attempt to plane away small perturbations and capture only the most significant trends of the line (McMaster and Shea 1992). As with simplification, there are many approaches for the process. Research has shown that a careful integration of simplification and smoothing routines can produce a simplified, yet aesthetically acceptable, result.

Aggregation involves merging multiple point features, such as a cluster of buildings. This process involves grouping point locations and representing them as areal units. The critical problem in this operation is determining both the density of points needed to identify a cluster to be aggregated and specifying the boundary around the resulting cluster. The most common approach is to triangulate the points (create triangles among neighboring points) and determine the density of the triangles (a grouping of smaller triangles might represent a cluster for aggregation).
Amalgamation is the process of fusing together nearby polygons, and is needed for both noncontinuous and continuous areal data. A noncontinuous example is a series of small islands in close proximity with size and detail that cannot be depicted at the smaller scale. A continuous example is with census tract data, where several tracts with similar statistical attributes can be joined together. Amalgamation is a very difficult problem in urban environments where a series of complex buildings might need to be joined.

The collapse operation involves the conversion of geometry. For instance, it might be that a complex urban area is collapsed to a point due to scale change and resymbolized with a geometric form, such as a circle. A complex set of buildings may be replaced with a simple rectangle—which might also involve amalgamation.

Merging involves fusing together groups of linear features, such as parallel railway lines, or edges of a river or stream. This is a form of collapse, where an areal feature is converted to a line. A simple solution is to average the two or multiple sides of a feature, and use this average to calculate the new feature’s position.
**Refinement** is another form of resymbolization, much like collapse. However, refinement is an operation that involves reducing a multiple set of features such as roads, buildings, and other types of urban structures to a simplified representation rather than a conversion of geometry. The key with refinement is that complex geometries are resymbolized to a simpler form, there is a “typification” of the objects. The example of refinement shown in the illustration is a selection of a stream network to depict the “essence” of the distribution in a simplified form.

**Exaggeration** is one of the more commonly applied generalization operations. Often it is necessary to amplify a specific part of an object to maintain clarity in scale reduction. The example in the illustration depicts the exaggeration of the mouth of a bay that would close under scale reduction.

**Enhancement** involves a symbolization change to emphasize the importance of a particular object. For instance, the delineation of a bridge under an existing road is often portrayed as a series of cased lines that assist in emphasizing that feature over another.

**Displacement** is perhaps the most difficult of the generalization operations, as it requires complex measurement. The problem can be illustrated with a series of cultural features in close proximity to a complex coastline. Assume, for example, that a highway and railroad follow a coastline in close proximity, with a series of smaller islands offshore. In the process of scale reduction, all features would tend to coalesce. The operation of displacement would pull the features apart to prevent this coalescence. What is critical in the displacement operation is the calculation of a displacement hierarchy because one feature will likely have to be shifted away from another.

### 2.0 GEOGRAPHIC VISUALIZATION

![Diagram](image)

(Parts taken from Chapter 26, Trends in Research and Development, *Thematic Cartography and Geographic Visualization*, Vol. 3, Slocum, McMaster, Kessler, and Howard)

The discipline of cartography has changed considerably since the 1960s, evolving from a discipline based on pen and ink to one based on computer technology. As the field continues to evolve, it is important to keep pace with ongoing research and new developments. The purpose of this chapter is to examine some of these developments; specifically, we will consider the
following: Daniel Carr and his colleagues’ work with linked micromap plots and conditioned choropleth maps, using senses other than vision to interpret spatial patterns, collaborative geovisualization, multimodal interfaces, information visualization, spatial data mining, visual analytics, and mobile mapping and location-based services (LBS). **Linked micromap plots (LM plots)** focus on a series of small maps termed micromaps, which divide a single spatial distribution into pieces. LM plots focus on local pattern perception as opposed to overall spatial pattern perception, and permit the display of statistical information such as confidence intervals for each observation. In a fashion similar to LM plots, **conditioned choropleth maps (CCmaps)** split a spatial pattern for a single attribute into a series of choropleth maps. In the case of CCmaps, rows and columns of maps are used, with the rows and columns corresponding to two attributes that might explain the attribute displayed in the choropleth map series. Throughout this book we have used vision to interpret spatial patterns, and a portion of the book’s title (geovisualization) reflects this. In this chapter, we will consider how other senses (e.g., sound, touch (or **haptics**), and even smell) might be used to assist in visualization, or be an alternative to visualization (for those who are blind or visually impaired). **Collaborative geovisualization** (or **geocollaboration**) refers to geovisualization activities in which more than one individual is involved in the visualization process. For instance, you might be located at one university and wish to discuss with a researcher at another university the visualization aspects of a climate-change model. Ideally, you should both be able to manipulate the model from your respective locations and see the results of what the other person is doing. We consider two applications of collaborative geovisualization: Nick Hedley and his colleagues’ work with augmented reality (AR) and a specialized projection table known as HI-SPACE; and Alan MacEachren and his colleagues’ work with ImmersaDesks and HI-SPACE. Although today, most of us are still working with Windows and associated mouse-based interfaces, researchers are experimenting with novel interfaces that utilize speech, lip movements, pen-based gestures, freehand gestures, and head and body movements. **Multimodal interfaces** are novel interfaces that involve two or more of these techniques; for instance, you might point to a location on a screen without touching the screen (a free-hand gesture) and say, “Show me all homes within 100 meters of this location.” Some of the most exciting work in multimodal interfaces is being done at the GeoVISTA Center at Penn State University, where researchers are developing collaborative geovisualization systems for emergency management situations. **Information visualization** involves the visualization and analysis of nonnumeric abstract information such as the nature of topics that are discussed on the front page of a newspaper over a month long period. **Spatialization** is the process of converting such abstract information to a spatial framework in which visualization is possible.

Two geographers who have done extensive work in information visualization. An example of Skupin’s work is his attempt to visualize the relationships among more than 2,000 abstracts submitted to the 1999 annual meeting of the Association of American Geographers (Skupin 2002, 2004). Given that the discipline of geography normally is divided into three broad areas (i.e., human geography, physical geography, and techniques), Skupin wondered how these three areas, and their relations, might be expressed in the relations among abstracts, and also whether other divisions of the discipline might be appropriate. One visualization that Skupin created appears in Figure 26.10, where we see two-dimensional “maps” of the relationships among
topics found in the abstracts. The leftmost map portrays the most generalized version, in which 10 regions arise, whereas the other two maps show the detail that is possible when maps of 100 and 800 regions are used, respectively.

A spatialization of topics found in abstracts submitted to the 1999 meeting of the Association of American Geographers: Map A represents the most generalized view, whereas maps B and C show the detail that is possible when zooming to various levels. (After Skupin 2002, “A cartographic approach to visualizing conference abstracts,” *IEEE Computer Graphics and Applications* 22, no. 1, p. 56; © 2002 IEEE.)

Visual Analytics
The term visual analytics was developed by the U.S. National Visualization and Analytics Center (NVAC; [http://nvac.pnl.gov/](http://nvac.pnl.gov/)) and was described in depth in the book documenting NVAC’s research agenda: *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (Thomas and Cook 2005). Visual analytics is defined as “the science of analytical reasoning facilitated by interactive visual interfaces” that should “detect the expected and discover the unexpected” (Thomas and Cook 2005, p. 4). In the Summer 2007 issue of the journal *Cartographica*, which focuses on visual analytics, Menno-Jan Kraak argues that visual analytics is essential for understanding and solving complex spatiotemporal problems. He states: Geovisualization offers interactive access to the data behind the map that is realized by combining graphics with geo-computational tools and database techniques. . . . However, to be able to deal with global challenges, we need more. Visualization has to be combined with analytics. (p. 115). The key is that large multivariate spatiotemporal data sets cannot be interpreted using just visualization; rather, interpreting them requires a synthesis of visualization and analytical methods. NVAC was formed in response to the September 11, 2001, terrorist attacks on the United States. Today, NVAC’s mission is to:

- Detect, prevent, and reduce the threat of terrorist attacks
- Identify and assess threats and vulnerabilities to our homeland
- Recover and minimize damage from terrorist attacks, should they occur ([http://nvac.pnl.gov/about.stm](http://nvac.pnl.gov/about.stm))
To get input from the academic community, NVAC has created five Regional Visualization and Analytics Centers (RVACS; http://nvac.pnl.gov/centers.htm). One of these centers, the North-East Visualization and Analytics Center (NEVAC), is coordinated with the GeoVISTA Center at Penn State, which we mentioned earlier in this chapter. NEVAC’s Web site (http://www.geovista.psu.edu/NEVAC/index.html) indicates that although homeland security is a focal point, advancements in visual analytics “will be relevant for all domains in which an ability to make sense out of complex information and apply knowledge to real-world decisions is needed,” as in disease epidemiology, environmental science and management, and regional planning. Examples of research involving visual analytics can be found in the January 2007 issue of Cartographica (as already mentioned), in the International Journal of Geographical Information Science (Vol. 21, Issue 8, 2007), in the Spring 2007 issue of Information Visualization, and as individual papers published elsewhere (e.g., Guo et al. 2006). Many visual analytics papers have a strong data-mining emphasis; for instance, Guo (2007) found that spatial data mining was essential to analyzing the pandemic spread of a population of 1.6 million people associated with more than 180,000 locations. Since visual analytics research often involves large multivariate spatiotemporal data sets, it can be challenging for a reader without a strong background in geovisualization and data mining to understand.

3.0 PUBLIC PARTICIPATION MAPPING

One of the most vibrant areas of GIScience, public participation GIS has developed research programs and applied projects around the world. Examples integrate researchers, practitioners, and citizens in myriad ways. For example, a recent project at Syracuse University, “The Syracuse Hunger Project” utilized GIS and mapping technologies to understand and help solve hunger-related issues in the city and to deal with nutritional inequalities. The researchers determined the location and operating hours of food pantries, the number of meals served, statistics on participation in government food aid programs, and census data about income, race, and housing quality. The study used spatial approaches to better understand the pattern of hunger, and where resources might be targeted. PPGIS research has made important advances in its engagement with issues for GIS and society research. Another examples of neighborhood mapping is depicted below, which illustrates the relationship between Toxic Release Inventory and other sites and geodemographic data for the Phillips Neighborhood or Minneapolis.
These examples show the possible relationship between geospatial technologies and social problems and raises as series of vexing questions on what broadly is called GIS and society. Some of these questions include: Who has access to the data and software? Are certain parts of society empowered, or disempowered, by the use of GIS.

As we think about the access to and use of spatial information by a public, however that is defined, the emerging field of public participation GIS will be very important. Recent research has identified a series of models by which the public can gain access to spatial information, maps, and in some cases analysis itself. These six models include:

1) Community-based (in-house) GIS
2) University/Community Partnerships
3) GIS facilities in universities and public libraries
4) “Map Rooms”
5) Internet map servers
6) Neighborhood GIS center

It is obvious that the most promising of these six is the Internet map server, where there have been a multitude of sites created over the recent five years, perhaps one of the best known of which is “Google Earth”. Such map servers are emerging at many scales, from the local (community-based mapping systems to look at neighborhood characteristics) to the state-level (sites showing the distribution of natural resources) to global scale sites such as Google Earth. Each of these six models has their own distinct advantages and disadvantages. As noted in the paper (Leitner, et. al, 2000) each these models of delivery can be differentiated along several dimensions, including: Communication Structures, Nature of Interaction with GIS, Location, Stakeholders, and the Legal and Ethical Issues.

A fundamental question in the maturation of PPGIS is what do we mean by a public? When the Minnesota DNR creates and distributes geospatial wetlands data, are these not a form of public GIS? When the Bureau of the Census distributes digital versions of boundary files and statistical
information, is this not public? The real meaning here is a public participation GIS provides access at a more grassroots level, a level of communities, of neighborhoods, of even the individual.

For the purpose of this discussion, PPGI Science represents a body of knowledge that parallels GI Science, but with specific concerns towards the broader utilization of these technologies by all involved people, especially the proverbial average citizen, often represented through community groups, grassroots organizations, and other such entities. What, then, are the fundamental questions that are emerging? Some might include:

- What technologies are most appropriate for community groups? How does the technological expertise get maintained in communities? How do community groups gain access to the appropriate data?
- What models of access are most appropriate?
- What are the methods of localized knowledge acquisition that are most appropriate? What are new methods—Primary survey work?
- How will these technologies fundamentally change the political/social structures of community groups?
- What forms, and new forms, of representation are best suited for public participation work?
- How do neighborhood groups deal with issues of scale (e.g., relationship with municipal and state regulations)?
- Is there a fundamental difference in the cognition of space and spatial principles among community groups?

Volunteer Geographies
The availability of user-friendly yet customizeable web services for on-line mapping has changed the ways that people can access and work with geographic information. Ultimately, this has utterly reconfigured the broader relationship between societal knowledge of place, geographic information resources, and the roles and uses of maps. Coupled with the boom in low-cost GPS equipment, which makes it very easy to collect locational data, what is becoming known as volunteer mapping has seen amazing growth in the last ten years. Volunteered geographic information includes a broad range of digital data including geotagged images online, wikipedia entries with geographic coordinates and more specialized place descriptions found in Wikimedia, mashups produced with GoogleEarth, NASA WorldWind or similar virtual earth and mapping applications, to volunteer efforts to create public-domain geospatial data layers, such as OpenStreetMap. The possibilities for creating user-generated content are endless.

While many of the potentials of volunteer mapping dovetail with PPGIS applications, it is actually distinct as the volunteer aspects focus more on producing geographic information and developing infrastructures, often commercially oriented, although some of the more interesting applications develop counter to governmental restrictions on the availability of GI. In fact, as in the case of OpenStreetMap, volunteer mapping is actually producing alternative data sets to government sanctioned, funded, and controlled data. As the acquisition and/or licensing costs may be exorbitant for most people, volunteer mapping has attracted a great deal of interest in European countries, where the costs of street and basic topographical data can be very high.

Taking off from this basic description of volunteer geographies, a useful division comes by distinguishing volunteer geographies underwritten by government bodies, on one hand, and
commercial grassroots organized collection and provision of geographic information on the other. Until now, most virtual geographic information fits the latter category, although examples of WaterShed Watches supplementing citizen collected data with geographical location information and then mapped certainly exist. Governments and non-governmental organizations already offer data for creating browsable and interactive maps of their data. More prevalent and the driving force behind much of volunteer geographic information are the multitude of citizen group and commercial applications. Examples abound from around the world use these online mapping packages to powerfully visualize their data to support citizen group activities, for example the Neighbors Against Irresponsible Logging (NAIL), documented online. More popular examples include Wikimapia, Flickr and scores of other volunteer geographic information applications.

A number of companies are now developing ways to take on roles in commercially oriented social networking and connect them to specific hardware platforms. For example, the application Twinkle, available on Apple's iPhone, allows contacts to follow updates about your location and brief messages that you wish to share. Loopt allows you to find out where people on a contact list are and what they are up to too. Twinkle and Loopt are just two of seven social networking application for the iPhone that allow people to track each other in various ways. While there are certainly some privacy concerns to be considered, at the moment these applications are all based on an opt-in licensing model that requires each user to specifically indicate how they want to make their locational information available to a list of contacts and the general public. These are becoming part of location-based social networking, expected to have worldwide revenues of $3.3 billion by 2013.

The ubiquity of GI in most parts of the world and its growing significance elsewhere continues to put GIS and society researchers before challenges and opportunities. The speed of these changes is astonishing and we often struggle to keep up with GI-based applications that a few years would have come across as science-fiction dreams lifted from the pages of futuristic comics. Fortunately there is a sound basis and continuation of activities. While the vibrancy of the activities is of great relevance for GIS and Society work, we should not lose sight of their origins and continue to follow core areas of GI science work on PPGIS and organizations.

As this discussion has shown, any new technology undoubtedly will have an impact on the society that it is embedded in. Conversely, the society itself has helped to determine the shape and success of that technology. Geographic information systems are now, more than ever, having a significant impact on most societies. Publicly-available mapping web sites, low-cost global positioning systems, vehicle navigation systems, and myriad other mobile mapping technologies (mobile devices and services using digital maps populated with rich attribute information such as restaurants, hospitals, schools, and retail), coupled with the wide dissemination of free spatial data have changed the way many humans now live. But the opportunities come with costs including differential knowledge that impacts decision-making, a loss of “spatial” privacy, and the increase in spatial surveillance such as the cameras now installed in urban areas at key locations. Research in maps, GIS, and society must continue to look all aspects of GIS on our society, and in particular how these sophisticated technologies can help “society” understand the complex social and natural environment in which they live.
REFERENCES


Forward

The early history of geodesy is well documented in An Introduction to GEODESY, The History and Concepts of Modern Geodesy by James R. Smith, and later history and geodetic technologies are covered in Geodesy: a Look to the Future (NRC, 1985) and Geodesy in the Year 2000 (NRC, 1990). The technical aspects of the Global Navigation Satellite Systems are discussed in the 20th Annual Proceedings from the Institute on Navigation, as well as in Inside GNSS and GPS World. I therefore decided to limit this work to those aspects of geodesy that pertain to positioning, datum and coordinate systems as they relate to surveying, mapping, and remote sensing because these topics are most likely appropriate to Defense research.

INTRODUCTION

The science of geodesy can be traced back to around 240 BC, when Eratosthenes determined the circumference of the Earth using the fact that during the summer solstice at the Egyptian city of Swenet, the sun illuminated the bottom of a deep well while at the same moment in Alexandria, a pole cast a shadow of 7.5° (1/50th of a circle) some 5000 stadia away (stadia being related to a camel’s steps). He calculated that 50 times 5000 stadia would represent the circumference of the Earth. His calculation was amazingly accurate, considering that most scientists at the time theorized that the Earth was flat. Since that time, geodesy has evolved as a science that requires state-of-the-art measuring equipment to determine precise distances and angles on the Earth’s surface. Each advance in technology and mathematics brings us closer to determining the true size and shape of the Earth. Perhaps the greatest advances in geodesy occurred shortly after 1957, when Russia launched Sputnik I and ushered in the space age.1 Until this time, all the accumulated science in geodesy since 240 BC relegated geodetic systems to unique country-sized areas on the Earth’s surface. The space age ultimately enabled geodesists all over the globe to use a single technology with positions determined relative to a single global coordinate system. Researchers at Johns Hopkins University Applied Physics Laboratory found a way to compute any accurate orbit for Sputnik based on the Doppler Shift of the transmitted signal. One month later, they reversed the problem and reasoned that if the orbit was known, the Doppler shift could be used to position an Earth-bound receiver. This early work led to the design and successful launch of TRANSIT 1B, the first navigation satellite, in 1960. In 1962, TRANSIT was declared operational for use by ballistic missile submarines.2 At the same time, the determination of worldwide gravity values became important for determining missile trajectories. As such, geodesists were called on to create gravity models using a consistent datum for worldwide consistency. This paper will cover some of the major aspects that have gotten the science of geodesy to where it is today since the early days of the space program.

1 See <http://history.nasa.gov/sputnik>.
Definitions

Geodesy is the science concerned with determining the size and shape of the Earth. In practice, geodesy involves defining a coordinate system that best fits the Earth’s surface over a given area where the curvature of the Earth is taken into account to avoid the accumulation of error in angles and distance measurements. The unique set of parameters which define the coordinate system is referred to as a datum. Datums can be local, national, or global. The larger the area to be covered, the more complex the size and shape of the surface becomes. A large area must also account for temporal deformations of the Earth’s crust. This is particularly problematic for geodetic networks created incrementally over long periods of time. Prior to the use of satellite technology, each country had a unique coordinate system derived from astronomic latitude and longitude. Today, Global Navigation Satellite Systems (GNSS) are used for positioning and, as such, worldwide datums have been defined.

Positions can be classified as “absolute,” “incremental” or “relative.” Absolute positions are determined directly from observations of stars or satellites. Incremental positions are derived from a series of measurements, generally forming geometric shapes whose interior angles and side lengths are mathematically combined to calculate positions. Relative positions, which are predominantly used today, are calculated from satellite data collected simultaneously at known and unknown locations.

DATUMS

Prior to the introduction of satellite techniques, most nations used separate datums for horizontal and vertical coordinates. Separate datums were used because horizontal geodetic positions are reduced to a common mathematical surface that closely represents the size and shape of the national territory at sea level to remove displacements in position caused by the topography. Removing the topography makes the determination of a position independent of the path taken in the survey. Vertical geodetic heights, however, are referenced to an adopted mean sea level that is closely associated with the gravity field and that follows an undulating surface known as the geoid. The undulations are caused by different mass distributions within the Earth. The vertical coordinate ensured that “water would always flow downhill.”

Horizontal and vertical observations used different techniques, instrumentation and data reduction. Horizontal coordinates (geodetic latitude and longitude) were derived from a combination of astronomic and terrestrial observations. Astronomic latitude and longitude were determined periodically at the same location as corresponding geodetic latitude and longitude to determine a correction (Laplace Correction) to the simplifying assumption that geodetic horizontal angles were measured orthogonal to the reference ellipsoid. In fact, the observations are orthogonal to the local gravity surface and the two surfaces are rarely coincident or parallel to each other. The Laplace Correction insured that the geodetic network maintained proper orientation with respect to the reference ellipsoid.
Vertical coordinates (orthometric heights\(^3\)) were derived from spirit level measurements parallel to the Earth’s gravity field. Orthometric heights are defined as the vertical distance between the elevated points on the surface of the Earth perpendicular to every equipotential surface until reaching the reference equipotential surface. An equipotential surface is a surface where the gravitational potential is the same everywhere on that surface (i.e., water will not flow on that surface). As a result, elevations are more correctly referred to as orthometric heights.

Today, with the use of GNSS for determining coordinates, there are no longer separate datums for horizontal and vertical coordinates. GNSS coordinates and their associated datum are determined on a smooth mathematical surface called an ellipsoid that most closely represents the shape of the Earth on a global scale. To convert ellipsoid heights to orthometric heights, a mathematical model of the difference between the ellipsoid and gravitational surface (geoid) must be determined. This has become one of the foremost tasks of modern geodesists.

Worldwide datums can be uniquely defined by eight parameters, including three parameters that define the origin or center of mass of the Earth, three parameters that define the orientation of the three coordinate axes, and two parameters that define the size and shape of the reference ellipsoid. The coordinates are Geocentric X, Y, and Z (having an origin as the center of mass of the Earth), which can be transformed into latitude, longitude and ellipsoid height. Ellipsoid height can be further transformed into Orthometric Height using a Geoid Model. The two most prominent world datums are the World Geodetic System 1984 (WGS84) and the International Terrestrial Reference Frame (ITRF). The ITRF has three additional parameters which account for the velocity of the coordinates, given that the different crustal plates on the Earth move at different rates and direction relative to one another.

**World Geodetic Reference Systems**

Prior to 1967, the defining parameters for a reference ellipsoid were chosen to best fit a particular geographic region. Having the surface of the reference ellipsoid close to the physical surface simplified computations. One of the first truly global ellipsoid models (the Hayford Ellipsoid) was developed in 1910 and adopted by the International Union of Geodesy and Geophysics (IUGG) in 1924. By 1967, it was apparent that a more accurate global ellipsoid was required for international geodesy, and the IUGG recommended using GRS 67 where a greater degree of accuracy was required. The formal adoption of an internationally agreed upon set of datum parameters set the stage for improving coordinate systems around the world.

The Geodetic Reference System of 1980 (GRS80) was approved and adopted at the 1979 IUGG General Assembly. Until this time, the semi-major axis and the flattening defined the size and shape of the reference ellipsoid. In the case of GRS 80, four parameters were adopted, each of which is absolute in value: the semi-major axis of the reference ellipsoid (a), the geocentric gravitational constant of the Earth including its atmosphere (GM), the dynamical form factor of the Earth (J2), and the angular velocity of the Earth (\(\omega\)). These four parameters define an ellipse in revolution, centered on the center of mass of the Earth, and having the same size and gravitational potential of the Earth. Based on these absolute values, eight parameters are commonly used to describe a datum and its coordinate system. Three parameters define the

---

\(^3\) Height is the more acceptable term when referring to geodetic coordinates. Elevation is generally reserved for height above a tidal datum such as Mean Lower Low Water in the US.
center of mass of the Earth as X=0, Y=0 and Z=0 (i.e., Earth-centered), three parameters define the orientation of the orthogonal axes (i.e., Earth-fixed), and two parameters define the size and shape of the reference ellipsoid centered on the origin. The volume of the reference ellipsoid is related to the accepted mass of the Earth. The Z-axis exits at the mean rotational axis (North Pole). X and Y axes are orthogonal to each other and to the Z-axis so their plane best describes the equatorial plane of the Earth. The X-axis exits the equatorial plane at zero degrees latitude and longitude and most closely approximates the longitude of Universal Coordinated Time. The Y-axis exits the equatorial plane at zero degrees latitude and 90 degrees east longitude. The semi-major axis of the “equipotential” ellipsoid is 6,378,137 m and the flattening is 1:298.257222101.

**International Earth Rotation Service (IERS)**

IERS was established in 1988 to replace the International Polar Motion Service (IPMS). Its creation marks the transition from classical astronomic techniques for determining Earth orientation to the use of satellites and galactic space technology. As positioning by satellites became a reality, so did the need to most accurately establish and maintain the most accurate worldwide reference frame. IERS is supported by the International Astronomic Union and the IUGG. IERS is responsible for (1) defining and maintaining a worldwide terrestrial reference system based on space geodesy, (2) defining and maintaining a conventional celestial reference frame, and (3) determining Earth orientation parameters, including terrestrial and celestial polar coordinates and universal time. To accomplish these objectives, IERS employs technologies such as VLBI, SLR and LLR.

**Very Long Baseline Interferometry (VLBI).** VLBI uses signals from distant extragalactic radio sources. Two or more radio telescopes simultaneously record the radio signals and accurate time ticks from different parts of the world on a magnetic tape. The signals on the recorded tapes are correlated at a processing center. The distance between the telescopes is related to the time difference of the correlated signals. Extremely long baselines can be measured with centimeter accuracy. Global coordinates, Earth orientation parameters and temporal variations in the orientation of the Earth (polar motion) can be determined with great accuracy. Regular observing campaigns are run to help determine the parameters necessary for global geodetic networks: (1) geocentric X and Y coordinates for the rotational axis of the Earth, (2) universal time (zero longitude), (3) celestial coordinates of the pole, and (4) coordinates of new radio telescopes.

**Satellite Laser Ranging (SLR).** A laser beam is transmitted from Earth to an orbiting satellite composed of retro-reflectors. The time of the beam’s round trip is a function of two times the range between the transmitted pulse and the satellite. With sufficient ground stations, the orbit of the satellite can be determined accurately enough to measure slight perturbations in the orbit resulting from anomalies in the gravity field, and therefore to better determine the size and shape of the Earth. SLR using the Laser Geodynamics Satellite (LAGEOS) is a source for rapid determination of polar motion and time.

**Lunar Laser Ranging (LLR).** Lunar Laser Ranging is similar technology to SLR, but the laser beam is aimed at retro-reflectors established on the moon. The extreme distance to the moon
(400,000 km) makes it more difficult to obtain a reflected signal. Nevertheless, the ranging is useful for studying the effects of tidal coupling. LLR has determined that the moon is moving away from the Earth at a rate of 3.7 cm per year, resulting in a decrease in the Earth’s rate of rotation (time).

**International Terrestrial Reference Frame (ITRF)**

The International Terrestrial Reference Frame is an international cooperative effort created and maintained under the International Earth Rotation Service. ITRF uses GPS, VLBI, SAR and LLR to construct a worldwide coordinate system that accounts for the continuous motion of the pole, varying rotational velocity of the Earth, and the movement of crustal plates with respect to one another. The ITRF includes geocentric coordinates and velocities for each coordinate. It assigns higher weights to seven SLR and three VLBI sites. Some 800 GPS sites are monitored around the world, contributing to a particular epoch solution. Updated coordinates for a station can be computed by multiplying the velocity times the time elapsed since the beginning of the epoch. A number of global and national datums (e.g., WGS84 and NAD83) have transformation parameters that enable the accuracy of their coordinate realization to be measured against ITRF.

**DEPARTMENT OF DEFENSE WORLD GEODETIC SYSTEM**

With the advent of the space age, a worldwide unified geodetic coordinate system became essential for numerous reasons, including:

1. Need to facilitate international space science and astronautics
2. Lack of inter-continental geodetic connectivity
3. Inability of large continental geodetic systems to provide a worldwide geo-data basis
4. Need for global charts for navigation and geography
5. Need for a standardized, NATO-wide geospatial reference system to support the Western Cold War.

Worldwide efforts, particularly within the U.S. military (DOD), led to the creation of the World Geodetic System of 1960. The DOD had to design a datum consisting of a smooth surface of zero elevation. This elevation was different from traditional elevations, which were computed separately from position and defined by a surface of constant gravitational potential (i.e., equipotential surface or geoid), which varied from nation to nation. The DOD understood that a smooth reference surface described by a single equation would be desirable for a worldwide datum. The task was to patch together many existing continental datums into a cohesive worldwide datum that best fit the Earth.

A combination of astronomic, geodetic and surface gravity data as well as surveys from long range (300 km) aircraft navigation systems developed during WWII was used to define a “best fitting” Earth-centered ellipsoid for a worldwide datum. The astronomic and geodetic data were used to determine accurate deflection of the vertical. The aircraft navigation systems (e.g., SHORAN [SHOrt RAange Navigation System]) proved useful in evaluating weaknesses in continental geodetic networks and, due to efforts by the Air Force, connected Florida to Puerto Rico to Trinidad and eventually to South America. This early form of electronic ranging systems
allowed the connection of smaller sub-networks and achieved 1:50,000 accuracy. Classical methods were more accurate but were limited by line of sight.

The Army and Air Force each developed an Earth-centered ellipsoid using different approaches to the gravimetric datum orientation method. The Army and Air Force results agreed quite well and with national datums. The resulting agreed upon datum became WGS 1960. WGS 1960 lasted about five years during which time satellite navigation systems (TRANSIT) were being developed. Also during this time, more gravimetric data were collected and continental geodetic networks were expanded as a result of improved technology such as electronic distance measuring equipment and satellite positioning. In 1966, a World Geodetic System Committee composed of the Army, Navy and Air Force was established to determine a more accurate worldwide datum to satisfy mapping, charting and geodesy. Using most recently collected geodetic and gravity data and astrogeodetic technology, WGS 66 was derived having a flattening of 1:298.25 and a semi-major axis of 6,378,145 m. WGS 66 was implemented in 1967. In addition to the datum definition, a 5°×5° mean free air gravity anomaly field provided basic data for producing a WGS 66 gravity geoid, the first practical global geoid model. The implementation of a gravity geoid would become paramount in the future to enable the transformation of accurate satellite derived ellipsoid heights to orthometric heights.

WGS 66 lasted for about six years before being replaced by WGS 72. The new datum used more surface gravity data and the TRANSIT navigation system was operated in a “geodetic” mode that extracted one-meter level horizontal accuracy worldwide. Least squares solutions were obtained for the geodetic data and parameters associated with the gravity field; one of the largest least squares solutions ever attempted. Separate solutions were obtained for the ellipsoid parameters, datum shifts and other associated parameters. The semi-major axis of 6,378,135 m and the flattening of 1:298.26 were adopted.

WGS 72 lasted approximately 12 years, but by 1980, more accuracy, inter-continental connections between national datums, crustal dynamics and improved gravity models demanded a more accurate mathematical representation of the Earth. The problem was no longer limited to DOD requirements. WGS 72 was replaced by WGS84, which used the GRS80 ellipsoid parameters along with astrogeodetic technologies such as VLBI, SLR and LLR.

The orientation for WGS 84 follows the criteria for GRS80 with an epoch date of 1984 with respect to the Conventional Terrestrial Pole and Zero Meridian. At this point, a distinction between the reference frame and coordinates must be made. The GRS 80 reference frame is absolute to the least significant digit obtainable. The reference frame, however, is realized or made operational by coordinates of geodetic points on the surface of the Earth. In the case of WGS 84, the realization is accomplished at five Air Force GPS tracking stations and seven National Imagery and Mapping Agency (NIMA) tracking stations situated around the world. The coordinates of these twelve stations determine the coordinates of all other points in geodetic and mapping programs. The five Air Force stations are located near the equator and are nearly equal distance apart. The seven NIMA stations were added to obtain more equally spaced north-south tracking stations. The implementation of WGS 84 began with a least squares solution involving DOD geodetic control stations throughout the world with coordinates established using TRANSIT satellite absolute positioning (a TRANSIT position has a 1-meter absolute accuracy).
The results of the least squares solution determined the coordinates for the tracking stations and, therefore, for the center of mass of the Earth. The solution also improved the accuracy of global coordinates for their geodetic network. The tracking station coordinates have been refined twice since 1984, which fixed a systematic ellipsoid height bias attributed to limitations in accuracy of the TRANSIT determined absolute positions.

WGS 84 continues to be refined, with the latest version (WGS84 [G1150]) effective in July 2005. Since WGS 84(G730), WGS84 has been tested against the International Terrestrial Reference Frame 1994 (ITRF94) at the 0.1-meter level of agreement using a seven parameter transformation. Corresponding precise ephemerides between ITRF94 and WGS84 reveal systematic differences no larger than 0.02 meters day to day variance than is statistically insignificant and much lower than is required for mapping and charting (less than 0.3 meters). The tracking stations have an absolute accuracy of ± 0.05 m (one standard deviation).

In addition to the changes in tracking station coordinates, an intensive effort by NIMA, NASA Goddard Space Flight Center and the Ohio State University led to a new global gravity model Earth Gravity Model 1996 (EDM96). A mathematical model of crustal motion (NNR-NUVEL1A) has been produced to account for crustal motion between the tracking stations. Although the accuracy of the tracking station positions may be a limiting factor for geodesy, their main purpose is to support absolute positioning for navigation. On the other hand, relative geodetic accuracy can be accomplished and is not dependent on the GPS orbits derived from the tracking stations positions. Since WGS 84 is very important because it is the datum used by the Global Positioning System. The flattening originally used by the DOD differed from GRS 80 in the sixth decimal place. It has been asserted that the difference was due to the number of significant digits used by the DOD in the computation. This discrepancy was resolved with WGS 84(G730).

**NORTH AMERICAN DATUM (NAD)**

The first continental North American Datum (NAD27) was produced by the Coast and Geodetic Survey using geodetic terrestrial and astronomic observations. The terrestrial observations formed chains of triangles in east-west and north-south directions across the North American continent. The least squares adjustment of the geodetic and astronomic data began in 1927 and was completed in the United States in 1932. Canada and Mexico eventually incorporated their networks into NAD27, making it a true “North American” datum. NAD27 is a horizontal datum defined using the Clarke Spheroid of 1866 and a terrestrial origin at Meades Ranch, the geographic center of the United States. The datum fits the North American continent well, but quickly diverges outside the continental limits. The accuracy of existing geodetic positions were improved by readjusting the triangulation of the entire network using Laplace azimaths, although the mix of loop sizes (ranging in length from a few hundred kilometers to 3000 km) and insufficient number and spacing of Laplace Corrections resulted in balancing errors and accumulated errors over long chains of triangles. NAD27 had a network relative accuracy of directly connected triangulation stations of 1:100,000 and absolute network accuracy (unconnected triangulation stations) of 10 meters.
Prior to about 1970, all geodetic instruments were mechanical and angular measurements were far more accurate than the limited set of distance measurements. By 1970, electronically measured distances were possible, enabling geodetic surveyors to obtain higher accuracy and increase the density of control points. When connections were made between the original chains of triangulation, large distortions were discovered. This problem was compounded by the introduction of the TRANSIT satellite positioning system, which provided absolute positioning capabilities of one meter. As discussed above, astronomic measurements had to be incorporated into a geodetic network to obtain an accurate knowledge of the deflection of the vertical, which was a significant limiting factor. Absolute positioning by satellites is unaffected by the deflection angles if the same Earth-centered ellipsoid is used for the geodetic network. Therefore, satellite positioning was the motivating force to move away from a national datum to an Earth centered datum. Although TRANSIT positions were accurate to only one meter at best, they uncovered scale errors in the network introduced before electronic distant measuring technology. To overcome these inherent weaknesses in NAD27, the North American Datum of 1983 (NAD83) was adopted.

In name and datum definition NAD83 is still used today. The greatest difference between NAD27 and NAD83 is that NAD83 is based on an Earth-centered Earth-fixed ellipsoid having a three dimensional right-handed orthogonal coordinate system. The NAD83 (and WGS84) datum has eight internationally agreed upon parameters (GRS80). The National Geodetic Survey performed a simultaneous adjustment of approximately 1.786 million observations, obtained at approximately 266,435 triangulation stations covering the United States, to determine the origin. The relative accuracy of directly connected triangulation stations remained 1:100,000, but the absolute accuracy improved to one meter. The new North American Datum was named NAD83 (86) because the final coordinates from the adjustment were not available until 1986. The readjustment was the largest simultaneous solution of equations at the time and was recorded in the Guinness Book of World Records. Today, this same solution could be done on a laptop computer. NAD83 is realized or made operational by the coordinates of 266,435 geodetic points on the surface of the Earth unlike WGS 84 where the realization is accomplished by 12 GPS tracking stations situated around the world.

NAD83 (86) lasted for only seven years. During these years the Global Positioning System (GPS) evolved into the tool of choice for geodesy. Techniques involving static observations in a relative position mode with post-processing were developed. Centimeter-level horizontal positioning became possible over distances of hundreds of kilometers. At the same time, however, the DOD encrypted the coded messages being broadcast by the GPS satellites, denying high accuracy positioning. Geodesists found a work-around methodology that used only a smooth carrier wave signal broadcasted by the satellites which was unaffected by encryption. One only had to develop a technique to figure out how many whole wavelengths existed between the various GPS satellites and the ground receiver. Now considering that the satellite makes a complete orbit every twelve sidereal hours and that the radius of its orbit is approximately 25,000 km, the satellite must travel at a velocity of 13,090 km/hour or approximately 2.3 miles/sec. It is no trivial matter to determine the exact number of 19 cm long wavelengths for each satellite. Work by geodesists at the National Geodetic Survey not only solved this problem, but also ushered in the possibility of obtaining centimeter level accuracy in real time.
NAD83 (86) was replaced by the High Accuracy Reference Network (HARN), a subset of the U.S. geodetic network. Each point in the sub network (19,000 stations) was occupied by a GPS geodetic quality receiver. The NGS used this information to obtain a better scale and rotation for NAD83. As a result of this adjustment, directly connected points had a relative accuracy of 1:1,000,000 and the absolute accuracy of 0.1 m. Even though ellipsoid heights were computed for this adjustment, GPS surveying techniques were not sufficiently sophisticated to yield high-accuracy vertical information.

The HARN lasted only seven years. In that interval, GNSS technology improved significantly. In 1994, the NGS began an effort to operate about 10 GPS reference stations on a 24/7 basis. These Continuously Operating Base Stations (CORS) were intended to provide archived data for users to download and use in post processing their GPS data. If there were a sufficient number of CORS nationwide, then surveys using the CORS data would automatically be tied to NAD83. NGS would use the CORS data to improve the accuracy of the CORS network, monitor crustal motion, and determine precise ephemeris of GPS satellites within 48 hours of data collection. The NGS now has over 1500 CORS sites and adds approximately 15 new sites per month. Geodesists essentially created their own network of GNSS tracking stations that enabled them to compute highly accurate positions.

The CORS network replaced the HARN in 2001. Today CORS coordinates are designated NAD83 (CORS96) POSITION (EPOCH 2002.0). The network and local accuracy are the same (0.01 m). The IRTF uses CORS stations around the world to account for crustal motion velocities. Every CORS and every geodetic control point has a velocity associated with its position. The velocity has an initial date/time associated with the velocity and the user can update the position by multiplying the velocity component by the time elapsed since the initial epoch date.

**NAD83 – WGS 84 Differences**

The Defense Mapping Agency (DMA) used TRANSIT observations at its worldwide satellite tracking stations in its least squares adjustment to best determine WGS 84. NAD 83 used millions of classical observations and limited TRANSIT observations across North America in its least squares adjustment to best determine NAD83. NAD83 and WGS84 use the same GRS80 definition but, due to different adjusted data, their locations for the center of mass of the Earth differ by approximately two meters. The difference in coordinates is noticeable only when an absolute and relative GPS position is obtained for the same point. An absolute position determined by GPS is limited to the code range accuracy of GPS and the satellite ephemeris data. A rule of thumb is that a coded range (ala GPS) is accurate to one percent of the length of an individual code chip (Van Sickle, *GPS for Land Surveyors*). That is 0.3 m for GPS Precise Code and 3 m for the Coarse Acquisition Code. Therefore, a position based on absolute positioning will seldom be more accurate than two meters, which corresponds to the uncertainty of the origin.

A relative position using GPS requires a differential solution involving satellite data collected simultaneously at a known point (either WGS 84 or NAD83) and an unknown point. A relative position using a known point on NAD83 results in NAD 83 coordinates for the unknown point.
The difference between a point positioned absolutely by GPS (relative to the center of mass defined by WGS 84) and one positioned relatively by GPS using NAD 83 (relative to the center of mass defined by NAD83) can be greater than three meters. Nearly all mapping and geodesy functions using GPS rely on relative positioning techniques. In 1995, the United States published a statement in the Federal Register proclaiming that NAD83 and WGS84 are functionally equivalent for national mapping and charting at scales larger than 1:5,000.\textsuperscript{4} This statement was safe, given that the difference between a WGS 84 and NAD83 derived position is insignificant compared with the accuracy of national maps and aeronautical and nautical charts at scales smaller than 1:5,000.

**GEODETIC INSTRUMENTATION AND TECHNIQUES**

**Traditional Surveying Techniques**

Prior to GPS, most horizontal angle observations were performed using a theodolite. This instrument had an accurately graduated horizontal circle, and the angular measurement was read directly by the observer. The instruments had a horizontal angle measurement accuracy of 0.2 arc seconds (1/6,480,000 part of a circle), which is equivalent to the width of a paper match at 6 miles. A network of known and unknown points would be designed to enable various combinations of adjacent points to be connected to form triangles; a process known as triangulation. The connected points had to be intervisible to enable the interior angles of the triangles to be measured with a theodolite. The curvature of the Earth and obstruction by objects limited the maximum distance between points. To extend the distance, the USC&GS built observing towers as high as 116 feet above the point. The field check for the accuracy of the angular measurements was $180^\circ + \text{spherical excess} \pm 3''$ (spherical triangles have more than $180^\circ$ determined as a function of the area of the triangle). Distance measurements were either taped on the ground or determined electronically using a microwave signal, white light or laser beam transmitted by the instrument and reflected off a prism on the other end of the line being measured.

Most height measurements are determined with a level, which instruments a line of sight that can be made perpendicular to the local gravity vector. Readings are then observed off two rods (calibrated rulers, usually two or more meters long). Measuring the change in elevation between two points required beginning with one rod being set vertically over a known elevation point (benchmark) and the placing the second rod vertically on a temporary point along the path to the point whose elevation is to be determined. Once the “back sight” on the rod located at the benchmark and the “foresight” on the temporary point were recorded, the difference between the readings was computed (back sight minus foresight). The rod on the benchmark was moved forward to a second temporary point and the process was repeated until a foresight was finally observed on the point whose elevation was to be determined. The sum of the back sights minus foresights was the change in elevation between the starting and ending points. The tolerance for the most accurate leveling performed by the USC&GS is 4 mm times the square root of the distance leveled between known benchmarks in kilometers (roughly 5mm/mile). The process for making both horizontal and height measurements is slow, labor intensive, and, as a result, expensive.

The introduction of electronics and microchips enabled horizontal and vertical measurements to be merged into a single instrument called a Total Station. Advances in the technology have enabled pointing and reading to be performed automatically as the telescope follows the reflector from point to point. The observer moves the reflector to a desired location and the Total Station follows the reflector to that location. This technology eliminates the need for the “rodmen.” The most precise vertical measurements still employ a level, but the readings are automated by use of bar coded level rods. GNSS technology has all but replaced the need for triangulation techniques and has eliminated many of the labor-intensive procedures for determining a point position. The classical horizontal positioning and leveling process are rapidly being replaced by GNSS technology, which drastically reduces the distance between points, time, labor, and expense.

The coordinates for a given control point are represented on the ground by a survey monument, most commonly a brass or aluminum disk set in concrete or bedrock. The stability and access to the survey monument is of prime importance. Many of the early triangulation stations were established on mountain tops to extend the line of sight above the horizon. These stations are not very accessible and their usefulness has diminished. With modern GNSS techniques, more effort is expended in finding an accessible location for the control point and as a result, many survey monuments are established along public right-of-ways. A comprehensive history of survey monumentation is given in Bottles, Pots, & Pans? – Marking the Surveys of the U.S. Coast & Geodetic Survey and NOAA (Leigh).

Early Satellite Technology

The U.S. Navy developed one of the first operational satellite systems known as the Navy Navigational Satellite System (NNSS) eventually renamed TRANSIT. Its purpose was to provide a worldwide positioning system and timing network for nuclear missile submarines. It was accurate to about 200 feet if the vessel’s velocity was accurately known. A 0.5 knot error in velocity could decrease positional accuracy to 600 feet. The orbit and range information was only adequate for providing a two-dimensional position; however, since the submarines were on the surface when collecting TRANSIT data, their elevation could be assumed to be zero. A satellite pass could last up to 20 minutes and during that time the vessel’s direction and speed had to remain constant.

In 1967, NNSS was declassified and became widely used by civilians for geodetic surveying. The system was comprised of six satellites weighing approximately 120 pounds, in low Earth orbit (1100 km) with a 12-hour period, distributed in space so that one satellite could be generally be tracked every 90 minutes. The TRANSIT satellites transmitted orbital information at two steady frequencies (150 and 400 MHz), which were Doppler shifted due to the high relative motion between the satellite and terrestrial receiver. TRANSIT transmitted on two frequencies to enable the removal of the effect of ionospheric refraction from the range information. The amount of Doppler shift was measured at the receiver, which compared the incoming shifted signal against accurate oscillators contained in the receiver. By integrating the Doppler shift over a two minute period, the range to the satellite could be computed. The position was derived from one satellite.
A commercial version of the TRANSIT receiver, the Magnavox Geoceiver, was introduced in 1971 and used extensively by NIMA and NGS to determine positions. The Geoceiver was the first man-portable satellite tracking station suitable for accurate mapping in relatively remote areas. Each “observation” required multiple satellite passes collected over a two- to five-day period (20 to 50 satellite passes) on a cassette tape which had to be changed every two to three satellite passes (many of these sites were in remote areas so the observer had to camp out). Tests showed that the Geoceiver could determine positions with an accuracy of better than 10 cm. The accuracy of the measured Doppler shift was limited to the stability of the receiver oscillator. Due to the satellite’s low Earth orbit, the orbit trajectory was susceptible to atmospheric drag and gravity anomalies that made accurate orbit determination difficult, and therefore, limited its use for high-accuracy geodetic work. Each Geoceiver unit sold for about $50,000. The system was eventually turned off in 1996.

“TRANSITing” to GPS to GNSS

TRANSIT was eventually replaced by the Navigation Satellite Timing and Ranging System (NAVSTAR), better known as the Global Positioning System (GPS). GPS is based on radio navigation systems, such as LORAN, where ranges are measured by determining the difference in time between transmission and reception of a coded message. GPS combined two design concepts, one from the Air Force (621B) and the other from the Navy (Timation). Both concepts would improve the global accuracy of positioning and time dissemination, but the Air Force requirement included three dimensional positioning in a highly dynamic situation.

The Air Force work began in 1963 with a study on the military potential for new technology particularly associated with rocket boosters and satellites. An early focus on space applications quickly developed into precise positioning of aircraft. The objective was to obtain 15-meter horizontal accuracy for aircraft with a variable course and speed. Tests to extend the navigation system to seven satellites in equatorial orbits began shortly after. The orbits would be known and the aircraft would carry a clock synchronized to the satellites prior to takeoff. Four simultaneous satellite ranges could be measured, solving for the unknown receiver clock bias and a three-dimensional position. A pseudo-random code was developed so that all satellites could broadcast on the same frequency and enable the terrestrial receiver to differentiate between the various satellite signals. Concept 621B eventually included a 20 satellite system in four oval orbits extending up to 30 degrees north and south of the equator. Meanwhile, the Naval Research Laboratory (NRL) continued investigating new technology for improved timing and navigation. The Navy concept, named Timation (for timing and navigation), used atomic time standards onboard 21 to 27 satellites in eight hour orbits inclined 55 degrees to the equator. Testing and evaluation of 621B and Timation continued through 1971.

In 1973, a system incorporating the best features from both concepts was designed. Both concepts used three-dimensional positioning in real time, and the final design used the signal structure and frequencies from the 621B project and similar satellite orbits to the Timation concept but at an altitude having a twelve-hour orbital period. This compromise became NAVSTAR GPS as we know it today. GPS is composed of 24 satellites, the five DOD and seven NIMA tracking stations mentioned above, and a command center. The satellites are monitored constantly for health and orbit determination. Every three hours, updated orbit
information (satellite ephemeris data) is transmitted to the satellites. The civilian community is the largest group of users.

The GPS requires a minimum of 24 satellites for worldwide coverage, with a minimum of four satellites being observed simultaneously to obtain a three-dimensional position. The satellites are in a nearly circular orbit at an altitude of 20,000 km above the Earth with a 12-hour (sidereal time) period. The GPS differs from TRANSIT by relying on accurate timing to determine the satellite range. The time of signal reception minus the time of signal transmission multiplied by the speed of light determines the range between satellite and receiver. GPS uses two frequencies for accurate positioning, L1 and L2, which are broadcast on 1575 and 1227MHz. From a geodetic prospective, it became practical (i.e., without radio telescopes) for the first time to measure three-dimensional transoceanic and continent-wide vectors at decimeter level accuracy. The GPS further reinforced the need for an international datum complete with crustal motion velocities to remove as much of the temporal crustal movement as possible. It was soon discovered that precise orbital information would be necessary as well.

The GPS constellation has been replenished and modernized since the Block I satellites were first launched in 1978. The original satellites were placed into orbital planes to provide optimal coverage for test and evaluation over the White Sands Proving Grounds. They carried two atomic and one rubidium time standards. Momentum problems required firing onboard rockets to keep the satellites pointed at Earth and prevent them from tumbling out of control. The Block I satellites lived well beyond their expected lifetime of 5 years, with the last Block I satellite being turned off in 1989. There was no encryption to the L1 and L2 codes, so full accuracy was possible for all users. Ten-meter accuracy for aircraft was achieved and the geodetic community was envisioning centimeter accuracy with post-processed data.

Block II satellites followed from 1989 to 1997. These satellites were radiation hardened, the momentum problem had been corrected, and improvements would allow the satellites to function for up to six hours without any intervention from the ground control. The coded signals were also degraded or encrypted. The code was degraded by dithering (slowing down or speeding up) the clock and truncating the orbit parameters. The degradation was known as selective availability (SA). The satellites also contained anti-spoofing (AS) capability, which meant that the satellite signals could not be corrupted by an external/unfriendly signal. The DOD guaranteed 100-meter absolute accuracy with the degraded messages. Military receivers were able to use the full absolute accuracy potential of 10 meters.

Geodesists soon found that centimeter accuracy with post processing could be obtained using the pure carrier wave signal instead of the encrypted code messages (see “North American Datum” above). However, SA remained an issue and a point of irritation between DOD and civilian users until it was finally turned off in 2000. A factor in the decision was an incident that took place on April 20, 1978. Korean Airlines flight 707 from Paris to Seoul strayed into Soviet airspace and was forced to make an emergency landing on a frozen lake after being damaged by a Russian missile. Two died and 14 were injured in the accident. This accident might have been avoided had more accurate GPS navigation been available.
Block IIR-M satellites have a new second coded civil frequency (L2C) and two new military frequencies (L1M and L2M), which will eliminate the military requirement for Selective Availability in the future. The message format in L2C contains more accurate information about the satellite orbit and clock parameters necessary for improved positional accuracy. This will give civil users two unencrypted signals to help solve for ionospheric refraction errors. The final satellites in the Block II family (Block IIF) have an additional civilian frequency (L5), which is designed for transportation safety-of-life issues. The signal will be broadcasted on the Aeronautical Radio Navigation Service band. The full 24-satellite capability will be achieved by 2018.

Block IIIA satellites will follow in 2014 and will have a fourth civilian frequency (L1C). These satellites will be backward compatible and will have increased accuracy and security with improved system survivability. The Block IIIA constellation of 24 satellites will be implemented by 2021.

**Global Orbiting Navigation Satellite System (GLONASS)**

GLONASS is the Russian version of GPS. The first satellites were launched in 1982 and a full constellation of 24 satellites was achieved by 1996. The number of satellites declined to 7 by 2001 because of collapse of the Soviet Union and the control of the Baikonur Cosmodrome for launching new GLONASS satellites by a newly independent Kazakhstan. At the end of 2009, 21 satellites were in orbit.

GLONASS has two versions: Uragan K and Uragan M. The Uragan M, which was launched beginning in 2003, has a 7 year life expectancy and broadcasts separate frequencies for civilian use. The Uragan K will have a life expectancy of 12 to 15 years and will be launched by the end of 2009. Traditionally, 3 new satellites are launched every Christmas. Unlike GPS, which uses a pseudo-random code generation so all satellites share the same frequency, the GLONASS system uses 12 satellite frequencies such that two satellites share the same frequency on diametrically opposite sides of the Earth so they do not interfere with one another. GLONASS has the advantage of higher real time autonomous position accuracy because the precise coding was no longer encrypted after 2004.

The integration of GPS and GLONASS into a single ground receiver is difficult due to time keeping differences. GPS time is based on the internationally accepted UTC but the transmitted time does not account for the leap second adjustment to UTC. The GLONASS system uses its own version of UTC regulated by its own time keeping service and not only accounts for the leap seconds but also makes periodic adjustments to its time keeping.

**GALILEO**

GALILEO is the European version of GPS, designed with a for profit business model to pay for the navigation satellite system. The first Galileo satellite, GIOVE-A (Italian for Jupiter), was launched in November 2005. GIOVE-B is also in orbit. The satellites are used for orbit validation experiments and to reserve the international radio spectrum necessary for a future full implementation of the system. GALILEO will ultimately have five levels of service. Open
Service will offer free access to basic signals for general navigation uses. Safety-of-Life Service is similar, but guarantees a higher level of integrity monitoring. Public Regulated Service is encrypted and is intended to be used by public security and civil authorities. Search and Rescue Service is intended to enhance other international space search and rescue services. Finally, Commercial Service will offer custom services for unique applications and is expected to generate the largest source of revenue.

**BEIDOU/COMPASS**

COMPASS is the Chinese navigation satellite system. Three BEIDOU geostationary satellites are on orbit and provide some level of navigational accuracy over China. Eventually the system will include up to 35 satellites at various altitudes to cover all of China and its neighboring countries. COMPASS has a business model similar to GALILEO in that a standard positioning service will be available to the general public and regulated (encrypted) service will be available for profit. There is a possibility that China will cooperate with GALILEO to develop a single system.

**The Quasi-Zenith System (QZSS)**

The QZSS system is composed of 3 geostationary satellites at high elevation angles over Japan. QZSS broadcasts GPS-like signals. The constellation will augment GPS, particularly in areas with significant signal obstructions such as canyons or tall buildings.

**How is GNSS Used?**

The geometry (Dilution of Precision [DOP]) of the visible satellites at the receiver’s location is key to obtaining a reliable position. The DOP is measured by a unitless number, with smaller numbers representing greater geometric strength. The DOP is computed using the first derivative of the range equations between satellites and receiver. One can imagine the volume of air contained within the shape formed by the satellites in space and the receiver on or near the ground. As an example, the shape produced by four satellites and one ground receiver might resemble an inverted pyramid. The more volume the pyramid contains the better the geometry. If the four satellites are nearly in line with one another and the receiver, the geometric shape would have very little volume and the resulting solution would be poor or impossible to compute. For most mapping and surveying purposes, a DOP of 4 or less is sufficient.

GNSS can be used for absolute, relative or incremental positioning. As stated earlier, an absolute position relies strictly on the data obtained from the satellites and has limited use for surveying or mapping. Relative positioning relies on differential solution involving some computed corrections or receivers used in pairs with one receiver or a known geodetic location. The U.S. Coast Guard operates a network of GPS receivers covering the United States and its territories. Each of the receivers is located over a geodetic control point established by the NOAA National Geodetic Survey. The positions computed from the satellite data at these locations are compared to the known coordinates and correctors are computed and broadcasted by the Coast Guard. The correction takes the form of range and range rate corrections. The user can receive these corrections via a signal broadcasted by the Coast Guard. The range rate is
multiplied by the number of seconds that has elapsed since the correctors were computed and applied to the range correction. The final range corrections are applied directly to the incoming range data at the user’s location. A user equipped with a GPS receiver that can receive the correctors in real time can achieve a positional accuracy of 1 meter or less. Accuracy of decimeters can be achieved if the DOP is sufficiently low. This type of relative positioning is frequently used for mapping applications. Mapping grade accuracy can also be obtained by post-processing using data from the National Geodetic Surveys CORS.

Relative positioning can also be performed by occupying known and unknown points simultaneously and collecting up to one hour of simultaneous data. The data are processed using a method known as double differencing, which mathematically removes the systematic errors of related to the satellite orbit, clock and atmosphere. Accuracies of a few centimeters are possible with this technique.

Incremental positioning uses a network of known and unknown points connected by a series of relative observations. The relative vectors between points are mathematically fit by least squares computations to improve the overall accuracy. Horizontal accuracy of 1 cm and vertical (ellipsoid) accuracy of 2 cm is possible when following the NOAA National Geodetic Survey’s guidelines for National Height Modernization. Accuracies of centimeters require the use of GNSS receivers that can take advantage of both L1 and L2 frequencies to remove the effects of ionospheric refraction.

More recently, surveyors and engineers have demanded real time centimeter accuracy for applications ranging from precision agriculture to machine-controlled road construction. The initial solution was to use Real Time Kinematic (RTK) surveying, in which one GNSS receiver was positioned over a known geodetic control point (the base station) and transmitted its raw satellite data via a VHF radio modem to a second receiver located over an unknown point (the rover). Double differencing was used to compute the vector solution between base and rover at the location of the rover in real time. This technique required considerable equipment and had a useful range of less than 5 miles from the base station. With the introduction of cellular digital technology, cellular modems could be used and the base and rover could be spaced approximately 10 miles apart. The distance between base and rover was now limited by the ability to remove ionospheric refraction. This approach was better, but it still required considerable equipment and a worker for security at the base location. Another limitation was the lack of redundancy or confidence in the position based on a single vector solution between the base and rover. The accuracy of rover coordinates was entirely dependent on the accuracy of the base station coordinates.

RTK has since matured into the technology of Real Time Networks (RTN). Like the CORS only with closer spacing of reference stations, centimeter level positions can be obtained anywhere within the RTN where cellular coverage is available. A single dual frequency receiver cell phone with CDMA capability is all the equipment that is necessary. RTNs are usually established by state governments, although some private networks exist. Nearly all RTNs use both GPS and GLONASS to increase the number of satellites available at any given moment.
1. Introduction

In this White Paper, I attempt to provide a snapshot of the state of the art in geospatial analysis, particularly as it pertains to the exploratory spatial analysis of crime events. Admittedly, this is only a sub-domain within geospatial analysis, but it is pertinent to the analysis of intelligence pertaining to security and violent conflicts. The bulk of the White Paper is an excerpt from a recent Report prepared for the U.S. Army Corps of Engineers, Engineer Research and Development Center, which dealt with the development of a methodological template for spatial decision support systems in support of tactical operations (Anselin et al. 2009). This, in turn, was derived from earlier overviews of the methodology behind spatial analysis of crime in Anselin et al. (2000, 2008) and Messner and Anselin (2004).

Spatial analysis (now often referred to as “geo”-spatial analysis) is broadly defined as a “set of methods useful when the data are spatial” (Goodchild and Longley 1999). More specifically, it encompasses a collection of techniques to add value to data contained in a geographic information system. As such, spatial analysis forms an important component of the evolving discipline of “Geographic Information Science” (Goodchild 1992). It encompasses many different methodologies that share the common characteristic that the result of an analysis is affected by the location of or distance between observations. The recent text by de Smith et al (2007) includes four main areas in addition to the basic analytic manipulations embedded in most GIS: data exploration and spatial statistics, surface and field analysis, network and location analysis, and geocomputational methods and modeling. In this paper, attention is focused on data exploration and spatial statistics.

An important reason for the growth of geospatial analysis over the past twenty-some years was the realization in the late 1980s that the technology of geographic information systems (and especially desktop systems) provided an excellent opportunity to operationalize and take advantage of the wealth of analytical techniques developed in the quantitative geography literature. In addition, the combination of the computing power in the GIS and advanced methods of spatial statistical analysis provided the opportunity to develop integrated systems that contributed not only to practice, but also led to scientific advances and new methods. Early discussions of the integration of spatial analytical methods with GIS can be found in Goodchild (1987), Goodchild et al (1992), and Anselin and Getis (1992), among others, and in Goodchild (2010) for a recent perspective. An important institutional factor was the establishment in the U.S. of the National Center for Geographic Information and Analysis (Abler 1987), which, through funding from the National Science Foundation provided a major impetus for the development and adoption of spatial analytical methodology. A similar role was played about ten years later by the NSF funded Center for Spatially Integrated Social Science (Goodchild et al 2000).
Early compilations of methods, applications and software tools for geospatial analysis can be found in Fotheringham and Rogerson (1994), and Fischer and Getis (1997), among others. More recent reviews include Fotheringham and Rogerson (2009), Anselin and Rey (2010), and Fischer and Getis (2010). Extensive technical detail can be found in those references. By design, the current paper is limited to a broad overview with a focus limited to exploratory data analysis.

From a methodological perspective, the study of the location of violent events, for example, as associated with an insurgency is a special case of “point pattern analysis.” Interest focuses on the extent to which such events cluster in space and on the locations where those clusters (or “hot spots”) may be found. Increasingly, this also includes attempts at explaining why the clusters are where they are as a function of covariates (explanatory variables) that can be readily measured. Point pattern analysis has seen extensive application in ecology, epidemiology as well as in crime analysis (a classic technical reference is Diggle 2003, a more introductory treatment and extensive references can be found in Waller and Gotway 2004). Such analyses of point events (or their aggregates by areal units) can be readily extended to applications in the context of military conflicts, such as IED attacks (e.g., McFate 2005, Riese 2006).

The remainder of the paper consists of a review of exploratory spatial data analysis, with particular reference to spatial crime analysis applications. The discussion is kept at a non-technical level, with references provided to the methodological literature for technical details and to specific applications in crime analysis for illustrations. The paper closes with some brief concluding remarks.

2. **Exploratory Spatial Data Analysis of Crime**

Arguably the first stage in a spatial analysis of crime is the exploratory stage. Exploratory data analysis (EDA) is a branch of statistics started by John Tukey (Tukey 1977), and stresses an inductive approach. As spelled out by the statistician I.J. Good (1983), it is a collection of techniques used to discover potentially explicable patterns. The emphasis is on discovery of interesting patterns, which may be amenable to explanation, but the explanation itself is not part of EDA. EDA consists of many different graphical devices, such as charts, tables, graphs and maps. These are referred to as views of the data, facilitating interactive discovery through a combination of graphical representations and summaries (Buja et al. 1996).

Exploratory Spatial Data Analysis (ESDA) is a superset of EDA that is focused on the spatial aspects of the data (Anselin 1999). This includes describing spatial distributions, identifying atypical spatial observations (spatial outliers, as distinct from regular outliers), discovering patterns of spatial association (spatial autocorrelation) and suggesting spatial regimes (spatial heterogeneity) (e.g., MacEachren and Kraak 1997, Andrienko and Andrienko 2005).

The techniques reviewed in this section are organized into four groups, starting with general crime mapping and geovisualization and moving on to traditional point pattern analysis, hot spot detection and space-time exploration.

---

5 This section is a slightly edited version of Section 4 in Anselin et al (2009).
2.1. Crime Mapping

In the late 1990s and early 21st century, the use of computerized crime mapping saw an explosive growth, reflected in several books and edited volumes devoted to the topic. Early examples include Block et al. (1995), Eck and Weisburd (1995), Weisburd and McEwen (1997), LaVigne and Wartell (1998), and Harries (1999). Increasingly, the traditional mapping (choropleth maps) and basic spatial analysis operations (buffering, distance measures) are viewed as integral parts of a geographic information system and extended with more sophisticated (statistical) techniques to identify hot spots, highlight outliers, and suggest patterns, as argued in Anselin et al. (2000). Extensive illustrations can be found in Block (2000), Goldsmith et al. (2000), LaVigne and Wartell (2000), Hirschfield and Bowers (2001), Leipnik and Albert (2003), Boba (2005), Chainey and Ratcliffe (2005), Eck et al. (2005), Wang (2005b) and Ratcliffe (2006).

Basic GIS use and computerized maps have become so standard in crime analysis that they will not be elaborated upon here. Some specialized maps warrant a brief mention, however. For example, in Poulsen and Kennedy (2004), so-called dasymetric maps are used to depict the spatial distribution of burglaries in an urban area. These maps use additional GIS layers (such as housing units and land use) as a filter to constrain the area of administrative areal units to reflect more realistic locations for the crimes. In other words, these maps provide a compromise between assigning the same rate to the full administrative unit (the standard approach in choropleth mapping) and depicting the individual point locations. This is especially useful when the latter are not available and it avoids the potentially misleading effect of the area and arbitrary boundaries of administrative units typical of choropleth maps. Additional statistical maps can be used to avoid this problem, such as cartograms, animation and conditional maps (see, e.g., Monmonier 1989, Takatsu and Gahegan 2002, Anselin et al. 2006). Also, specialized outlier maps can be employed to highlight locations with unusually high values (Anselin 1999, Anselin et al 2004), or to identify sharp gradients in crime rates, i.e., so-called spatial outliers. For example, in Harries (2006) neighborhoods (census block groups) are identified where high quintile values are adjacent to low quintile values, suggesting an extreme crime gradient.

Additional methods, where the GIS and mapping are combined with pattern analysis, hot spot detection and surveillance are treated in the next sections.

2.2. Pattern Analysis

In this paper, pattern analysis is used to designate traditional descriptive and exploratory methods of statistical spatial point analysis as well as data mining techniques that have evolved from the computer science literature. It is distinguished from hot spot detection (section 2.3), where the focus is on the use of indicators for spatial autocorrelation and clustering methods to identify regions of elevated crime incidence or risk.

2.2.1. Statistical Analysis of Point Pattern

Descriptive statistics for point patterns include mean and median location and the standard deviational ellipse, which give an indication of the central tendency in the spatial distribution of
the points and the spread and orientation of points around this center. For example, these methods have been implemented in the widely adopted CrimeStat software package (Levine 2006, 2007) as well as in a number of other software tools and have been used in many applications. For example, LeBeau (1987) applied this technique to track the changes in the spatial pattern of rapes. These methods can also be readily incorporated into a GIS system in support of policing actions (e.g., to track the spatial dynamics of 911 calls).

A more refined technique to describe the spatial distribution of points is kernel smoothing, which creates a smooth surface representing the density of the points. In essence, this is a weighted moving average of the count of points within a circle of a given bandwidth, where the weights are given by the chosen kernel function (for detailed illustration, see, e.g., Levine 2007). Some examples of the application to spatial crime analysis are Steenberghen et al (2004) who use it to describe the distribution of road accidents, and Corcoran et al. (2007) who include it into their review of spatial analytical methods applied to the study of fires.

Perhaps the most commonly used statistic to assess the absence of complete spatial randomness in a point pattern is Ripley’s (1976) K function. The K function focuses on so-called second order properties of a point pattern, which are similar to the notion of a covariance. The first order characteristic is simply the intensity of the process, or, the average number of points per unit area, for example, as summarized in a kernel density function. The second order characteristic is then some measure of covariance between intensities at different locations. More precisely, the K function is the ratio of the expected number of additional events within a given distance from an arbitrary event to the intensity of the process. It is readily calculated by counting the number of points within an increasing radius from each event in the pattern. It is typically computed for a number of distance ranges and plotted against distance. It is included as a function in the CrimeStat software and has seen many applications (see also Anselin et al. 2008).

While the K function focuses on the overall patterning of points (“clustering”), interest often centers on specific locations of “clusters.” As such, the K function is not able to provide this information. An extension of the notion of local indicators of spatial association (Anselin 1995) to identify local clusters by means of the differential of the K function, the so-called product density function, is advanced in Cressie and Collins (2001a, b). A slightly different approach was recently presented in Mateu et al (2007).

One limitation of the K function as traditionally applied is that it is best suited for a situation of an isotropic plane, in which an event can be located anywhere. However, in practice, there are often limitations to the possible locations. For example, when events occur on a street network, the space in between the network links and nodes becomes impossible as a location. Recent work by Okabe and co-workers has extended the K function to events on a network, using shortest path distances on the network instead of the traditional omni-directional “as the crow flies” distance. The basic methodology was established in a series of papers by Okabe et al. (1995), Okabe and Kitamura (1996), Okabe and Yamada (2001), and Okabe and Satoh (2005), and it has been implemented in the SANET toolbox for network spatial analysis (Okabe et al. 2006a, b).

The network K function has seen applications in a number of areas, such as the location of acacia plants (Spooner et al 2004) and accidents on a road network (Yamada and Thill 2004; e.g.,
contrast with a traditional K function analysis of traffic accidents in Jones et al 1996). Yamada and Thill (2004) also carry out a comparison of the results of the traditional (planar) K analysis with the network K function. Similarly, in Lu and Chen (2007), the results of a planar and network K are compared for urban crime on a street network. The planar K tends to result in false positives for a less dense street network and low crime density, in contrast, dense street and dense crime lead to more false negatives. In other words, the performance of the network K function relative to the planar K is related to the structure of the street network and the density of point events. Further work is needed to establish the degree of generality of the findings in this case study.

2.2.2. Data Mining

Parallel to the attention paid to pattern recognition from a statistical viewpoint, developments in computer science have yielded methods of machine learning and knowledge discovery that are designed to recognize patterns in multivariate data sets. In crime analysis, this begins with automatic information extraction from various records and incident reports and the application of machine learning (such as text mining) and rule based expert systems to ultimately yield an operational decision support system. A recent overview of the application of data and text mining in crime analysis, with an emphasis on risk and threat assessment, and the use of predictive analytics to obtain operationally actionable output is given by McCue (2007). Discussions of different approaches can also be found in Brown and Hagen (2003), Chen et al (2003) and Yang and Li (2007). Arguably the best-known system in operational use to date is the COPLINK system, adopted by police departments in the major metropolitan areas in the U.S. (e.g., Chung et al 2005).

2.3. Hot Spot Detection

Specialized techniques for the detection of hot spots follow a number of different logics. Three different categories are distinguished here: scan statistics, methods based on spatial autocorrelation statistics, and generic cluster detection techniques. They are briefly reviewed in turn.

2.3.1. Scan Statistics

So-called scan statistics consist of counting the number of events in a geometric shape (usually a circle) and comparing those to a reference pattern of spatial randomness. Early examples are the Geographical Analysis Machine (GAM) of Openshaw et al (1987), and the space-time analysis of crime (STAC) of Block (1995, 2000). Both of these methods consist of counting the number of points in a series of overlapping circles and labeling them as significant when the observed count is extreme relative to a reference distribution of simulated spatially random points. The STAC method is implemented in the CrimeStat software package, in which an identified cluster of points is represented by their standard deviational ellipse (see centrography in 2.2.1).

These early scan statistics suffer from the problem of multiple comparisons (overlapping circles) and are sensitive to parameter settings (radius of circle, etc.). The Kulldorff (1997, 1999) scan statistic and its later refinements address some of these concerns by using a likelihood criterion.
to identify clusters. In essence, the scan statistic considers circles of increasing radius and identifies that circle that maximizes the probability of having events inside the circle exceeding that outside the circle. Kulldorff’s scan statistic is implemented in the specialized SatScan software package (http://www.satscan.org). A recent generalization to the detection of arbitrarily shaped hotspots is the so-called upper level set (ULS) scan statistic (Patil and Taillie 2003) and its extension to bivariate data contexts in Modarres and Patil (2007).

An alternative extension is the augmentation of the likelihood idea of the scan statistic with an optimization procedure using simulated annealing to detect spatial clusters of arbitrary shape by Duczmal and Assuncao (2004). This is applied to the identification of clusters in the spatial distribution of homicides in Belo Horizonte, Brazil.

2.3.2. Spatial Autocorrelation Statistics

A second broad category of approaches bases the identification of clusters and spatial outliers on the results of a statistical test for spatial autocorrelation. These methods pertain to data that have been aggregated into areal units, such as administrative units or artificial grids, so called lattice data (contrasting with point patterns). For example, in spatial crime analysis, this often pertains to the count of events by spatial unit, or to a rate (the count of events divided by the population at risk).

A spatial autocorrelation statistic is a formal test of the match between value or attribute similarity and locational similarity. The statistic summarizes both aspects and is deemed to be significant if the probability (p-value) that the statistic would take this value in a spatially random pattern is extremely low. Measures of attribute similarity summarize the similarity (or dissimilarity) between the values observed at two locations. Three popular formal expressions for this are the cross product, as a measure of similarity, and the squared difference and absolute difference, as measures of dissimilarity. Locational similarity is formalized through a spatial weights matrix, which expresses the notion of neighbor. Spatial weights are not necessarily geographical, but can incorporate social network structures as well (for a classic treatment of spatial autocorrelation, see Cliff and Ord 1973, 1981).

Similar to the K function for point pattern analysis (section 2.2.1), a global spatial autocorrelation statistic (like Moran’s I or Geary’s c) is not appropriate for the identification of local clusters or hot spots. To that end, a local version of the statistics needs to be employed, a so-called Local Indicator of Spatial Association or LISA (Anselin 1995). Significant LISA statistics suggest locations where the value of the variable of interest is more grouped with that of its neighbors than likely under spatial randomness. Therefore, such locations become identified as local clusters, either hot spots (high values surrounded by other high values), or cold spots (low values surrounded by low values). Alternatively, in some instances spatial outliers may be identified by significant LISA statistics indicating negative local spatial autocorrelation, where low values are surrounded by high values, or vice versa.

A commonly used LISA statistic is the local Moran, a location-specific version of the familiar Moran’s I statistic for spatial autocorrelation (Anselin 1995). This has been applied to the identification of high homicide county clusters in Messner et al (1999), for example (for more
extensive overviews, see also Messner and Anselin 2004 and Anselin et al 2008). A related application is to the identification of so-called black zones, or road segments that exhibit an extreme number of vehicle accidents (for an early approach, see Black and Thomas 1998). Local Moran statistics are used to identify significant concentrations of high accident numbers in Flahaut et al (2003) and Steenberghen et al (2004) (see also Geurts et al. 2004, for an assessment of methods to identify and rank black zones). A related approach is the extension of the network K function and the LISA statistic to local indicators of network constrained clusters (LINCS) in Yamada and Thill (2007). This is also used to identify segments on a road network with elevated numbers of vehicle crashes.

A slightly different local statistic is the Gi (and Gi*) test developed by Getis and Ord (1992) (see also Ord and Getis 1995). Similar to the local Moran, this statistic identifies locations of local hot spots and local cold spots (but not spatial outliers). It has been applied to the study of burglaries in urban areas by Craglia et al (2000). Interestingly, in that study, the Gi statistic is compared to the more traditional STAC approach and found to be superior in identifying true clusters. Ratcliffe and McCullah (1999) use the Gi statistic in combination with a global moving window to distinguish between hotspots and hotbeds in residential burglary and motor vehicle crime. They suggest that some of the problems caused by the modifiable areal unit problem (MAUP) are avoided by changing the search area of the moving window.

Local spatial autocorrelation measures are included in the software GeoDa (Anselin et al 2006), STARS (Rey and Janikas 2006), CrimeStat (Levine 2006), the ArcGIS spatial statistics toolbox (Allen 2009), the open source R spdep package (Bivand et al 2008), as well as several others.

2.3.3. Generic Cluster Detection

A third category of methods to detect hot spots uses heuristic methods from the discipline of operations research to construct clusters of areas that are similar with respect to some characteristic. These techniques can be applied to individual points or to aggregate spatial units. Specifically, clusters are formed such that the similarity of the cluster members within the same cluster is greater than between clusters. Similarity can be based on distance or on a multivariate characterization (as in k-means clustering). Applications of these techniques to urban crime in Queensland are illustrated in Murray et al (2001) (see also Murray and Estivill-Castro 1998).

A recent article by Grubesic (2006) suggests that fuzzy clustering techniques may be superior in some respects relative to the standard hierarchical clustering techniques. Such fuzzy methods do not yield “hard” membership in each partition but a degree of fuzziness. This creates some challenges for the visualization of the results, e.g., by means of membership probability surfaces. Grubesic (2006) illustrates this with an application to crime events in a neighborhood in Cincinnati, Ohio. Related approaches are so-called contiguity-constrained clustering methods (Duque et al 2007), where it is guaranteed that the identified clusters consist of connected spatial units, which is not always the case when using standard clustering algorithms, such as the k-means clustering contained in CrimeStat.
2.4. **Space-Time Exploration**

Many techniques to explore patterns that occur both across space and over time are straightforward generalizations of pure cross-sectional methods. For example, the scan statistic (2.3.1) can be extended to identify space-time clusters, the local Moran statistic (2.3.2) can be applied to compare patterns of occurrence with that of neighbors at a different point in time, etc. The research question at hand is very similar to that employed in epidemiological studies of the spread of disease. In spatial crime analysis, the counterpart of this is the notion that the risk of a particular criminal event spreads over time to nearby locations. Space-time exploration has many commonalities with the surveillance and forecasting methods. I first consider the more exploratory techniques and then briefly review surveillance methods.

A commonly used procedure inspired by the statistical point pattern literature is the Knox test to identify space-time clusters (see Diggle 2003). For example, this was applied in a wide ranging comparison of space-time patterns in burglaries across 10 urban areas in Johnson et al (2007). A similar extension is the random point nearest neighbor technique of Ratcliffe (2005), which is applied to the change in the spatial distribution of burglaries in Canberra, Australia.

A related interest in the study of the “contagion” of crime risk is whether some type of displacement may occur, particularly due to a previous police intervention. This focus on displacement is the topic of a number of efforts, such as the so-called aoristic signatures of Ratcliffe (2000, 2002) and the weighted displacement quotient of Bowers and Johnson (2003). Aoristic signatures are a method to deal with the imprecision in the recorded time of the criminal event. A temporal weight is constructed to reflect the probability that an event occurred in a given period. These weights can be attached to the spatial locations of the events and yield different visualizations (e.g., the cylinders used in Ratcliffe 2000) and surface representations. The weighted displacement quotients uses a similar rationale as local space-time autocorrelation quotients in that changes in the crime rate in a buffer zone are examined around the original location of criminal events. This yields some sort of location quotient that incorporates a measure of change over time (see Bowers and Johnson 2003).

Other approaches consist of creative extensions of cartographic techniques to capture the spatial dynamics of criminal events. As reviewed in Brundson et al (2007) exploratory space-time visualization can be carried out by means of map animation, creative use of so-called comaps, isosurfaces and linked plots.

Griffiths and Chavez (2004) use an innovative combination of local spatial autocorrelation statistics (ESDA) with the trajectory method proposed by Nagin (1999) to study the space-time dynamics of crime in Chicago neighborhoods. Applying the trajectory method to the crime patterns over time for each neighborhood studied yields a grouping of neighborhoods by trajectory type. This is then examined by means of local spatial autocorrelation statistics to assess the extent to which neighborhoods with similar trajectories also cluster in space.

A similarly creative combination of techniques is the use of circular statistics to compare the dynamics of criminal events outlined in Brundson and Corcoran (2006). The circular statistics
(originally developed to analyze directional patterns) are adapted to assess and model geographical patterns in the daily cycles of events. Specifically, Brunsdon and Corcoran (2006) apply this to study criminal damage in the city of Cardiff, Wales and use a kernel smoothing technique to visually represent the distribution by time of day. This is then applied to a geographical comparison between the city center and the rest of the city.

Several of the techniques reviewed under the heading of space-time exploration (2.4) as well as regression analysis have been and could be implemented as part of surveillance systems aimed at detecting important changes in patterns over time. Such systems have a strong tradition in epidemiology and public health analysis, where they are used to detect the advent of a new epidemic or to identify an unusual outbreak of a disease. The ultimate goal of surveillance methods is to develop an automated decision support system that provides “alerts” when needed.

A number of point pattern techniques have been suggested specifically in the context of surveillance. For example, the spatial scan statistic of Kulldorff (2001) can be readily implemented to accomplish this. Also, Rogerson (2001) and Rogerson and Sun (2001) track the change over time in the spatial pattern of point events by combining a nearest neighbor statistic and a cumulative sum method. Porter and Brown (2007) suggest a method to detect the change in the distribution of point process by constructing an intensity function that depends on features (such as distance to landmarks) as a special case of marked point pattern analysis.

An alternative perspective represented by so-called surveillance methods is based on the time domain and uses forecasting techniques. This is more appropriate in allocating future crime fighting resources, for example, future deployment of police forces. In the context of a spatial analysis of crime, forecasting is relevant when a locational component is preserved. In order to have sufficient statistical validity, the spatial units of analysis will typically be fairly aggregate. In many instances, this precludes a meaningful spatial analysis.

A special issue of the *Journal of Forecasting* (Gorr and Harries 2003) considers a number of methodological issues pertaining to crime forecasting, such as the accuracy for small areas (Gorr et al 2003). A number of novel combinations of techniques are suggested as well, such as the use of “features” to model the transition density between patterns of events over time in Liu and Brown (2003), and the combination of cluster detection with an artificial neural network forecasting routine in Corcoran et al (2003).

One common characteristic of crime forecasting techniques is the need for considerable data points, both over time and across space, which is not required of the exploratory techniques.

3. **Concluding Remarks**

Given the space constraints, one cannot do full justice to the breadth and depth of geospatial analysis. However, the references included in this paper provide an entry into the range of methods, theories and software tools available to date. Some interesting perspectives and thoughts on the future of geospatial analysis are provided in Golledge (2009), Goodchild (2009) and Anselin (2010).
References


Lu, Yongmei and Xuwei Chen (2007). On the false alarm of planar K-function when analyzing urban crime distributed along streets. Social Science Research 36, 611-632.


Porter, Michael D. and Donald E. Brown (2007). Detecting local regions of change in high-dimensional criminal or terrorist point processes. *Computational Statistics and Data Analysis* 51, 2753-2768.


