FUSION OF REMOTE SENSING AND NON-AUTHORITATIVE DATA FOR FLOOD DISASTER AND TRANSPORTATION INFRASTRUCTURE ASSESSMENT

by

Emily K. Schnebele
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Science

Committee:

Guido Cervone, Dissertation Director
Dr. Nigel Waters, Committee Member
Dr. Richard Medina, Committee Member
Dr. Monica Gentili, Committee Member
Dr. Sheryl L. Beach, Acting Department Chair
Dr. Richard Dieckio, Interim Associate Dean for Student and Academic Affairs, College of Science
Dr. Peggy Agouris, Interim Dean, College of Science

Date: November 25, 2013

Fall Semester 2013
George Mason University
Fairfax, VA
Fusion of Remote Sensing and Non-authoritative Data for Flood Disaster and Transportation Infrastructure Assessment

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Emily K. Schnebele
Master of Arts
University of Maryland at College Park, 1994
Bachelor of Science
University of Maryland at College Park, 1992

Director: Dr. Guido Cervone, Associate Professor
Department of Geography and GeoInformation Science

Fall Semester 2013
George Mason University
Fairfax, VA
Acknowledgments

Although I am listed as the solitary author of this work, its completion would not have been possible without the insight, guidance, and encouragement from my advisor, Dr. Guido Cervone. I have been privileged to have such a talented scientist and researcher as my advisor and mentor. No matter how challenging the task, from technical help to moral support, he always came to my aid. I have truly enjoyed being his student and I am eternally grateful for everything he has done for me.

I also would like to thank the members of my committee, Dr. Nigel Waters, Dr. Richard Medina, and Dr. Monica Gentili. I have been extremely fortunate to have a committee who cares so much about me and my work. Without fail, they responded to my questions and requests for help by providing insightful and valuable advice. In particular, I would like to thank Dr. Waters for hiring me as his Research Assistant; the experience has been invaluable and it has been a pleasure to work alongside such a generous and talented professor. Dr. Medina is a true scholar and has been a source of trusted and valuable advice. Finally, I would like to thank Dr. Gentili for her thoughtful questions and kind support.

I am also very grateful to Mr. Caesar Singh and the United States Department of Transportation for funding my education and research. The financial support has allowed me to attend school full-time and focus on the completion of this dissertation.

Finally, I would like to thank my family for their endless patience and support during the long process of completing this degree.

Work performed under this project has been partially supported by US DOT’s Research and Innovative Technology Administration (RITA) award # RITARS-12-H-GMU (GMU #202717). DISCLAIMER: The views, opinions, findings and conclusions reflected in this presentation are the responsibility of the author(s) only and do not represent the official policy or position of the USDOT/RITA, or any State or other entity.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>Abstract</td>
<td>x</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation and Problem Statement</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Traditional Flood Assessment</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Non-authoritative Data</td>
<td>2</td>
</tr>
<tr>
<td>1.4 Scope of Dissertation</td>
<td>4</td>
</tr>
<tr>
<td>1.5 Dissertation Organization</td>
<td>5</td>
</tr>
<tr>
<td><strong>2 Literature Review</strong></td>
<td>7</td>
</tr>
<tr>
<td>2.1 Disasters and Risk</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Technology as a Resource</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Flood Assessment</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1 Hydrologic modeling</td>
<td>14</td>
</tr>
<tr>
<td>2.3.2 River stage/DEM</td>
<td>15</td>
</tr>
<tr>
<td>2.3.3 Remote sensing</td>
<td>15</td>
</tr>
<tr>
<td>2.4 Volunteered Geographic Information and Disasters</td>
<td>17</td>
</tr>
<tr>
<td>2.5 Disaster and Transportation Analysis</td>
<td>20</td>
</tr>
<tr>
<td>2.6 Data Fusion</td>
<td>21</td>
</tr>
<tr>
<td><strong>3 Methodology</strong></td>
<td>24</td>
</tr>
<tr>
<td>3.1 Geospatial Methods</td>
<td>24</td>
</tr>
<tr>
<td>3.2 Data Fusion</td>
<td>25</td>
</tr>
<tr>
<td>3.2.1 Overview</td>
<td>25</td>
</tr>
<tr>
<td>3.2.2 Data pre-processing</td>
<td>27</td>
</tr>
<tr>
<td>3.2.3 Data integration</td>
<td>29</td>
</tr>
<tr>
<td>3.2.4 Road hazard map</td>
<td>30</td>
</tr>
<tr>
<td><strong>4 Application of Non-authoritative Data for Flood Estimation</strong></td>
<td>32</td>
</tr>
<tr>
<td>4.1 Data</td>
<td>32</td>
</tr>
<tr>
<td>4.2 Data Analysis</td>
<td>34</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Number and percentages of pixels classified as water.</td>
<td>41</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparison between non-authoritative and authoritative data.</td>
<td>53</td>
</tr>
<tr>
<td>7.1</td>
<td>Data sources and availability for Calgary floods.</td>
<td>76</td>
</tr>
</tbody>
</table>
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Spectrum of confidence associated with authoritative and non-authoritative data sources.</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>Cumulative damage from floods in the United States (a) and globally (b) in US Dollars, 1900-2012.</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Flowchart illustrating model for the fusion of remote sensing and non-authoritative data.</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>Layers generated from multiple sources of remote sensing, authoritative, and non-authoritative data.</td>
<td>28</td>
</tr>
<tr>
<td>4.1</td>
<td>Maximum daily precipitation rate and accumulated precipitation for the period ranging from 1 April to 31 May 2011.</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>Digital Elevation Model of Memphis and the surrounding area.</td>
<td>35</td>
</tr>
<tr>
<td>4.3</td>
<td>Year 2011 water height profile for Mississippi River at Memphis, TN.</td>
<td>35</td>
</tr>
<tr>
<td>4.4</td>
<td>Water pixel classification using a digital elevation model (DEM) and Landsat data.</td>
<td>39</td>
</tr>
<tr>
<td>4.5</td>
<td>Flood hazard map indicating the probability of flood in percentage using DEM, Landsat, and ground data.</td>
<td>40</td>
</tr>
<tr>
<td>4.6</td>
<td>Histogram of pixels classified as water.</td>
<td>42</td>
</tr>
<tr>
<td>5.1</td>
<td>Crowsourced assessments for the Civil Air Patrol data.</td>
<td>46</td>
</tr>
<tr>
<td>5.2</td>
<td>Storm surge extent generated by FEMA and the locations of Civil Air Patrol photos and geolocated videos.</td>
<td>50</td>
</tr>
<tr>
<td>5.3</td>
<td>Classification of damage within FEMA surge extent using non-authoritative sources.</td>
<td>51</td>
</tr>
<tr>
<td>5.4</td>
<td>Example of YouTube video documenting flooding.</td>
<td>51</td>
</tr>
<tr>
<td>5.5</td>
<td>Designated areas ranging from medium to severely damaged based on non-authoritative data.</td>
<td>52</td>
</tr>
<tr>
<td>5.6</td>
<td>Progression of tweets mentioning the word “flood” in the New York City area.</td>
<td>54</td>
</tr>
<tr>
<td>5.7</td>
<td>Agreement between Civil Air Patrol photos and FEMA evaluation for Hurricane Sandy.</td>
<td>56</td>
</tr>
</tbody>
</table>
5.8 Disagreement between Civil Air Patrol photos and FEMA evaluation for Hurricane Sandy. .............................................. 57
6.1 Water depth measured by FEMA MOTF at public schools in New York City. 59
6.2 Location of Volunteered Geographic Information for October 29-30, 2012. . 60
6.3 Depiction of an artificial neural network. .................................. 63
6.4 Modeled surge extent for October 30 at 1:00pm. .......................... 64
6.5 Classification of flooding in New York using an artificial neural network. . 66
6.6 Classification of flooding in New York using an artificial neural network with an added layer of classified roads. ............................... 66
6.7 Damage assessment from FEMA based on aerial photographs and flood depth. 67
7.1 Digital Elevation Model for Calgary. ...................................... 70
7.2 Mean daily water height for June 2013 on the Bow River in downtown Calgary. 71
7.3 Water classification and DEM flood extent estimation for June 22. ........ 74
7.4 Water classification from SAR data for Calgary floods. ................... 75
7.5 Distribution of non-authoritative data collected for the Calgary floods. . 77
7.6 Flood extent predicted for June 21 as compared to closed areas in Calgary. 78
7.7 Flood extent estimation and road assessment for Calgary. ................ 79
7.8 Progression of tweet volume and flooded area over time. ................. 80
Abstract

FUSION OF REMOTE SENSING AND NON-AUTHORITATIVE DATA FOR FLOOD DISASTER AND TRANSPORTATION INFRASTRUCTURE ASSESSMENT

Emily K. Schnebele, PhD
George Mason University, 2013
Dissertation Director: Dr. Guido Cervone

Flooding is the most frequently occurring natural hazard on Earth; with catastrophic, large scale floods causing immense damage to people, property, and the environment. Over the past 20 years, remote sensing has become the standard technique for flood identification because of its ability to offer synoptic coverage. Unfortunately, remote sensing data are not always available or only provide partial or incomplete information of an event due to revisit limitations, cloud cover, and vegetation canopy. The ability to produce accurate and timely flood assessments before, during, and after an event is a critical safety tool for flood disaster management. Furthermore, knowledge of road conditions and accessibility is crucial for emergency managers, first responders, and residents.

This research describes a model that leverages non-authoritative data to improve flood extent mapping and the evaluation of transportation networks during all phases of a flood disaster. Non-authoritative data can provide real-time, on-the-ground information when traditional data sources may be incomplete or lacking. The novelty of this approach is the application of freely available, non-authoritative data and its effective integration with established data and methods. Although this model will not replace existing flood mapping and disaster protocols, as a result of fusing heterogeneous data of varying spatial and tem-
poral scales, it allows for increased certainty in flood assessment by “filling in the gaps” in the spatial and temporal progression of a flood event.

The research model and its application are defined by four case studies of recent flood events in the United States and Canada. The model illustrates how non-authoritative, authoritative, and remote-sensing data can be integrated together during or after a flood event to provide damage assessments, temporal progressions of a flood event, and near real-time flood estimations.
Chapter 1: Introduction

1.1 Motivation and Problem Statement

Flooding is the most frequently occurring natural hazard on Earth. Catastrophic, large scale floods cause immense damage to people, property, and the environment. Recovery costs associated with large flood disasters such as Hurricane Sandy often range in the hundreds of millions of dollars and in the some cases, tens of billions. The ability to produce accurate and timely flood assessments before, during, and after an event is a critical safety tool for flood disaster management. Furthermore, knowledge of road conditions and accessibility is crucial for emergency managers, first responders, and residents but can be difficult to accomplish in a timely manner for large or remote areas. Over the past two decades, remote sensing has become the standard technique for flood identification and management because of its ability to offer synoptic coverage (Smith, 1997). Unfortunately, remote sensing data may be insufficient as a function of revisit time or obstructed due to clouds or vegetation. Thus, it is difficult to generate a complete portrayal of an event.

1.2 Traditional Flood Assessment

Although remote sensing has become the standard technique for flood assessment data, the quality and availability of information collected will vary by sensor. RADAR data, in particular, are a good resource for flood identification because of the capability to distinguish water bodies from other land cover while penetrating through vegetative canopy and cloud cover (Laura et al., 1990). But the application of RADAR data can be difficult due to limited swaths and long revisit times. The integration of additional data, such as multiple imagery, digital elevation models (DEM), and ground data (river/rain gauges) is often used
to augment flood assessments or to combat inadequate or incomplete data. For example, Townsend and Walsh (1998) combined synthetic aperture radar (SAR) with Landsat TM imagery and a DEM to derive potential areas of inundation. Wang et al. (2002) proposed the integration of Landsat TM data with a DEM and river gauge data to predict inundated areas under forest and cloud canopy. Aerial platforms, both manned and unmanned, are particularly well suited after major catastrophic events because they can fly below the clouds, and thus acquire data in a targeted and timely fashion. Although these methods can improve data availability or assessment, there are still often gaps in the spatial and temporal data infrastructure of a flood event.

1.3 Non-authoritative Data

Traditionally, geographic data are collected, produced, and managed by trained professional cartographers, geographers, and government agencies and are afforded credibility because of the authority of these sources (Flanagin and Metzger, 2008). Furthermore, information which comes from these official, authoritative agencies or persons carries certain a level of trust (Goodchild and Glennon, 2010). This trust is garnered from our belief that the information is verified, reliable, and confirmed to the best of their abilities. By its very definition, the word “authority” describes “an accepted source of information” or is “a person or body of persons in whom authority is vested, [such] as a government agency” (Dictionary.com, 2013). By defining what authoritative data are, non-authoritative geographic data can therefore be defined as their converse. These are data which are not collected and distributed by traditional, authoritative sources such as government emergency management agencies or trained geographers. Non-authoritative data can be collected from a wide variety of sources and the credibility or level of trust will vary with source characteristics (Figure 1.1). Sources can range from those considered to be somewhat “authoritative” such as power outage information garnered from local power companies to what is clearly non-authoritative, such as texts posted anonymously on social media. Even data which
Authoritative
Non-Authoritative

Government Agencies
Public Sources
User Generated Content

Ground data    remote sensing    traffic cameras    power outages    local news    cell phones    photos    microblogs

High confidence                         Low confidence

Figure 1.1: Spectrum of confidence associated with authoritative and non-authoritative data sources.

learn toward the authoritative, such as information provided by power companies, should be categorized as non-authoritative. This is because the intent of the power company who lists, or even maps, areas of power outages is not for these data to be used for geographic research, in all likelihood the person passing the data to community is not a trained geographer, and it is unknown whether this power outage map is verified or authenticated. The power company is providing a service to its customers, how many people in my area are without power? and where are the outages generally located? Even if illustrated on a map, these data are still non-authoritative because a knowledge of how the information is produced and whether it has been verified is lacking.

One of the largest sources of non-authoritative data today are volunteered geographic data (VGI), a specific type of user generated content, which is voluntarily contributed and contains temporal and spatial information (Goodchild, 2007). Because of the massive amount of real-time, on-the-ground data generated and distributed daily, the utilization of VGI during disasters is a new and growing research agenda. For example, VGI has been shown to document the progression of a disaster as well as promote situational awareness during an event (Sakaki et al., 2010; Vieweg et al., 2010; Oxendine, 2013). Twitter, a popular microblogging service, has been used by local citizens to provide disaster information as much as an half an hour before “official” reports have been available (Stollberg and De Groeve, 2012). These tweets provide useful information to first responders when other sources are unavailable. But, these data have yet to be regularly and systematically applied in large scale disaster relief situations for multiple reasons, including difficulties of
authentication and confirmation, questions of quality and reliability, and difficulties associated with harvesting data from heterogeneous and non-structured sources (Flanagin and Metzger, 2008; Schlieder and Yanenko, 2010; Tapia et al., 2011).

Another example of non-authoritative, volunteered information harnesses the power of group contribution, or the “wisdom of crowds” (Surowiecki, 2005). Crowdsourcing, a process where a task is undertaken by a large group of people rather than by a single individual or expert, can result in successful problem solving (Howe, 2006). Examples of successful crowdsourcing include Wikipedia and Open Street Map, where information is voluntarily contributed and the public manages content and errors. Non-authoritative data are not limited solely to volunteered information and may include data which were originally intended for other purposes, but can be employed to provide information during disasters. For example, traffic cameras provide reliable, geolocated information regarding traffic and road conditions in many cities worldwide, but have yet to be applied as a data source during flood events.

1.4 Scope of Dissertation

This research extends the concept of employing multiple data sources for improved identification or performance by leveraging non-authoritative data to augment flood extent mapping and the evaluation of transportation infrastructure. The novelty of this approach is the application of freely available, non-authoritative data and its integration with established data and methods. Despite their non-scientific nature, non-authoritative data offer new and additional information which harnesses the power of “citizens as sensors” and “wisdom of crowds” to fill in gaps in the data infrastructure (Surowiecki, 2005; Goodchild, 2007; Sui and Goodchild, 2011). The application of non-authoritative data provides an opportunity to include real-time, on-the-ground information when traditional data sources may be incomplete or lacking.

The model presented is based on the fusion of data from various and multiple data
sources. The integration of data from multiple sources, which may have varying resolutions, sparse data, or different levels of uncertainty can provide information when they would not do so if used in isolation. The model yields a flood extent map generated from the integration of non-authoritative, authoritative, and remote sensing data which can be applied for transportation assessment. Although not considered ground truth, the fusion of multiple non-authoritative data sources helps fill in gaps in the spatial and temporal coverage of an event. In addition, the ability to identify potential areas of road damage or inaccessibility from flooding can optimize response initiatives by identifying areas of severe damage. This dissertation addresses the following science questions:

1. Does the fusion of non-authoritative data with remote sensing data improve flood inundation mapping?

2. Are there specific data required? When, during the progression of the event, are specific data needed?

3. Can these flood maps be used to define obstructed/impassable roads?

4. Does the model provide actionable information in near real-time?

1.5 Dissertation Organization

This research contributes to flood disaster and transportation management by providing a model which leverages new and previously under used data to augment and fill in gaps in the spatial and temporal data infrastructure. Specifically, non-authoritative are fused collectively and with remote sensing and other sources of authoritative data to produce maps of flood extent estimation. The flood analyses can then be utilized for transportation infrastructure assessment. The remainder of the dissertation is organized as follows.

Chapter Two provides a review of literature related to research in flood disasters and risk, how non-authoritative data are used in our society, and the traditional methods for flood assessment. In addition, this chapter also discusses current applications of volunteered
geographic information during disasters, as well as the state of the art in remote sensing for transportation infrastructure assessment. The chapter concludes with a discussion of data fusion applications and methods. Chapter Three provides a theoretical framework and the general methodology developed for this dissertation. Chapters Four through Seven are case studies which illustrate how the methodology defined in Chapter Three is applied. Specifically, Chapter Four provides a proof of concept demonstrating how even a small amount of volunteered information can be used to refine the results of a traditional flood estimation. In Chapter Five, crowdsourced and volunteered data are utilized to create an estimation of flood extent and damage which is employed to create a road damage map. In Chapter Six, a machine learning classifier is used to illustrate how multiple data sources can be integrated and applied for near real-time flood assessment. In Chapter Seven, non-authoritative data are combined to create a time series of a flood event. Finally, Chapter Eight provides a discussion of the contributions of this research.
Chapter 2: Literature Review

2.1 Disasters and Risk

Disasters cause widespread devastation and adversely affect every part of society. They cause loss of life, injuries, and damage to property and the environment. They strain economies by requiring tremendous capital for response and recovery efforts. The United Nations Office for Disaster Risk Reduction defines a disaster as “a serious disruption of the functioning of a community or a society involving widespread human, material, economic or environmental losses and impacts, which exceeds the ability of the affected community or society to cope using its own resources” (UNISDR, 2013). Disaster risk is a function of the hazard, or triggering event, as well as a population’s or region’s vulnerability. This vulnerability is a consequence of social, economic, and political factors that define how well a person, or people, can endure and recuperate from the effects of a hazard (Cutter, 1993; Wisner, 2003).

Disasters can be triggered by natural and/or anthropogenic hazards. Natural hazards have a natural source, such as flooding resulting from excessive precipitation or hurricane, or tectonic activity triggering a volcanic eruption or tsunami. Anthropogenic hazards, on the other hand, can be more complex to define. Broadly speaking, these are hazards caused, produced, or influenced by humans. Within this framework, examples range from societal (such as acts of terrorism) to the technological (such as a catastrophic failure of a bridge or the emission of radiation from a nuclear power plant) (Alexander, 2002). Natural and anthropogenic hazards can also occur in tandem or in a cascading sequence of events. These combined hazards result in complex disaster situations, where a natural hazard triggers an anthropogenic hazard. Examples include the breached levees of New Orleans following Hurricane Katrina in August 2005, the damaged Fukushima Daiichi reactors in Japan following
a tsunami in March 2011, and dike failures along the Elbe River in Germany in June 2013. Although hurricanes, tsunamis, and flooding are clearly natural in origin, the resulting disasters were exacerbated by the failure of human built infrastructures (Freudenburg et al., 2008). The construction of infrastructure in areas susceptible to natural hazards increases vulnerability by allowing people to live in hazardous areas as well as creating a false sense of security (Proverbs et al., 2011). In a recent survey, residents of a new subdivision in California were unaware of their flood risk within levee-protected areas and under federal flood insurance rules, these areas were not considered as being in the floodplain although their elevation was below sea level (Ludy and Kondolf, 2012). It is areas such as these where hazards can be elevated to disasters. There is difficulty in calculating and preparing for complex disasters because of the lack of an historical record by which traditional statistical methods can be used to calculate risk. Casti (2012) suggests utilizing the discrepancy between the hazard (ie. the magnitude of an earthquake and subsequent tsunami) and the human built infrastructure (ie. the height of a seawall or location of a nuclear power plant) as a risk measurement tool for extreme and complex disasters.

The frequency of disasters also seems to be increasing. Although this may be an effect of better reporting, improved sensing technologies, and changes in global population distribution, the reports are still compelling. For example, from 1961-2011, the average number of reported natural disasters was approximately 230 a year. This is in stark contrast to the previous 50 years, 1910-1960, when the annual average was approximately 15 a year. In 2011 alone, there were approximately 359 reported natural disasters, with victims totaling over 200 million and economic damage estimated at over $360 billion (EMDAT, 2013). Often a risk matrix is utilized to assess the level of risk. This is accomplished by taking the product an event’s probability of occurrence and its severity level.

Over the past fifty years, 1961-2011, the world has moved from a predominantly rural to predominantly urban society (UNDESA, 2013). Densely populated mega-cities are becoming more common as global population continues to grow and people migrate from rural to urban environments (Mitchell, 1999; Chamie, 2004). The human-built environment can
exacerbate an urban population’s lack of understanding regarding natural hazards as well as increase their vulnerability. “Often the construction of defenses allows development in what are now regarded as safe areas. The result of a catastrophic flood will consequently affect far more property and can more than offset the gains by protecting the original settlement” (Smith and Tobin, 1979; Proverbs et al., 2011, pg.4).

Flood hazards, in particular, are a global problem and are the most common natural hazard on Earth. Flood disasters cause widespread destruction, loss of human lives, and extensive damage to property and the environment (Jha et al., 2012). Flood management spans multiple disciplines with goals ranging from prediction to response for areas ranging in size from local to global. Flooding is not limited to a particular region or continent and varies in scale from creek and river flooding to tsunami or hurricane driven coastal flooding. As shown in Figure 2.1, the damage and recovery costs associated with flooding have increased substantially in the United States and globally over the past century (EMDAT, 2013). In addition, over the past 50 years, 1961-2011, approximately 70 million people annually were affected by flooding. In 2011, over 130 million people were affected by floods and $70 billion were spent on flood recovery (EMDAT, 2013). Furthermore, flooding is only expected to increase. Precision of regional regression equations for predicting annual streamflow increases significantly when climate variability is included (Vogel et al., 1999). Recent research indicates that climate change will result in a global increase in the frequency and severity of flooding. In the later part of the 21st Century, the 100 year flood is projected to occur every 10-50 years in the majority of rivers in South and Southeast Asia, Australia, Africa, and Northeast Eurasia (Hirabayashi et al., 2013).

2.2 Technology as a Resource

Although the general threat of natural or anthropogenic hazards can be somewhat alleviated by technological advances, such as building codes, it is not possible to mitigate for every conceivable hazard. Advances and changes in how technology is used can serve as valuable resources. For example, the spread of the internet to mobile devices and the advent of web
Figure 2.1: Cumulative damage from floods in the United States (a) and globally (b) in US Dollars, 1900-2012. (Source: EM-DAT: The OFDA/CRED International Disaster Database, Université catholique de Louvain, Brussels, Belgium )
2.0 and user generated content provides new and unique data sources (O’Reilly, 2007). In the past, authoritative (top down) data were the only information source available during disasters and maps were created by trained geographers. Today, vast amounts of information are reported by the general population (ground up) in real-time. These “ground up” data present not only vast resources of information, but a new class of data, one created without scientific intent but which researchers are now attempting to use scientifically.

These non-scientific data come in all forms, from text messages and images posted on social media, to data collected by the general public. For example, in 1993 a unique program was initiated to install professional-grade meteorological equipment at schools so teachers could utilize real-world data in their math, science, and geography curriculum. In addition to providing over 250,000 students and teachers today with real-world data, the WeatherBug program currently provides data to local news stations, the National Weather Service, and hosts approximately 24 million visitors/month on its website (WeatherBug, 2013). The Voluntary Observing Ships’ (VOS) scheme is a program used by the World Meteorological Organization to collect data from the world’s oceans using voluntary observing ships. These data afford information regarding large regions of the world’s oceans which are otherwise unmonitored and are used for weather forecasting, global climate research, and as ground truth for remote sensing (NOAA, 2013). Both programs are highly successful, although data collection is not performed by trained scientists.

In the field of disaster management, where up-to-date information is critical for the allocation of resources and emergency personnel, new technologies and data resources can be used to support all phases of disaster management. Specifically, the enormous amount of non-scientific, user generated content produced provides a new source of real-time, on-the-ground information. “Traditional mapping agencies cannot find the resources to field the teams of experts that would be needed to create maps of rapidly evolving events such as [the Haitian earthquake or Santa Barbara wildfires], whereas volunteers empowered by modern technologies and equipped with senses and intelligence can provide a very effective alternative” (Goodchild, 2007, 2011, pg 5). In addition to the massive amount of user generated...
content available on social media platforms, there is also a vast array of other information available as well. This includes authoritative data from news sources or government agencies, traditional sources such as satellite imagery, and novel sources such as traffic cameras. How to integrate these data effectively during disasters becomes the question. Specifically, “what protocols and procedures can be developed to link asserted, crowd-sourced social-media data with authoritative data to fill gaps in spatial data infrastructure?” (Sui and Goodchild, 2011, pg.1742).

The application of non-scientific, user generated content for disaster management is a new research agenda, where handling data volume as well as coping with its unstructured, heterogeneous nature are active questions. Because these data are not from traditional or authoritative sources, questions of quality and reliability are also of concern. Even data considered traditional or authoritative may have varying levels of certainty. Techniques originated in the mathematic and computer science disciplines such as data mining, machine learning, and data fusion provide quantitative methods for managing and analyzing large data sets (Quinlan, 1986; Agrawal et al., 1993; Hall and Llinas, 1997). Although not specifically developed for geographic inquiry, these tools can be used to allow for data of varying certainties as well as to integrate heterogeneous, multi-source data to “fill in the gaps” for improved disaster management.

Remote sensing has been used for decades as a resource for flood identification and management because of its ability to offer synoptic coverage. Unfortunately, there are often gaps in either the availability of images as a function of revisit time, or obstruction in images from clouds or vegetation making it difficult to get a complete time series of an event. The integration of additional data, such as multiple imagery, digital elevation models (DEM), and ground data (river/rain gauges) is often used to augment flood assessments. Because of the uncertainty associated with simulating natural earth processes such as flooding, models and maps can still lack in accuracy or precision. In addition, flood modeling and assessing is further complicated in areas of high topographic relief, for ungauged basins, as well as regions where the human-built environment has altered run-off characteristics.
Volunteered Geographic Information (VGI), a specific type of user generated content, is voluntarily contributed data which contains temporal and spatial information (Goodchild, 2007). The integration of VGI with authoritative and traditional data sources offers an opportunity to improve flood extent mapping by including a new source of real-time, on-the-ground information. Systematically integrating VGI with traditional data sources such as remote sensing images or DEMs can improve flood extent mapping by filling in spatial and temporal gaps. Specifically, remote sensing data may provide information over large areas, but is limited temporally by satellite revisit times. Volunteered geospatial data may have continuous temporal coverage over the period of the interest, but consist of only sparse point data.

2.3 Flood Assessment

The ability to produce accurate and timely flood assessments before, during, and after an event is a critical safety tool for flood disaster management. Flood forecasting and estimating flood risk utilizes multiple data types including precipitation, channel bathymetry, river stage, river discharge, land use/cover, and estimates of evapotranspiration. For example, Special Flood Hazard Areas are defined by the Federal Emergency Management Agency (FEMA) based on historical river flow, storm tides, rainfall, topography, and hydrologic/hydraulic models (NFIP, 2013). Although the specific intent for the designation of Special Flood Hazard Areas is for use in FEMA’s National Flood Insurance Program (NFIP), the level of risk defined on FEMA’s flood hazard maps is also utilized by many state and local agencies to determine building zones and construction specifications. During and after flood disasters the prediction and identification of flood waters is crucial for determining road accessibility, evacuations, emergency response, etc. Techniques for flood mapping can be categorized into three different approaches: hydrologic modeling, the combined use of river stage and topographic data, and the classification of remote sensing images.
2.3.1 Hydrologic modeling

Flood disaster forecasting and planning often utilize estimates generated by hydrologic and hydraulic models. Models simulate channel flow and surface and subsurface runoff using mathematical modeling of physical parameters such as elevation, precipitation, soil and surface characteristics to produce hydrographs for a given basin. Hydrographs can then be used to produce inundation maps for flood forecasting. Most models incorporate multiple types of data such as point source river and rain gauges as well as remotely sensed raster data, such as topographic (LiDAR), precipitation (RADAR), landcover/use, and soils to try and improve accuracy, especially for ungauged basins or those with sparse coverage (Sun et al., 2000; Neary et al., 2004; Khan et al., 2011; Wang et al., 2011).

Difficulties with model accuracies result from data with only partial spatial or temporal coverage or in basins with large heterogeneity in its land cover/use, soils, and topography (Colby et al., 2000; Bates, 2004; Knebl et al., 2005). For example, although rain gauges have been shown to provide reasonably accurate measurements of precipitation, they contain errors when averaged for ungauged regions (Xie and Arkin, 1997). Even in areas with moderate numbers of gauges, interpolation is needed to obtain values between sites (Ahrens, 2006). Because of the difficulties associated with modeling natural systems as well as complications associated with interpolating point data over large areas, the resolution and predictive abilities of the models become more homogenized as basin size increases.

The gap between the hydrologic/hydraulic engineers and the GIS community is being broached by the U.S. Army Corps of Engineers Hydrologic Engineering Center (HEC) models. These models (HEC-HMS and HEC-RAS) utilize GIS tools along with raster and vector data to establish watershed characteristics as well as connectivity between hydrologic elements. Multiple applications for water resources management can be developed by employing national hydrologic and elevation data, among others, within a GIS framework (Maidment and Djokic, 2000).
2.3.2 River stage/DEM

The combined use of river stage with topographic information is another method for estimating flood extent as well as inundation depth. Digital elevation models (DEM) can provide very accurate digital topographic information. But the accuracy of DEM itself will depend on the source and resolution of the data samples. The use of LiDAR has been shown to be a highly effective source for the production of high resolution DEMs (Liu, 2008). River stage information is collected from point measurements at gauging stations. In the United States, the US Geologic Survey (USGS) provides daily stream data for over 26,000 sites (USGS, 2013).

When the water elevation is known, areas at elevations less than or equal to that height can be considered underwater. These areas are then compared to known water boundaries to determine flood extent. Although this method is only applicable for gauged rivers, it has been shown to provide reasonable results even without up-to-date data (Longbotham et al., 2012). The interpolation between gauges can be coarse for areas of high topographic relief where rapid changes in river characteristics can occur over relatively short distances.

In addition to estimating flood extent, flood depth can also be approximated using this method. Flood depth information is crucial for flood hazard mapping because as the depth of flood waters increases, so does the potential for damage. The combined use of DEM elevation data with flood inundation area, either estimated from river height or remote sensing data, are regularly used to estimate depth and is considered highly effective (Sanyal and Lu, 2004).

2.3.3 Remote sensing

Space and airborne imagery are utilized for flood assessment because of their high spatial resolution and capacity to provide information for areas of poor accessibility or lacking in ground measurements (Smith, 1997; Cobby et al., 2001). Unmanned Aerial Vehicles (UAV), in particular, are capable of providing high resolution, near real-time imagery often with less expense than manned aerial or spaceborne platforms. Their quick response times,
high maneuverability and resolutions make them important tools for disaster assessment (Tatham, 2009).

Unlike hydrologic models or topographic/river height methods which are used for prediction and forecasting, remote sensing images identify current water/flooding. These data provide continuous spatial coverage (as opposed to point measurements taken from rain or river gauges), provide information for ungauged basins, and are unaffected by the local, on-the-ground hazard. High resolution data are particularly useful for the spatial analysis of water pixels. When data before and after a flood event are available, it is possible to classify land cover change, and thus identify which areas are flooded. Because the spectral response of water varies along the electromagnetic spectrum, different attributes are utilized to identify its surface extent.

In the Near Infrared (NIR) (0.8-1.1 μm) water is easily distinguished from soil and vegetation due to its strong absorption (Smith, 1997). Frazier and Page (2000) utilized single-band density slicing and multispectral classification of Landsat TM data to detect water bodies with over a 96% accuracy. However, the ability to detect water in the NIR as well as in the visible (0.4-0.7 μm) parts of the EM spectrum is hampered by obstructions from cloud cover and vegetative canopy. Wang et al. (2002) proposed the integration of Landsat TM data with a digital elevation model (DEM) and river gauge data to predict inundation areas under forest and cloud canopy. Wang’s methodology employed Landsat, DEM, and river gauge data collectively in an attempt to improve flood analysis using a rule based approach.

Active microwave sensing (RADAR) (0.75-100 cm) is utilized to detect water by taking advantage of the low backscattered signal from smooth water surfaces (Schumann and Di Baldassarre, 2010). RADAR has the ability to penetrate through clouds and forest canopy and does not require daylight (Laura et al., 1990). Water turbulence from wind, rain, or emergent vegetation will increase the backscatter from the water surface making the delineation between water and land more difficult. Synthetic Aperture Radar (SAR) has been shown to be highly effective for the detection of water and classification of flooding.
Kussul et al. (2008) used a neural network approach to classify flood extent from multi-source SAR imagery and had classification accuracy results ranging from 85.4%-95.99%. Many studies combine SAR with other data sources to increase classification accuracy. Townsend and Walsh (1998) found the combined use of SAR and multi-spectral Landsat TM data to be highly effective in identifying flooded and non-flooded areas. Mason et al. (2010) used TerraSAR-X data with LiDAR to identify urban flooding with 76% of the urban water pixels correctly classified. Martinis et al. (2009) used unsupervised thresholding and segmentation classifications to detect water in TerraSAR-X data for near-real time flood detection. The process included the integration of a DEM after classification and achieved an overall accuracy of over 95%.

Passive microwave sensing (0.15-30cm) records energy naturally emitted from the Earth surface. Because land and water have different thermal inertia and emission properties it is possible to distinguish between land and water using their brightness temperatures. Passive microwave is used for water detection at continental and global scales. De Groeve et al. (2006) developed a methodology for global scale flood alerts using data from the Advanced Microwave Scanning Radiometer-Earth Observing Satellite (AMSR-E). The low resolution of the data does not allow for precise inundation mapping, but can indicate flooded regions. The launch of NASA’s Suomi National Polar-orbiting Partnership (NPP) in 2011, provides a new wealth of passive microwave data collected by the Advanced Technology Microwave Sounder (ATMS), one of five sensors included in its payload (NASA, 2013).

2.4 Volunteered Geographic Information and Disasters

An emerging and quickly growing data source not yet fully utilized with respect to natural hazards is volunteered geographic information (VGI) (Goodchild, 2007). This general class of data, voluntarily contributed and made available, contain temporal and spatial information. Data sources include pictures, videos, sounds, text messages, etc. Due to the spread of the internet to mobile devices, an unprecedented and massive amount of ground data
have become available, often in real-time. Some data are geolocated automatically, while others can be geolocated by analyzing content.

Although volunteered data are often published without scientific intent, and usually carry little scientific merit, it is still possible to mine mission critical information. For example, during hurricane Katrina, geolocated pictures and videos searchable through Google provided early emergency response with ground-view information. These data have been used during major events, with the capture in near real-time the evolution and impact of major hazards (De Longueville et al., 2009; Pultar et al., 2009; Heverin and Zach, 2010; Vieweg et al., 2010; Acar and Muraki, 2011; Verma et al., 2011; Earle et al., 2012; Tyshchuk et al., 2012).

Volunteered data collected during and after disasters are quickly emerging as a data source during crisis and hazardous events. Liu et al. (2008); Hyvärinen and Saltikoff (2010); Zhang et al. (2012) show how photos from Flickr have been used to derive local meteorological information, capture and record the physical features of an event, and identify and document flood height. Following the Fukushima nuclear disaster in 2011, the Japanese public supplemented authoritative government sensors with user generated content. Individuals throughout the country bought personal geiger counters and contributed to a crowd-sourced geigermap (SAFECAST, 2013). Recently, volunteered data have been evaluated for the use in estimating flood inundation depth and for mapping flood extent. For rapid flood damage estimation, Poser and Dransch (2010) interpolated flood inundation depth from VGI and found estimates to be comparable to interpolated in situ measurements as well as model predictions. McDougall (2011) estimated flood extent by using VGI and river gauge data to create a DEM which was then compared to the natural topographic surface. Crowdsourced maps have shown to provide near real-time information of flooding, street closures, and support systems such as shelters (McDougall, 2012). The information they provide also evolves with the disaster, for example, first including locations of active flooding while later providing response phase information such as the locations for trash drop off or clean bottled water. In this study, when possible, reports were verified with
approximately 99% confirmed.

These potentially valuable, real-time data have yet to be systematically applied in large scale disaster relief situations for multiple reasons, including difficulties of authentication and confirmation, questions of quality and reliability, and difficulties associated with harvesting data from heterogeneous and non-structured sources (Flanagin and Metzger, 2008; Schlieder and Yanenko, 2010; Tapia et al., 2011). Recently, research is focusing on ways in which we can attempt to measure quality and accuracy of VGI. For example, Goodchild and Li (2012) explore the merits of three approaches to quality assurance in VGI: crowdsourced, social, and geographic. While each have advantages, for example, crowdsourcing can be useful for deriving information that may be difficult or time-consuming to obtain or informal “gatekeepers” can be used to monitor or authenticate, they have distinct disadvantages as well. Crowdsourcing is most successful for well known areas, but less so for remote or obscure areas or when there is a lack of willing and knowledgeable participants. “Gatekeepers”, or trusted individuals, are just another, less formal model, of traditional authoritative monitoring. Research also focuses on the question of the quality of the individual within the “crowd”. For example, Foody et al. (2013) using a latent class analysis of crowdsourced land cover classifications in remote sensing imagery, were able to evaluate the relative accuracy of specific volunteers. Quality assurance is not a simple matter, and research will likely continue to focus on developing ways in which to assign value to VGI. In the meantime, social media platforms are increasingly utilized by the public as well as authorities to disseminate VGI during crises. Although the transmission of data through channels like Twitter and Facebook can be informal and unplanned, they often become coordinating environments for residents and authorities and provide an avenue for two way communication (Bruns et al., 2012).

The geolocation of these data is also a paramount concern for effective disaster support applications. Although some VGI contain geolocation information offered by the user in the form of latitude/longitude coordinates, the vast majority do not; thus spurring a new research agenda dedicated to developing effective methods and tools for geolocating VGI.
The querying of content using keywords and the assignment of location based on named locations, people, or hashtags is one approach to geolocating Tweets (MacEachren et al., 2011). Cheng et al. (2010) also employ the content of Tweets to estimate a user’s location. They automatically identify and classify words with a local geo-scope and are able to estimate 51% of Twitter users within a 100 miles of their true location. Other techniques for geolocating Twitter data include extracting the user location from the location field of their profile (Kumar et al., 2013). In addition, others utilize a user’s social network as a means to classify their location based on the hypothesis that the more friends a user has in a given location, the more likely they are to reside at that location. Using this approach, Rout et al. (2013) illustrate how 50% of Twitter users could be geolocated at the city level based solely on their social connections. Backstrom et al. (2010) developed an algorithm to estimate the location of Facebook users using the location of their friends. Although using friends as a method of geolocation improves as the number of friends a person has increases, they correctly identified, within a 25 mile radius, the location for 69.1% of users who had 16 or more friends containing location information. They found this approach to outperform the use of IP addresses for geolocation, which had a 57.2% success rate. While these methods can be adept in varying resolutions at geolocating users, results can be still be relatively coarse with many geolocation practices successful at the city level. However, continuing research will likely improve and refine our ability to geolocate VGI at even finer scales.

2.5 Disaster and Transportation Analysis

While a functioning transportation network is essential in day-to-day life, it is particularly critical during and after hazard events. Information regarding accessibility or obstructed and damaged roadways and bridges is imperative for emergency responders. Remote sensing has been used for decades to detect and locate floods, forest fires, landslides, and many other hazards of varying type and spatial scale (Chuvieco and Congalton, 1989; Mantovani
et al., 1996; Tralli et al., 2005; Voigt et al., 2007). In addition to capturing the location and progression of a hazard, remote sensing data also catalogs damage to the built environment. These data can also be utilized to determine road accessibility during and after flooding, although image resolution, cloud/vegetation cover, or revisit times can limit data availability. For the evaluation of transportation infrastructure following Hurricane Katrina, a variety of assessment techniques were utilized including visual, non-destructive, and remote sensing. However, the assessment of transportation infrastructure over such a large area could have been accelerated through the use of high resolution imagery and geospatial analysis (Uddin, 2011).

Recent studies have focused on the application of remote sensing data after earthquakes or flooding specifically to assess transportation networks. Butenuth et al. (2011) used multi-sensor, multi-temporal imagery to identify flooded roads. Ehrlich et al. (2009) identified, using pre- and post-disaster very high resolution (VHR) optical imagery (1m or better), infrastructure and road damage after the 2008 Wenchuan earthquake. The combination of optical satellite imagery with a DEM to assess roads for accessibility after flooding was used to create a model for application in near-real time for emergency managers (Frey and Butenuth, 2011).

2.6 Data Fusion

Data fusion is a process by which data from multiple sources are integrated together for increased accuracy, improved performance, or a more complete description of a phenomena. Multi-sensor data fusion is utilized by many disciplines including robotics, weather forecasting, transportation management, medical diagnosis, and military applications such as target classification and tracking (Hall and Llinas, 1997; Klein, 2004; Siciliano and Khatib, 2008).

Although a wide variety of algorithms and approaches can be used to fuse data, all techniques employ an ordered, hierarchical framework to transition from a collected group
of multi-source or multi-sensor data to an assimilated interpretation of a study area (Hall and Llinas, 1997). The actual fusion of data can occur at a variety of levels. If data are commensurate, the physical parameters of the data can be merged at the pixel, or low, level. Feature level fusion fuses the features extracted from the different data sources. High, decision level fusion combines data using a decision process after initial classifications of the individual data sources. Regardless, the process involves a transition from raw data to knowledge through increasing levels of abstraction (Das, 2008). The evolution from raw data to knowledge is complicated by problems associated with noisy data, missing values, and differing resolutions.

Data fusion is often employed with geospatial data to integrate information with varying spatial, temporal, and spectral resolutions as well as to reduce uncertainties associated from using a single source (Simone et al., 2002; Zhang, 2010). The fused data then provides new or better information than what would be available from one source (Pohl and Van Genderen, 1998).

**Statistics**

Statistical-based data fusion employs techniques such as classical inference and probability, Bayesian inference, evidential theory or fuzzy logic. Data fusion has also been used extensively for the mapping of hazards and flood inundation (Lu et al., 2010). For example, (Amici et al., 2004) classified flooded areas by fusing multi-temporal ERS-1/2 and Radarsat SAR images using fuzzy logic. (Xie et al., 2012) estimated flood risk by fusing precipitation, DEM, and Landsat TM land use data using random set theory. Butenuth et al. (2011) fused multi-temporal optical and RADAR using probabilities for infrastructure assessment after flooding.

**Machine Learning**

Machine learning algorithms are widely used for the fusion of remote sensing data and have been shown to be highly effective for land cover classification. Cervone and Haack (2012)
tested three different supervised machine learning algorithms with a fused RADAR/optical data set. They found the fused data set improved land cover classification compared to RADAR or optical data alone with a decision rule classifier yielding the best results. Singh et al. (2012) fused LiDAR and Landsat TM data for land cover classification using a supervised maximum likelihood classifier and classification trees. The fused data improved the classification by 8%, when compared to the Landsat alone data, and by 32%, when compared to only using LiDAR.

Artificial neural networks (ANN), in particular, are often utilized for combining multiple data sources for improved classification and prediction and have recently been applied in flood forecasting and planning research. For example, Kia et al. (2012) employed an ANN technique along with a GIS to model flood-prone areas in Malaysia. By integrating multiple data layers such as rainfall, slope, and land use they developed flood maps which could be used by local or national government to protect lives and property during flood events. Ghalkhani et al. (2012) tested simulated flood hydrographs from the Hydrologic Engineering Center HEC-RAS model against hydrographs simulated by an ANN and an adaptive neuro-fuzzy inference system (ANFIS). They found that the ANN and ANFIS results coincided with the HEC-RAS simulated results, therefore offering a more rapid alternative for flood forecasting and warning. Zounemat-Kermani et al. (2013) used upstream flow records along with an ANN to predict daily watershed runoff and found that the ANN approach more accurate than a multiple linear regression analysis. Utilizing river sensor data such as water velocity, water level, air moisture, wind velocity, rainfall, and atmospheric pressure as input into an ANN, Roy et al. (2012) developed an early warning flood system.
Chapter 3: Methodology

3.1 Geospatial Methods

Tobler’s First Law of Geography tells us, “everything is related to everything else, but near things are more related than distant things” (Tobler, 1970, pg.236). This allows us to assume some level of dependence among spatial data as well as to closely examine spatial information found to be inconsistent with its surroundings (Sui, 2004). By their very nature non-authoritative data are often perceived to contain a higher level of uncertainty than authoritative data, but actually can provide local, on-the-ground information to residents during disaster or crisis situations. During the Southern California wildfires of 2007, residents reported that local news had a hard time keeping up with rapid changes and that correct, localized information pertinent to residents were available through non-authoritative sources such as social media (Sutton et al., 2008). Non-authoritative data which have a spatial component are uniquely suited to provide useful information. Spatial dependence allows for estimations at unknown points to be derived from their surrounding location. In addition, outliers are viewed with more scrutiny, therefore providing an increased confidence in data which are consistent with their environment.

Spatial data are also unique in that it is not possible to measure every point on the Earth’s surface. Interpolation can be used to “fill in the gaps” between data points to create an estimated surface of whatever spatial phenomena are being studied. Spatial interpolation methods employ the spatial dependence between variables to create estimates at unobserved locations.

Although the assumption of spatial dependence allows for estimates to be derived for unobserved locations, there is always the possibility for error. The application of data from multiple sources can work to further reduce uncertainty. The utilization of data from
multiple sources can help provide a more complete description of a phenomena. For example, data fusion is often employed with remote sensing data to combine information of varying spatial, temporal, and spectral resolutions as well as to reduce uncertainties associated from using a single source (Zhang, 2010). The fused data then provides new or better information than what would be available from a single source (Pohl and Van Genderen, 1998). The incorporation of multiple data sources or methods for improved performance or increased accuracy is not limited to the field of remote sensing. Boosting, a common machine learning technique, has been shown to be an effective method for generating accurate prediction rules by combining rough, or less than accurate, algorithms together (Freund et al., 1999). While the individual algorithms may be singularly weak, their combination can result in a strong learner. Furthermore, redundancies in observations provides an increase in the confidence of observations or estimates, while data from multiple sources can provide information when they might not do so if used in isolation.

3.2 Data Fusion

3.2.1 Overview

This model is based on the fusion of different layers generated from various data sources. This integration of multiple layers, which may have varying resolutions, sparse data, or different levels of uncertainty can provide more information when they would not do so if used singularly. The general methodology for fusing non-authoritative, authoritative, and remote sensing data from multiple sources for flood damage and road assessment is illustrated by a flowchart in Figure 3.1. While the precise definition of data fusion will vary by discipline, for example, in computer science the process of data integration is considered to be the “data fusion”; in this dissertation data fusion encompasses the entire methodology, from pre-processing through integration.

Layers are created using available remote sensing data, digital elevation models, or ground information as traditionally used for flood assessment. Figure 3.2 illustrates the
Figure 3.1: Flowchart illustrating model for the fusion of remote sensing and non-authoritative data.
layering of data from different sources. The novelty of this methodology is the use of non-authoritative data to create additional data layers which are used to augment traditional data sources when they may be lacking or incomplete. The result is shown in the bottom layer, where a flood hazard map is generated.

The resulting flood hazard map is then employed to create a road damage map. The creation of a fused data product for flood and road damage assessment utilizes a three stage approach:

1. Data pre-processing
2. Data integration
3. Road hazard assessment

3.2.2 Data pre-processing

Pre-processing methods are data dependent and can be accomplished using any method best suited for a particular combination of data and location. This stage consists of three steps: (1) processing (2) water identification and (3) layer creation. The first step includes the acquisition of data, filtering, calibration, and georectification or geolocation. The second step involves the identification of water. For data from authoritative sources this may include a machine learning classification of a remote sensing image or pairing river gauge data with a digital elevation model. Crowdsourcing is an example of a method which could be utilized to identify water from non-authoritative data. Because layers are created from multiple sources, pixel values need to be commensurate for all layers. This is accomplished in the third step which consists of normalizing and generating layers from each source. The maximum and minimum values are set the same for all layers, for example, every layer would have values ranging from 0-5, which 0 representing no water and 5 representing pixels which have the most certainty of water, such as from remote sensing data. The intermediate pixel values would be the result of kernel smoothing and/or confidence in data source (i.e. photo
Figure 3.2: Layers generated from multiple sources of remote sensing, authoritative, and non-authoritative data.
versus tweet). The generation of multiple, commensurate layers identifying water completes the data pre-processing stage.

3.2.3 Data integration

Once comparable layers are produced they are integrated to create a flood hazard map. In this research, two methods are illustrated as integration tools: (1) a geostatistical interpolation and (2) a machine learning classification.

Geo-statistical interpolation

The integration of non-authoritative data is accomplished by interpolating to create a damage assessment surface. There are multiple methods which can be utilized to interpolate values between points. For example, a spline interpolation generates a smooth surface which passes through all data points. Another alternative is an inverse weighted distance technique, a deterministic method which creates a surface based on the distance between points. It assigns more weight to closer points by assuming points closer together are more alike than those further apart. These can be simpler techniques compared to more complex methods and therefore can be computationally fast (Lu and Wong, 2008).

Geostatistical methods such as kriging, use statistical methods to estimate the variance between points to create interpolated surfaces. Kriging allows for spatial correlation between values (i.e. locations/severity of flooding) to be considered and is often used with Earth science data (Oliver and Webster, 1990; Olea and Olea, 1999; Waters, 2008). It utilizes the distance between points, similar to an inverse weighted distance method, but also considers the spatial arrangement of the nearby measured values. A semivariogram is generated to estimate the spatial autocorrelation which is then applied to predict values. Unlike deterministic methods such as inverse distance, a kriging interpolator is capable of providing some measure of error associated with the predicted values (Stein, 1999). New methods also incorporate semantics of the surrounding data points with ordinary kriging interpolation for improved results (Bhattacharjee et al., 2013). Multiple, evenly distributed points are
required for the most accurate interpolation results (Patel and Waters, 2012). Ordinary kriging has been used to interpolate non-authoritative data for flood damage assessment to augment traditional assessment methodologies (Schnebele et al., 2013). The general formula for a kriging interpolator is:

\[
\hat{Z}(x_0) = \sum_{i=1}^{n} w_i Z(x_i)
\]  

(3.1)

where \(\hat{Z}(x_0)\) is an estimator for unknown value \((x_0)\), \(n\) is the number of observed values, \(w_i\) is the weight assigned by the variogram, and \(Z(x_i)\) are the observed values.

**Machine learning**

Using a non-linear method, the integration on non-authoritative, authoritative, and remote sensing data can also be accomplished using a machine learning classification method. Artificial neural networks (ANN) are inspired by logical processes using interconnected elements (neurons) to solve a problem in parallel. They utilize non-linear partitioning with training data error output to adjust the corresponding weights of input data. They are particularly appropriate when the distribution functions of the data are unknown and have been shown to perform well at pattern recognition and classification of multisource remote sensing data as well as for applications in flood forecasting and planning (Benediktsson et al., 1990; Ghalkhani et al., 2012; Kia et al., 2012; Roy et al., 2012; Zounemat-Kermani et al., 2013). The integration stage is completed by the creation of a flood extent map generated from multiple data sources.

**3.2.4 Road hazard map**

The identification of affected roads is accomplished by pairing a road network with the flood extent map. The road network is layered over the flood extent map. Roads are then identified as potentially compromised or impassable based on the underlying flood extent
map. Road damage severity can also be classified based on an underlying flood damage assessment.
Chapter 4: Application of Non-authoritative Data for Flood Estimation

The May 2011 flooding of the Mississippi River was one of the worst floods since the Great Flood of 1927. In Memphis, TN the Mississippi River crested at 14.6m, the highest crest since 1937, which caused the evacuation of approximately 1,300 homes.

Using the 2011 flooding of Mississippi River as an example, this case study illustrates how authoritative data are traditionally used to estimate flood extent and how the application of even a small amount of non-authoritative data can be used to refine flood estimations.

4.1 Data

Non-Authoritative Data

Non-authoritative data were downloaded using the Google search engine through their photos, videos and news portal. They included sources from Flickr, YouTube, Weather Underground, Wikipedia, and abc24.com. In particular, videos (n=6) and photos (n=8) from the first two weeks of May 2011 which documented the flooding were selected. A list of Memphis road closures on May 12, 2011 (n=37) was collected from an on-line news source. Some of the data contained geolocation information, while others were geolocated using the Google API.

Remote Sensing Data

Full-resolution GeoTIFF Multispectral Landsat ETM+ images for 2 January and 10 May 2011 are used. The data were downloaded from the USGS Hazards Data Distribution System (HDDS). Landsat data are comprised of seven spectral bands: optical (0.45-0.52,
0.52-0.60, 0.63-0.69μm), near-IR (0.77-0.90μm), mid-IR (1.55 - 1.75, 2.09 - 2.35μm) and thermal-IR (10.40 - 12.50μm) with a spatial resolution of 30 meters. The images were georeferenced to UTM coordinates in ArcGIS and an area encompassing Memphis and its greater metropolitan area was selected at a scale of 1:145,000.

**Meteorological Data**

Meteorological data relative to maximum daily precipitation rate and total daily precipitation were obtained from the NCEP CPCMorphed Precipitation Model and Weather Underground (WU), respectively (Joyce et al., 2004; WeatherUnderground, 2011). Figure 4.1 shows the NCEP daily precipitation rate (bars) and the WU accumulated precipitation (solid line) for the period ranging from 1 April to 31 May 2011. The acquisition time for the May Landsat data is shown, and it occurs after the period of intense rainfall during the end of April. These meteorological data are used to identify appropriate dates for the remote sensing data. It is desirable that a scene is selected after a period of intense rainfall in order to identify the maximum flood extent.

**Digital Elevation Model Data**

A USGS Seamless Data Warehouse DEM with a 30m resolution was used in this study. The DEM is georeferenced to UTM coordinates in ArcGIS and exported at the same 1:145,000 scale as the Landsat data (Figure 4.2).

**River Gage Data**

River gage data for the Mississippi River in Memphis was collected from the US Army Corps of Engineers RiverGages website (RiverGages.com, 2011). The data used for this study were collected from gage MS126 located at longitude: 90.07667000 W, latitude: 35.12306000 N. Data were selected in elevation (meters) format so they could effectively be used in conjunction with the DEM.
Figure 4.1: Maximum daily precipitation rate and accumulated precipitation for the period ranging from 1 April to 31 May 2011.

Figure 4.3 shows the height information for MS126 for the entire year 2011. The acquisition time for the January and May Landsat data are indicated, and they correspond, respectively, to the almost minimum and maximum water heights for the entire year. The river gage height information is paired with the DEM to derive the approximate flood extent.

4.2 Data Analysis

4.2.1 Identification of flood extent

Different methodologies can be used to identify the extent of water over the geographical region of interest. The goal of this step is to generate one or more maps using the input layers which identify regions where water is detected. The task is method-independent, and it can use any method that is best suited for a particular combination of data and location.

In this study, two different methods are employed to identify flood extent. The first
Figure 4.2: Digital Elevation Model of Memphis and the surrounding area.

Figure 4.3: Year 2011 water height profile for Mississippi River at Memphis, TN.
involves the use of remote sensing data and machine learning classification, and the second the use of a DEM and river gage data.

4.2.2 Generation of flood hazard maps

After one or more flood extent maps are generated, a flood hazard map is created by computing the probability for each pixel to be flooded. This probability map is generated by applying a kernel density smoothing operation over the 2D data, and then by normalizing the result. Let \((w_1x_1, w_1x_2, \ldots, w_1x_n)\) be weighted samples drawn from a distribution with an unknown density \(f\), the goal is to estimate the shape of this function. The general kernel density estimator is:

\[
f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{w(x - x_i)}{h}\right)
\]

(4.1)

where \(K\) is the kernel function, \(h\) is the bandwidth, and \(w\) is a user selected weighting scalar. The weight \(w\) describes the importance of a particular observation, or the confidence associated with the flood extent map. In this application, using a weighted kernel function is paramount because ground observations cannot be considered “ground truth” proper, since non-authoritative data carry intrinsic uncertainties due to their generally non-scientific nature. Therefore, using different \(w\) values properly assigns levels of confidence to the various observations.

The identification of \(w\) weights is problem specific and domain dependent, but most importantly, it is dependent on data quality. The \(w\) weights are used to include the concept of “significance” of the data in the algorithm and the analysis. It is assumed when working with such heterogeneous data, the information might vary significantly, and therefore decisions should be made, when possible, using better data.

Quantitative measures to define the \(w\) weights can be established. For example, when using satellite data the pixels along the center of the swath or those that are cloud free, are preferred in the analysis. Most satellite products have a quality index associated with each
pixel that can be used to set the appropriate $w$ weight.

For non-authoritative data, the $w$ weight may vary depending on the characteristics of the source. For example, the volume of the data can be used to assign higher $w$ weight to data with dense spatial coverage and numerous observations. Higher $w$ weight can also be dependent on the source itself. For example, observations published by local news are assumed to have been validated more than points volunteered anonymously. Finally, there is also a subjective component that can be taken into account, assigning different $w$ weight to specific users or regions.

The output of the kernel smoothing is a map with contour lines illustrating the probability that specific regions are flooded. The specific methodology used for Kernel smoothing and its R implementation used in this study is described by (Wand and Jones, 1995).

### 4.2.3 Ground data integration

The last step consists of modifying the flood hazard map by the integration of the ground data. The nature of this data is different from the data used to generate the previous flood extent maps. It is usually comprised of sparse point data, identifying the presence or absence of flooding in a specific region.

For this study, weight $w$ values are assigned experimentally. They are first equally assigned and then adjusted based on characteristics and confidence in the data source. By assuming the machine learning tree induction and the DEM/river gage approach are equally adept at classifying water, their weight $w$ values are kept constant while values of the non-authoritative ground data are adjusted. The flooded roads documented by local news sources are assumed more reliable than the sparse, point data of the videos and the pictures. Based on this assumption, the weights were set to 3 for the news data, and to 2 for the pictures and videos. Equal $w$ values of 1 were assigned for both the the DEM and Landsat data.
4.3 Results and Discussion

4.3.1 Flood classification using DEM and river gauge data

A DEM and river gage data are used to classify water pixels for 2 January and 10 May 2011. Pixels in the DEM with a height less than or equal to the river gauge height are set as water pixels. Specifically, heights of 56m and 70m are used for January and May, respectively. Figure 4.4a,c show the areas identified as water for January and May dates, imposed over the DEM. The scale information is the same as in Figure 4.2.

4.3.2 Flood classification using machine learning tree induction

Water pixels are identified in both the January and May Landsat images by using a machine learning tree induction classifier. Ripley (2008) describes the general rule induction methodology and its implementation in the R statistical package used in this study.

For the machine learning classification, 4 control areas of roughly the same size are identified, 2 over the Mississippi river as examples of water pixels and 2 over different regions with no water pixels as counter-examples. Landsat multispectral data relative to these regions are used as training events by the decision tree classifier. The learned tree is then used to classify the remaining water pixels in the scene. This process is repeated for both the January and May scene, and is illustrated in Figure 4.4b,d.

About 1% of the total number of pixels are used as training pixels (events), and the remaining 99% are classified according to the induction tree generated.

4.3.3 Flood hazard maps and ground data integration

The methods described in Section 4.2.2 are employed to generate flood hazard maps using both the DEM and Landsat pixel classifications. The goal is to assign each pixel a probability of being part of the flooded area. Figure 4.5a,c show the probability contour lines for January and May, respectively.

Additional data with ground information is then used to refine the January and May
Figure 4.4: Water pixel classification using the DEM (a and c) and Landsat (b and d) and for January (top) and May (bottom) data. The background (b and d) is from Landsat band 3.
Figure 4.5: Flood hazard map indicating the probability of flood in percentage using DEM, Landsat, and ground data for January 2011 (a and b) and for May 2011 (c and d).
Table 4.1: Number and percentages of pixels classified as water.

<table>
<thead>
<tr>
<th>P(w)</th>
<th>Jan</th>
<th>Jan+Ground</th>
<th>May</th>
<th>May+Ground</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-20%</td>
<td>12,6567 (51%)</td>
<td>108,932 (44%)</td>
<td>11,4104 (46%)</td>
<td>99,368 (40%)</td>
</tr>
<tr>
<td>20-40%</td>
<td>53,718 (21%)</td>
<td>57,475 (23%)</td>
<td>40,190 (16%)</td>
<td>30,373 (12%)</td>
</tr>
<tr>
<td>40-60%</td>
<td>21,136 (08%)</td>
<td>27,601 (11%)</td>
<td>21,235 (08%)</td>
<td>26,997 (11%)</td>
</tr>
<tr>
<td>60-80%</td>
<td>12,660 (05%)</td>
<td>18,969 (08%)</td>
<td>23,260 (09%)</td>
<td>29,368 (12%)</td>
</tr>
<tr>
<td>80-100%</td>
<td>35,919 (14%)</td>
<td>37,023 (15%)</td>
<td>51,211 (20%)</td>
<td>63,894 (26%)</td>
</tr>
</tbody>
</table>

flood hazard maps (Figure 4.5b,d). The images show both the location and type of the ground information (Video, Photos, News), and the resulting probabilities when these data are taken into account. The data are imposed on both the January and May hazard maps, although all ground data are collected from the May flood. The image generated for the May flood, Figure (4.5d), shows modifications to the flood hazard map after the incorporation of the ground data. The ground data are also incorporated with the January (non-flood) image (Figure 4.5b) to illustrate how a preliminary hazard map could be generated if current satellite data are not available. In both instances, the addition of supplemental information in the form of non-authoritative ground data alters the flood map by expanding the area of possible inundation and by adjusting pixel values.

The pixel classifications are summarized in Table 4.1 and by a histogram in Figure 4.6. As expected, when generally comparing flood versus non-flood scenes as in Figure 4.5a,c, more pixels have a higher probability (60-100%) of being flooded and less pixels have a lower probability (0-40%) of being flooded in the May (flooded) image as compared to the January (non-flooded) image.

When the ground data are incorporated into the hazard maps (Figure 4.5b,d), a spatial analysis shows noteworthy changes. The incorporation of ground data yields enhancements to the May flood hazard map (Figure 4.5d) which are evident by the progression of contour lines and reclassifications of pixels. Higher value contour lines, indicating a greater probability of a region being flooded, progress toward the northeast, where the majority of ground information are located. Examining the differences between the two May scenes in Table
4.1 The percentage of pixels classified as having a low probability (0 - 20%) of being flooded as well as the pixels classified as having a high probability (80-100%) of flooding decreases or increases 6 percentage points, respectively, after the incorporation of ground data. These changes illustrate that although both the DEM/river gage and Landsat classification techniques can be highly accurate in identifying flooded areas, the addition of real-time on the ground data, verifying the presence of water in a specific area, can augment an inundation map.

Applying the layer of ground data to the January Hazard Map (Figure 4.5b) illustrates how a small amount of real-time non-authoritative ground data could be integrated with an historical image to identify possible flooded regions. The number of non-water pixels (0-20%) is reduced by 7 percentage points and reclassified to higher probability classes (Table 4.1).

Figure 4.5b,d shows while non-authoritative ground data does modify the flood hazard
maps, the amount of modification is limited by the spatial distribution of the ground data. The evolution of the contour lines, or areas of change, in both images are restricted to regions where the non-authoritative data are located. This illustrates while the incorporation of non-authoritative ground data does affect a change in both flood hazard maps, the areas of change are controlled by the quantity and distribution of the non-authoritative data.

4.4 Conclusions

This work illustrates that even a small amount of ground data points can alter the flood assessment when compared to satellite and DEM data alone. Although the best interpolation results will be achieved using large data sets of multiple, spatially distributed points, sparse data are still capable of providing important information. Furthermore, the spatial distribution of the non-authoritative data limits the areas where change is detected. This methodology is particularly useful when satellite data is limited or of poor quality. Additionally, the ground data can add a time component which can help determine the change occurring between remote sensing observations.
Chapter 5: Crowdsourced Data for Flood and Damage Assessment

Crowdsourced aerial remote sensing data along with volunteered geographic data are employed for flood damage assessment and the identification of road damage in the New York City area following Hurricane Sandy. Hurricane Sandy was a major storm which impacted a large portion of the US East coast in October 2012 with damage and recovery costs estimated to be over $40 billion in New York alone (NewYorkTimes, 2012).

Crowdsourced photos and volunteered geographic data are fused together using a geostatistical interpolation to create an estimation of flood damage in New York City following Hurricane Sandy. This damage assessment is utilized to augment an authoritative storm surge map as well as to create a road damage map for the affected region.

This case study illustrates how crowdsourced data can augment an authoritative flood assessment and how the use of multiple sources can provide non-redundant information.

5.1 Data

5.1.1 Non-authoritative data

Volunteered geographic data

Geolocated videos (n=16) which documented flooding and damage were collected from a Hurricane Sandy Google Earth website where YouTube videos supplied by Storyful could be accessed (Storyful, 2012). Although the number of videos used in this work may be considered a small sample size, the application of even a small number of volunteered data have been shown to improve flood extent mapping (Schnebele and Cervone, 2013). YouTube, a video-sharing website, is utilized by millions of people for the sharing of videos
covering a wide range of topics and experiences (YouTube, 2013). Through this site the public voluntarily shares information, often documenting damage resulting from natural hazards.

Twitter, a social networking site, is often utilized by the public to share information about their daily lives through micro-blogging (Twitter, 2013). Arizona State University’s TweetTracker provided Twitter data for this project (Kumar et al., 2011). Tweets generated in the New York City area extending from 40.92N to 40.54N latitude and 73.75W to 74.13W longitude from October 26 - November 3, 2012 containing the word “flood” were used to provide a temporal framework.

**Crowdsourced data**

The Civil Air Patrol, the civilian branch of the US Air Force, was tasked with collecting aerial photos of the US East Coast following the impact of Hurricane Sandy. Within days of the storm making landfall, hundreds of missions were flown by volunteers from Cape Cod, MA to Cap May, NJ. From these missions, thousands of aerial photos of the coastline were generated, including those documenting heavily flooded areas.

The photos were placed on a Hurricane Sandy Google Crisis Map website (Figure 5.1) for the public to assess visible damage through a crowdsourcing portal supported by MapMill (Google, 2012; Cotner, 2012). This yielded a large damage assessment data set generated from crowdsourced, non-authoritative, non-traditional sources. The photos were also made available online through a Federal Emergency Management Agency (FEMA) website for residents to search by street address to see what, if any, damage their homes may have sustained (FEMA, 2012a).

**5.1.2 Authoritative data**

**Federal Emergency Management Agency (FEMA) surge map**

The FEMA Modeling Task Force (MOTF) created storm surge maps for the US East Coast following Hurricane Sandy from field-verified high water marks and storm surge sensor data.
Figure 5.1: Crowsourced assessments for the Civil Air Patrol data. Damage assessment: red=high, yellow=medium, green=none.
FEMA employed these data along with a digital elevation model (DEM) to create a surge boundary for each state.

A FEMA MOTF shapefile was downloaded from FEMA’s GeoPlatform website and imported into ArcGIS 10 for analysis (FEMA, 2012b). The GeoPlatform site supplies data and analytics for emergency management. The shapefile utilized for this research was the finalized version (dated February 14, 2013) for New York City with a 1 meter horizontal resolution and a New York State Plane coordinate system (Figure 5.2a).

Road layer

A 2012 TIGER/line® shapefile of road networks for the New York City area was downloaded from the US Census Bureau (USCensus, 2012). The layer was georeferenced to New York State Plane coordinates in ArcGIS 10.

5.2 Data Analysis

5.2.1 Non-authoritative damage assessment

Non-authoritative data are integrated by interpolating to create a damage assessment surface. As illustrated with Equation 3.1 and discussed in Section 3.2.3, the geostatistical technique of kriging creates an interpolated surface from the spatial arrangement and variance of the nearby measured values.

5.2.2 Integration with authoritative data

After a damage assessment surface is created from non-authoritative data, it is integrated with available authoritative information. For this research, authoritative data in form of a storm surge map created by FEMA MOTF is utilized as a comparison of flood extent as well as to illustrate how non-authoritative data can provide a range of damage estimations, enhancing traditional storm surge products.
5.2.3 Generation of road damage map

Road damage is determined using the damage assessment surface created from the fusion of non-authoritative sources. Utilizing ArcGIS 10 software, a road network is layered over the damage assessment surface. Roads are classified based on the underlying damage assessment surface.

5.3 Results and Discussion

5.3.1 Damage assessment and authoritative data

Civil Air Patrol damage assessments for the area from 33N to 26N latitude and 90W to 84W longitude were downloaded directly from MapMill. The photographs were collected by the Civil Air Patrol between October 31-November 11, 2013 (within days of Hurricane Sandy impacting the New York City area). The photos were aggregated into a 500m grid structure. The value for each grid point is a function of the number of images present in each grid and their average crowdsourced damage assessment. As a result, each grid has a value from 1 to 10, with 1 representing no damage and 10 severe damage/flooding.

The videos were provided with geolocated information, and were visually assessed by the author. The small number of videos (n=15) did not require any crowdsourcing or automated assessment. Furthermore, it is shown in Schnebele and Cervone (2013) that even a small number of properly located VGI data can help improve flood assessment. Each video point was assigned a value of 10 (severe damage/flooding).

The Civil Air Patrol and YouTube data were fused together using a kriging interpolation as described in Section 3.2.3, resulting in a damage assessment surface generated solely from non-authoritative data. Kriging allows for spatial correlation between values (ie. locations/severity of flooding) to be considered and is often used with Earth science data (Oliver and Webster, 1990; Olea and Olea, 1999). Ordinary kriging generated a strong interpolation model. Cross-validation statistics yielded a standardized mean prediction error of 0.0008 and a standardized root-mean-squared prediction error of 0.9967. Figure 5.2c
illustrates the damage assessment within the boundaries of the FEMA surge extent. A histogram (Figure 5.3) shows the ranges in these damage assessment values, located within the FEMA surge boundary. The peak in medium/severe damage values (7-8) illustrates how non-authoritative data can provide damage information not conveyed in the FEMA map.

Ground information in the form of geolocated videos (Figure 5.4) enhances the non-authoritative data set by providing flood information not conveyed in the Civil Air Patrol photos. As illustrated in (Figure 5.2b), the locations of the videos (green triangles) did not coincide with locations of photos rated as medium/severe damage (larger orange circles, values 7-10). Reasons for this disparity may include flooding captured on video had receded before the Civil Air Patrol flights or were captured at night, or flooding may have occurred in areas which were not in a flight path or were unable to be seen from aerial platforms (ie. flooding in tunnels, under overpasses). By using multiple data sources, flood or damage details not captured by one source can be provided by another.

A comparison of flood surface area between the two maps was also conducted. The storm surge area on the FEMA map is approximately 121 km$^2$. Using the higher rated areas of damage (regions with values from 7-10) from the non-authoritative assessment yielded an approximate surface area of flooding and damage of 157 km$^2$ (Figure 5.5). Using only the areas classified as medium-severely damaged, the surface area generated from non-authoritative sources is within 23% of FEMA’s surge extent for New York City.

Overall, there is a very good agreement between the flood extent from FEMA and the assessment generated with the proposed methodology. Figure 5.7 are examples of agreement between photos identifying flooding/damage and the FEMA generated flood extent while Figure 5.8 include examples where the locations of flooding or damage did not agree between the Civil Air Patrol and the FEMA data. These areas were located along coastal edges and therefore precision is most likely the cause of the discrepancies.

Sources of error in non-authoritative data, such as incorrect information (false positive/negative) or improper geolocation needed to be considered. Incorrect information can
(a) Storm surge created by FEMA MOTF for New York City.
(b) Locations of Civil Air Patrol photos and geolocated videos documenting flooding.

(c) Damage assessment generated from non-authoritative data within FEMA surge boundary.
(d) Road damage assessment based on analysis of non-authoritative data.

Figure 5.2: Storm surge extent generated by FEMA and the locations of Civil Air Patrol photos and geolocated videos (a and b). Flood damage assessment generated from non-authoritative data and the subsequent classification of potential road damage (c and d).
Figure 5.3: Classification of damage within FEMA surge extent using non-authoritative sources.

Figure 5.4: Example of YouTube video documenting flooding.
Figure 5.5: Designated areas ranging from medium to severely damaged (medium=7,8 severe=9,10) based on non-authoritative data.

be mitigated by including visually verified photos/videos and the application of multiple sources. Crowdsourcing, in particular, can increase accuracy and enhance information reliability compared to single source observations (Giles, 2005). Geolocation errors can be reduced with automation.

Sparse data or data skewed in favor of densely populated or landmark areas makes the use of non-authoritative data sources especially challenging. Increasing data volume and integrating authoritative data into the methodology can yield increased confidence and include underrepresented areas. Table 5.1 compares and summarizes some features of each type of data. Although non-authoritative data can provide timely, local information often in large volume, they are often viewed with uncertainty. Conversely, the verification and authentication of authoritative data yield trusted results at the cost of time.
Table 5.1: Comparison between non-authoritative and authoritative data.

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Non-authoritative Data</th>
<th>Authoritative Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>volume</td>
<td>real-time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>citizens as sensors</td>
<td></td>
</tr>
<tr>
<td>Challenges</td>
<td>sampling bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td>unconfirmed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>reliable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>verified</td>
<td></td>
</tr>
<tr>
<td></td>
<td>authenticated</td>
<td></td>
</tr>
<tr>
<td></td>
<td>slow</td>
<td></td>
</tr>
<tr>
<td></td>
<td>unavailable</td>
<td></td>
</tr>
</tbody>
</table>

**Temporal assessment**

For this study, Twitter data were used to provide a temporal rather than spatial assessment. Although Tweets were geolocated using TweetTracker, uncertainty in their location did not allow for a study at a street resolution. However, they provide precise temporal information that can be used to understand the progression of the surge extent over time. To understand the temporal progression is crucial during and after flood events, and is very hard to understand using remote sensing instruments, due to their inherent carrier limitations. Tweeter data can effectively be used to overcome this limitation. For example, Figure 5.6 illustrates how the peak in the number of tweets containing the word “flood” occurs October 29 and 30, 2012 coinciding with the majority of flooding when Hurricane Sandy made landfall the night of October 29th and the continued flooding on the 30th.

**5.3.2 Road damage map**

In Figure 5.2c, the damage assessment is limited to the FEMA generated surge extent for the sake of comparison. For the classification of road damage, the non-authoritative assessment is not limited by the FEMA boundary. The fusion of the non-authoritative data predicted flooding and damage outside the FEMA surge boundary, so the full damage assessment was utilized for the road classification. A road network from the TIGER/line® shapefile was layered over the damage assessment surface. Road damage was then classified based on the underlying flood damage assessment (Figure 5.2d).

By using the damage assessment surface along with a high resolution road network
Figure 5.6: Progression of tweets mentioning the word “flood” in the New York City area.
layer, roads which may have severe damage can be identified at the street level. This allows authorities to prioritize site inspections, task additional aerial data collection, or identify routes which may compromised.

5.4 Conclusions

The application and integration of non-authoritative data offers opportunities to augment traditional data and methods for flood extent mapping and damage assessment. Crowdsourcing harnesses the power of “citizens as sensors” and “wisdom of crowds” to define a new method for data classification (Surowiecki, 2005; Goodchild, 2007). Although questions of reliability and validity are of concern when utilizing non-authoritative data, especially during natural disasters, these data can be employed along with traditional authoritative data and methods to enhance our knowledge of ground conditions. Although not considered ground truth, the fusion of multiple non-authoritative data sources helps fill in gaps in the spatial and temporal coverage of an event. In addition, the ability to identify potential areas of road damage or inaccessibility from flooding can optimize response initiatives by identifying areas of severe damage.
Figure 5.7: Agreement between Civil Air Patrol photos and FEMA evaluation for flooded (a and b) not flooded (c and d).
Figure 5.8: Disagreement between Civil Air Patrol photos and FEMA evaluation for flooded (a and b) not flooded (c and d).
Chapter 6: Real-Time Flood Assessment using Crowdsourced and Volunteered Data

This chapter presents a machine learning classification for the integration of authoritative, non-authoritative, and remote sensing data. Using the flooding caused by Hurricane Sandy in New York as a study area, an artificial neural network algorithm uses data from multiple sources to create a network for pattern recognition which is used to predict flooding on a subsequent day. It is possible to extend this approach to create near real-time estimation of flooding for future events.

6.1 Data

6.1.1 Authoritative data

Ground Data

Authoritative ground data consists of data collected by FEMA and the USGS. Two FEMA data sets are utilized. The first, as described in Section 5.1.2, is a surge extent map created from verified high water marks and storm surge sensor data (Figure 5.2a). The second set comprises water depth data collected from inundated New York City public schools. The water depth at schools were ascertained from water marks taken from on-site structures (Figure 6.1). Both data sets were collected by the FEMA Modeling Task Force (MOTF) which consists of experts in hazard assessment and the modeling of hazard losses. The goal of FEMA MOTF is to combine observed information from disasters to confirm and support impact assessments (FEMA, 2013). Water height collected by the United States Geological Survey (USGS) storm-tide monitoring provides an additional source of authoritative ground data.
Figure 6.1: Water depth measured by FEMA MOTF at public schools in New York City.

**Surge Model**

Wind data collected during Hurricane Sandy are used as input data for the coupled SWAN + ADCIRC model to simulate hurricane storm surge along the coast (Dietrich et al., 2011). Daily surge extent layers generated by the model are used as estimations of flood extent for October 29 and Oct 30, 2012 in New York City.

**Remote Sensing**

Extensive cloud cover inhibited the collection of high resolution satellite remote sensing data during and immediately after Hurricane Sandy. Because aerial platforms have the ability to fly below cloud cover and can be tasked rapidly, they provide an important platform for remote sensing data collection. Civil Air Patrol photos as described in Section 5.1.1 provide valuable remote sensing data collected within days of Hurricane Sandy impacting the US East Coast.
6.1.2 Non-authoritative data

Two types of non-authoritative data are utilized, remote sensing data classified by crowdsourcing and volunteered geographic information. As described in (Section 5.1.1), the Civil Air Patrol photos were placed on a Hurricane Sandy Google Earth site for crowdsourcing, yielding a large damage assessment data set classified by the general public.

The volunteered geographic information consists of tweets, videos, and photos documenting flooding in New York (40.92N to 40.54N latitude and 73.75W to 74.13W) October 29-30, 2012. Section 5.1.1 describes the tweets and geolocated videos utilized. For this research, the data are limited to those generated for the two days of this study, October 29-30, 2012 (tweets, n=1604; videos, n=16). An additional data set of photos (n=25) which documented flooding within the study domain for the two days were downloaded using the Google search engine (Figure 6.2).
6.2 Data Analysis

6.2.1 Data layer generation

As described in Section 3.2.2, layers which identify water are created from each data source. The method for the identification or classification of water is source dependent. For this study, data layers were generated from each source for October 29-30, 2012, the two days of maximum flooding in New York City. Once water is identified, the layers are converted to the same projection, normalized, and converted to a raster format in preparation for the data integration.

Authoritative data

Multiple authoritative data sets are used to create individual layers identifying the presence of water. The FEMA surge map is available as a shapefile indicating maximum flood extent and was imported directly into ArcGIS for analysis. The other data sets, inundated schools, USGS storm-tide monitoring, and surge model output, consist of water height at point locations. Two different methods were employed, a kernel density smoothing operator as described in Section 4.2.2 by Equation 4.1 for the inundated schools data and an interpolation for the surge model and USGS data to create water surface layers from point data.

Non-authoritative data

The non-authoritative data consists of point data (crowdsourced remote sensing data, tweets, videos, photos) identifying the presence of water. The input layers were created by plotting and georeferencing the data and then applying a kernel density smoothing operation as previously discussed. In addition to a Civil Air Patrol layer generated from smoothed points, a road network layer classified in Section 5.3.2 by the crowdsourced photos is employed as an additional data layer (Figure 5.2d).
6.2.2 Data layer integration

In Chapter 7 a weighted sum overlay is used to combine the layers. An alternative approach to layer integration is presented in this chapter. Once the individual flood layers are generated, they are integrated together using an artificial neural network machine learning algorithm. Artificial neural networks are non-linear data modeling tools for discovering patterns in data from a series of inputs (Atkinson and Tatnall, 1997). The network consists of interconnected nodes comprising of an input layer, a hidden layer, and an output layer (Figure 6.3). In this research, the nodes of the input layer consist of the flood identification layers created during pre-processing and the output layer is a flood assessment surface. The hidden layer nodes, or neurons, are the computational units of the network. The neuron receives the inputs and produces responses. Benediktsson et al. (1990) defines the simplest formal model of the neuron, where the output value is approximated by the function:

\[ o = K \phi \left( \sum_{j=1}^{n} w_j x_j - \theta \right) \] (6.1)

where \( K \) is a constant, \( \phi \) is a non-linear function, \( w_j \) are the weights assigned by the network, and \( \theta \) is a threshold. The network takes inputs \( x \) and produces a response \( o_i \) from the output units \( i \). The outputs are either \( o_i = 1 \) if the neuron \( i \) is active for the input \( x \), or \( o_i = 0 \) if it is inactive. The network learns the weights through iterative training and will converge when there is no change from one iteration to the next.

The trained network can then be used for the classification of a new data set. A feed-forward artificial neural network was implemented for this work using the R statistical package (Venables and Ripley, 2002).
6.3 Results and Discussion

6.3.1 Flood extent identified by authoritative sources

Flood extent surfaces are generated from four different sources of authoritative data. (1) The FEMA surge layer is available in a shapefile format and contains a polygon defining the estimated maximum flood extent for Hurricane Sandy. Because this layer represents the maximum extent, only one layer is available. (2) The storm surge model predicted hourly water depth at point locations over the entire study area. Buoy data were used to approximate the time (1:00pm) of maximum flood extent for each day, October 29 and October 30, 2012. The water height modeled for each point is interpolated using a spline function to create a water height surface. Next, a USGS seamless DEM is subtracted from the water height surface to create a water depth layer (Figure 6.4). This procedure was performed for both days at 1:00pm, generating two surge layers from the model. (3) The USGS storm-tide monitoring program also provided point data of water height. The same interpolation and DEM subtraction procedures are applied to create a water depth layer from the USGS data points. The points from the USGS are also for maximum water height, and therefore, only one layer was generated for the event. (4) The water height measurements collected at inundated structures at York City public schools are used to create a forth layer identifying
water. Because these points are sparse and not distributed over the study domain, there are not interpolated to create an entire flooded surface. The points are smoothed using a kernel density operator as described in Section 4.2.2.

6.3.2 Flood extent identified by non-authoritative sources

Although some of the non-authoritative data sets used in this study are previously applied in Chapter 5 a different data integration method is presented in this Chapter. Section 5.3.1 discusses the initial filtering and geolocation of the Civil Air Patrol and videos data sets. Photos are added as an additional data source for the analysis illustrated in this Chapter. The small number of photos allowed for visual assessment by the author. The Civil Air Patrol photos were taken after the hurricane had passed New York City and so represent the cumulative effects of the storm; therefore one layer is available for use for both days. The other sources, tweets, photos, and videos are geolocated and plotted for each day (October 29 and 30). All the individual point layers, Civil Air Patrol crowdsourced data and VGI,
are then individually smoothed using the kernel smoothing as previously discussed. The identification of water from non-authoritative sources is often sparse and not uniformly distributed, therefore layers representing water will vary greatly from one source to the next as well as from one day to the next.

6.3.3 Layer integration and generation of flood map

The methodology described in Section 6.2.2 is employed to generate flood hazard map using multiple input layers identifying the presence of water. The goal is to classify each pixel as being flooded or not flooded. The neural network classifier is trained using the data layers from October 29 and tested on the October 30 layers. Because the inundated schools, USGS, and Civil Air Patrol data represented maximum flood extent, it was possible to only generate one layer from each data set, therefore these data were used for both days.

The initial training and testing data sets produced results indicating flooding along the coast lines of New York City (Figure 6.5). Often during flood events, road closures provided by authorities can be used as an additional source of information. Because of the predicted severity of Hurricane Sandy, large portions of lower Manhattan and other coastal areas were evacuated prior to the arrival of the storm. Therefore, few roads were documented as closed during the storm. As a substitution, the road damage layer generated in Section 5.3.2 was applied in following training and testing of the neural network (Figure 6.6). This resulted in a broader classification of flooded areas.

The classification was then compared to a damage assessment generated by FEMA based on damage surveyed from aerial photos and water depth (Figure 6.7). Similar areas along the coast, such as lower Manhattan, and the entire Breezy Point area are indicated as flooded, or damaged by flooding, by both the FEMA assessment (Figure 6.7) and the second neural network which incorporated a roads layer (Figure 6.6)
Figure 6.5: Classification of flooding in New York using an artificial neural network.

Figure 6.6: Classification of flooding in New York using an artificial neural network with an added layer of classified roads.
6.4 Conclusions

This research illustrates how multi-source data can be integrated to provide an estimation of flood extent using a neural network classification model. Although in this specific work, some layers (such as the USGS data) were required for training and testing because they represented maximum extent for the event, future implementations of this model could utilize real-time data as they become available. More specifically, in this work VGI are aggregated by day and only the daily maximum extent predicted by the surge model was used. Hourly modeled surge estimates from the model in conjunction with hourly collected VGI, road closures, traffic camera data, etc. could be processed into layer formats and used as inputs into a neural network providing near real-time estimates of a flood event.
Chapter 7: Time Series of Flood Extent using Non-authoritative Data

In June of 2013, the combination of excessive precipitation and saturated ground caused unprecedented flooding in the Canadian province of Alberta. The City of Calgary, in particular, experienced sudden and extensive flooding causing the evacuation of more than 100,000 people (Upton, 2013). The damage and recovery costs for public buildings and infrastructure in the City of Calgary are estimated to cost over $400 million (Fletcher, 2013).

Because of extensive cloud cover and revisit limitations, remote sensing data of the Calgary flooding in June 2013 were extremely limited. This paper illustrates a methodology where the incorporation of freely available, non-authoritative data of various resolutions and accuracies are integrated with sparse traditional sources data to provide enhanced flood assessment. In addition, the length of the flood event allowed for the creation of a time series of the flood’s progression.

7.1 Data

7.1.1 Non-authoritative data

Volunteered Geographic Information

Geolocated photos (n=39) which documented flooding within the study domain (51.064456N to 51.013595N latitude and -114.136188W to -114.003663W longitude) were downloaded using the Google search engine.

Arizona State University’s TweetTracker provided Twitter data for this project (Kumar et al., 2011). Geolocated tweets (n=63) generated in the study domain during June 21-26,
Traffic Cameras

The City of Calgary maintains 72 traffic cameras which provide real-time traffic conditions for major roads around the city. The images collected by the cameras were manually inspected on the beyond.ca website on June 26, 2013. At that time all of the cameras were offline with time stamps for 8:30am, June 21, 2013. A few cameras (n=7) provided information regarding the state of the roads (clear/flooded) on the morning of June 21, while the majority did not have imagery available.

Road Closures

A list of Calgary road and bridge closures on June 21, 2013 (n=36) were collected from an on-line news source. Using a road network of Calgary downloaded from the OpenStreetMap website, the data were digitized in ArcGIS 10 to recreate road closures for June 21 (OpenStreetMap, 2013). Road closures for June 26, 2013 were downloaded from a Google Crisis map accessed from The City of Calgary website (The City of Calgary, 2013). The data were downloaded into ArcGIS 10 and converted from a KML format to a GeoTiff layer.

Authoritative data

Digital Photo

A RGB composite photograph of Calgary was captured by the International Space Station’s Environmental Research and Visualization System (ISERV) on June 22, 2013, one day after the flood peaked in the downtown area. The image was imported into ArcGIS 10 where it was georectified to a UTM coordinate system.

RADAR

Synthetic Aperture Radar (SAR) imagery of Calgary and the surrounding High River area were collected by RADARSAT-2 on June 22, 2013. MDA Geospatial Services processed the
SAR data and provided a flood analysis in shapefile format. The downtown Calgary area was selected from the available data and georeferenced to a UTM coordinate system.

**Digital Elevation Model**

An AltaLIS LiDAR DEM with a 30cm vertical and 50cm horizontal accuracy was provided by the University of Calgary. The data were converted from an ESRI Arc Grid ASCII format into a GeoTiff layer with UTM coordinates in ArcGIS 10 (Figure 7.1).

**River Gauge**

River gage data for the Bow River in Calgary were downloaded from the Environment Canada website. The data are provided by the Water Survey of Canada, the national authority for water levels and flow data in Canada. The data used for this study were collected in downtown Calgary from the Bow River station 05BH004 located at longitude: -114.051389W, latitude: 51.05N. Mean daily water height for 2012 as well as mean daily water height for June 2013 were utilized for the study (Figure 7.2). The primary water level
7.2 Data Analysis

7.2.1 Data layer generation

Water identification and flood extent mapping can be accomplished using a variety of methodologies. The goal of this step is to generate multiple data layers which identify areas where water is detected. The task is method-independent and can be accomplished using any method best suited for a particular combination of data and location. For this research, multiple methods were required and utilized to create individual layers from seven different data sources.
Traditional data

Traditional methods of flood classification are employed for three data sources and include a supervised classification of a remote sensing image, a flood extent product generated from SAR, and the pairing of a DEM and river gauge data.

Non-authoritative data

The non-authoritative data consists of point and line data identifying the presence of water. These layers are created by plotting and georeferencing the data and then applying a kernel density smoothing operation as described by Equation 4.1 in Section 4.2.2. The kernel smoothing was accomplished using ArcGIS 10 which employs a quadratic kernel function as described by (Silverman, 1986).

The density smoothing is employed with the point and line data to spatially extend their areal representation in preparation for the integration of layers. This is a necessary step because point data can become insignificant when combined or merged with data from other sources, such as flood extent estimated from a DEM and river gauge data. Following the kernel smoothing, the layers are converted to raster format to facilitate layer integration.

7.2.2 Layer merge

Following the generation of individual data layers, a weighted sum overlay application is utilized to merge them together. The use of a weighted sum overlay approach allows for two processes to be accomplished in one step: (1) weights are assigned to each data layer based on data characteristics (2) multiple data layers are integrated into a single comprehensive layer per time interval. The presence of water ($W_i$) at cell $i$ is given by:

$$W_i = \sum_{i=1}^{n} w_i x_i$$

(7.1)
where weight $w$ is a user selected weighting scalar chosen for each data layer and $x$ is the value of a cell $i$. The weight describes the importance of a particular observation, or the confidence associated with a data source. Following the addition of a weight to each layer, the layers are summed together yielding a comprehensive merged data layer for each specified time interval. Although the presence of water from at least one data source will classify any cell $i$ as water, the value of each cell $i$ in the merged layer results from two factors: (1) the number of layers where water is indicated and (2) the weight of each layer.

7.2.3 Prediction map

A flood prediction map is then generated for each time interval using the comprehensive merged layer. This may be accomplished using a variety of mathematical, statistical, or machine learning approaches. For this work, kriging (Equation 3.1) is used to create a geostatistical representation from the merged layer. The geostatistical interpolation yields a predicted flood extent product for each time interval based on the fusion of traditional and non-authoritative data.

7.3 Results and Discussion

7.3.1 Flood determination using supervised classification

When data before and after a flood event are available, it is possible to classify land cover change, and thus identify which areas are flooded. A RGB composite photograph of Calgary was captured by the International Space Station’s Environmental Research and Visualization System (ISERV) on June 22, 2013. Although the image was captured at almost the peak of the flood event, the classification of water in RGB composite photos is not optimal because of the difficulty distinguishing between water and land in the visible spectrum. In the ISERV image of Calgary, it is difficult to differentiate between the flood water, which appears as very brown, from roads and concrete in the urban areas. Although not ideal and containing noise, a supervised classification of the ISERV image is able to correctly identify
large areas of water and was employed as a data layer for June 22 (Figure 7.3).

**7.3.2 Flood extent identified by SAR**

Flood extent generated from data collected by RADARSAT-2 on June 22, 2013 was made available to the public by MDA Geospatial Services. The scenes were originally planned for a separate purpose and were obtained using a wide beam, covering an area of 150km$^2$ with a 30m resolution. Consequently, the ground resolution was lower than would be optimally employed when tasking specifically for flood analysis. In addition, the lack of RADAR return off the water mixed with an oversaturated return from buildings made it difficult to accurately identify flood extent in the urban downtown area. As a result, the SAR data layer significantly underestimates the flood extent when compared to photos which document the presence of water in large areas of downtown Calgary (Figure 7.4). Regardless, the SAR data is included in this research because any information documenting the presence of water contributes and further strengthens the flood extent estimation as a whole.
7.3.3 Flood classification using DEM and river gauge data

A DEM and river gage data are used to classify water pixels daily from June 21-26, 2013. Pixels in the DEM with a height less than or equal to the river height for each date are set as water pixels. Because of local topography, the elevation of the river drops approximately 23m across the study area. Consequently, when using water height data from the Bow River at Calgary station (located approximately in the center of the domain), the western and eastern reaches of the Bow River under or over flood, respectively. A normalized DEM was created by incrementally decreasing the elevation west of the gauge as well as increasing east of the gauge. The new DEM was calibrated using the mean water height from 2012.

7.3.4 Non-authoritative data layers

The technique described by Equation 4.1 is employed during the generation of individual layers from the point and line non-authoritative data for each day of readily available data. In the process of performing the smoothing operation, the bandwidth was varied by data source. By increasing the bandwidths during the smoothing operation the density values
were not significantly changed, but by incorporating a large number of surrounding cells, it is possible to create a more generalized grid.

7.3.5 Layer merge and flood extent estimation

Using the methodology described in Section 7.2.2 the data layers are merged together for each date, June 21-26, 2013, yielding 6 layers for geostatistical interpolation. The layer weights are assigned heuristically based on the characteristics of the data sources. Specifically, non-authoritative data by its very definition comes with more uncertainty than data from authoritative sources and were weighted less than the layers created from the traditional data and methods.

Because this research extended over a 6 day period, there were more data available on some days compared to others (Table 7.1). This did not affect the actual methodology for layering, as the layers are weighted and summed together in one step regardless of the number of layers used.

The locations of the non-authoritative data were generally well distributed across the domain (Figure 7.5). Although the volume of non-authoritative data varied from day to day with some days only having a sparse amount, it has been shown that even a small amount of properly located VGI data can help improve flood assessment (Schnebele and Cervone, 2013).

Following the merging of layers, prediction maps are generated as discussed in Section

<table>
<thead>
<tr>
<th>Data</th>
<th>6.21</th>
<th>6.22</th>
<th>6.23</th>
<th>6.24</th>
<th>6.25</th>
<th>6.26</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEM/River Gauge</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Street Closures</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tweets</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Photos</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RADAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Traffic Cameras</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
3.2.3. Figure 7.6 is a comparison of the maximum flood extent which was predicted on June 21 (Figure 7.7a) and areas indicated as closed from a Google Crisis map available for this flood event. The predicted flood extent appears to estimate flooding well in some areas and over estimate it in other areas. Some of the areas of over estimation are likely due to the DEM utilized which had been manually normalized to account for changes in elevation in the scene. Figure 7.7 illustrates daily predicted flood extents for June 21-26, 2013. The daily series demonstrates a progression from severely flooded, June 21, through a flood recession. The quantity of data available each day does appear to affect the prediction map results. For example, only two days had road closure data, June 21 and June 26. Because of the quantity and variety of the data for June 21, the road closure layer is well assimilated with the rest of the data (Figure 7.7a). This is not the case for June 26, where a much smaller amount of data were available. This results in the road closure layer being evident as indicated by the horizontal flood prediction in the center of the image (Figure 7.7f). An assumption was made that the road categorized as flooded on June 26 was likely flooded on
previous days as well, but because of a lack of road data for June 22-25, it was not included in the original analysis. Therefore, a decision was made to include the closed road on June 26 into the data sets for previous days. This results in the horizontal flooded area in (Figure 7.7b, c, d, e). The sparseness of data is also evident by the circular areas of higher flood prediction. These are the result of individual tweets which are located too far from the the majority of data in the scene to be properly integrated (Figure 7.7b, c, f). By comparing these flood prediction maps to the one created for June 21 (Figure 7.7a) it is clear that a smoother and richer prediction can be accomplished as data volume and diversity increases.

The overall tweet volume corresponds well to the progression of the flood event (Figure 7.8). The maximum number of tweets are posted during the peak of the flood and then reduce as the flood recedes. It is unclear why there are small increases in the number of tweets during the later days of the flood event. These tweets may be related to flood recovery with information regarding power outages, drinking water, or closures/openings of
Figure 7.7: Flood extent estimation and road assessment.
Figure 7.8: Progression of tweet volume and flooded area over time.

public facilities. Figure 7.8 also illustrates the area of the flood as a function of time. By using the flood extent estimations created with this methodology, flood area is represented as the percentage of pixels classified as flooded each day in (Figure 7.7a-f). Flood area does increases slightly the last day of the study. This is likely the result of a corresponding increase in tweets for the same day and not an actual increase in flood area.

The flood extent predictions can be further processed by applying an additional kernel smoothing operation. This may be necessary for layers with lower data quantities. For this research, a smoother flood extent was desired. The flood prediction maps were exported from ArcGIS as GeoTiff files and then smoothed using R statistical software. The same kernel density estimator as in Equation 4.1 was applied. The specific methodology used for kernel smoothing and its R implementation is described by (Wand and Jones, 1995).
7.3.6 Road assessment

The identification of affected roads can be accomplished by pairing a road network layer with the flood extent estimation. Roads which are located within the areas classified as flooded are identified as regions which are in need of additional evaluation and are possibly compromised or impassable (Figure 7.7a-f). Roads can be further prioritized as a function of distance from the flood source (i.e. river or coastline) or distance from the flood boundary. This can aid in prioritizing site inspections and determining optimal routes for first responders and residents. In addition, pairing non-authoritative data with road closures collected from news and web sources provides enhanced temporal resolution of compromised roads during the progression of the event.

7.4 Conclusions

The June 2013 flooding in Calgary serves as a good example of how remote sensing data, although a reliable and well tested data source, are not always available or perhaps cannot provide a complete description of a flood event. The utilization and integration of multiple data sources offers an opportunity to include real-time, on-the-ground information. Although not considered ground truth, non-authoritative data provides information in areas where there might otherwise be none. The fusion of data from multiple sources can yield a more robust flood assessment than that which is generated from a single source.
Chapter 8: Discussion and Summary

Flooding is a global hazard with large scale flood disasters responsible for extensive damage to people, property, and the environment. In addition, researchers predict an increase in the frequency (decrease in return period) of severe flood events (i.e. 100 year flood) in coming decades (Hirabayashi et al., 2013). Although remote sensing is the de-facto standard for flood assessment, data may be unavailable or sparse due to revisit limitations, cloud cover, or vegetation canopy. Hydrologic models are often used for prediction and forecasting, but they can be lengthy to implement and are not routinely used for real-time flood assessment. Timely information regarding the location of flooding as well as identifying compromised or affected roads is crucial for residents, first responders, and emergency managers.

With the advent of web 2.0 and mobile devices, an enormous amount of real-time data is available, from user-generated content to traffic information. The application of these real-time, on the ground data sources provides an opportunity to fill in the gaps in the traditional data infrastructure. But, the large volume of available data and the heterogeneity of sources makes their application during disasters a complex and non-trivial problem. Traditionally, geographic data are produced and managed by trained geographers from official, authoritative sources. A level of trust is associated with these data as well as a confidence in their credibility (Flanagin and Metzger, 2008; Goodchild and Glennon, 2010). In contrast, non-authoritative data describes novel sources of data collected freely from the web which carry no assertion of correctness and are frequently created by unknown producers. Although there can be uncertainty in these non-authoritative sources, they can provide real-time, local information which may be valuable when authoritative data, are lacking, incomplete, or lengthy to produce.

The fusion of remote sensing data, a reliable and trusted source, with non-authoritative data provides an opportunity to integrate high spatial, low temporal resolution (remote
sensing) data with high temporal, low spatial resolution (non-authoritative) data, thereby providing a novel model for flood assessment applications. Furthermore, this flood information can be used to infer knowledge of road conditions for the optimization of response initiatives. The model presented in this dissertation illustrates how the fusion of multiple sources of remote sensing, authoritative, and non-authoritative data of varying scales and resolutions can be integrated to enhance estimations of the flood extent (Chapter 4), augment traditional flood assessments (Chapter 5), provide real-time flood assessment (Chapter 6), and create a time series of a flood event (Chapter 7) (Schnebele and Cervone, 2013; Schnebele et al., 2013).

8.1 Non-authoritative Data Characteristics

Data distribution

The application of high temporal, non-authoritative, ground data adds a component which can help determine the change occurring between remote sensing observations and/or provide information when remote sensing observations are sparse or lacking. However, the areas of change are controlled by the distribution and quantity of the data. For example, landmark areas are more likely to receive public attention and have flooding documented than other less notable areas. Therefore researchers should be aware of, and recognize, the potential for skewness in the spatial distribution of the available data, and thus the information garnered from it. Moreover, a lack of ground data can be simply an indication of no flooding or can be the result of differences in the characteristics of places within the domain.

Data quantity

In addition to the issues of spatial distribution of the available ground data, the quantity of data also changes as the disaster evolves over time. Tweet volume, for example, provides a sense of the temporal progression of an event, with a spike in activity at the beginning or
peak of a flood event. This is illustrated in Figures 5.6 and 7.8 where the majority of tweets were captured on Oct 30 and June 22, the peak of the flood events in New York City and Calgary, respectively. However, as the flood event progresses, there is a marked decline in tweet volume which may be a result of both the evolution of the actual event as well as a loss of interest, especially from people not directly affected by the event. The importance of data quantity is evident in Figure 7.7 where a decrease in quantity and variability of data during the progression of the event creates a less consistent flooded surface, with single tweets standing out in isolation on days when the quantity of data is low.

8.2 Model Characteristics

Model validation

Early testing of this model using the Memphis 2011 floods (Chapter 4), shows how the addition of ground data can refine flood extent estimations. The inclusion of the non-authoritative ground data reconfigures what is classified as flooded, increasing the number of flooded pixels and decreasing the non-flooded pixels along the edges of the prediction map based on local observations of flooding which were not captured by the Landsat/DEM/river gauge data alone (Figure 4.5 c and d).

The application and results of the model for the Hurricane Sandy flooding in New York City agrees well with the “official” estimation from FEMA (Chapter 5). The FEMA estimation was generated using interpolated high water marks collected in situ and took weeks to achieve whereby the model was able to generate similar results using crowdsourced and VGI data. In addition to a good agreement, this model provides a richer assessment by offering a range of damage values compared to the binary assessment produced by FEMA (Figure 5.2 a and c). Furthermore, the flood extent and damage assessments could be accomplished in near real-time with the use of an artificial neural network as the classification and integration method (Chapter 6).
Model settings

The scale of the domain and quantity of ground data will affect the choice of smoothing parameters. In addition, smoothing, as well as the application of weights, allows for some data uncertainty to be considered during the analysis. By increasing the smoothing parameter, or bandwidth, the location of point data can be generalized over a range of cells. This helps to accommodate for a potential lack of precision in the geolocation of point observations and also allows for the representation of flooded locations to be expressed over a larger area. Generalizing the area over a range of cells is especially important as the size of the domain increases. Without providing some smoothing mechanism, the point data become insignificant in relation to other areal data sets (ie. remote sensing classification). In general, as the size of the domain increases, or as the map scale decreases, the smoothing parameter needs to be increased. It is possible that as the volume of point data increases, less smoothing would be required because of a greater representation of the point data.

The addition of weights allows for variations in source characteristics and uncertainties to be considered. For example, the volume of the data can be used to assign higher weight to data with dense spatial coverage and numerous observations. The weight can also be dependent on the source itself. For example, observations published by local news are assumed to have more credibility than points volunteered anonymously. The timing of the data could also be used as a metric for quality. As shown, tweet volume decreases during the progression of the event, with perhaps non-local producers dropping out as interest fades. This would then yield a more valuable data set of tweets, those just generated by local producers, which could be inferred to be of higher quality and thus garner a higher weight. However, it is not possible to set the weights as an absolute because each flood event is unique and there will be differences in data sources, availability, and quantity. In this work, the weights are assigned linearly with the highest weight given to sources believed to have the most credibility following the scale depicted in (Figure 1.1).
Model accuracy

Geolocation accuracy, especially of tweets, as discussed in (Section 2.4), is accomplished using a variety of techniques with the majority of methods being able to successfully identify the locations of producers at the city level. Accurately geolocating these data becomes more difficult at finer scales. A potentially more accurate source of data for large scale domains would be cell phone data which can provide near real-time locations of citizens as the population fluctuates temporally and spatially (Oxendine et al., 2012). The inclusion of cell phone data has the potential to not only increase accuracy, but also the quantity and distribution of available data which could further improve model results.

Although the distribution and quantity of non-authoritative data affect the estimation of flood extent, these data can provide a unique knowledge that is not possible to acquire using remote sensing instruments alone. First, non-authoritative sources can be used to provide ground information when remote sensing data may be unavailable or limited (Chapter 7). Second, they can provide information not captured by remote sensing such as flooding at a micro-level. Producers often document flooding they are directly affected by, such as in their backyard or in their home. At such scales, this is often difficult to captured by remote sensing. Third, non-authoritative data provides information in places which are difficult to sense remotely such as flooding under bridges, in subways, and at night (Figure 5.2 b).

8.3 Economic Viability

Accessible and reliable transportation networks are necessary for moving people and goods during our daily lives and are critically important during disasters when evacuations and response initiatives are paramount. Following the Colorado floods of September 2013, many major roads were still under repair 2 months after the event (CODOT, 2013). The event required the inspection of over 1000 bridges and destroyed approximately 200 miles of highway and 50 bridges (Whaley, 2013; Zellinger, 2013). During the flooding, non-authoritative data, such as unmanned aerial vehicles, provided rapid information of flooding and road

86
conditions (FALCON, 2013). Rapid and directed identification of affected areas allows for Departments of Transportation and Emergency Management Departments to prioritize their site visits and response initiatives. Because of the large number of inspections required, the Colorado Department of Transportation used a triage system to determine which bridges to inspect first (Zellinger, 2013). By applying non-authoritative data sources, flood identification and road assessments can be generated to give officials information for remote areas, such as flash flooding on mountain roads in Colorado, or for micro-locations, such as within neighborhoods. The automatic collection of non-authoritative data and fusion of multiple sources only requires a few personnel to implement and can provide rapid flood and road assessments. The costs (monetary and time) associated with sending out road crews for site inspections can be decreased when specific areas are identified and prioritized.

8.4 Conclusions

The model presented in this work introduces a systematic way to incorporate novel and non-authoritative data sources with traditional data sources and methods for improved flood and transportation assessment. With the massive amount of volunteered, user generated, and non-authoritative data generated everyday, new systems and procedures should be developed to apply these data in meaningful ways. It is only when we can harness and make meaning from the vast amount of available data can we begin to improve when and how we respond to disasters.
References


Bruns, A., Burgess, J. E., Crawford, K., Shaw, F., 2012. #qldfloods and @QPSMedia: Crisis communication on Twitter in the 2011 South East Queensland floods: Research Report. ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology.


URL http://www.coloradodot.info/travel/colorado-flood-highway-updates


URL http://queens.brownstoner.com/tag/google-crisis-map


URL http://dictionary.reference.com/


EMDAT, 2013. EM-DAT: The OFDA/CRED International Disaster Database, Université catholique de Louvain, Brussels, Belgium.
URL http://www.emdat.be/database

FALCON, 2013. Falcon UAV Supports Colorado Flooding Until Grounded by FEMA.
URL http://www.falcon-uav.com/

URL http://fema.apps.esri.com/checkyourhome/

FEMA, 2012b. FEMA GeoPlatform.
URL http://fema.maps.arcgis.com/home/

URL http://fema.maps.arcgis.com/home/item.html?id=307dd522499d4a44a33d7296a5da5ea0


URL http://metronews.ca/?s=Calgary+flood+costs+now+total+%25C%24460+million+%25A+Report

Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. Transactions in GIS. Wiley Online Library.


URL http://google.org/crisismap/sandy-2012


Kumar, S., Morstatter, F., Liu, H., 2013. Twitter Data Analytics. Springer.


URL http://npp.gsfc.nasa.gov/


URL http://www.floodsmart.gov

NOAA, 2013. The WMO Voluntary Observing Ships (VOS) Scheme.
URL http://www.vos.noaa.gov


OpenStreetMap, 2013. OpenStreetMap.
URL http://www.openstreetmap.org/


URL http://rivergages.mvr.usace.army.mil/


URL http://japan.failedrobot.com/


URL https://storyful.com/


Twitter, 2013. The fastest, simplest way to stay close to everything you care about. URL https://twitter.com/about


UNDESA, 2013. World Urbanization Prospects, the 2011 Revision.
   URL http://esa.un.org/unpd/wup/index.htm

UNISDR, 2013. Terminology on DRR.
   URL http://www.unisdr.org/we/inform/terminology

Upton, J., 2013. Calgary floods trigger an oil spill and a mass evacuation.
   URL http://grist.org/news/

   URL http://www.census.gov/cgi-bin/geo/shapefiles2011/main

USGS, 2013. USGS Water Data for the Nation.
   URL http://waterdata.usgs.gov/nwis

   URL http://www.stats.ox.ac.uk/pub/MASS4


URL http://weather.weatherbug.com

URL http://www.wunderground.com/history/

Whaley, M., 2013. Colorado to drain its $100 million road contingency fund to pay for flood fixes.
URL http://www.denverpost.com/breakingnews


YouTube, 2013. About YouTube. URL http://www.youtube.com/


Curriculum Vitae

Emily Schneble received her B.S in Geography from the University of Maryland at College Park in 1992 and her M.A. in Geography from the University of Maryland at College Park in 1994. She entered the PhD program in Earth Systems and GeoInformation Science at George Mason University in spring 2010 and received her PhD in the fall 2013.