COMPUTING WITH BIG SPATIAL DISASTER DATA FOR COASTAL RESILIENCE DECISION SUPPORT

by

XUAN HU

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Civil and Environmental Engineering

Written under the direction of

Dr. Jie Gong

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

MAY, 2018

ABSTRACT OF THE DISSERTATION

COMPUTING WITH BIG SPATIAL DISASTER DATA FOR COASTAL RESILIENCE DECISION SUPPORT

By XUAN HU

Dissertation Director:

Dr. Jie Gong

Severe weather events such as hurricanes, ice storms, surge, and flooding have been occurring across the U.S and around the world, threatening places where economic and industrial activities are heavily concentrated. These extreme events are now increasing observed and monitored with a loosely coupled network of geospatial sensors. Analysis of these datasets offers tremendous opportunities in improving the resilience and adaptability of coastal communities in the face of future natural disasters. Despite the high values in these data sets, the vast size and complex processing requirements of these new data sets make it challenging to effectively use them in coastal community management applications, in particular emergencies. Yet, unprocessed data are intangible and non-consumable, which is often resulting in ‘data-rich-but-information-poor” situation.

The overarching goal of this research is to research, develop, and evaluate a data processing framework that is capable of efficiently processing the emerging large geospatial data sets and extract crucial information to enhance disaster management during large-scale extreme events. This research systematically studied the fundamental aspects of big spatial disaster data including the anatomy of big spatial disaster data, data processing patterns, data quality issues, uncertainty propagation along the analytics pipeline, and adaptive processing in time-sensitive environments. More specifically, this dissertation addresses the following research questions.

1. What is the basic anatomy of big spatial disaster data?
2. What are the core operation categories and processing patterns with big spatial disaster data?
3. How does the uncertainty associated with spatial disaster data sets propagate through a given processing pipeline?
4. How to adequately represent users' dynamic and complex information needs and processing requirement during coastal resilience investigations in a unified framework?
5. How to dynamically adapt 3D disaster data analytics given user information needs and processing requirements and algorithm and dataset descriptions?

In Chapter 2, I characterized the basic anatomy of big spatial disaster data to highlight the challenges and opportunities in using these emerging data sets in coastal community management applications during extreme events. I also characterized data processing patterns associated with the emerging big spatial disaster data sets and

abstracted these patterns into core operation categories. These work laid the foundation for realizing cloud-based computing of these data sets for disaster response applications.

In Chapter 3, I used a case study based approach to demonstrate approaches for quantifying uncertainty propagation in processing geospatial data sets. More specifically, I proposed a method to identify the optimal strategy for approximation parameter selection in interpolating Light Detection and Ranging (LiDAR) data into Digital Elevation Models (DEMs). The method is developed to address the need to model accuracy loss in rapid generation of DEMs, which are essential pieces of information used in disaster response and flooding simulation.

In Chapter 3, I proposed a DEA based information salience model to prioritize the sequence of the information processing tasks. The model provides a unified way of representing user information needs and balancing these needs to realize optimized data processing sequences. More specifically, this model integrates the DEA efficiency score with linguistic group decision process. The proposed model is tested against a hurricane sandy based case study in the Barnegat Peninsula, New Jersey. The results indicate that the proposed model prototyped a framework for information articulation between decision-makers and the data processing team. The proposed model will help to accelerate the data-information transliteration and reduce the possible ‘data-rich-but-information-poor” situation

Based on Chapter 3, I proposed in Chapter 4 a stream data processing approach that realized accelerated information extraction from large quantities of geospatial data given various user information needs. The approach is capable of representing complex spatial data analytics into a workflow centric data analysis representation and levering the

flexible computing resources in the cloud and at the edge to improve information extraction from these large data sets.

Throughout this dissertation research, I used extensively Hurricane Sandy related data sets as use cases to evaluate the proposed approaches. The results demonstrated the proposed approaches provide a scalable approach for information extraction from spatial disaster data within a realistic time bound. It is important to recognize that this research does not focus on developing algorithms for data processing tasks such as segmentation and object recognition. Instead, it focuses on formulating mechanisms to integrate existing spatial data analytics into the emerging big data processing frameworks and to address the particular challenges in using the big spatial disaster data for coastal resilience decision support. In terms of future research, it is beneficial to investigate the development of dedicated disaster data processing algorithms and integrate them into the framework developed in this research.

ACKNOWLEDGEMENTS

I have worked with a great number of people whose much-appreciated contributions in assorted ways to the research and the making of the dissertation deserve special mention. First and foremost, I would like to express my sincere gratitude to my advisor Dr. Jie Gong. I would not start Ph.D. in the first place if it weren't him. He has been an excellent mentor to me. His advice, encouragement, and support are invaluable.

Besides my advisor, I would like to thank the rest of my thesis committee--Dr. Manish Parashar, Dr. Frankliin Moon, Dr. Peter Jin --for their insightful comments, which incented me to widen my research from various perspectives. I would also like to thank faculty and staff Department of Civil & Environmental Engineering, Center for Advanced Infrastructure and Transportation (CAIT) at Rutgers.

This dissertation benefited greatly from the input and support of several collaborators. Eduard Renart and Dr. Manish Parashar from Rutgers Discovery Informatics Institute (RDI2) were critical in helping me implementing the stream computing code. Adam Hatzikyriakou, Siyuan Xian and Dr Ning Lin from Princeton University are nice people to work with and they shared me insightful knowledge on building damage assessments. Finally, Dr Bingsheng Liu from Tianjin University provided invaluable guidance on modelling the decision-making process.

The burden of writing this thesis was lessened substantially by the support of my friends. I deeply indebted to Yi Yu, Zixiang Zhou, Guangji Xu and many others for being dear companions during the past five years.

My family deserve special mention for their inseparable support. I cannot imagine where I would be without my family. Their love and support have been unquestioning and essential. I dedicate this dissertation to them.

Dedication

To my family…

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Problem Statement

In the past decades, the coastal communities in the United States have sustained tremendous damages as the results of several severe coastal storm events such as Hurricane Katrina and Hurricane Sandy. These extreme events have led to substantial loss of livelihoods, costing billions of dollars in the form of destroyed private and public property and mobilization of emergency response personnel and resource. For instance, it is estimated that the Federal Emergency Management Agency (FEMA) has spent over $4 billion on response missions and $14 billion more on recovery activities during Hurricane Sandy. Volumes of scientific evidence and data suggest extreme weather events will continue to multiply and intensify. Nevertheless, it seems that the resilience of natural systems in coastal communities are prone to worsen in the face of threats from climate change and sea level rise and evolving societal pressures from the growing coastal population and needs for construction development (Wright et al. 2015). Without actions and investment to strengthen the resilience of communities, these extreme events will continue to drain federal, state and local budgets, hurt businesses' bottom lines, and threaten the prosperity of future generations.

Community resilience has two key elements: "capability" and "adaptability" Table 1-1. The first term refers to the inherent characteristic (such as strength, diversity, etc.) of a community. Bruneau et al. (2003) interpreted this term as "Robustness" and "Redundancy". The second term "adaptability" refers to the quality of being able to adjust to new conditions. Bruneau et al. (2003) define this term as "Resourcefulness" and "Rapidity". The research focus of the "capability" analysis is on the pre-event conditions

and inherent characterizes of a system. It has significant overlaps with the term "vulnerability", which defined by Adger (2006) as "the state of susceptibility to harm from exposure to stress associated with environmental and social change and from the absence of capacity to adapt". The study of "adaptability" emphasis on ability of a system to respond and adapt to changes.

Table 1-1 Key Elements of community resilience (Bruneau et al. 2003)

Elements	Definition	Properties	Content
capability	The inherent power or ability (such as strength, diversity, etc.) of a community	Robustness	Strength, or the ability of elements, systems, and other units of analysis to withstand a given level of stress or demand without suffering degradation or loss of function;
		Redundancy	The extent to which elements, systems, or other units of analysis exist that are substitutable.
adaptability	The quality of being able to adjust to new conditions	Resourcefulness	The capacity to identify problems, establish priorities, and mobilize resources when conditions exist that threaten to disrupt some element, system, or other unit of analysis
		Rapidity	The capacity to meet priorities and achieve goals in a timely manner in order to contain losses and avoid future disruption.

Extensive research efforts have been devoted to the "capability" aspect in community resilience research. Pioneering studies attempted to promote the understanding of community resilience by proposing conceptual frameworks and defining resilience indicators. For instance, Cutter et al. (2008) proposed a disaster resilience of place (DROP) model based framework. They proposed a six-dimensional community resilience system including ecological, social, economic, institutional, infrastructure, and community competence. In another study, Longstaff et al. (2010) designed a similar five-dimensional assessment framework. Additionally, Renschler et al. (2010) established a holistic PEOPLES ("PEOPLES" is the acronym of the seven-dimension community resilience indicators) framework to define and measure disaster

resilience. While these studies have a major emphasis on the definitional debate of long-term "capability" of community resilience, few of them focused on the need for "adaptability" as a critical component in community resilience.

It is important to recognize that although improving "capability" is a fundamental step to improve community resilience, a community also needs "adaptability" to cope with disasters especially during the disaster response phase. For instance, Bruneau et al. (2003) suggested that to cope with earthquake impacts, a community also need "adaptability" to absorb a shock if it occurs. To this end, adaptability is a key enabler to maintain essential community functionalities and reduce unnecessary damage. In contrast to the good "adaptability", lack of adaptability during disaster response can result in substantial losses of livelihoods and properties. For instance, lack of adaptive mechanisms in identifying problems during Tianjin explosion (2015) has, to certain degree, led to the shocking death toll of 173, being a majority of them firefighters at the disaster scene. In another example, it is reported that FEMA has wasted over $3 million on poor adaptability in setting up New Jersey field office alone. If FEMA were able to make arrangements in advanced rather than waiting until hurricane sandy hit, they could save $1.5 million in renovation and leasing, while on the other hand, another $1.5 million would be save if they were able to reduce the work space as the disaster response staff shrunk. Therefore, improving the "adaptability" during disaster response will be one of the best investments in improving the overall community resilience. The detailed requirements of building "adaptability" can be described as follows:

- Enumerating all possible scenarios in a disaster is under normal circumstance an impossible undertaking (Han et al. 1998) as the disaster situation may

change dramatically as time goes by. Frequently updated situation awareness is the key information required to reduce the number of scenarios to be evaluated.

- Since the updated information can provide new insights, a system needs the capability to modify its strategy immediately according to the real-time disaster situation.

As stated above, the first requirement for building "adaptability" is the need for understanding the situation; and the second one can be concluded as mechanisms to adapt according to situations. Situation awareness is the starting point of the adaptive behaviour of a community in the face of extreme weather events. When a disaster strikes, it is critical that stakeholders such as emergency response organizations and residents have the access to real-time information in order to assess the situation and respond appropriately. The need for information is amplified during extreme events as precise predictions of these events are challenging to achieve based on expert opinions or historical event data. Take Hurricane Sandy as an example, the observed surge was estimated to occur every 400-800 years (Lin et al. 2012) or even over 1000 years (Brandon et al. 2014).

Advanced geospatial sensing technologies are playing an increasingly important role for decision support in disaster preparation, response, and recovery operations as they greatly expand our ability of collecting disaster data during large-scale extreme events. For instance, in recent years, because state and federal agencies have made airborne LiDAR (Light Detection and Ranging) data collection a priority, post-storm LiDAR collection is now routine after large surge event and vast amounts of disaster data

are now freely available online (e.g. NOAA's *Digital Coast*). In another example, emerging high resolution sensing systems such as terrestrial/mobile LiDAR have also been deployed for damage data collection during recent events such as Superstorm Sandy, generating an unprecedented amount of 3D geospatial data. Lastly, volunteered geographic information (VGI), such as geo-tagged disaster photos, is a new breed of disaster data which further enriches but also complicates disaster data analysis. Analysis of these data sets offer tremendous opportunities in improving community resilience by enhancing community capabilities and adaptability during extreme events.

Geospatial data have been frequently used in vulnerability assessment, which is closely related to identifying gaps in a community's capabilities to cope with extreme events. Although the use of emerging large geospatial data sets in these types of analysis is difficult, the analysis can be eventually accomplished given sufficient time. What most challenging is to use these data sets to obtain better situation awareness in time-sensitive applications. Current geospatial data analysis frameworks are inadequate in handling these large data sets especially during large-scale extreme events. Figure 1-1 depicts a simplified map of current data analytics for disaster response practice that essentially involves three key processes including data acquisition, data processing, and decision-making. In the first step, data are collected from multiple resources such as mobile platforms, airborne systems, and social media outlets. The next step is data processing, which analyses the data according to the particular goals such as detecting the morphology changes of the dune using LiDAR digital elevation map (DEM) data or identifying the disaster-prone area using twitter data, etc. The final step is decision-making, in which experts identify and choose alternative response operations including

mobilizing resources or planning search and rescue operations, etc. Notably, in practice, the connections between these three processes are one-directional: the data acquisition step collects data and pushes them to the data processing step, and the data processing step delivers processing results to the decision making step. In other words, there is no formal feedback loop in the system. The lack of feedback mechanisms causes the decision makers to have little control over the specific tasks to be processed and the corresponding time requirement. In normal situations, the lack of feedback mechanism between these steps can be compensated by performing data collection, data processing, and decision making in an iterative fashion until all information are obtained. This is highly infeasible in disastrous situations because of the necessity of making quick decisions in a dynamically changing environment. In contrast to normal situations, collection and processing of large geospatial data sets during extreme events require careful coordination and integration with decision making processes.

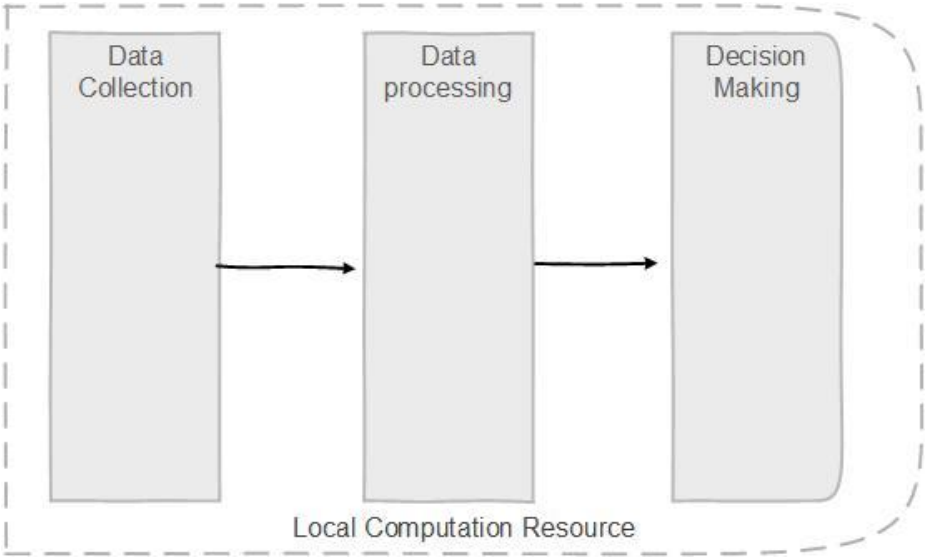


Figure 1-1 A simplified map of current data analytics for disaster response practice

It is reasonable to expect that without efficient data analysis frameworks, the growing quantities of geospatial data sets collected during extreme events actually create

significant challenges instead of benefits to disaster management. All data or information at some point has to be assessed and interpreted. One simple point is that the greater the volume of such data, the more difficulty any efficient and effective assessment will be. Therefore, a fundamental consideration in deploying advanced data collection systems in extreme events is the balance of data quantity and data quality. Today's geospatial data sets collected during extreme events tend to be heterogeneous, in large quantity, and with less qualitative certainty. The lack of an efficient data analysis framework could cause information overload and workload skewness (Bruinsma 2010). Information overload refers to the flooding in data exceeds the inherent processing capability of the computational infrastructure, resulting in the information loss. On the other hand, workload skewness refers to the depreciation of information because of poorly leveraging of computation resource: too many computation efforts are investing to the process with less value. The reality is that while we are capable of deploying a growing number of spatial sensing technologies for data collection during natural disasters, the vast quantities of collected data still have to go through painstakingly manual analyses and crucial information can often no way be extracted from these data sets in time to support critical decision makings.

1.2 Research Objectives and Research Questions

The overarching goal of this research is to research, develop, and evaluate a data processing framework that is capable of efficiently processing the emerging large geospatial data sets and extract crucial information to enhance disaster management during large-scale extreme events. This research will systematically study the fundamental aspects of big spatial disaster data including the anatomy of big spatial disaster data, data processing patterns, data quality issues, uncertainty propagation along

the analytics pipeline, and adaptive processing in time-sensitive environments. More specifically, the goals of this project include:

- Explain the feasibility and challenges of using big data for situation awareness
- Design an uncertainty quantification framework that enables the computers to balance between processing time and quality of results to meet stakeholders' information requirement in time-sensitive applications.
- Formally represent information needs in disaster response to assist workflow composition
- Design fast computing techniques to promote "adaptability" performance

To realize such research objectives, this research must address the following questions (Figure 1-2):

(1) What is the basic anatomy of big spatial disaster data?

(2) What are the core operation categories and processing patterns with big spatial disaster data?

(3) How does the uncertainty associated with spatial disaster data sets propagate through a given processing pipeline?

(4) How to adequately represent users' dynamic and complex information needs and processing requirement during coastal resilience investigations in a unified framework?

(5) How to dynamically adapt 3D disaster data analytics given user information needs and processing requirements and algorithm and dataset descriptions?

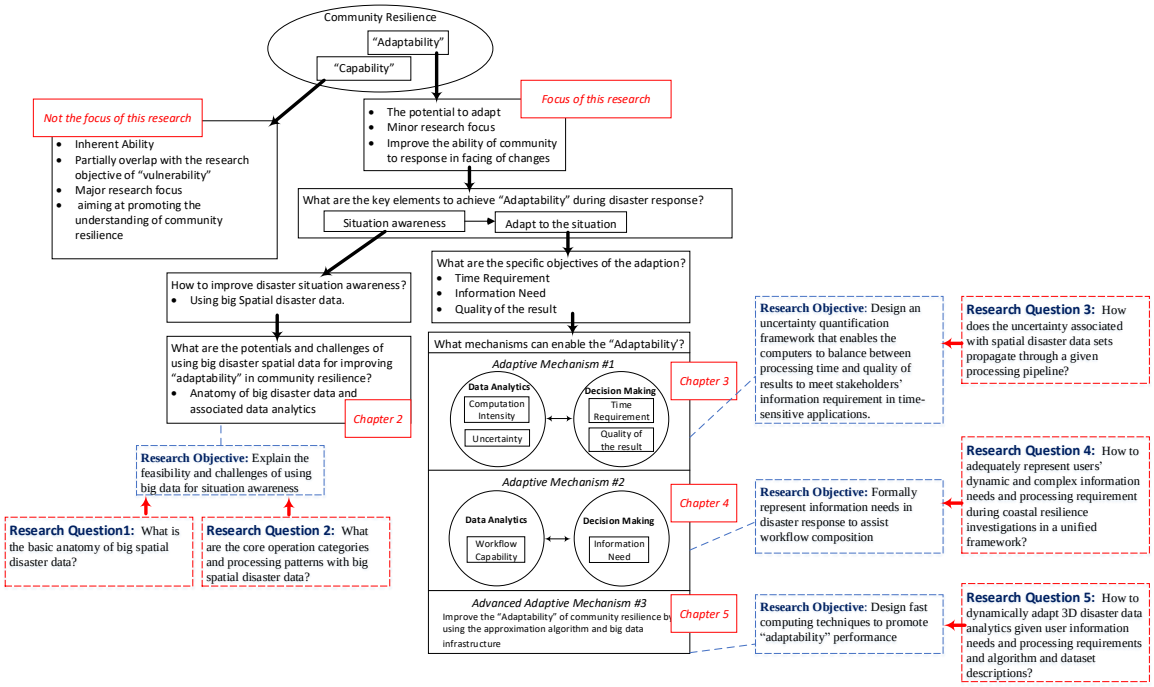


Figure 1-2 Research Questions

My envisioned framework is shown in Figure 1-3.

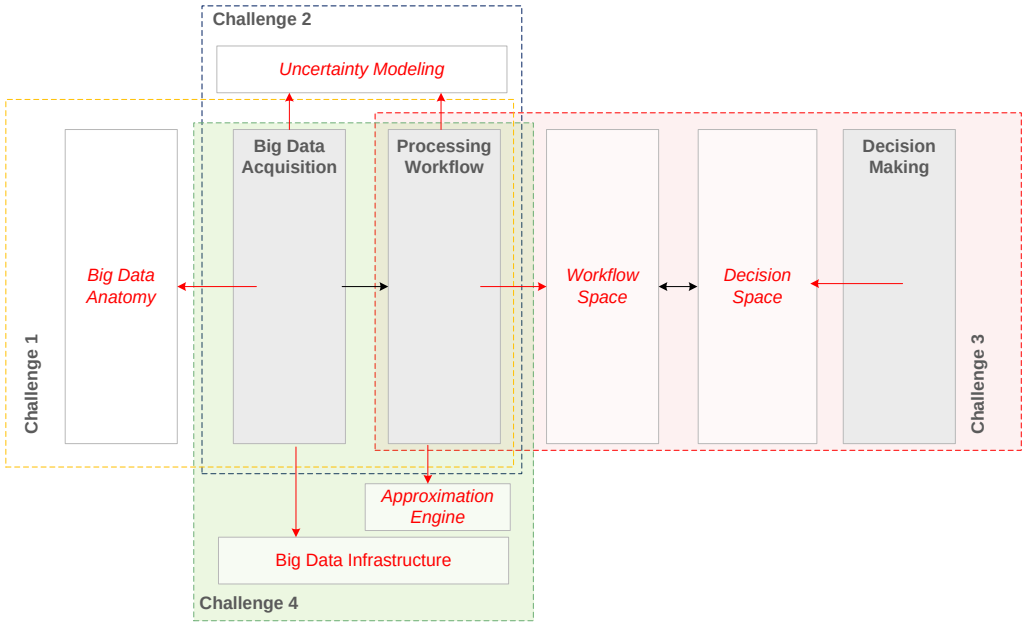


Figure 1-3 A process Map of resilient data analytics methodology in disaster response

1.3 Challenges and Research Contributions

The proposed framework addresses four challenges related to efficient use of large geospatial data sets during extreme events.

Challenge 1: Lack of clear understanding on the basic structure of big visual disaster data and their role in disaster management

Analysis of big visual disaster data offers tremendous opportunities in improving our understanding, modelling, and prediction of the impacts of coastal hazards on communities and ecosystems. While the big visual disaster data has widely adopted in routine coastal resilience applications, the role of this data in disaster management is controversial. It is widely agreed that big visual disaster data contains indispensable disaster information that can be integrated into ongoing decision processes, however, it is arguable whether such data is central or peripheral because the unclear structure of big visual disaster data cast doubt in the effective and efficient interpretation. To fully exploit the merit of big visual disaster data, it is necessary to revisit and characterize and the structure of the big spatial data in disaster situation awareness to make information available on time and at relevant level of decision-making in disaster management.

Challenge 2: Lack of understanding on the quality of big visual disaster data

The quality of big visual disaster data is rife with uncertainty (Fisher 1999). This uncertainty in data quality not merely refers to data accuracy (or error), but also includes other characteristics such as lineage, goodness of fit for designated applications, etc. Data quality is always an issue in big visual disaster data related disaster response applications. Therefore, handling big visual disaster data requires the proper accommodation of uncertainty in data quality. Poorly handling of uncertainty, at best, result in inaccuracy of the information, and at worse, result in fatal errors. Awareness of data quality is principal

for both data processing and decision-making. In data processing, it requires data quality analysis to provide prior knowledge of whether the processing results are good to use or just a waste of time. On the other hand, in decision-making, the data quality determines how much trust decision-makers or experts are able to place in the information, and consequently determine the merits of the information. To this end, performing a comprehensive data quality analysis is equally important as processing big visual disaster data.

Challenge 3: Lack of formal modelling of processing goals, computational workflows in a distributed computing environment, and the coordination of decision making and computational workflow

One of the significant challenge for using big visual disaster data in coastal resilience application is coordination of decision making and computational workflow. This coordination requires a closing loop between decision-making and data processing. From the experts or decision-making perspective, it needs insights: key signals and tightly packaged summaries of relevant, intriguing disaster information. On the other hand, from the data processing perspective, it urges a clarified, well-defined goal, which they can convert to a series of feasible computation tasks. Currently, there is a huge shaded area between this decision-making and data processing: there is lacking of formal modeling of processing goals, computational workflows in a distributed computing environment, and the coordination of decision making and computational workflow. Understanding of both technology and vernacular of decision-making is difficult. Mapping technology capabilities to vernacular of decision-making goals is even more complicated.

Challenge 4: Adaptive processing in time-sensitive applications

Disaster response involves making difficult decisions within a short time window. This determines that information extraction from big visual disaster is time sensitive in the same manner (Lippitt et al. 2014). The importance of getting timely information during natural disasters has recently motivated a nationwide survey of many U.S. emergency response organizations to understand the relationship between value and lag-time of the information in disaster response (Hodgson et al. 2014). In disaster, the merit of big visual disaster data diminishes rapidly as time goes on. Different from the routine processing that emphasis on maximizing performance (e.g. accuracy), disaster response applications allows a sacrifice of performance in trading for speed to meet the strict time budget. To this end, anticipating adaptive mechanism that could adjust the processing to the time budget remains has profound meaning.

In summary, the abovementioned challenges can be concluded as two primary limitations in current data analytics methodology (1) there is lacking of clear understanding in the inherent characteristics (data structure and data quality) of big visual disaster data; (2) there is lacking of adaptive mechanism that can adjust the processing workflow to time-sensitive decision makings. Therefore, to address these limitations, the central motivation of this study is to develop a resilient data analytics methodology that could fully exploit two major needs of "adaptability": (1) adapt big visual disaster data to time-sensitive application; (2) adapt data processing workflow to decision-making needs. In doing so, such research will have far-reaching implications on operationalizing the community resilience concept and eventually promoting the community resilience in the real world.

1.4 Dissertation Organization

Two key elements of the “adaptability” during disaster response including (1) situation awareness; and (2) adapt to the situation. For the first element, the using of big visual disaster data is proposed to improve situation awareness. To clearly explore the potential of big visual disaster data, chapter 2 investigates the feasibility of using big data for situation awareness. For the second element, chapter 3, 4, 5 will look deep into what mechanism can enable the “adaptability” to disaster situation. Chapter 3 address the question of How does the uncertainty associated with spatial disaster data sets propagate through a given processing pipeline? The objective of chapter 3 is to design an uncertainty quantification framework that enables the computers to balance between processing time and quality of results to meet stakeholders’ information requirement in time-sensitive applications. Chapter 4 seeks to formally model stakeholder information needs during disaster response such that data processing workflow can be autonomously composed for spatial big disaster data. Chapter 5 designs and implements fast computing techniques to promote “adaptability” performance. In each chapter, the proposed methodology will be validated using the empirical data sets collected during Hurricane Sandy in selected disaster analytics applications.

Chapter 2 Big Spatial Data in Coastal Resilience Applications: Anatomy, Scientific Workflow, and Processing Requirement

2.1 Introduction

Nearly 500 natural disasters occur each year in the world. While the causes of them vary from event to event, what they share in common is that each time they strike, these unexpected events cause loss of lives, profound economic loss, social disruption, and so. Coastal communities possess one of the most dynamic interfaces between human civilization and the natural environment. In the worldwide, over 38% of the human population lives in the coastal zone, with over half of the human population live in coastal counties in the United States. The coastal communities in the United States have been recently severely impacted by several major hurricane events including Hurricane Katrina (2005), Hurricane Sandy (2012), Hurricane Arthur (2014) and the recent Hurricane Mathews (2016). For example, Hurricane Sandy (2012) wreaked havoc and destruction across the Atlantic coastal area spanning 24 states, flooding and destroying residential houses, knocking down power lines systems, paralyzing transportation system – not to mention the shattered lives of hundreds of thousands of people.

With the climate system becoming increasingly aggressive, future storm events are likely to raise the cost in coastal areas in a variety of ways. First, the impacts of climate change are likely to worsen the problems that coastal areas already face. Sea level rise, increase in precipitation and changes in frequency and intensity of storm events are increasingly exposing the vulnerability of coastal areas. Second, future coastal development may reduce the resilience of natural systems to respond to storm events. Increase in population, construction of physical infrastructures as well as growing human activities are prone to disrupt natural coastal and marine ecosystems. Eventually, the

fundamental weakness of humankind comes to the fore every time a storm comes, and the formidable force of nature topples the strongest of a person.

To cope with the continuing storm events and the inherent vulnerabilities in coastal communities, enhancing community resilience is essential. Building resilience requires improvement in “adaptability.” Improving the “adaptability” of community resilience needs real-time situation awareness to deal with dynamic threats. It is critical for experts, decision makers, and emergency personnel to have access to real-time information in order to assess the situation and respond appropriately (NSF 2015). The need for information is intensified in these small probability events because predicting the impact of these events is under normal circumstance impossible using either human experience or historical data. Han et al. (1998) highlight the importance of real-time information. They stated that to cope with disasters, a system needs to modify its strategy immediately according to real-time disaster situations. Nevertheless, disaster scenarios are under normal circumstance impossible to be fully enumerated by domain experts without rapidly updated information.

While there is a global consensus on the need for timely information, the merit of spatial data is not fully recognized until recently as more spatial data are being collected and analyzed during extreme events. Though the phrase “80% of data is geographic” is arguable (Garson and Biggs 1992), most disaster-related data have some kind of geospatial features. Due to the lack of efficient methods to extract information from these data sets, traditional ways of using these data sets often give up the spatial components and attempt to tailor the data to transactional database structure despite the fact that these spatial components may contain essential information for decision making.

There are fundamental challenges in managing, analyzing, and interpreting the growing size and complexity of spatial disaster datasets. The vast size and complex processing requirements of these new data sets make it challenging to be utilized effectively in real-life scenarios. For example, during large-scale coastal storm events, crucial information is often hidden in these data sets and is in no way integrated into ongoing decision processes. To fully exploit their potential, we need to revisit existing spatial data analytics and develop new capabilities to rapidly synthesize information from data at rest (i.e., data already stored in the system) and data-in-transit from sensors and make information available on time and at the relevant level of decision-making. This study uses immense quantities of spatial disaster data collected during Hurricane Sandy as the empirical datasets to characterize the anatomy of big spatial disaster data and to analyze the processing patterns in using these data sets in disaster management applications. This chapter addresses the following critical research questions: (1) What is the basic anatomy of spatial disaster data? ; and (2) What are the core operation categories and processing patterns with spatial disaster data? The answers to these questions advance our understanding on the basic principles and processes involved in big spatial disaster data-driven disaster response and recovery as well as the critical research needs for improving data support for disaster recovery decision-making.

2.2 The Role of Spatial Data in Coastal Resilience Applications

2.2.1 Disaster management cycle

In general, disaster management, structured as disaster management cycle in Figure 2-1, consists of four phases including mitigation, preparation, response, and recovery. Although, the cycle can be considered as a continuum (Joyce et al. 2009). Usually, the mitigation phase is considered as the first phase. Mitigation incorporates all

possible activities to enhance the inherent “capability” of a community through reducing the chance of a hazard from happening, preventing a hazard from forming a disaster, and minimizing the damaging effects of the inevitable disaster. This is achieved through the process of risk identification, assets structure modification, etc. Preparedness is another phase that emphasizes on development of community “capability” to deal with the potential hazard. This phase includes readiness planning and activities to handle a disaster in the realization of residual risks. Objectives in the preparedness phase include identification and development of necessary systems, skills, and resource before hazards occur. The next phase is response phase, which is closely relevant to the “adaptability” behavior of a community. This phase focuses primarily on enforcing operations to protect life and property during disasters. Typical activities during the response phase include evacuation, search and rescue, establishing immediate emergency shelters, etc. Finally, recovery is the phase that deals with the aftermath of a disaster. Recovery phase includes both short-term tasks such as restoration of lifeline essentials and the longer-term tasks such as rebuilding of communities. Moreover, the recovery phase can be considered as the summary of a previous disaster that will have far-reaching implication in building the long-term “capability” of the next coming hazard (Becker et al. 2008).

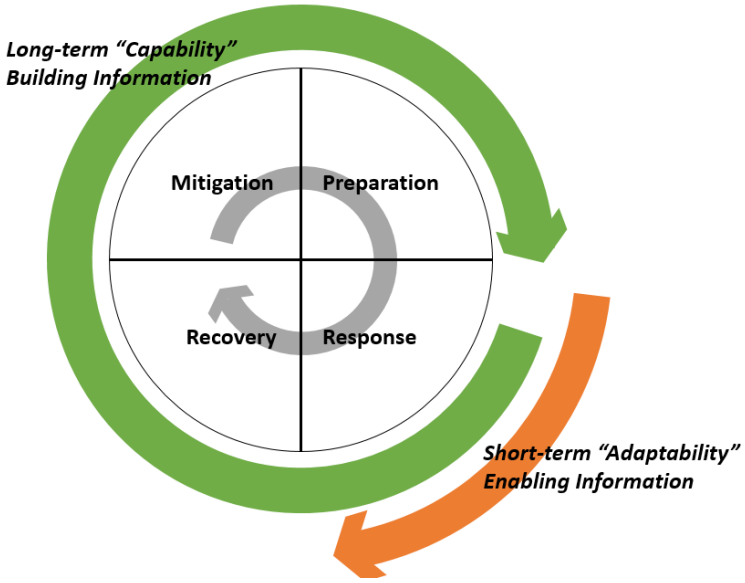


Figure 2-1 Disaster Management Cycle

Table 2-1 Key categories of information need in disaster management

Category	Applicable Phases	Key features	Objective
Long-term "capability" building information	Mitigation, Preparation, Recovery & response	Resolution sensitive and spatially relevant	Provide baseline information for risk identification, damage impact prediction, community system monitoring, community recovery planning, and disaster response
Short-term "adaptability" enabling information	Response	Time sensitive, spectrally-relevance, acceptable spatial relevance	Provide immediate damage assessment information for disaster operations

What these four disaster management phases share in common is that they all need some kind of information and knowledge to reinforce corresponding actives and operations. In today's increasingly networked, digitized, sensor-laden, information-driven world, big data provide new insights into disaster management. The objective of using big data in a disaster is to provide the appropriate information in a temporally, and spatially relevant context (Joyce et al. 2009). In the light of this, big data need to be adapted to meet the requirements of all four phases. In general, the need for information

in disaster management cycle can be divided into two aspects (Table 2-1): long-term "capability" building information and short-term "adaptability" enabling information.

2.2.2 Data acquisition

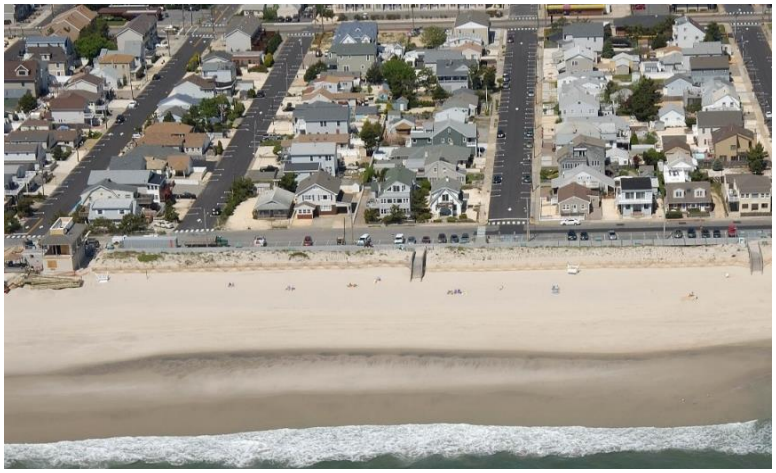
To cope with the information need for the disaster management cycle, different data acquisition and processing technologies are often deployed to gather information. Baseline ancillary geospatial data are historical geospatial data that previously collected and stored in transactional databases. These data provide essential baseline information including but are not limited to, demography information, terrain elevation, land use, building footprints (Figure 2-2), critical infrastructure information (Figure 2-3), and so on. The value of ancillary geospatial data in resilience analysis is widely recognized such as in flood inundation modeling (Sanders 2007) and behavior models for disasters evacuation (Chen et al. 2011). However, these data are static and carefully prepared for general purposes: they sacrifice processing efficiency for details, and they do not convey real-time situation information about the impacted areas. Therefore, under normal circumstances, these ancillary spatial data are potent in predictive analysis, but not equally effective in validation or interactive analyses. Validation models and interactive decision-making often urges the acquisition of data with high frequency and rapidness (Miyazaki et al. 2015). To this end, many remote sensing techniques are now capable of facilitating the capturing of real-time spatial data.

Figure 2-2 Ancillary geospatial data showing building outline polygons

Figure 2-3 Ancillary geospatial data showing main and distributed pipeline network

Imagery data is one of the most common types of data collected during natural disasters. Real-time or near real-time observations from satellite, aerial, and ground platforms serve as essential means for imagery data collection for enhancing resilience (Gillespie et al. 2007), vulnerability analysis (Walker 1996), and decision support (Voigt et al. 2007). Satellite imagery data (Figure 2-5) is most efficient in terms of capturing the terrain condition of a spacious area (Miyazaki et al. 2015). Since the repeat interval of most satellites is often daily to monthly (Gillespie et al. 2007), it is most suitable for monitoring and modeling relationship between human activity and long-term environmental or climate impact (Miyazaki et al. 2015). Compared to the lengthy revisit

time of satellite imagery methods, aircrafts or drones are often ready to be deployed to capture either vertical or oblique imagery (Figure 2-4) in a timely manner. Such rapidness contributes to the prompt situation awareness in early impact analysis (Hirokawa et al. 2007). Meanwhile, oblique imagery often has a higher resolution than the satellite imagery data, making it more accurate for supporting post-disaster assessment (Ezequiel et al. 2014).



a. Oblique Imagery in Ortley Beach Before Hurricane Sandy

b. Oblique Imagery in Ortley Beach After Hurricane Sandy

Figure 2-4 Oblique Imagery in Ortley before and after hurricane sandy (Data from USGS)

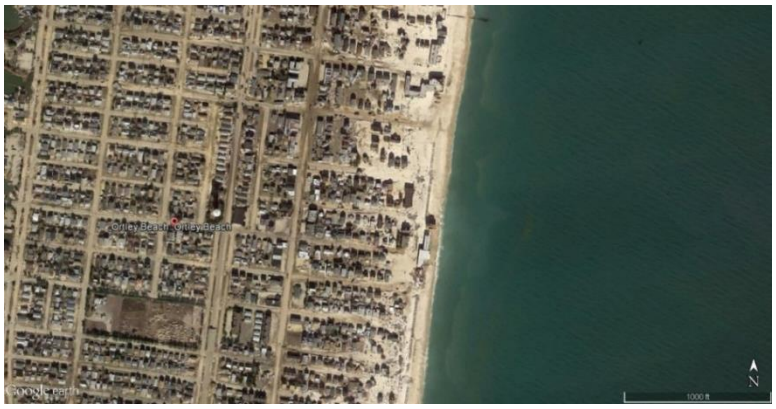


Figure 2-5 Satellite Imagery in Ortlety Beach after Hurricane Sandy

Besides 2D image processing applications such as segmentation (Shi and Malik 2000), another process called photogrammetry (Figure 2-6) attempt to extract the third dimension or depth information from the 2D images (Zhu and Brilakis 2007). It has been adopted to recovery the 3D spatial information of buildings (Suveg and Vosselman 2004), infrastructures (Brilakis et al. 2011), underwater structures (Beall et al. 2010), and indoor scenes (Izadi et al. 2011). Furthermore, a method derived from photogrammetry method is videogrammetry in which cameras continuously record video frames in a given time span and the sequential characteristic such a recording mechanism allows feature points in consecutive frames to match automatically. Notwithstanding the advantages of using imagery-based data, they have their limitations. In particular, the quality of imagery-based information analysis is sensitive to environmental conditions such as light, weather or relative positions as well as the functionality of processing software programs (Dai and Lu 2010). Suitability of such techniques in terms of spatial accuracy is still a major concern (Bhatla et al. 2012).

Figure 2-6a. High-resolution 2D imagery data by ground survey team

Figure 2-6b. 3D model (SURE) reconstructed based on the High-resolution 2D imagery

Figure 2-6c. 3D model (123D CATCH) reconstructed based on the High-resolution 2D imagery

Figure 2-6 3D Reconstruction using 2D imagery data

(Adopted from Zhou et al. (2015))

Light Detection and Ranging (LiDAR) is another emerging technology that facilitates the collection of spatial data in a disaster environment. This technology is built on the principle of the time of flight: a laser scanner automatically records the scanning angels and the travel time (from the emission of a pulse to its return) of each probed objects. Hundreds of thousands of point measurements in the form of x, y, z coordinates and intensity values can be saved each second, and accumulatively forming 3D shapes as point clouds. LiDAR data have multiple benefits. First, LiDAR uses active sensing mechanisms, eliminating the need for ambient light to operate. As a result, it can be functional at night time or in other not so ideal light conditions. Second, it provides data in better spatial accuracy compared to imagery approaches (Csanyi and Toth 2007). Third, the LiDAR instruments can be mounted onto different platforms such as mobile

and aerial platforms, and their data coverage rate can reach hundreds of miles per hour. On account of these three advantages, LiDAR technologies accommodate the needs of mapping both large-scale objects such as terrain (Lefsky et al. 2002) (Figure 2-7) and small-scale objects such as buildings and vegetation (Gong et al. 2012). This technology has been widely deployed in an earthquake (Olsen et al. 2012), flooding (Poulter and Halpin 2008) and forest fire (Morsdorf et al. 2004), and hurricane (Gong et al. 2012). Apart from these benefits, one significant concern with LiDAR data is their complex and intensive computation needs in data indexing and storage (Schön et al. 2009), visualizations, and feature extraction (Rottensteiner and Briese 2002). The complexity of these algorithms often results in lengthy computation time, which will eventually hamper the effective utilization of LiDAR data.

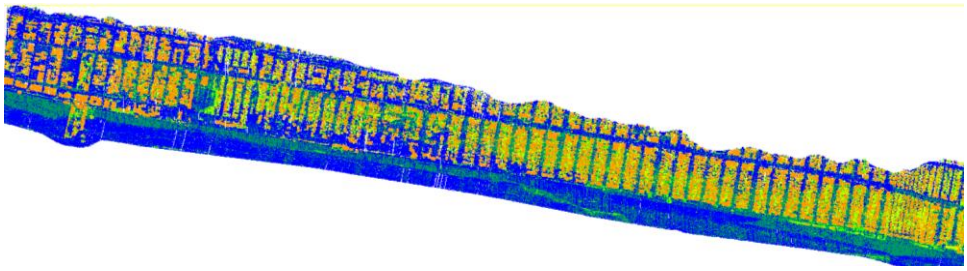


Figure 2-7 Airborne LiDAR data
(color by classification, USGS)

More recently, mobile devices, as well as advancement in social media, have provided new paths for citizens or individuals to generate non-expert spatial data. Average citizens even without domain knowledge are able to record geotagged disaster situation (Goodchild and Glennon 2010) and contribute to the collection of a new type of spatial disaster data called Volunteered Geographic Information (VGI). This has increasingly become a primary communication channel between citizens and public authorities (Miyazaki et al. 2015). It empowers authorities or experts to hear from

"people's voice" though "crowd voting." "People's voice" provides a unique gateway for decision makers and specialists to aware the disaster situation through feeling sentiments of victims. For example, by analyzing the tweets, researchers are able to promptly identify the disaster-prone area (Earle et al. 2012) and conduct early impact analysis (Sakaki et al. 2010). Despite the merit of using crowdsourcing data, the challenge remains significant. Because, these spatial data are from non-authority sources, the quality of such data remains questionable (Allahbakhsh et al. 2013). For example, researchers have studied the impacts of fake and incorrect information during hurricane sandy (Gupta et al. 2013). They concluded that the artificial error could pose significant threats to the credibility of the "crowd voting."

2.2.3 Challenges and opportunities

Geospatial data have frequently been used in vulnerability assessment, which is closely related to identifying gaps in a community's capabilities to cope with extreme events. Although the use of emerging large geospatial data sets in these types of analysis is difficult, the analysis can be eventually accomplished given sufficient time. What most challenging is to use these data sets to obtain better situation awareness in time-sensitive applications. Current geospatial data analysis frameworks are inadequate in handling these large datasets, especially during large-scale extreme events.

Figure 2-8 depicts a simplified map of current data analytics for disaster response practice that essentially involves three key processes including data acquisition, data processing, and decision-making. In the first step, data are collected from multiple resources such as mobile platforms, airborne systems, and social media outlets. The next step is data processing, which analyzes the data according to the particular goals such as

detecting the morphology changes of the dune using LiDAR digital elevation map (DEM) data or identifying the disaster-prone area using twitter data, etc. The final step is decision-making, in which experts identify and choose alternative response operations including mobilizing resources or planning search and rescue operations, etc. Notably, in practice, the connections between these three processes are one-directional: the data acquisition step collects data and pushes them to the data processing step, and the data processing step delivers processing results to the decision making step. In other words, there is no formal feedback loop in the system. The lack of feedback mechanisms causes the decision makers to have little control over the specific tasks to be processed and the corresponding time requirement. In normal situations, the lack of feedback mechanism between these steps can be compensated by performing data collection, data processing, and decision-making in an iterative fashion until all information is obtained. Waiting for all information to be ready is highly infeasible in disastrous situations because of the necessity of making quick decisions in a dynamically changing environment. In contrast to normal situations, collection and processing of large geospatial datasets during extreme events require careful coordination and integration with decision-making processes. The overall challenges related to efficient use of large geospatial datasets during extreme events are summarized as below:

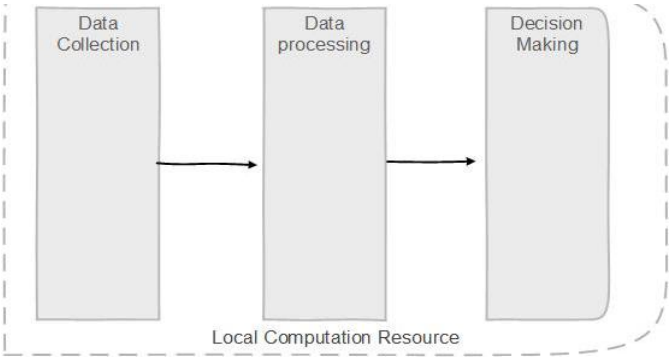


Figure 2-8 A simplified map of current data analytics for disaster response practice

Challenge 1: Lack of clear understanding on the basic structure of big visual disaster data and their role in disaster management

Analysis of big visual disaster data offers tremendous opportunities in improving our understanding, modeling, and prediction of the impacts of coastal hazards on communities and ecosystems. While the big visual disaster data has widely adopted in routine coastal resilience applications, the role of this data in disaster management is controversial. It is widely agreed that big visual disaster data contains indispensable disaster information that can be integrated into ongoing decision processes, however, it is arguable whether such data is central or peripheral because the unclear structure of big visual disaster data cast doubt in the effective and efficient interpretation. To fully exploit the merit of big visual disaster data, it is necessary to revisit and characterize and the structure of the big spatial data in disaster situation awareness to make information available on time and at relevant level of decision-making in disaster management.

Challenge 2: Lack of understanding on the quality of big visual disaster data

The quality of big visual disaster data is rife with uncertainty (Fisher 1999). This uncertainty in data quality not merely refers to data accuracy (or error), but also includes other characteristics such as lineage, the goodness of fit for designated applications, etc. Data quality is always an issue in big visual disaster data related disaster response applications. Therefore, handling big visual disaster data requires the proper accommodation of uncertainty in data quality. Poorly handling of uncertainty, at best, result in inaccuracy of the information, and at worse, result in fatal errors. Awareness of data quality is principal for both data processing and decision-making. In data processing, it requires data quality analysis to provide prior knowledge of whether the processing

results are right-fit to use or just a waste of time. On the other hand, in decision-making, the data quality determines how much trust decision-makers or experts can place in the information, and consequently determine the merits of the information. To this end, performing a comprehensive data quality analysis is equally important as processing big visual disaster data.

Challenge 3: Lack of formal modeling of processing goals, computational workflows in a distributed computing environment, and the coordination of decision making and computational workflow

One of the significant challenges for using big visual disaster data in coastal resilience application is coordination of decision-making and computational workflow. This coordination requires a closing loop between decision-making and data processing. From the experts or decision-making perspective, it needs insights: key signals and tightly packaged summaries of relevant, intriguing disaster information. On the other hand, from the data processing perspective, it urges a clarified, well-defined goal, which they can convert to a series of feasible computation tasks. Currently, there is a huge shaded area between this decision-making and data processing: there is lacking of formal modeling of processing goals, computational workflows in a distributed computing environment, and the coordination of decision making and computational workflow. Understanding of both technology and vernacular of decision-making is difficult. Mapping technology capabilities to vernacular of decision-making goals are even more complicated.

Challenge 4: Adaptive processing in time-sensitive applications

Disaster response involves making difficult decisions within a short time window. This determines that information extraction from big visual disaster is time sensitive in the same manner (Lippitt et al. 2014). The importance of getting timely information during natural disasters has recently motivated a nationwide survey of many U.S. emergency response organizations to understand the relationship between value and lag-time of the information in disaster response (Hodgson et al. 2014). In disasters, the merit of big visual disaster data diminishes rapidly as time goes on. Different from the routine processing that emphasis on maximizing performance (e.g., accuracy), disaster response applications allow a sacrifice of performance in trading for speed to meet the strict time budget. To this end, anticipating adaptive mechanism that could adjust the processing to the time budget remains has profound meaning.

2.3 A Hurricane Sandy Inspired Big Data Framework for Coastal Resilience Investigations with Heterogeneous Spatial Data

2.3.1 Geospatial Response to Hurricane Sandy

Driven by the abovementioned three primary needs, this study presents a Hurricane Sandy Inspired Big Data Framework for Coastal Resilience Investigations with heterogeneous spatial Data. Like during many other extreme events, geospatial products and tools are an essential part of every stage of disaster management during Hurricane Sandy, from planning through response, and recovery to mitigation of future events. However, unlike many other extreme events where the available spatial data are often limited in size and type, Hurricane Sandy has seen a surge of massive spatial data sets (Figure 2-9). Table 2-5 shows the type of spatial disaster data, either collected or identified during hurricane Sandy. These datasets are imagery and point cloud data and

can be more broadly defined as low-dimensional, spatiotemporal datasets, in which data elements are defined at points in a 2D/3D spatial coordinate system and over time.

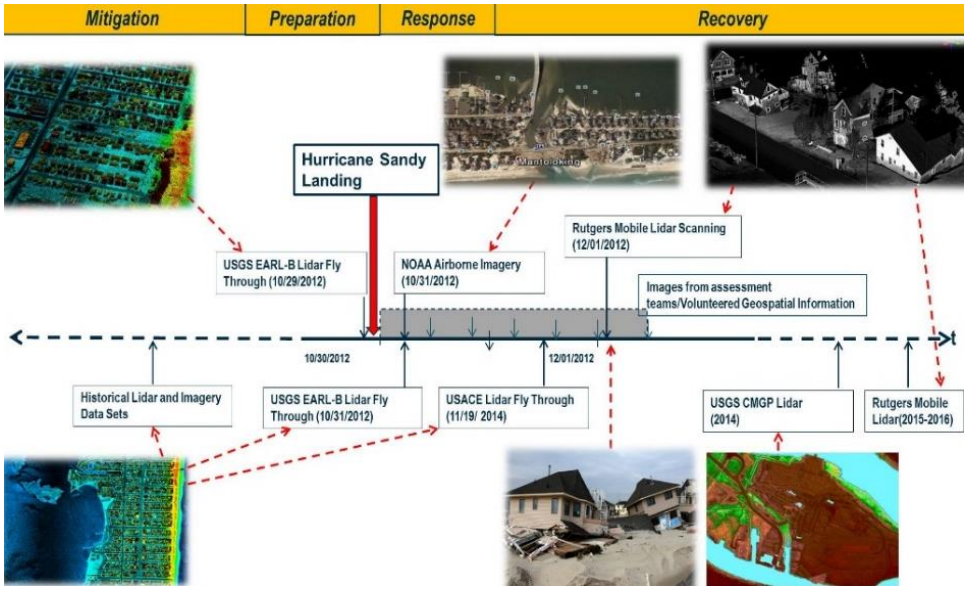


Figure 2-9 Hurricane Sandy related 3D disaster

The specific data sets considered in this study include various airborne LiDAR data sets collected at different points of time before and after the landing of Hurricane Sandy (Table 2-5). First, airborne LiDAR data dated back to 2010 exist for the most of the New York-New Jersey metropolitan area and are archived in data repositories including Digital Coast and USGS Click. Second, on October 29, 2016, the day before Hurricane Sandy landed in New Jersey, the USGS Coastal and Marine Geology Program collected airborne LiDAR data along the New Jersey Coast using its Experimental Advanced Airborne Research LiDAR-B (EAARL-B) system. Immediately after the landing of Hurricane Sandy, NOAA collected airborne imagery followed by USGS EARL-B airborne LiDAR data collection. During the period of November 11-24, 2012, USACE conducted another wave of airborne LiDAR data collection along the New Jersey and New York coastal line. During the period of December 5-9, 2012, Rutgers

conducted mobile LiDAR scanning of severely impacted coastal communities in the state of New Jersey and New York City. Throughout the disaster response period, street-level images of storm damage have also been collected by various damage assessment teams and citizens. Some of them were distributed through social media channels such as Facebook and Twitter. During the disaster recovery stage, more geospatial data sets have also been collected to assess recovery progress and future vulnerability. These data sets include 2014 USGS airborne LiDAR data collection along the coastal lines in the northeast region and mobile LiDAR data collection in Ocean County, New Jersey in 2016. Collectively, these data sets are too massive to be efficiently managed and processed to derive scientific insights into ways of improving coastal resilience. In the following, this study characterizes the basic anatomy of big spatial disaster data to highlight the big data challenge in using these data sets in coastal resilience applications.

2.3.2 Data analytic framework

Existing analytical frameworks for interpreting 3D disaster datasets are insufficient for time-sensitive applications. For example, immediately after the landing of hurricane events, there is a great urgency to process the extensive and heterogeneous point cloud data, dynamically evolving in time as more data come in, and make sound decisions given limited and sparse resources. In this kind of scenarios, emergency response organizations would seek near real-time information about the extent of damages to homes and critical infrastructure systems such as transportation network, healthcare facilities, energy infrastructures, and wastewater treatment facilities. Data analysts in support of these organizations would construct a workflow of analytics steps consisting of selecting available point cloud data sources, querying point cloud data in

geographic regions of interest by defining customized boundaries or overlaying vector data describing the locations of existing facilities, fusing point cloud data from different sources, selecting damage assessment analysis methods such as change detection with pre- and post-event data or classification of facility damages based on mono-event data, and aggregating extracted damage information into inputs to various decision planning models and tools that support search and rescue operations and restoration planning for critical infrastructures. They would also monitor information from social media outlets and potentially use them to prioritize data processing in specific regions. These types of analyses require efficiently querying of giant point cloud datasets, co-registration of heterogeneous point cloud data from different sensors on the fly, and near real-time detection and classification of damages to infrastructure systems and buildings.

Current ways of point cloud data management and analysis also pay little attention to the coordinated use of many interrelated analysis pipelines. For various point cloud data consumers, choices of analytic workflows have to factor in time, stakeholder information needs, and tolerance of errors. In large urban communities, people and infrastructure systems are often highly concentrated to achieve productive proximity. In many post-disaster scenarios, infrastructure stakeholders may have overlapping or interdependent information needs due to the geographic interdependency of infrastructure systems. For example, to reduce the risk of secondary catastrophic events, natural gas pipeline operators would seek information on where is the debris and overwash deposited as this will impact their accessibility to critical gas shutoff valves (Zhou et al. 2016). At the same time, this information is also critically sought by other organizations like FEMA and local emergency response organizations. Therefore, it is clear that analytics

applications should not be designed in a silo. Instead, there is a great need for mechanisms to design flexible and concerted geo-data workflows to maximize their utilities. Unfortunately, there appears to be no framework existing for tailoring and optimizing spatial analytics to needed capabilities in various phases in disaster management.

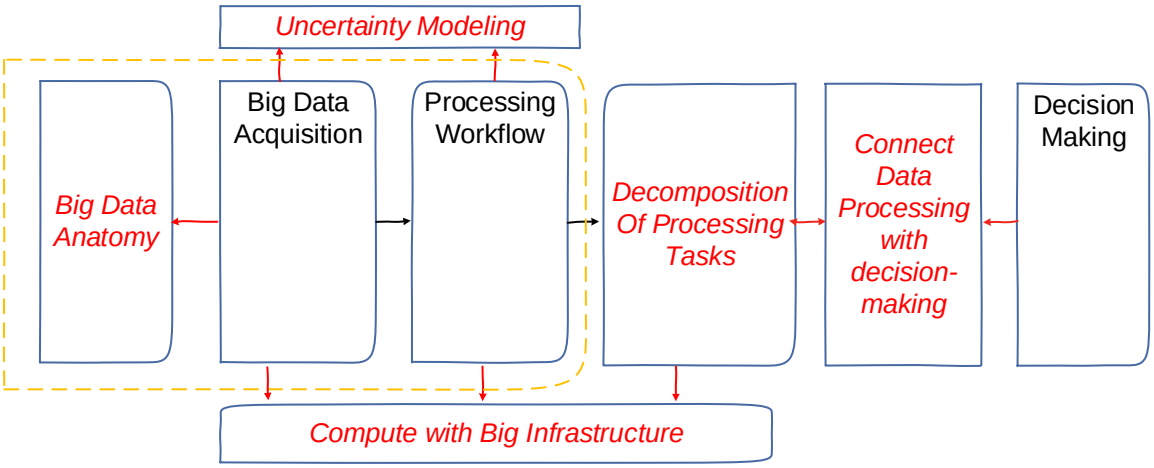


Figure 2-10 Data analytic framework

In addressing to the abovementioned research gap, we proposed a data analytic framework (Figure 2-10) for the timely applications. In addition to the tradition data analytic framework described in Figure 2-8, we added time-sensitive features to the existing data analytic framework. In the following sections, we will describe the five essential elements in the data analytics framework.

2.3.3 Anatomy of Big Spatial Disaster Data

Volume

Although big data do not purely mean the large volume of data, data volume remains a major concern in disaster response and recovery missions, where how large the amount of data generated often determines what kind of protocols to be used for storage

and transferring and how much computation resources are required to process it. Advanced sensors such as mobile LiDAR systems can generate a large amount of data (Table 2-2) each day. For example, the Rutgers Mobile LiDAR survey system, each day (considering 8 hours of data collection), can generate around 43 gigabytes of LiDAR data in the format of pcap file and approximately 1 gigabyte of position file. Another 300 gigabytes of 360-degree imagery data would be captured along with LiDAR scanner. Thus, the sheer volume of the data generated each day from the mobile system alone is around 350 gigabytes. This study systematically analyzed the volume of the Hurricane Sandy related spatial data sets. Table 2-2 provides a quick summary of the volume of these data sets used in disaster response and recovery phases during Hurricane Sandy.

Table 2-2 Volume of Hurricane Sandy related spatial disaster data sets

Survey Type	Data collection date	Data Volume
Archived airborne LiDAR	Archived	29.6GB
USGS EARL-B LiDAR	10/29/12	2.1GB
USGS EARL-B LiDAR	10/31/12	2.1GB
USACE LiDAR	11/19/14	21.2 GB
Rutgers mobile LiDAR	12/01/12	575GB
USGS CMGP LiDAR	2014	105GB
Photos for SFM reconstruction	Streaming	20GB
Rutgers mobile LiDAR	2015-2016	15TB

Data Structure

Data structure type refers to the structure or organization that data come with. In general, there are three primary data structure types including structured data, semi-structured data, and unstructured data. Structured data refers to the data that stored in pre-defined data models such as relational databases or spreadsheets. Most ancillary data are structured data, e.g., Excel spreadsheets, ArcGIS shapefile, etc. Unstructured data refer to data that do not have a pre-defined data model for information extraction. Most of the on-

site remote sensing data are fallen into this category such as LiDAR data. Semi-structured data lies in between structured data and unstructured data. More specifically, it can be considered as a type of structured data, but lack a strict structure imposed by an underlying data model. One example is social media data. The text itself is structured data, but the time tag and location, as well as other information, add to the complexity. Given the uncertainty of disaster environment, situation awareness for disaster response requires the integration of data with different structured types.

Table 2-3 Resolution and vertical accuracy of Sandy related 3D disaster data sets

Data Type	Resolution (pts/m2)	Vertical Accuracy (cm)
Archived airborne LiDAR	1-4	36.6
USGS EARL-B LiDAR	1-2	20
USGS EARL-B LiDAR	1-2	20
USACE LiDAR	1-4	8.2
Rutgers mobile LiDAR	1000 - 8000	5
USGS CMGP LiDAR	1-4	6
SFM reconstruction (Zixiang et al. 2015)	500-2000	20
Rutgers mobile LiDAR	2000 - 4000	5

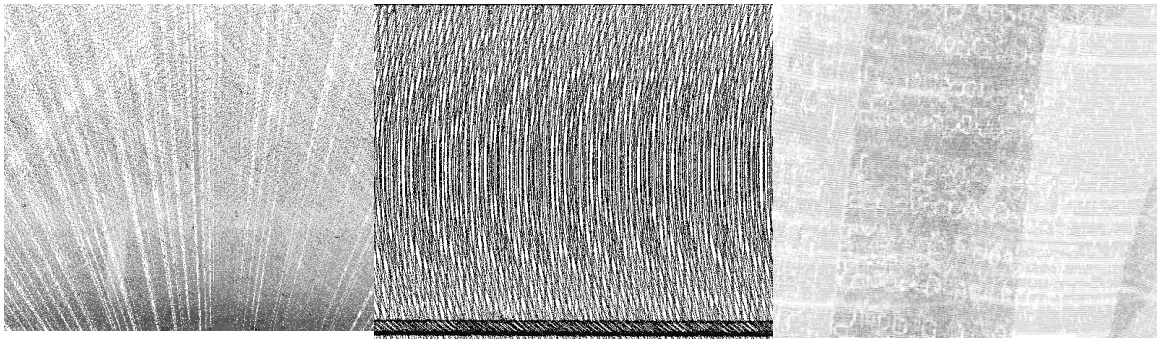


Figure 2-11a. Static Terrestrial LiDAR

Figure 2-11b. Mobile LiDAR System

Figure 2-11c. Airborne LiDAR System

Figure 2-11 Comparison of scanning pattern from different scanning platforms

Spatial Completeness

Beyond volume, variety, and variability, the challenge also arose from dealing with incompletion/inaccurate information. For example, one primary issue in using airborne LiDAR for building damage assessment is their vertical perspectives, which strictly limits their sensor readings to building roofs (Olsen 2013). In contrast, ground-based spatial sensing methods such as mobile LiDAR mostly capture data from the horizontal perspective and inevitably miss part of objects that are not visible from the driving paths. Figure 2-12 depicts a coverage analysis of roof and wall respectively using the Rutgers mobile LiDAR system (MLS).

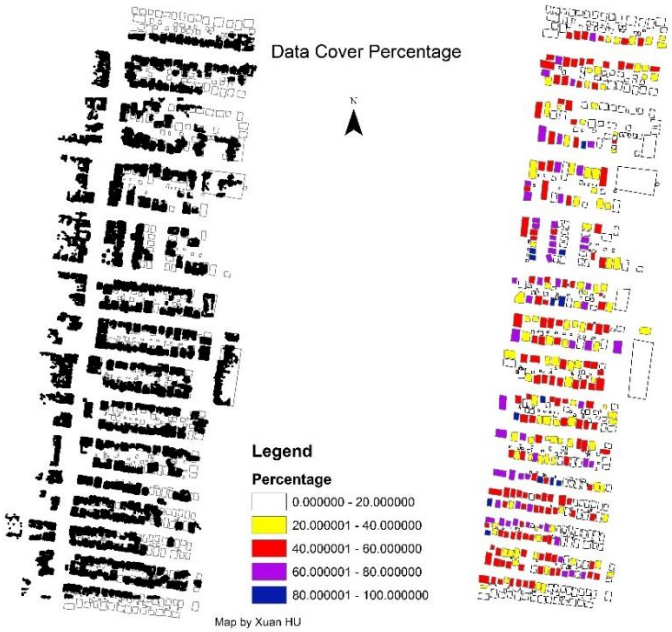


Figure 2-12a. Roof coverage analysis results

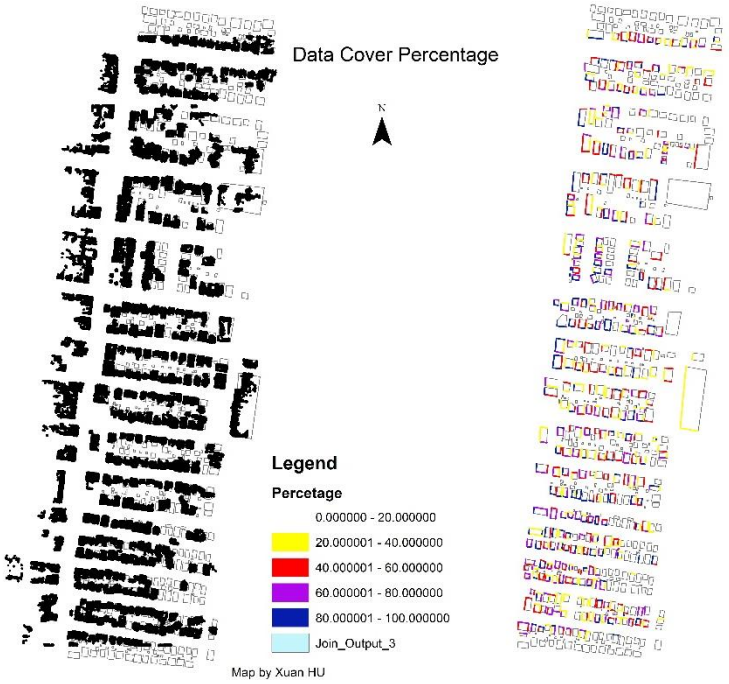


Figure 2-12b. Wall coverage analysis results

Figure 2-12 Roof and wall coverage analysis results

Veracity

The difference in data accuracy is also a common problem in conducting comparative analysis among various spatial data sets. For example, Table 2-4 lists the accuracy of various spatial disaster datasets expressed in terms of Residual Mean Square Error (RMSE). The issue of accuracy also exists in single data sets. For example, this study analyzed to determine the consistency of vertical accuracy in mobile LiDAR datasets. The author compared the accuracy of two datasets: (1) USGS airborne LiDAR dataset; (2) Rutgers Mobile LiDAR dataset. The accuracy of USGS airborne LiDAR data is controlled via a network of ground control points. On the other hand, no control points were used in mobile LiDAR data collection. The accuracy of the mobile LiDAR data is calculated as the difference between the ground surface elevation detected using the airborne LiDAR and the ground surface elevation detected by mobile LiDAR. The comparison is conducted in various environmental conditions including urban

environment (New Brunswick downtown), shoreline residential communities (Normandy Beach and Ortley Beach) and important infrastructures (Rockaway Bridge), which are representative for the dense population area. The analysis results show that mobile LiDAR performs better in open shoreline area (Table 2-4). The urban high-rise buildings, vegetation, traffic signals are potential obstructions that shade the GPS signals. Use of these data with different vertical accuracies in disaster response and recovery is a challenging endeavor as many analyses involve change detection between these data sets to detect damaged structures.

Velocity

Velocity is another challenge associated with big spatial disaster data. It does not only refer to how fast data are generated, but also refer to the need of speed in data analysis. Nowadays, spatial data can be collected at an almost unimaginable speed. For example, many mobile LiDAR systems can collect point measurement at 1 million points per second. More tremendous amount of data can be generated in the social media: every minute, more than 100 hours of video are uploaded to YouTube, 5 thousand tweets are sent, and the number appears on the growing. For major disasters, real-time or near real-time spatial data acquisition is critical and feasible. For example, USGS was able to collect the pre- and post-event data for the entire New Jersey and New York shoreline area within a day or two. During Hurricane Sandy, there is also a constant stream of social media based image data capturing the ever dynamic disaster impact. While the importance of real-time data collection is well recognized during disastrous events, extracting meaningful information in a timely manner remains one of the most significant challenges.

2.3.4 Decomposition of processing tasks

While there appear to be a variety of ways to process the large volume of spatial data, the most common ones include generation of digital surface models, feature extraction, and change detection.

Digital Elevation Models

Table 2-4 Accuracy of mobile LiDAR data in different environments

Environment	Locations	Airborne Data	Mean (m)	Standard Deviation (m)
Urban& Residential	New Brunswick Downtown	USGS NJ	-0.073	0.051
Urban& Schools	Rutgers Busch Campus	USGS NJ	-0.123	0.021
Shoreline & Residential	Normandy Beach	USGS NJ	-0.069	0.037
Shoreline & Residential	Ortley Beach	USGS NJ	-0.103	0.02
Urban & Infrastructure	Rockaway Bridge	USGS NY	0.023	0.044

Disaster response can be impeded by the lack of precise terrain information. LiDAR and photogrammetry approaches have the capability to capture elevation data in large areas immediately, and the collected point cloud data can be used to create highly accurate 3D representations of the impacted terrain. This cartographic information plays a vital role in assessing damage, analyzing potential risks. The DEMs can be generated by eliminating the non-ground objects (Meng et al. 2009). A detailed review of different ground filtering algorithms can be found in Meng et al. (2010). Numerous studies have attempted to improve the performance of ground filtering algorithms by deploying approaches such as interpolation-based (Briese and Pfeifer 2001) or morphology-based methods (Kobler et al. 2007). However, these algorithms often require searching for

neighbors, which are computationally expensive (Meng et al. 2010). The time sensitivity issue of the algorithms is only considered by few studies (Liu et al. 2007).

Table 2-5 Different features extracted from LiDAR

Feature Category	Feature type	Reference
Building	3D Models	Rottensteiner and Briese (2002); Verma et al. (2006)
	planar (Building footprint/ Roof Polygons)	Rottensteiner et al. (2005); Awrangjeb et al. (2013) Henn et al. (2013)
Planimetrics	Roadways	Peterson et al. (2008); Olsen (2013)
Infrastructure	Transmission Lines	McLaughlin (2006); Jwa et al. (2009)
	Pipelines	Son et al. (2014)
	Storage Tanks	Fernández-Lozano et al. (2015)
Street 'Furniture'	Street Light	Yu et al. (2015)
	Power/Telco Pole	Jwa et al. (2009)
	Fire Hydrant	Korah et al. (2011)
	Debris	Labiak et al. (2011)
Hydrologic Features	Channel Network	Passalacqua et al. (2010)

Feature Extraction

Object or feature extraction is much more complicated than generating DTMs (Mayer 2008). Prior to feature extraction, segmentation is the fundamental step for exploitation of 3D point clouds (Yang et al. 2015). One of the earliest studies in segmentation of point cloud data is conducted by Henderson and Bhanu (1982), who developed a region growing algorithm using a spatial proximity graph. Rabbani et al. (2006) introduced a widely used k-nearest neighbors method for point cloud applications. However, it is not until random sample consensus (RANSAC) (Schnabel et al. 2007) that the time efficiency is taken into serious consideration. In another attempt to reduce the computation cost, voxels based segmentation algorithm is developed by Lim and Suter (2008). Distinctive features are further extracted by tilling with different parameters using fitting based on algorithms such as local fitting surfaces (LoFS) (Mongus et al. 2014) or

support vector machines (SVM) (Mountrakis et al. 2011). A list of different features extracted from LiDAR datasets is shown in Table 2-5.

Change Detection

Change detection refers to the process of identifying meaningful difference over different observations (Singh 1989). As a result, change detection often requires at least two datasets with overlaps. Among the ten common change detection techniques described in Singh (1989), image differencing method is the most widely used one for LiDAR applications (Trinder and Salah 2012). In this method, two datasets are spatially registered, subtracted, and then pass through a user-specified threshold so that the significance can be identified. This procedure has frequently been employed in disaster situations such as landslide (Hsiao et al. 2004), earthquake (Zhang et al. 2006), and hurricane (Hatzikyriakou et al. 2015).

In this research, major processing tasks are decomposed into different operations. We will target core operation categories employed in applications that synthesize information from spatiotemporal sensor data in our research. The core operations produce different levels of data products that can be consumed by client applications. For example, a client application may request only a subset of satellite imagery data covering the east coast of the US. The operations can also be chained to form analysis workflows to create other types of data products. An example workflow could be a pipeline of: [data cleaning → subsampling mapping → object segmentation → object classification → change detection] operations. Each operation's data access and processing patterns, as well as the composition of the analytics application, are important factors in I/O, communication, memory, and processing overheads. The data access and processing

patterns range from local and regular, to indexed data access, to irregular and global access to data -- please see the third column in Table 2-6. Local data access patterns correspond to accesses to a single data element or data elements within a small neighborhood in a spatial and temporal region (e.g., data cleaning and low-level transformations). Regular access patterns involve sweeps over data elements, while irregular accesses may involve accesses to data elements in a random manner (e.g., certain types of object classification algorithms, morphological reconstruction operations in object segmentation). Some data access patterns may involve generalized reductions and comparisons (e.g., aggregation) and indexed access (e.g., queries for data subsetting and change quantification).

Core Operation Categories

The composition of the analytics applications encapsulates several application-level data processing structures as well. First, original datasets can often be partitioned into tiles or chunks, and several categories of operations in Table 2-6 can be executed on each chunk independently. Spilting large datasets into chunks leads to a bag-of-tasks processing pattern. Second, processing of a single chunk or a group of chunks can be expressed as a hierarchical coarse-grain data flow pattern (Beynon et al. 2001; Plale and Schwan 2000; Tan et al. 2010). For example, transformation, filtering, mapping, and segmentation operations can be composed as a workflow. The segmentation operation itself may consist of a pipeline of lower level operations as well. Third, several types of operations such as aggregation and classification can be represented as MapReduce style (Dean and Ghemawat 2008; Dean and Ghemawat 2010) computations. The detail descriptions of each core operation categories are summarized as in Table 2-6.

Table 2-6 Core operation categories in spatial disaster data processing

Core Operation Categories	Example Operations	Data Assess pattern	Computation Complexity
Data Cleaning & Quality Control	Transformations to reduce effects of sensor/measurement artifacts. Transform sensor acquired measurements to domain specific variables.	Mixture of local and global pattern	Moderate computational complexity.
Low-Level Transformations	Transformations of a dataset to another format. E.g., coordinate transformation (such as UTM to GCS), value conversion (such as. RGB to grayscale conversion), or as geometry transformation (3D to 2D projection).	Mainly local pattern	Low to moderate, mainly data-intensive computations
Data Subsetting, Filtering, Subsampling	Select portions of a dataset corresponding to regions in the atlas and/or time intervals. Select portions of a dataset based on value ranges. Subsample data to reduce resolution and data size.	Local as well as indexed pattern	Low to moderate, mainly data-intensive computations
Spatio-temporal Mapping & Registration	Create composite dataset from multiple spatially co-incident datasets. Create derived dataset from spatially co-incident datasets obtained at different times.	Irregular local and global data pattern	Moderate to high computational complexity.
Object Segmentation	Segment “base level” objects such as ground, road, dune, vegetation, and buildings. Extract features from “base level” objects.	Irregular, but primarily local data pattern	High computational complexity.
Object Classification	Classify “base level” individual objects at finer details such as utility poles, building types, and transportation assets through a possibly iterative combination of clustering, machine learning and human input (active learning).	Irregular local and global data patterns	High computational complexity.
Change Detection, Comparison, and Quantification	Quantify changes over time in domain-specific low-level variables, base level objects, and high-level objects. Construct “change objects” to describe changes in low-level domain specific variables, base level, and high-level objects. Spatial queries for selecting and comparing segmented regions and objects.	Mixture of local and global data patterns as well as indexed	High Complexity and data-intensive computations.

Table 2-7 Example application scenarios mapped to the core operation categories

Operation Category	Weather Prediction	Monitoring and Change Analysis	Pathology Image Analysis
Data Cleaning & Quality Control	Remove anomalous measurements from MODIS and convert spectral intensities to the value of interest.	Remove unusual readings. Convert signal intensities to color and other values of interest.	Color normalization. Thresholding of pixel and regional grayscale values.
Low-Level Transformations	Spatial selection/crossmatch to find the portion of a dataset that is corresponding to a given geographic region.	Spatial selection/crossmatch to find portion of a dataset corresponding to a given geographic region	Selection of regions within an image. Thresholding of pixel values.
Data Subsetting, Filtering, Subsampling	Mapping tiles to map projection. Generation of a mosaic of tiles to get complete coverage.	Registering low and high-resolution images corresponding to same regions.	Deformable registration of images to an anatomical atlas.
Object Segmentation	Segmentation of regions with similar land surface temperature.	Segmentation of buildings, trees, plants, etc.	Segmentation of nuclei and cells. Compute texture and shape features
Object Classification	Classification of segmented regions.	Classification of buildings, trees, plants.	K-means clustering of nuclei into categories.
Spatio-temporal Aggregation	Time-series calculations on changing land and air conditions.	Aggregation of labeled buildings, trees, plants into residential, industrial, vegetation areas.	Aggregation of object features for per image features.
Change Detection, Comparison, and Quantification	Spatial and temporal queries on classified regions and aggregation to look for changing weather patterns.	Characterize vegetation changes over time and are	Spatial queries to compare segmented nuclei and features.

Table 2-8 Corresponding tools for the seven core operation categories

Core Operation Categories	Tools in Lastool	Tools in Cloud Compare	Tools in Terrasolid
Data Cleaning & Quality Control	LasControl, LasDuplicate, LasInfo, LasNoise, LasPrecisiom, LasReturn, LasThin, LasValidate, LasView,	Noise Filter, SOR (Statistical Outlier Removal) filter, Remove Duplicate Points, Hidden Points Removal	TerraMatch: Calibration and Strip Adjustment, Tie Lines tools, Match tools (e.g., apply correction, find intensity correction,), etc.;
Low-Level Transformations	Blast2Dem, Blast2Iso, Las2Dem, Las2Iso, Las2Las, Las2Shp, Las2tin, las2Txt, Las2Zip, Shp2Las, LasPublish,	Fit Tool (plane, sphere, 2D polygon, 2.5D quadric), Unroll, Rasterize and Contour Plot, Contour Plot to Mesh	Projection tool (Coordinate Transformations, Geoid adjustment), Convert Storage Format (to kmz, dgn, etc.)
Data Subsetting, Filtering, Subsampling	Las2Las, LasCanopy, LasClip, LasGrid, LasIndex, LasCoverage, LaSsort, LasSpilt,	Subsampling Tool (by random, space, octree)	Point Filtering Tools (by classification, intensity)
Spatio-temporal Mapping & Registration	LasColor, LasTrack, LasPlane	Align (point pairs picking), Match Boundary Box Centers, Match Scales, Fine Registration	TerraPhoto: Camera Calibration Tool, Color Correction Tool, Improving Image Positioning Tool, Color Points, and Selection Shapes Tools, Manage Trajectories Tool.
Object Segmentation	Las2Boundry, LasClassify, LasHeight, LasGround,	Label Connect Component, Cross Section/ Unfold, Section, Facet Detection, RANSAC shape detection	TerraScan: Macro Classification tool (Classify / By intensity; Classify / Surface Points, Classify Using Brush, etc.), Power Lines using Least Squares Fitting, TerraModel (Surface Modeling)
Object Classification	lasclassify, lasheight, lasground,	CANUPO Classification, Cloth Simulation Filter (CSF)	TerraScan: Macro Classification tool (Classify / By intensity; Classify / Surface points, Classify Using Brush and etc.),
Change Detection, Comparison, and Quantification	-	Compute 2.5D Volume	TerraScan Change Detection Tool

Table 2-7 lists examples of application-specific operations as mapped to the core operation categories defined in Table 2-6. All three applications have similar operations, although they use spatiotemporal datasets for different purposes or may handle different data types. Thus, we argue that an efficient framework that can support the core operation categories can benefit a wide range of applications. Table 2-8 depicts Corresponding tools for the seven core operation categories in three major cloud computing software: Lastool, Cloud Compare, and Terrasolid respectively.

2.3.5 Identify the uncertainty associated with big data acquisition and processing

While spatial information (e.g. LiDAR, high-resolution imagery, etc.) facilities the rapid collecting of spatial disaster data for situation awareness, the uncertainty associated with the spatial data could be a defect that devalue the merit of the spatial data. Poor handling of the uncertainty can, at best, lead to errors in the knowledge that the data represent and at worst can bring fatal consequences (Fisher 1999). Therefore, without clear understanding of the uncertainty in spatial data, the value of the big spatial data will be discounted as experts and decision makers may question about the accuracy of spatial data and especially their derived products using fast computing techniques and even reluctant to use it for their judgments (Gahegan and Ehlers 2000). In general, there are two types of uncertainties (Table 2-9).

Raw data uncertainty refers to the inherent uncertainty of the raw datasets. The objective of studying raw data uncertainty is to identify the uncertainty in data acquisition and derive methods for data quality control. Hasselman et al. (2005) defined the studying of raw data uncertainty as the research question of "How accurate is the data perfect representation of the real world?" There are numerous researches on raw data uncertainty

often in the form of uncertainty (Beard et al. 1991; Fisher 1999), error (Goodchild 1994), accuracy (Goodchild 1994), data quality (ISO 2002), etc. Known factors contributing error in LiDAR data include, but are not limited to, errors in LiDAR systems, navigation systems, and system calibration. Mezian et al. (2016) studied the impact factors for system accuracy on two aspects: laser scanner accuracy and navigation system accuracy. The accuracy of the laser scanning accuracy is determined by properties of targeted objects (roughness, reflectivity), scanner mechanism precision (mirror center offset) as well as weather conditions (temperature, humidity) (Mezian et al. 2016; Soudarissanane et al. 2008).

Table 2-9 Summary of different uncertainty analysis needs in different phases

Type	Data Acquisition Phase	Application Phase
Uncertainty Type	Raw Data uncertainty	Uncertainty Trade-off
Objective	Identify the uncertainty in data acquisition and derive method for data quality control	Leverage between data characteristic and algorithm performance (time, accuracy)
Research question	How accurate is the data prefect representation of the real world? (Hasselman et al. 2005)	Is the data accurate enough for the specific application? (Hasselman et al. 2005)
Methods	Data Accuracy/ error analysis	Sensitivity analysis / Uncertainty (error) modelling
Uncertainty Representation	Value (Root Mean Square Error, Standard deviation)	Trade-off

The other type of uncertainty is uncertainty trade-off, which is focused on leveraging between data characteristic and algorithm performance (time, accuracy). Hasselman et al. (2005) defined the study of uncertainty trade-off as “Is the data accurate enough for the specific application?” In real-world applications, it is essential to have a comprehensive understanding on uncertainty trade-off. Understanding the uncertainty

trade-off is crucial in disaster environments where time is limited and computational resources are often constrained. Disaster applications desire not only computational speed but also acceptable uncertainty level. Without properly addressing either the timely or the uncertainty issues, the value of the LiDAR data-derived products will be demerit as experts and decision makers may cast doubt and even reluctant to use them for their judgments (Gahegan and Ehlers 2000). In the context of LiDAR data-derived DEMs, larger interpolation cell size results in DEMs with smaller file sizes and less computation time. Meanwhile, the larger interpolation cell size also potentially increases the information loss while generating these DEMs. Achieving an ideal trade-off between computation time and uncertainty level requires a "frugal" interpolation cell size. In literature, most of the studies focus on the uncertainties of different interpolation methods. A comprehensive study comparing errors caused by different interpolation methods can be found in Bater and Coops (2009). However, very little work can be found on how interpolation cell size is related to the information loss in generating DEMs with LiDAR data. To this end, this study presented the studying of uncertainty-trade off as an essential element in the data analytics framework.

2.3.6 Computing with big data infrastructure

With the time requirement in disaster response, computing infrastructure systems are undergoing fundamental transformations. The first transformation is the shift from CPU computing to GPU computing (Liu 2013). CPU and GPU have rudimentary different design architectures. A CPU is designed for sequential serial processing with sophisticated control logic. Architecturally, a CPU has only a few cores that can handle a limited amount of tasks simultaneously with lots of cache memory. In contrast, a GPU is

designed for massively parallel processing with simple control logic. The ability of a GPU with 100+ cores to process thousands of threads can scale of the processing ability by 100 times over a CPU alone. Therefore, GPUs are especially well-suited for arithmetic parallel tasks. GPU-accelerated computing has now grown into a mainstream movement (Owens et al. 2008). One the one hand, hardware manufacturers such as NVIDIA and AMD, have been transforming GPUs into a computational powerhouse. On the other hand, GPU computing is supported by prevailing operation systems such as Apple (with OpenCL) and Microsoft (using DirectCompute). Concurrently, the exploding GPU capability has attracted more and more researchers to use it to cope with big spatial data. For instance, Yuan (2012) examined the performance of CUDA-enabled Graphics Processing Unit (GPU) for LiDAR processing. The author concluded that the proposed GPU method could scale up a thirty fold speed increase over a similar sequential algorithm. Similar scalability can be found in (Lukač and Žalik 2013) who developed a GPU-based roofs' solar potential estimation algorithm. More systematically, Plaza and Chang (2007) investigated insights as well as the challenge of GPU-based time-Critical Remote Sensing Applications.

The second transformation is the shift from processing batch data to streaming data (or synonym for real-time and near real-time data) and interactive analysis. Batch processing is designed for "data at rest" as a result, might have "medium to high latency" (or a response time from seconds to a few hours). MapReduce is a typical framework for batch processing. The Apache Hadoop framework enables the execution of applications on large computer cluster systems through the implementation of the Map/Reduce computational paradigm. Because of the "medium to high latency," stream processing

comes into being to satisfy fast data needs. Wähner (2014) stated that stream processing is most suitable for processing streaming sensor data. Typical computing framework for stream processing includes Apache Spark and Apache Storm. Another trend for big data analytics is interactive analysis. Interactive analysis or sometimes refers to “human in the loop,” is a set of techniques that combining computation power of machines and with the perceptive and cognitive capabilities of humans, in order to extract knowledge from large and complex datasets. The anticipation of human interaction can be more effective in dealing with unscheduled tasks and unpredictable disturbance. Moreover, it will go beyond the bottleneck of fully automation algorithms. Typical interactive analysis tools include Google’s Dremel, Apache Drill, etc. A summary of the big data analytic tools is shown in Table 2-10.

Table 2-10 A Catalog of Big Data Analytic Tools

Category	Tools	Application Area
Batch processing Tools	Apache Hadoop & MapReduce	Data-intensive distributed applications
	Dryad	Parallel and distributed processing of large data sets using a very small cluster or a large cluster
	Apache Mahout	Large-scale data analysis applications with scalable and commercial machine learning techniques
	Apache Spark	Batch and stream processing of large data sets
	Jaspersoft BI Suite	Report generation from columnar databases
	Pentaho Business Analytics	Report generation from both structured and unstructured large volume of data
	Talend Open Studio	Visual analysis of big data sets
Real-time stream processing tools	Storm	A distributed and fault-tolerant real-time computation system for processing limitless streaming data
	S4	A general-purpose, distributed, scalable, fault-tolerant, pluggable computing platform for processing continuously unbounded streams of data
	SQLstream s-Server	Processing of large-scale streaming data in real-time
	Splunk	A real-time and intelligent big data platform for exploiting information from machine-generated big data
	Apache Kafka	A high-throughput messaging system for managing streaming and operational data via in-memory analytical techniques for obtaining real-time decision making.
	SAP Hana	An in-memory analytics platform aimed for real-time analysis on business processes, predictive analysis, and sentiment data.
Interactive analysis tool	Google's Dremel	A system for processing nested data and capable of running aggregation queries over trillion-row tables in seconds
	Apache drill	A distributed interactive big data analysis tool capable of supporting different query languages, data formats, and data sources.

The third transformation is anticipating cloud computing to cope with big disaster data. As the development of sophisticated sensors tends to generate more data, the capability to process this vast and complex data could easily go beyond the power of regular desktop computers (Dorband et al. 2003). As mentioned in previous paragraphs, GPU computing facilities the fast computing. However, there are some limitations for GPU computing in local computers such as small onboard texture memory. To boost the

capability of GPU computing, one solution is to deploy high-performance clusters. Supercomputers, hosted by either universities or research institutes (e.g., Rutgers Caliburn) or commercial cloud computing companies (e.g., Amazon EC2, Microsoft Azure) can assist in reaping the benefits of GPU computing while avoiding its limitations (Fan et al. 2004). In traditional, high-performance computing (HPC) was the exclusive domain of government agencies. Yet, because cluster computing is proven to have the robust computational capability, relatively low-cost budget (Yang and Chen 2010), more and more agencies begin to implement HPC for their applications especially in dealing with massive data in the time-sensitive environment (Yang et al. 2011). Another method to handle the overwhelming spatial data is edge computing. Different from supercomputers, edge computing, is an optimising cloud computing system aiming at distributing the computation workloads from centralized points to the logical extremes of a network. The edge computing paradigm makes it feasible for the enormous amount of data that collected at the edge devices to be seamlessly processed at the edge as well in an efficient and timely manner. Moreover, state of the art data-driven edge processing framework such as the one proposed by Renart et al. (2017), allows users to define data-driven reactive behaviors that can efficiently exploit data content and location to dynamically and autonomously decide the way data are processed. In all, emerging needs for fast processing big spatial data fostered the development of cloud computing, which is evolutional in accelerating the computations related with information extraction in remote sensing (Lee et al. 2011).

2.3.7 Connecting data processing with decision making models

Disaster response involves humanitarian tasks that require a comprehensive understanding of disaster situation. These humanitarian tasks cannot be accomplished by either decision makers or data processing teams alone. Ideally, decision-makers need to be fully aware of what information can be processed and what kind of information cannot be processed within specific time constraints so that they could determine the optimal job sequence. However, decision-makers or experts might not necessarily have the domain knowledge in data processing; and there exists no cognition model in connecting the data processing tasks and decision-making processes. Notably, for large image datasets, it is difficult to provide an insight of what the information details and uncertainty would be like prior to the data processing. On the other hand, without specification of processing tasks form decision-makers, data processing teams may generate abundant or even worthless information, resulting in wasting of time and computational resource. In sum, an efficient response requires actions of decision-makers, data processing teams as well as collaboration between them. Without losing generify, primary complexes of the collaboration workflow is shown in Figure 2-13.

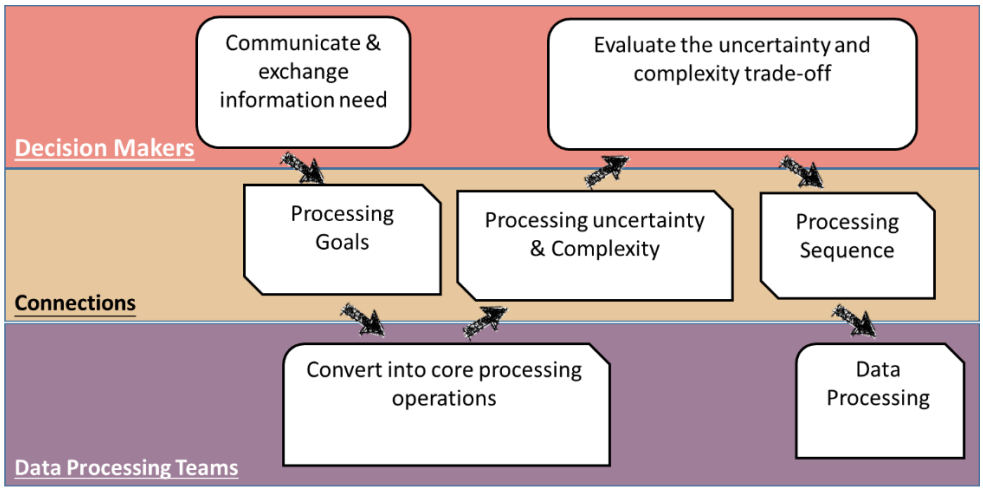


Figure 2-13 Connecting data processing with decision-making models

First, experts or decision-makers from different agencies need to settle down a list of core processing goals and hand it to the data processing teams. Disaster decision-making is a complex system that involves different level information. Estimating the cascading effect of a disaster does not only need information on the destruction of critical infrastructures (e.g., pipeline, power line, dunes, etc) and communications lines, but also intelligence on social, organizational and, economic structures that support the normal functioning of a community (Comfort et al. 2004). There could be conflicting roles structures in the experts and decision-makers (Bharosa et al. 2010). Therefore, it is critical to identify what kind of information are essential for response operations.

Second, data processing teams need to present the uncertainty and complexity of each processing goals to the decision-makers in a way that decision-makers even without domain knowledge can easily understand. One of the significant barriers for experts for job scheduling during a response is lacking sufficient knowledge in the uncertainty and processing complexity. In addressing this issue, the abstract processing goals are broken down into the core operation categories described in Table 2-6. In each operation categories, the uncertainty and complexity are provided as an index system, which enables experts to have a taste of what are the uncertainty and complexity associated with each processing tasks.

Finally, decision-makers need to send feedback to the data processing teams on the data processing sequence. The vast volume of available during a disaster often overwhelms the processing capability of computation infrastructure resulting in information overload. On the other hand, information for disaster response is time-critical (Horan and Schooley 2007). The merit of information in humanitarian relief is not only

determined by the content of the information, but also by the time value of the information. To achieve an effective and efficient response, the data processing need to fulfill the data processing tasks according to the information demand orders. (Starcke and Brand 2012). Under such circumstance, it is indispensable for decision-makers to establish a chronological sequence of data processing jobs (Bharosa et al. 2010).

2.3.8 Discussion

This chapter discussed the four basic challenges related to efficient use of large geospatial datasets during extreme events. The more advanced question is how the decision-making and data processing in extreme events shape the big 3D spatial data processing.

The first issue that needs to explore further is how to extend the crowdsourcing data for extreme events. The power of crowdsourced data may be an asset that could lead to substantial discoveries during such events. However, the use of crowdsourcing during remains regarded with skepticism. Many challenging research questions remain to be addressed. The spread of noise or fake data can have a negative impact on the individual and the decision-making during extreme events (Gupta et al. 2013; Shu et al. 2017). Therefore, fake data detection and information filtering on crowdsourced data become an emerging research topic. Second, handling crowdsourcing information requires unifying different data in a various format (e.g., text, images, 3D spatial data, .etc.) into a standard "readable" format. Most of the crowdsourcing-related studies are focused on Natural Language Processing (Callison-Burch and Dredze 2010; Sabou et al. 2012). It is desired that the crowdsourced data be made more accessible and enrich information as an integrated platform such as web-based (Barbier et al. 2012). Third, the computing and

delivery cost of crowdsourced data remains relatively high. Edge computing could be potentially a solution to lower the computation cost. Data collected at the emulated edge devices can be seamlessly processed at the edge as well in an efficient and timely manner. More importantly, Renart et al. (2017) proposed a content-driven edge computing framework that crowdsourced data quality control is performed in the edge level in a way that only critical information (the information rules that satisfied the rules by users) are submitted to the core.

The second issue is how to deploy advanced machine learning algorithms for extreme events. The recent Alpha go and Alpha go zero have proven the power of the emerging machine learning algorithms, either supervised or unsupervised, in solving the complicated, realistic problems. There are numerous attempts in deploying such machine learning algorithms to handle the spatial data such as LiDAR (Gleason and Im 2012), imagery (Marjanović et al. 2011), and text (Gupta et al. 2013). The remaining problem is how to transplant these advanced algorithms to deal with extreme events. Unsupervised algorithms, such as parametric Classification (Charaniya et al. 2004), clustering (Gupta et al. 2010), are extensively studied in dealing with geospatial data. However, it is arguable that the supervised algorithms are outperformed the unsupervised algorithms (Thomson 1998),. For supervised algorithms, one apparent defect is lacking properly annotated databases to train the data. Geospatial database such as ImageNet (Deng et al. 2009), Semantic 3D (Hackel et al. 2017) does not offer the feasibility of analyzing the disaster-related information. Other extreme events related datasets, either LiDAR or Image, are not annotated. Therefore, establishing a well-annotated extreme events-related database is essential. Last but not least, the existing machine learning algorithms either supervised or

unsupervised are only focused on a small piece of information extraction. Formally formulate information needs during extreme events and identify the decision-making objectives requires further exploration.

2.4 Conclusion

Severe weather events such as hurricanes, ice storms, surge, and flooding have been occurring across the U.S and around the world, threatening places where economic and industrial activities are heavily concentrated. In the face of these natural disasters, building community resilience is essential to reduce the loss of livelihoods, economic cost, and social disruption. These extreme events are now increasing observed and monitored with a loosely coupled network of geospatial sensors. For instance, in recent years, because state and federal agencies have made airborne LiDAR data collection a priority, post-storm LIDAR collection is now routine after large surge event, and the vast amount of disaster data are now freely available online. In another example, emerging high resolution sensing systems such as terrestrial/mobile LiDAR have also been deployed for damage data collection during recent events such as Superstorm Sandy, generating an unprecedented amount of visual disaster data. Lastly, volunteered geographic information, such as geo-tagged disaster photos, is a new breed of disaster data which methods have produced large and heterogeneous spatial disaster datasets spanning multiple spatial and temporal scale and with varying levels of confidence. Analysis of these datasets offers tremendous opportunities in improving the resilience and adaptability of coastal communities in the face of future natural disasters. Despite the high values in these data sets, the vast size and complex processing requirements of these new data sets make it challenging to efficiently use them in coastal community management applications, in particular,emergency situations.

In the second section, we identified two significant purposes for big spatial disaster data: Long-term “capability” building information and Short-term “adaptability” enabling information. To cope with these two different purposes, we revisited the applications of four major types of spatial disaster data: (1)ancillary geospatial, (2) imagery data, (3) LiDAR data and last Lastly, (4)volunteered geographic information. We then summarized the four challenges of using big spatial data for disaster response including. First, there is lacking clear understanding on the basic structure of big visual disaster data and their role in disaster management. Second, the knowledge on the quality and uncertainty associated with the big spatial disaster data remains insufficient. Third, there are few studies on formal modeling of processing goals, computational workflows in a distributed computing environment, and the coordination of decision-making and computational workflow. Finally, specifically for short-term “adaptability” enabling, it urges an adaptive process to adjust data processing to the time bounds requirement.

Driven by the growing needs of deploying spatial disaster data for more efficient response, in section three, we presented our research progresses in designing data analytics frameworks during extreme events to integrate, share, and process these large data sets for an array of critical disaster management tasks. We first characterized the basic anatomy of big spatial data from six aspects: Volume, variability, velocity, data structure, Spatial Completeness, and Veracity. Moreover, to standardize the data processing, we targeted core operation categories employed in applications that synthesize information from spatiotemporal sensor data in our research. The core operations produce different levels of data products that can be consumed by client applications. Then we highlighted the uncertainty issues associated with both big data

acquisition and application phase. A central component of our study is on how to use big data infrastructure to accelerate the processing of the massive amount of geospatial data, in particular streaming data, such that crucial insights can be extracted from the data within a realistic time-bound and time-sensitive decisions can be made to optimize coastal community operations during extreme events. We revisited three undergoing fundamental transformations in computing. Another missing puzzle in the current using of big spatial disaster data is the collaboration between decision-makers and data processing team. We proposed a collaboration workflow to connect the data processing with the decision-making process.

Last but not least, the discussion section pointed two advanced issues in deploying big 3D data during extreme events. The first issue is to promote the using of crowdsourced data for extreme events. The second issues are transplanting the advanced machine learning algorithms for extreme events. We identified two critical needs for this issue including (1) establish well-annotated database; and (2) formulate the information needs.

Chapter 3 Modelling Accuracy Loss in LiDAR-derived Digital Elevation Models

3.1. INTRODUCTION

Digital Elevation Models (DEMs) are 3D representations of the natural and built features on the Earth's surface. DEMs and their derived products provide vital information for terrain analysis based assessments or modeling (Wolock and Price 1994) and applications such as flood simulation and management (Qi and Altinakar 2011) and route modeling (Romanowicz et al. 2008). DEMs are also a critical piece of information sought in disaster response. For instance, major hurricane events often cause massive storm surges, beach erosion, levee breaches, scour, and damage to buildings and infrastructure. Many of these impacts are manifested in dramatic topographical changes, which can be readily identified and quantified from DEMs or DEM-derived products. The data used to generate DEM can be captured using a wide variety of techniques such as Global Positioning System (GPS), topographic survey (Wilson and Gallant 2000), interferometry (Kervyn 2001), and Airborne LIDAR point cloud (Liu 2008). Among these methods, airborne LiDAR is capable of scanning large geographic areas in a short amount of time. Because of this, airborne LiDAR systems are routinely deployed after large storm surge events.

While airborne LiDAR systems have been playing an increasingly important role in disaster events, there are fundamental challenges in deploying LiDAR for DEMs generation in disasters. One frequently asked question is how accurate are the LiDAR-derived DEMs. Like any kind of geospatial data, the value of the LiDAR-derived DEMs and their derived products are heavily dependent on their accuracy (Bater and Coops 2009; Mukherjee et al. 2013). LiDAR-to-DEM conversion often requires a unique

process called interpolation to infill the gaps and handle the duplicate points. Such process introduces non-negligible errors to the original LiDAR data. Without good understanding of the accuracy loss can, at best, lead to errors in the knowledge that the data represent and at worst can bring fatal consequences (Fisher 1999). The importance of studying the accuracy of LiDAR-derived DEMs is amplified when approximation techniques are adopted. Approximation mechanisms, either explicitly or implicitly, are often used to accelerate data processing speed (Pandey and Pompili 2016). Unfortunately, a common side effect of approximation is that it can bring significant accuracy loss to the original dataset. A crucial question in using approximating techniques is that after using approximate computing, will the LiDAR-derived DEMs maintain an acceptable level of accuracy. Without quantification of the accuracy loss, the value will be discounted as experts and decision makers may question about the merit of the LiDAR-derived DEMs and even reluctant to use it for their judgments (Gahegan and Ehlers 2000). As a result, deriving the accuracy loss associated with LiDAR-to-DEM interpolation is equally important as identifying the error source of the LiDAR data.

The goal of this study is to model the accuracy loss in the LiDAR-to-DEM conversion and consequently provide a methodology for choosing the optimal approximation strategy. The focus is on establishing a relationship between the LiDAR interpolation grid size and the accuracy loss so that end-users can select the proper interpolation parameter. The remainder of the article is organized as follows: Section 2 introduces the related work. Description of the model development is described in section 3. In section 4, results and discussion are presented. Section 5 provides conclusions.

3.2 related Work

Errors in LiDAR-derived DEMs can be introduced in two primary phases including the data acquisition phase and the LiDAR-to-DEM conversion phase. In the data acquisition phase, the study of LiDAR accuracy is motivated by the research question: “How accurately does the captured data represent the real world?” (Hasselman et al. 2005). Factors affecting the accuracy of the LiDAR data include (1) the performance of the LiDAR acquisition system and (2) the environmental factors. An airborne LiDAR system consists of two principal parts: a scanner and a navigation system. The scanner determines the precision of the LiDAR point clouds, and the navigation system (e.g., Differential Global Positioning System, Inertial Measurement Unit, and Distance measurement) determines the accuracy of the forward motion trajectory. The accuracies of both parts are often provided by the manufacturers. Prevailing airborne laser scanners such as Rigal (VQ-1560i) can achieve an accuracy of 20 mm. The position accuracy of the commercially available navigation system such as APPLANIX system ranges from 20 to 50 mm after post-processing. In addition to inaccuracy caused by the system hardware design, LiDAR data accuracy is also subject to changes from the environment such as the properties of targeted objects (roughness and reflectivity) and weather conditions (temperature and humidity) (Mezian et al. 2016). Leslar et al. (2014) concluded that environmental conditions could severely affect the accuracy of the LiDAR data. For instance, in cloudy conditions, shading of GPS signal can increase the trajectory error from centimeters to meters. Moreover, to minimize environmental impacts, many methods such as forward-backward processing (Kalal et al. 2010) and strip adjustment (Csanyi and Toth 2007) are proposed. There are extensive research efforts on quantifying the accuracy of airborne LiDAR in different environments

during the data acquisition phase. Many studies agreed that the vertical accuracy of airborne LiDAR could reach 20mm (Hodgson and Bresnahan 2004; May[1] and Toth 2007), though the exact performance varies with the studies. In current airborne LiDAR data acquisition projects, a specification regarding the accuracy of Airborne LiDAR data is often a required component in delivered data sets.

In contrast to the numerous studies devoted to investigating the accuracy in LiDAR data acquisition phase, Cooper et al. (2013) argued that a significant barrier for extending LiDAR in disaster applications is the lack of well-established error standards in LiDAR-derived products. In particular for generating LiDAR-derived DEMs, it often requires a process called interpolation to infill gaps and handle duplicate points during LiDAR-to-DEM conversion (Aguilar et al. 2010). This interpolation process often introduces non-negligible error into the final DEM products. Without a clear understanding of the loss in accuracy during the LiDAR-to-DEM conversion process, end-users may be reluctant to use LiDAR-derived products because they do not have the confidence in whether the DEM is accurate enough for specific applications (Hasselman et al. 2005). The importance of studying the accuracy of LiDAR-derived DEMs is amplified when approximation techniques are adopted. In disaster response or other time-sensitive applications, it is often necessary to deploy approximate computing techniques (Pandey and Pompili 2016) to sacrifice accuracy for computation efficiency. In another word, it is more desired to use a minimum amount of time to generate DEMs with an acceptable level of accuracy rather than to generate highly accurate DEMs using computationally expensive methods (Aguilar et al. 2010). This leads to the great need to

understand to what extent is approximating techniques affect the accuracy during LiDAR-to-DEM conversion.

For LiDAR-to-DEM conversion, two important choices in choosing approximation strategies are related to interpolation methods and interpolation parameters. Many studies conducted comparative analyses on how different interpolation methods affect the accuracy of resulted data products (Arun 2013; Bater and Coops 2009; Polat et al. 2015). However, based on these studies, interpolation methods are not an appropriate indicator for developing approximation strategies. First, all the studies pointed out that nearest-neighbor (NN) method performs better than others in terms of accuracy, but the computational complexity of NN methods has not been studied in detail. Second, the errors caused by different interpolation methods are less pronounced than the errors caused by interpolation parameters (point density, Bater and Coops (2009)). Furthermore, the difference in interpolation methods would not alter the resolution of the DEM, and therefore would not result in changes in data intensity. Therefore, it makes more sense to investigate the approximation strategy based on interpolation parameters.

In literature, among all interpolation parameters, grid size is heavily investigated because it not only affects the DEM accuracy (Bater and Coops 2009; Gao 1997) but also influences the data intensity. For instance, in the study by Bater and Coops (2009), the authors pointed out that there is an apparent increase in the amount of error at coarse resolution (large grid size) than at a finer resolution (small grid size). Nevertheless, prevailing studies provided limited guidance on how to appropriately select the grid size. Hengl (2006) suggested that grid size selection should be based on a number of factors

such as point density, distribution, horizontal accuracy, etc. Similarly, Liu et al. (2007) suggested that DEM resolution (interpolation grid size) must match with point density. Both studies, however, do not provide a connection between the grid size and the accuracy loss. Other studies statistically summarized the possible achievable accuracy given certain interpolation grid size (Aguilar et al. 2010; Erskine et al. 2007). However, such relationship might not necessarily hold for other grid size values. Very little work has been conducted on establishing the relationship between accuracy loss and interpolation grid size. Two exceptions are the works by Aguilar et al. (2010) and Huang (2000). Aguilar et al. (2010) modeled the propagation of error (information loss) due to different point densities. However, it is arguable that the error propagation is only the indicator for precision rather than accuracy. Huang (2000) provided a useful guideline for relating the grid interval with information loss. Nevertheless, it remains unclear on how to link these models to approximation strategies. To this end, this study proposed a model for evaluating the accuracy loss during LiDAR-to-DEMs conversion.

3.3. Research Methodology

This study aims to model the accuracy loss associated with the LiDAR-to-DEM conversion. The detailed methodology of the proposed model is presented in this section. First, built on top of the existing accuracy loss model, this study incorporates the accuracy loss model with two independent variables i and X. Table 3-1 summarizes the detail descriptions of these two variables as well as a list of notations used in this study. Then, the 2010 pre-Hurricane Sandy airborne LiDAR data was adopted to generate both the ground truth data and the test data. The Monte Carlo Simulation (MCS) is adopted for random checkpoints generation in ArcGIS. Then, based on the previous result, a multiple regression analysis is then performed in Matlab to derive the accuracy loss model. Last

but not least, the model built on 2010 dataset is validated utilizing the other three datasets.

Table 3-1 List of the notations

Notation	Meaning
i	Approximation strategy artifact, describes the parameter for approximation
X	Minimum averaging spacing
k	Interpolation method
c	Interpolation Cell Size
ρ	point density
a	Area of the dataset
s	Data size of the LiDAR-derived DEM
z_{LiDAR}	Vertical elevation extracted directly form LiDAR data
$z_{DEM,_{iX}}$	Vertical elevation extracted from LiDAR-derived DEM with strategy artifact i and Minimum averaging spacing X
$RMSE_{interpolate}$	Interpolation Root Mean Square (RMSE)
$RMSE_{interpolate,NN}$	Interpolation Root Mean Square (RMSE) using nearest neighbor (NN) method
$RMSE_{iX}$	Root Mean Square (RMSE) of the model based on strategy artifact i and Minimum averaging spacing X

3.3.1 Accuracy loss model

The accuracy of a DEM is defined as the root mean square errors (RMSE) between the all-possible interpolated heights and the ground truth (ASPRS 2004). In literature, the interpolation accuracy is reported to relate to factors such as interpolation method (Bater and Coops 2009), interpolation grid size (Aguilar et al. 2010; Erskine et al. 2007), point density (Bater and Coops 2009; Gao 1997), terrain slope (Gao 1997), and etc. In this study, in order to cope with the approximation need during time-sensitive applications, the authors introduced an approximation strategy indicator, denoted as i This indicator. In general, i can be calculated as the ratio between the absolute grid size c and the average point spacing ρ.

For analysis, assumed that the interpolation error is a function of three inputs: interpolation method k, point density ρ and interpolation cell size c. Then the relationship is expressed as below:

$$RMSE_{Interpolate} = f(k, \rho, c) \quad (1)$$

However, Bater and Coops (2009), suggested that the accuracy is less sensitive to interpolation methods than the interpolation cell size and proposed that nearest neighbor (NN) is preferred to other methods. Similar studies can be found in Arun (2013) and Polat et al. (2015). Both studies concluded that the NN method is favorite for interpolation of geo-morphologically smooth areas. It should be pointed out that the error caused by the interpolation method k is non-negligible. In this study, it is of particular interest to investigate how the approximation parameter (coarse interpolation cell size) can affect the accuracy performance of LiDAR-derived DEMs. To address this issue, the model is isolated by using the same NN. Then, the new model is represented as:

$$RMSE_{Interpolate,NN} = \varphi(\rho, c) \quad (2)$$

3.3.2 The explanatory variable

In literature, most researchers estimated the accuracy or RMSE of the DEMs based on the equation (2). They derived the accuracy loss function based on point density (Bater and Coops 2009; Gao 1997) and interpolation grid size (Erskine et al. 2007; Ziadat 2007). However, there could be considerable drawbacks when incorporating point density ρ and grid cell c into the RMSE estimation.

First, ρ describes the average level of the data resolution rather than the worst scenario. Nonetheless, it is the worst scenario rather than the average level that limits the overall interpolation accuracy. The average level and the worst scenario can be

equivalent only in datasets that with homogeneously distributed points. For example, there exist two areas a_1 and a_2. Let us assume that a_1 is with extremely low resolution ρ_1 while a_2 has a much higher resolution ρ_2 ($\rho_1 < \rho_2$). Then the point density ρ can be computed as the weighted average density ρ of a_1 and a_2, as $(\rho_1 a_1 + \rho_2 a_2)/(a_1 + a_2)$.

It should be noted that ρ can be significantly larger than ρ_1. Significant errors could be introduced when the accuracy loss is estimated based on the average case ρ instead of the worst case ρ_1. Therefore, there could be an apparent deviation when using point density ρ as a variable to describe the interpolation accuracy loss.

Second, most of the existing studies (Erskine et al. 2007; Ziadat 2007) evaluated the interpolation accuracy loss based on the absolute value of the grid size c. Nevertheless, selecting the absolute grid cell size c might not reflect the discrepancy among different datasets. For instance, if there exist two datasets with different density ρ_1 and ρ_2 ($\rho_1 < \rho_2$) respectively. It is often the case that when interpolating with the same absolute grid size c, the accuracy loss of these two datasets are prone to be different because the resolution of the dataset 2 is higher than that of dataset 1. It is arguable that all datasets with different resolutions would have the same performance when interpolating at the constant absolute grid size.

Third, the two variables ρ and c might not necessarily be independent. Hengl (2006) emphasized that grid cell selection should be based on the point density. In that sense, the absolute cell size c is a function of the ρ. In the light of this, it makes more sense that the approximation strategy is separated from the absolute cell size c in a way such that this approximation strategy is independent of the point density ρ.

In addressing the abovementioned limitations, this study proposed to incorporate two independent variables i and X into the model. Then the interpolation accuracy function based on the NN methods can be expressed as:

$$RMSE_{Interpolate,NN} = \varphi'(i, X) \quad (3)$$

Here, X describes the worst case of the point density. i describes the relative interpolation ratio. Then the interpolation cell size c can be computed as iX. The scope of this study is to identify optimal approximation strategy for interpolation.

3.3.3 Study area and datasets

This study selected the Lavallette borough in Ocean County, New Jersey as the study site. Lavallette is located in the Barnegat Peninsula, a long, narrow barrier peninsula that separates Barnegat Bay from the Atlantic Ocean. Hurricane Sandy took a major toll on the Lavallette borough. The sand dune was swept away during Hurricane Sandy. The entire community sustained significant damage. During Hurricane Sandy, the water level rose to more than three feet above the ground. Most of the oceanfront structures and buildings were severely damaged. Debris formed as a result of the eroded sand dunes and damaged structures move through the inland area of the community and cause additional damage to the properties that are farther away from the ocean. At least four LiDAR surveys (Table 3-2) were carried out in the study area, and these datasets were used for the performance validation purpose in this study. It is worth mentioning that these four datasets were collected using three different LiDAR systems. All four datasets are publicly available from the Digital Coast. Table 3-2 gives a detail description of the four datasets and its data specifications including vertical RMSE accuracy (95% confident interval), data collection time, LiDAR acquisition system, and the estimated

average point spacing. In this the following section, the 2010 airborne dataset is utilized to develop the model. In order to verify the versatility, the other three datasets are used to validate the proposed model in the later result in discussion section.

Table 3-2 List of datasets

Datasets	Vertical Accuracy RMSE (95% confidence) (m)	Data collection date	LiDAR acquisition system	Estimated minimum average point spacing (m)
2010 USACE NCMP LiDAR: Atlantic Coast (NJ)	0.2	20100828 to 20100911	CHARTS	0.85
2012 USACE NCMP LiDAR: Post-Sandy (NJ & NY) Point Cloud	0.125	20121116	CZMIL	0.75
2013 USACE NCMP Topobathy LiDAR: Barnegat Bay and Seaside (NJ)	0.196	20130903 to 20101002	CZMIL	0.5
2014 NOAA NGS Topobathy LiDAR: Post Sandy (SC to NY)	0.214 (0.057 – 0.221 depending on land cover type)	20140108 and 20140109	Riegl VQ-820G	0.25

3.3.4 The prediction model

This study utilized Z_{LiDAR} as the ground truth, which is extracted directly from the LiDAR data. The test data is elevation value after the LiDAR-to-DEM conversion, denoted as $Z_{DEM,iX}$. Then accuracy loss of LiDAR-derived interpolating at cell size iX can be computed as:

$$RMSE_{iX}^2 = \frac{\sum(Z_{LiDAR} - Z_{DEM,iX})^2}{n} \qquad (4)$$

A multiple regression model is designed to incorporate two independent variables (i, X) in estimating the accuracy loss in LiDAR-derived DEMs. In practice, it is difficult to compute the difference between observation value and the ground truth of the entire dataset. Instead, Huang (2000) suggested that using checks points to derive RMSE is most objective as long as two conditions are satisfied: (1) well distributed; (2) sufficient

checkpoints. To this end, this study devised a Monte Carlo Simulation (MCS) based method to generate random points that meet the abovementioned two conditions.

The MCS is conducted in ArcGIS. Each time, 100 random checkpoints in the study area were generated. Figure 3-1 depicts the distribution of 100 random checkpoints on a single MCS. At each checkpoint, the ground truth value is defined as the average elevation of all points extracted directly from the LiDAR data. Then the original LiDAR dataset is interpolated with different cell size iX using the nearest neighbour interpolation to generate a raster data. This study only considers the nearest neighbour interpolation because this method is favorited for its overall parsimonious nature such as ease of use, consistent accuracy, and high computation efficiency (Bater and Coops 2009). Then at each checkpoint, the observed elevation is extracted as the average elevation value in the same place from the raster data. It should be pointed that each checkpoint is buffered with a radius of $5X$ to ensure that a minimum of 20 points were within a single checkpoint.

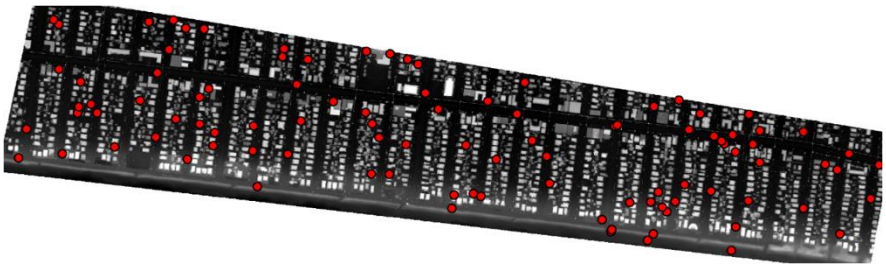


Figure 3-1 Random checkpoints (100) generated using Monte Carlo Simulation.
(The terrain map was generated by Xuan HU in ArcGIS using post-Sandy USGS EAARL-B data.)

The model is calibrated using the 2010 LiDAR data with different i. For each i, five replicated random MCS was replicated as illustrated in Figure 3-2. In each of the MCS, the minimum average spacing is summarized, with a range between 0.542m to 1.324m. The residuals between the raw data and the LiDAR-derived DEMs are

calculated over a sample of 100 checkpoints for each simulation. The results of the multiple-regression model are expressed in equation 5.

$$RMSE^2_{Interpolate,NN} = 0.126826i + 0.108581X - 0.09063 \quad (1 \le i \le 5) \qquad (5)$$

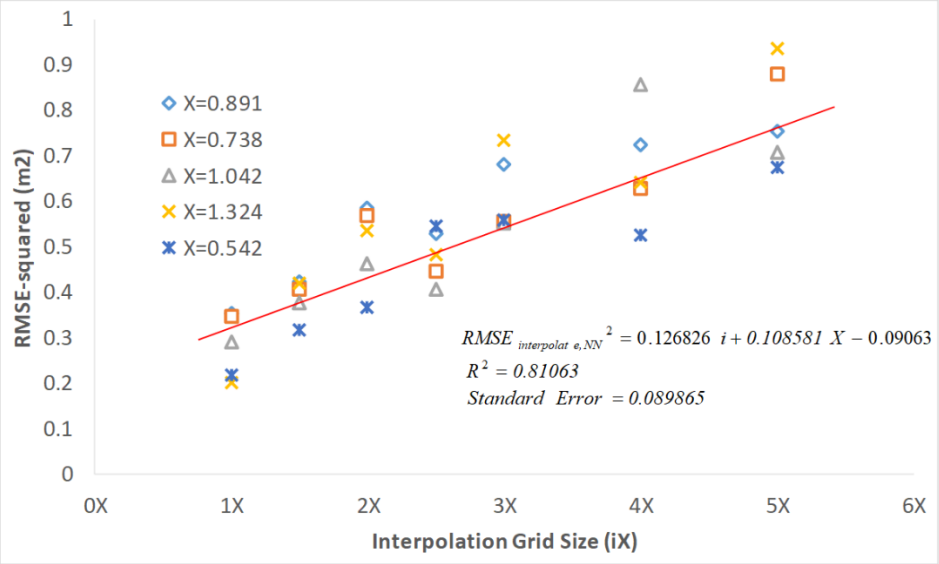


Figure 3-2 Model Calibration Result from five MCS based on 2010 dataset

Figure 3-2 depicts the model calibration result. The regression indicates an R-squared value of 0.8106 and a standard error of 0.08986. The p-value for the prediction variable i is lower than 0.001 and the p-value for the prediction variable X is less than 0.05 (0.0463), which indicate that both prediction variables are significantly correlated to the RMSE-squared. Then, the information loss function due to interpolation based on NN method is described in equation 6.

$$RMSE_{Interpolate,NN} = \sqrt{0.126826i + 0.108581X - 0.09063} \quad (1 \le i \le 5) \qquad (6)$$

3.3.5 Model Validation

Figure 3-4 displays the performance results of the aforementioned model based on four different types of LiDAR survey datasets. The blue lines are the prediction value of the proposed accuracy loss model as described in Equation (9). The orange lines are the

observed value based on the previous experiment design section. The observed accuracy loss is calculated as the RMSE between the LiDAR data and the LiDAR-derived DEMs. It is noteworthy that these four datasets were collected from three different LiDAR systems at completely different times. As a result, the reporting vertical errors of the datasets are significantly distinct, ranging from 0.125 meters to 0.20 meters and the minimum average spacing varies from 0.259 meters to 0.935 meters. In sum, the validation from four separate datasets suggests that the presented model (equation 6) can adequately predict the accuracy loss at the linear stage ($1X$ to $5X$).

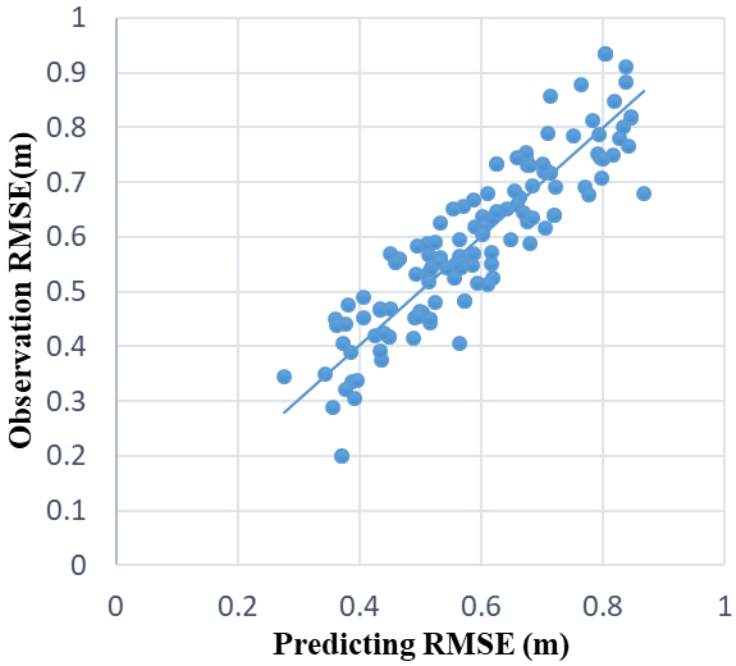


Figure 3-3 Goodness of fit of the RMSE Model

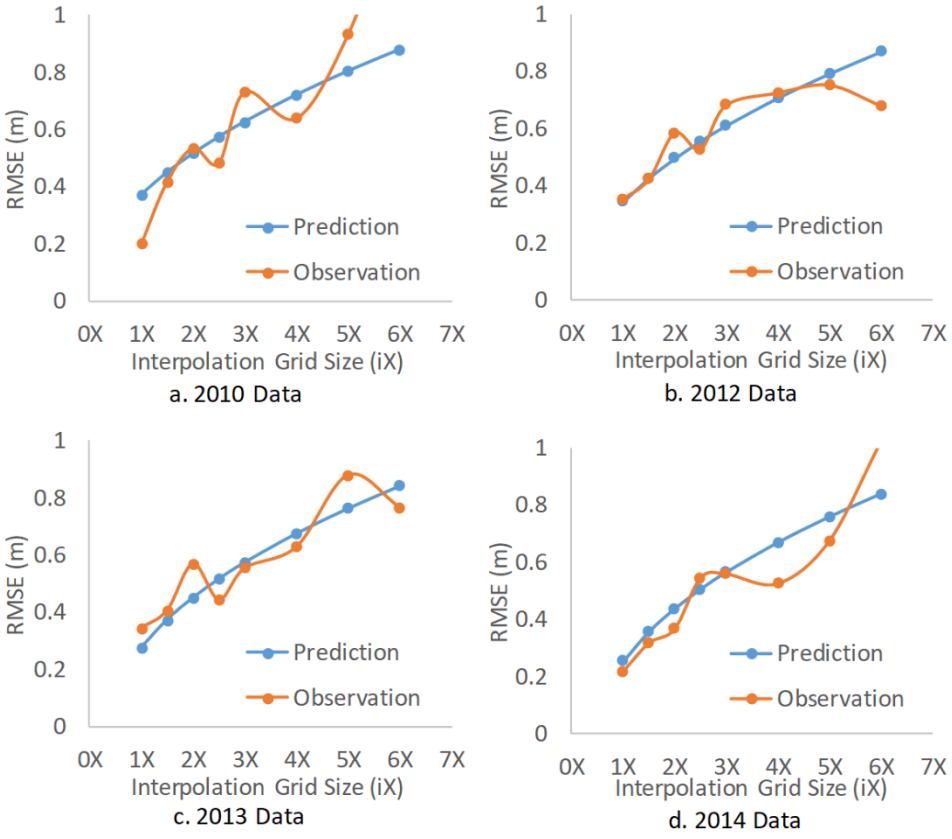


Figure 3-4 Model prediction performance on four LiDAR survey datasets. (a) Model prediction performance on 2010 data. **(b)** Model prediction performance on 2012 data. **(c)** Model prediction performance on 2013 data. **(d)** Model prediction performance on 2014 data.

Table 3-3 illustrates the mean error, standard error and R-squared for all four datasets, respectively. Compared to the minimum of 0.10 m system RMSE as shown in the third column, the mean prediction errors of four datasets (0.05 m) are acceptable. The R-squared values for the validation datasets are higher than 0.786, which means that the prediction model covers at least 78.6% of the variance. Figure 3-3 depicts a plot of the observed value and the proposed RMSE prediction model. The R-squared value is

0.8348, indicating a statistically significant correlation between the purposed model and the observed RMSE.

Table 3-3. Model Validation results using different datasets

Data	X (m)	RMSE @ $1X$ (m)	Mean Prediction Error (m)	Standard Prediction Error (m)	R-Squared
2010 Data	0.935	0.371	0.054	0.209	0.852
2012 Data	0.751	0.343	-0.010	0.079	0.786
2013 Data	0.364	0.275	0.015	0.075	0.830
2014 Data	0.259	0.254	-0.019	0.092	0.859

3.4 Results and discussion

3.4.1 Accuracy of LiDAR-derived DEM

Many studies have reported that the accuracy has a positive correlation with the increased of interpolation grid size (Aguilar et al. 2010; Erskine et al. 2007; Ziadat 2007). Results from abovementioned MCS indicated similar but slightly different results. In general, the RMSE increases when the strategy artifact i increases. Figure 3-5 depicts the estimated accuracy loss from five MCS. It should be pointed out that in practise, more than 50 MCS were conducted using the 2010 dataset and the results are similar. Based on the results, the authors suggested that the proper range of the strategy artifact i is 1 to 5. This suggestion is based on the following reasons.

First, when interpolating at a grid size less than one times the minimum averaging spacing $1X$ (e.g., 0.7 meters in this case), there is a fluctuation in the RMSE-squared. In practice, there is no need to interpolate at grid size less than $1X$, because the investment in computing resource at interpolating at smaller grid size does not necessarily guarantee a decrease in accuracy loss. Second, when interpolation grid size is greater than 5 times the minimum averaging spacing, there is an exponential growth in the accuracy loss

(RMSE-squared). It should be noted that terrain attributes that derived from DEMs are sensitive to the accuracy (Erskine et al. 2007). Moreover, Kienzle (2004) emphasized that the use of low accuracy DEMs may have a significant limitation in deriving terrain attributes. If the time is sensitive and the user has no option but to interpolate at larger grid size, caution must be practiced to use such DEMs products for quantitative analysis. Finally, when interpolating grid size is from $1X$ to $5X$, the relationship between approximation strategy i and RMSE-squared is almost linear, which makes it ideal to model the estimation loss (in terms of RMSE-squared) as a function of i. Granted that there could exist more complicated relationship than linear, such non-linear relationship is often sensitive to outliers. This study assumes a simple linear relationship between approximation strategy i and RMSE-squared at range $1X$ to $5X$. This range also represents the common choices in selecting the approximation strategy.

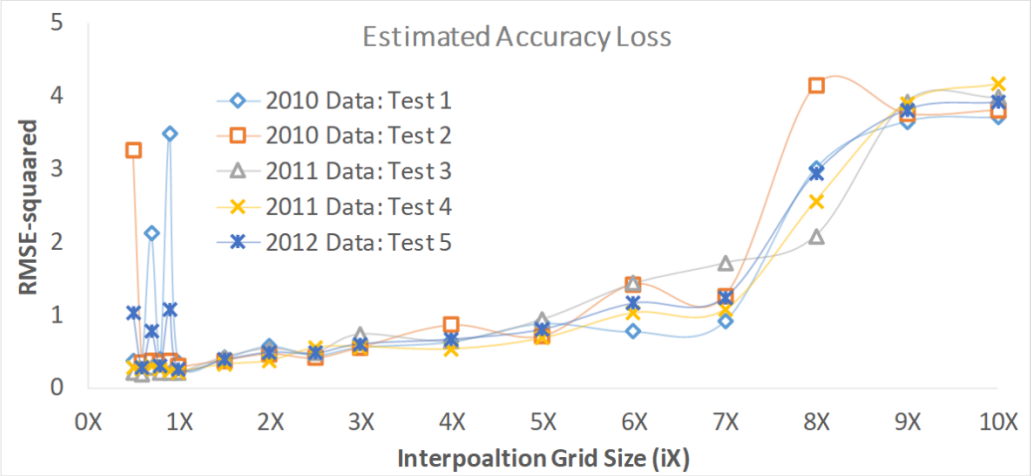


Figure 3-5 Estimated Accuracy Loss from 5 MCS (2010 data)

The benefits of the above accuracy loss model can be summarised as two-fold. First, the model could be deemed as a reference on how the approximation strategy i would affect the accuracy loss in the DEM data. In time-sensitive applications,

knowledge of accuracy loss associated with data processing is of remarkable interests in LiDAR-derived DEMs, especially if approximation techniques (e.g., interpolation at larger cell size) are adopted. The outcomes of terrain-derived attributes are strongly dependent on how much accuracy maintained after interpolation process (Kienzle 2004). Figure 3-6 depicts a visualization of the different accuracy loss because of different approximation strategy i. It can be easily obtained that when the approximation strategy value i increases, the DEM data becomes less clear. The above model demonstrates a mathematical relationship between accuracy loss in terms of RMSE and two independent variables (i, X), which allows users to access the prior knowledge on what is the possible RMSE when deploying particular approximation strategy i. Moreover, the visualization of the accuracy loss can also facilitate end-users to have a tangible understanding of the outcomes of DEM corresponding to different RMSE values.

Second, the model also enables a tool for end-users to determine the approximation strategy. In disasters, end-users often desire an optimal strategy that requires less computational effort while maintaining at least minimum user requirement. The present model facilitates a tool for end-users to balance the accuracy loss and computational intensity trade-off. In general, the data size can be easily computed using based on the approximation strategy i. For instance, let's supposed that the data size interpolating at grid size at $1X$ is s_1, then the data size interpolating at grid size iX is calculated as s_1/i^2. The relationship between interpolation grid size and data sizes, s_1/i^2, indicates that by deploying more aggressive approximation strategy (larger i value), the data size is considerably reduced. Moreover, data size is closely linked to computation intensity. A smaller data size often indicates less computation intensity and likely fewer

computation resources. The present model (equation 6) establishes a relationship between accuracy loss and data size, which facilitates end-users to evaluate the accuracy and computational intensity trade-off. In Figure 3-6, the four images depict DEM data generated based on different approximation strategy i using the 2010 dataset. For each DEM, the data size is summarized. To this end, end-users can reduce the computation intensity by selecting a coarse interpolation grid size (larger i value) while the accuracy maintains the minimum requirement of the application.

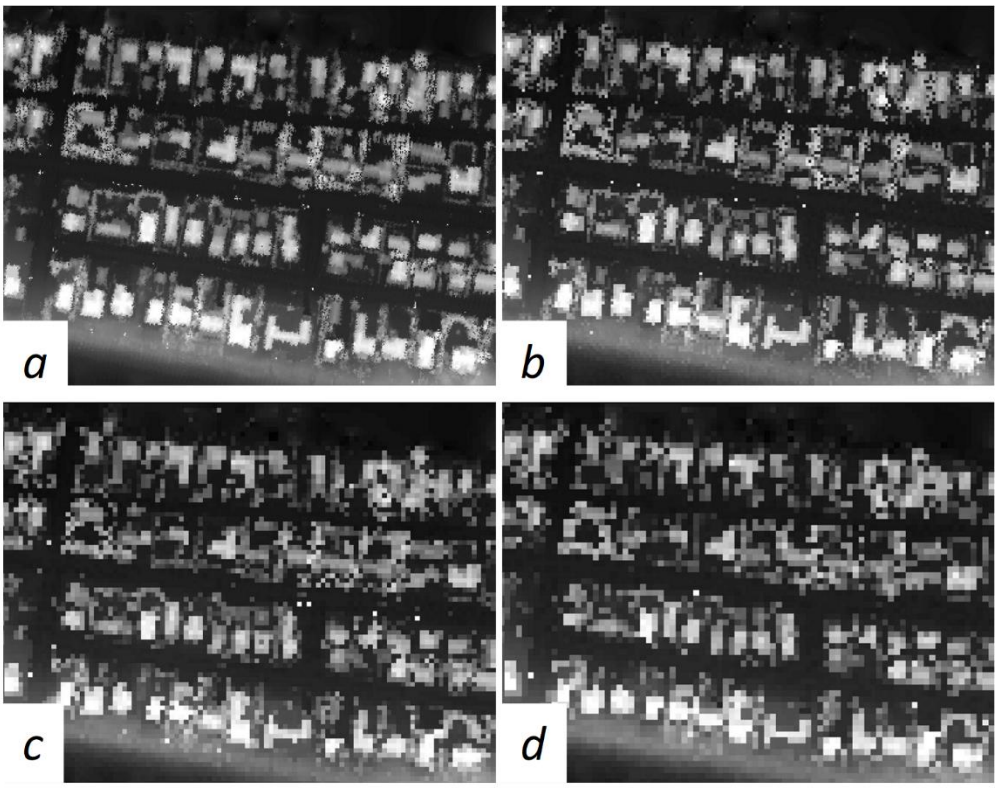


Figure 3-6 Comparison of LiDAR interpolation accuracy loss.

(The figure was generated by Xuan HU in ArcGIS using pre-sandy data collected by USACE)

(a) Interpolating at $1X$ (RMSE=0.37m, Data Size=20.70Mb);

(b) Interpolating at $2X$ (RMSE=0.56m, Data Size=6.24Mb);

(c) Interpolating at $4X$ (RMSE=0.72m, Data Size=852Kb);

(d) Interpolating at $5X$ (RMSE=0.80m, Data Size=531Kb

3.4.2 Comparison of dual input model and previous model

The introduction of minimum averaging spacing X and approximation strategy indicator i will have two major impacts on understanding the accuracy loss associated with interpolation. From the statistical point of view, the approximation strategy indicator i is independent with the data characteristic (point density ρ or minimum averaging spacing X). By separating the approximation strategy indicator i from the absolute cell size c, it facilitates the investigation on how artefact variable i and inherent data characteristic X affect the interpolation accuracy. From the application point of view, both variables have their meanings in time-sensitive applications. The scope of this study is to identify optimal approximation strategy for interpolation. The minimum average spacing X, reflects the characteristic of the LiDAR dataset. The approximation strategy indicator i directly represents what kind of strategy for the grid size approximation.

To verify that the combination of approximation strategy i and minimum average spacing X performs better than variables used in previous studies (e.g., point density ρ., or absolute interpolation grid size iX), four linear regression analyses are performed. First, three single regression analyses with a single variable are carried out to compare the significant level of three variables i, X, and iX respectively. Then a multiple regression analysis with two independent variables and X is performed to compare whether the RMSE is more sensitive to one single variable or a combination of two independent variables. The results are summarized in Table 3-4.

Table 3-4 Summary of the different regression model for variables selection

Test	Method	Variable	p-VALUE	R-Squared	Standard error
1	Single Regression with One Variable	X	0.421224	0.015466	0.272091
2	Linear Regression with a single Variable	i	1.35E-14	0.759957	0.100831
3	Linear Regression with a single Variable	iX	2.18E-10	0.620963	0.126703
4	Multiple regression with Two Independent Variable	i	5.86E-15	0.794789	0.094358
		X	*0.054617*		

First, the results of test 1 show that X is a poor indicator of the accuracy loss in the interpolation stage. The p-value (0.42) indicates that there is no significant correlation between the prediction variable X and the outcome RMSE-squared. Second, test 2 and test 3 compare the performance of the absolute value iX and the relative value i in predicting the interpolation accuracy. Both analyses have p-values less than 0.001, which means that both of variables are statistically significantly correlated with the RMSE-squared. However, the relative value i outperforms the absolute value iX when considering both the R-squared and the standard error. The R-squared value suggests that when predicting with the variable i, 75.9% of the variance can be explained compared to only 62% of the variance being explained when only using the absolute value iX. The above results suggest that when predicting the interpolation accuracy with a single variable, the variable i is a better prediction variable. Also, test 4 was performed to compare whether the relationship can be better predicted using two independent variables rather than a single variable. By comparing with test 2 or test 3, the results in test 4 indicate that the model with two-independent variables performs better than the models with a single variable. Therefore, it is logical to consider both the approximating strategy choice i and the minimum averaging spacing X for estimating accuracy loss. These

results align with what was found in previous studies: point density (Bater and Coops 2009; Gao 1997) and interpolation grid size (Erskine et al. 2007; Ziadat 2007) could affect the interpolation accuracy.

3.4.3 Impact of i and X on predicting the interpolation information loss

The proposed regression model indicates that the accuracy loss in terms of RMSE can be predicted by two independent variables i and X. In general, RMSE will increase when both i and X increases. However, these two variables will have different impacts on the accuracy loss. To better interpret the impact of i and X on predicting interpolation accuracy loss, accuracy loss at different locations are summarised in Table 3-5.

Table 3-5 Accuracy loss result of different datasets

Data	X(m)	RMSE (m)					Average Accuracy Loss Gradient (From 1X to 5X)
		@1X	@2X	@3X	@4X	@5X	
2010 Data	0.935	0.37	0.51	0.63	0.72	0.80	0.11
2012 Data	0.751	0.34	0.49	0.61	0.71	0.79	0.11
2013 Data	0.364	0.28	0.45	0.57	0.68	0.76	0.12
2014 Data	0.259	0.25	0.44	0.56	0.67	0.76	0.13

First, comparing the second and third column in Table 3-5, it can be obtained that when the minimum spacing X decreases, the initial interpolation accuracy loss ($@\ 1X$) decreases as well, regardless of the selection of strategy i. It stems from the fact that X is the inherent feature of the LiDAR data. The larger the X value is, the more coarse in the original LiDAR dataset. This characteristic will determine the maximum quality (initial accuracy) that the DEMs can be obtained from the given LiDAR dataset. Therefore, for applications that require high accuracy DEMs, special caution must be pay on this initial

accuracy loss. Otherwise, no matter what the approximation strategy value i is chosen, the accuracy loss will be beyond the limit of the application.

Second, the approximation strategy iX directly reflects the approximation strategy. A larger value of i means more aggressive approximation strategies and often result in less computation resource or time. For less accuracy sensitive applications, the data processing teams can opt for more computation efficiency strategies by deploying larger i value. Considering that for each individual dataset, the minimum averaging spacing is a fixed value, accuracy loss would be mostly dependent on the option of i. It indicates the gradient of accuracy loss when more aggressive approximation strategies are adopted.

Table 3-6 Comparison of the impact on accuracy loss of i and X.

Variable	Meaning	Description	Impact on RMSE
i	Relative ratio	User-defined artifact	Indicates the tendency of accuracy loss when the grid cell size is increased
X	Minimum average spacing	Inherent feature of the LiDAR data	Majorly affect the initial Accuracy loss (system) of the LiDAR-derived DEMs

In sum, the impact of two independent variables i and X on the accuracy loss are summarized in Table 3-6. The approximation strategy i, describes the relative ratio between the interpolation grid size and the point spacing of the LiDAR data. It indicates the gradient of accuracy loss when the grid size increases. On the other hand, the minimum average spacing, X, represents the inherent characteristic of the original LiDAR data. It majorly determines the upper bound performance that a DEM can be achieved from the given LiDAR data (with a minimum average spacing X).

3.5 Conclusions

The purpose of this study is to develop a method to determine the optimal strategy for selecting approximation parameter for the LiDAR-derived DEMs. In addressing to this need, the author introduces two independent variables (1) minimum average spacing; (2) approximation strategy artifact. Minimum average spacing, denoted as X, describes the characteristic of the LiDAR data. Approximation strategy artifact, denoted as i, indicates the approximation strategy. Typically, a larger value of i indicates more aggressive approximation strategy and less computation resource and time.

Then a multiple-regression model is proposed to incorporate two independent variables (i, X) for estimating the accuracy loss at the second linear increasing stage. MCSs are carried out to compare the RMSE-squared between observation value (elevation extracted from LiDAR-derived DEMs) and ground truth value (elevation extracted from LiDAR). Finding from the simulations indicate that the proposed model can act as a good approximation for the accuracy loss with a p-value less than 0.06 and an R-squared value larger than 0.78.

Compared to the previously accuracy loss model that based on absolute grid size, the regression analysis result suggests that the two-independent-variables (i, X) model are better performed regarding significance level (p-value), coverage (R-squared) and prediction error (standard error). The performance purposed model is then validated using four completely separate datasets. Results show that prediction values form all four datasets align with the observed values.

Finally, the impacts of variables i, X on the accuracy loss are summarized. The minimum average spacing X reflects the characteristic of LiDAR data. It determines the initial accuracy performance in the LiDAR-to-DEM conversion. The approximation

strategy artifact, i, indicates the options for approximation. It majorly determines the accuracy loss gradient at the linear interpolation stage.

It should be noted that the development model is based on the assumption that the relationship between approximation strategy i and accuracy loss at range $1X$ to $5X$ is linear. There could be a more complicated non-linear relationship. Moreover, this study was intended to determine the optimal approximation strategy. This study majorly considered the impact of two independent variables (i, X). In realistic, the interpolation method could pose significant on the accuracy loss. Future study would be likely to include the distinct interpolation methods into the estimation model. Another limitation of this study is the availability of high-resolution datasets. Due to the low resolution (X ranges from 0.25 to 0.9m), this study suggests the range of the approximation strategy artifact, i is from 1 to 5. However, it is believed that future datasets are prone to have higher quality (resolution) because of advancement in Laser scanner and navigation system. Therefore, the approximation strategy i is expected to have more potential options (values larger than 5).

Chapter 4 A Framework for Prioritizing Geospatial Data Processing Tasks to Maximize Information Efficiency during Time-Sensitive Disastrous Events

4.1. Introduction

Decisions during disaster response cannot be made without reliable information because decision-makers seek information as a process of sense-making to form their point of view (Devin 1983). The information is served as evidence to assimilate into what is already known (e.g., experience) (Kuhlthau 1991). Meanwhile, disasters are small probability events that occur so infrequently. For such small probability events, it is difficult to have the particular experience to draw on. For instance, FEMA issued National Flood Insurance Program (NFIP) to protect flooding damage. The primary focuses are placed on 1-percent-annual-chance floodplains (e.g., Zone A, Zone V, Zone VE) or at most 0.2-percent-annual-chance floodplains (Zone B, Zone X). Anything outside those zones is considered as a minimal risk of flooding. However, during Hurricane Harvey, most of the inundation area in Houston is actually located outside 1-in-500 chance of flooding area. In the light of this, the impact of a disaster could easily go beyond one's experience based judgment. While handling disasters, the incorrect or biased experience can cause considerable damage. In 2015 Tianjin explosion, relied on previous experience, a wrong call was made to allow fighters to enter the explosion site after the first explosion. 95 firefighters were killed by the second more powerful explosion. Such a tragedy could have been avoided if that decision was made based on more accurate disaster information. Moreover, decision-making consists of finding the best option from all feasible solutions (Herrera et al. 1996). The best option cannot be

derived without trustworthy information. Decision-making during disasters, therefore, calls for Information-based judgments.

Information-based judgments are often dependent on the availability of data and the processing algorithms to deal with these data. Modern technology provides unprecedented new opportunities for fulfilling the information needs during a disaster. Among these techniques, light detection and ranging (LiDAR), especially carried by airborne platforms, receives a growing concern and even becomes a routine survey in dealing with major disasters. For instance, there are at least five Airborne LiDAR survey regarding Hurricane Sandy. Among them, USGS conducted two airborne LiDAR survey before and immediately after Hurricane Sandy strikes Atlantic shoreline. While disaster data are more abundant than required, the remaining question is how to relate data processing to information required by decision-making. Unprocessed data are intangible and non-consumable, yet a plentiful resource that can be refined. To explore the potential of Airborne LiDAR, there are numerous studies investigating in processing algorithms in terrain mapping (Moore et al. 1993; Reutebuch et al. 2003), building damage assessment (Dong and Shan 2013; Li et al. 2008), flood simulation (Haile and Rientjes 2005; Webster et al. 2004), debris flow analysis (Scheidl et al. 2008), etc. Yet, none of these research is emphasized on disaster response or the time-sensitive environment. More often than not, the information processing effort is motivated by the argument that "what-is-going-on" (Vos et al. 2000). Such a vague argument cannot derive a clear objective and is prone to result in "data-rich-but-information-poor" situations (Timmerman et al. 2010). Such situations, to a great extent, depreciate the value of data. The value of data or information is best determined by value-in-use (Repo 1986). The salience of information

is directly related to the level of satisfaction directly by decision-makers rather than data processing teams (Vugteveen et al. 2014). Unfortunately, decision-makers often lack in-depth knowledge of data processing. The central motivation for information processing remained driven by data processing team. In the light of this, the real value of data is not fully exploited. There are significant gaps in satisfying decision-makers' information need.

The first gap is to formulate the value of information in disaster environment formally. Decision-making during disaster environment is multiple objectives (Barbarosoğlu et al. 2002; Hwang et al. 1993). It involves varying data processing tasks and integrates multiple algorithms. For instance, during disaster response, humanitarian relief often consists of a series of tasks such as minimizing human suffering and death (Beamon and Balcik 2008; Lwin and Murayama 2011), designing routes for search and rescue (SAR) operations (Kwan and Lee 2005; McCarthy et al. 2007), allocating disaster relief resources, assessing the risk of critical infrastructures to prevent secondary hazards (Ouyang 2014; Pederson et al. 2006; Tolone et al. 2004), and etc. In disaster response, the usefulness of information is not determined by how the collected data fulfill any single processing task, but by how the processed information from these tasks supports decision making. Thus, information for decision support is a selection process based on information salience. Salience deal with how relevant and usable information is to decision making bodies (Vugteveen et al. 2014). Nevertheless, there is limited research in identifying the salience of information studies (Cash et al. 2002; McNie 2007; Timmerman et al. 2001).

The second research gap is to model the information articulation process. Considering this process as a function, then the domain space (information offered by data processing teams) should match with the range (decision support information needs). Decision-making consists of finding the best option from all the feasible solutions (Herrera et al. 1996). Compared to data processing that targets at a clearly-defined objective, the information need for decision support is vague. Such a vague criteria cannot be used to derive precise numerical values (Delgado et al. 1998). In the real world, many decision processes take place in which goals, constraints, and consequence of possible action are not precisely known. Then a more realistic approach may be to use linguistic assessments instead of numerical value. Second, decision making is a group decision process. The complicated disaster situation often requires integrating experiences and judgments from different stakeholders, experts for decision-making. Based on different backgrounds, these decision-makers will often have disagreeing opinions and judgment (Herrera et al. 1996). Decision-making process, therefore, calls for a consensus reaching process that ensures all the preferences of different individuals have the maximum possible consistency.

The third research gap is to dynamically balance the conflicts in information processing needs. Dimensions of data processing need could vary significantly and even in a contradicting way. One emphasis for data processing is accuracy (Hodgson and Bresnahan 2004; Reutebuch et al. 2003), which is often motivated by the argument that "how perfect is the algorithm outcome represents the real world scenarios" (Hasselman et al. 2005). Nevertheless, temporal-resolution is another critical concern. The value of information will drop significantly as time goes by (Hodgson et al. 2014; Horan and

Schooley 2007). Unfortunately, accuracy and temporal resolution are two disagreeing criteria. Pursuing high accuracy information often requires high complexity processing algorithms, high-resolution datasets, and as a result more computation time. Moreover, the multiple objectives during disaster response could be transferred into opposing data needs. For example, building assessment can be performed by "change detection" using Digital Elevation Maps (DEMs). On the other hand, to prevent secondary hazard, detail visualization data are required for critical infrastructure damage assessment. DEMs and visualization data are distinctive that require particular data processing tasks. Therefore, it demands a mechanism to leverage the computation resource dynamically in a way that those disagreeing processing needs can maintain balanced.

In sum, a structure process between decision makers on the one hand, and data processing teams on the other hand, is essential to ensure that information supplied by the data processing is tailored to the needs of users (Sutherland et al. 2011; Timmerman et al. 2000; Vaughan et al. 2007). Therefore, this study proposes a data envelopment analysis (DEA) based model integrating with linguistic term to identify the information salience during a disaster. The proposed method will have three major contributions. First, this study presents an innovative perspective based on the "information efficiency" for tasks prioritization during disaster response. Second, the information needs preferences are modeled as cone constraints using the linguistic term, which can more objectively reflect opinions from the decision makers. Finally, The group decision process is embedded in the traditional DEA model, which distinguishes the specific features of disasters. The remaining part of this research will be organized as follows. Section 4.2 will introduce the related study on identifying the salience of information. Section 4.3 will describe the

proposed methodology for information articulation. A hurricane Sandy based case study and discussion will be presented in section 4.4, and section 4.5 will denote the final conclusions of this research.

4.2. Related Work

Salient information has repeatedly been identified as essential in evidence-based decision-making (Cash et al. 2002; Honig and Coburn 2008; Todd and Benbasat 1992). It is typical because there exist overwhelming data that may not be relevant or useful. Without clearly specifying the salience of different information, data suppliers could generate an excessive amount of data that might fail to provide timely and relevant information. Ward et al. (1986) described such situation as "data rich but information-poor syndrome". Even after three decades, this syndrome still exists. Identifying the salience of information remains challenging (Sutherland et al. 2011). Information providers, on the one hand, may wish to inform decision-makers through processing salience information, but they often lack knowledge in understanding the objective and priority of the information need. Decision-makers, on the other hand, may reluctant to incorporate information because they might doubt the effectiveness, relevance, and salience of the information. To support evidence-based decision making, it is essential to ensure that information supplied by the data processing team is tailored to the needs of users (Sutherland et al. 2011; Timmerman et al. 2000; Vaughan et al. 2007). In literature, to improve the usefulness of information, there are numerous studies (Cash et al. 2002; Dervin and Nilan 1986; Kuhlthau 1991; Sutherland et al. 2011; Timmerman et al. 2001; Timmerman et al. 2000) on designing theoretic framework to identify information needs. Most of these studies focused on two aspects to improve the information salience: (1) collaboration; (2) users' perspective paradigm. For instance, Timmerman et al. (2001)

clarified the process and participant for the process of identifying information need. In another example, Kuhlthau (1991) proposed that information seeking should be based on users' perspective. The author augured that information system should center on the users' problem rather than on the capability of p. In general, the research mentioned above all have profound influences in providing theoretical backgrounds for the need of identifying information salience, yet application-wise, it is still difficult to achieve (Timmerman et al. 2001).

Most of the emphases of computing focus on resource allocation. For instance, one classic model for computing architecture to evaluate Quality of Service (QoS) is based on queuing theory (Khazaei et al. 2012; Ma and Mark 1995; Vilaplana et al. 2014; Xiong and Perros 2009). The performance of the queue is evaluated based on the overall wait time and service time (Vilaplana et al. 2014). However, the emphasis of queuing theory is to predict the overall response time (wait time and service time) for computation resource management. Other researchers investigated the leverage of resource based on game theory (Wei et al. 2010), Virtual Machine (Xiao et al. 2013), etc. For the abovementioned studies, the primary focus is placed on the resource allocation, and each task is assumed to be identical with the same arrival and service distribution. The salience of each task, therefore, is often ignored. Another formulation is optimization problem. Most of the optimization problems that dedicated to computing are based on a cost function. The objective is to either maximize or minimize the cost function (outcome) while preserving specific constraints (e.g., time, resource) (Chapin et al. 1999). The salience of each task is implemented as the weight function of each input. Yet, this weight function is often assumed as known or derived based on other means.

Nevertheless, these studies emphasize from the data processing perspective. Kuhlthau (1991) stressed the importance of formulating the problem from user's problem. Data envelopment analysis (DEA) is a nonparametric method in operations research proposed by Charnes et al. (1978). DEA has been credited for deriving relationship between multiple outputs and inputs in decision making. It has been proven success in productivity analysis (Banker and Natarajan 2008; Eilat et al. 2008; Green et al. 1996; Sherman and Zhu 2006; Vitner et al. 2006). The model present in this study to determine the salience of each processing task was constructed based on the DEA model.

4.3. Methodology

4.3.1 Problem setting

It is assumed that this process to determine the information processing salience occurs during initial disaster response. Supposed that there is a group of d decision makers, $DM = \{DM_1, DM_2, \ldots DM_d\}$ going to determine the prioritization of information processing tasks. The entire task ranking process is based on information salience or information efficiency, denoted as $\theta_{i,j}$, defined by a DEA model. This study considered decision making unit (DMU) as a tuple(O_i, P_j). O denotes a set of site candidates for information processing. P denotes a set of processing tasks that has already been implemented for information generation. Then each DMU (O_i, P_j) is considered as triggering task P_j on siteO_i. Le X , Y represent the set of input and output variables respectively. Then the information salience can be regarded as a function of the weighted sum of the output variables Y and the weighted sum of the input variables X. The group of decision makers pairwise comparison of importance of output variables Y and input variables X, respectively using linguistic value. Based on their preferences, two fuzzy

preference matrix A and B are constructed for input and output variables, respectively. For the output variables, this study selected four disaster response functions proposed by Lindell and Perry (1992). For the input variables, two cost indicators, computational intensity and computational complexity, are selected. The entire information articulation process is considered as dynamic by changing two fuzzy preference matrix A and B. For illustration, considering two output variables { Y_1 =Emergency assessment, Y_2=Population protection}. At the very early stage of disaster response, the emphasis on humanitarian tasks can be expressed as a less preference value b_{12} of Y_1 over Y_2 such as 0.1, which means the population protection is much important than emergency assessment. In the later stage, when humanitarian tasks are less a concern, the preference value b_{12} of Y_1 over Y_2 can be adapted to a larger one such as 0.7. By assigning different value to the preference matrix, information salience of each DMU will alter accordingly.

4.3.2 Preliminaries

Definition 1 Group Decision (Herrera et al. 1996): A decision situation in which :

1) there are two or individuals, each of them characterized by her or his own presentations, attitudes, motivations, and personalities;
2) who recognized the existence of a common problem;
3) attempt to reach a collective decision.

Definition 2 Linguistic scale: Considering a finite and totally ordered label set $S = \{s_i | i = 0,1, \dots T\}$, with an odd cardinality. s_i represents a possible value for a linguistic real variable. The set S is a linguistic scale as long as the followings are met:

1) Ordinarily: $s_i > s_j$, if $i > j$ and $i, j \in S$.
2) Negation Operator: $neg(s_i) = s_j$ such that $i + j = T$.

3) Let I be the intensity function of the linguistic set S, such that $I: S \to [0, T]$, then the subscript of the elements $s_i \in S$ can be obtained as $I(s_i) = i$).

Definition 3 Linguistic judgment preference matrix (Herrera et al. 1996): Assuming a linguistic framework and a finite set $X = \{x_{1,} x_{2,} ... x_n\}$, the experts preference attitude over X can be defined a $n \times n$ linguistic preference matrix R, such that $R = (r_{ij})_{n \times n}$, r_{ij} represents the preference degree of alternative $x_{i,}$ over $x_{j,}$, and $\forall i, j \in \{1,2,...n\}$, $r_{ij} \in S$

Definition 4 Fuzzy preference judgment matrix (Orlovsky 1978): For a Linguistic judgment preference matrix $A = (a_{ij})$, a_{ij} is represented by a fuzzy value such that $a_{ij} + a_{ji} = 1$ and $a_{ii} = 0.5$, then A is a fuzzy preference judgment matrix,

Definition 5 Consistency preference matrix (Alonso et al. 2004): For a Linguistic judgment preference matrix $A = (a_{ij})_{n \times n}$, $\forall i, j, k \in \{1,2,...n\}$, if $a_{ij} = a_{ik} - a_{jk} + 0.5$, then matrix A a consistent fuzzy preference judgment matrix, The relationship $a_{ij} = a_{ik} - a_{jk} + 0.5$ is called additive consistency.

Theorem 1 Linguistic judgment preference matrix R to Fuzzy preference judgment matrix A conversion. R can be converted to A based on the following equation:

$$a_{ij} = \alpha \frac{I(s_k)}{T} + \frac{1-\alpha}{2},$$

Where $\alpha \in [0,1]$ is a distribution indicator such that the fuzzy value $a_{ij} \in [\frac{1-\alpha}{2}, \frac{1+\alpha}{2}]$.

Proof of Theorem 1:

$$a_{ij} + a_{ji} = \alpha \frac{I(s_k)}{T} + \frac{1-\alpha}{2} + \alpha \frac{I(neg(s_k))}{T} + \frac{1-\alpha}{2} = \alpha \frac{k}{T} + \frac{1-\alpha}{2} + \alpha \frac{T-k}{T} + \frac{1-\alpha}{2} = 1$$

Let $i = j$, then $a_{ii} + a_{ii} = 1, a_{ii} = 0.5$.

Definition 4 is satisfied.

Theorem 2 Aggregation of fuzzy judgment preference matrix into composite group fuzzy judgment preference matrix (or composite matrix). A composite matrix is the combination of all individuals' judgement matrix based on the weight of each individual. Assuming a group of d decision-makers $D=\{DM_1, DM_2, ...DM_d\}$, and for each individual decision maker $DM_k \in DM$, the fuzzy judgment preference matrix is $A^k=(a_{ij}^k)_{n\times n}$. Then the composite group fuzzy judgment preference matrix (or composite matrix) of d decision-makers $A=(a_{ij})_{n\times n}$ can be computed as $A=\sum_{k=1}^{d} w(k)A^k$, and $\sum_{k=1}^{d} w(k)=1$, $w(k)$ is the weight function of each individual decision maker k.

Theorem 3 Consistency of the composite fuzzy preference judgment matrix. Given that the composite matrix is the combination of the opinions from each individual, inconsistency could arise. Particular for the preference matrix, if $a_{ij}+a_{ji}\neq 1$, then the composite preference matrix is inconsistent, and therefore there exists no solution for the system. To ensure that the system of equations has at least on solution (or consistent), the consistent matrix $\overline{A}=(\overline{a_{ij}})_{n\times n}$ is obtained based on the following relationships."

$$\overline{a_{ij}}=\frac{\overline{a_i}-\overline{a_j}}{2(n-1)}+\frac{1}{2}, \overline{a_i}=\sum_{j=1}^{n} a_{ij}, i=1,2,...,n.$$

Proof:

Because

$$\overline{a_{ik}}=\frac{\overline{a_i}-\overline{a_k}}{2(n-1)}+\frac{1}{2}, \overline{a_{jk}}=\frac{\overline{a_j}-\overline{a_k}}{2(n-1)}+\frac{1}{2}, \overline{a_{ij}}=\frac{\overline{a_i}-\overline{a_j}}{2(n-1)}+\frac{1}{2}, \overline{a_{ji}}=\frac{\overline{a_j}-\overline{a_i}}{2(n-1)}+\frac{1}{2}$$

Hence

$$\bar{a}_{ik} - \bar{a}_{jk} + 0.5 = \frac{\overline{a_i} - \overline{a_k}}{2(n-1)} + \frac{1}{2} - \frac{\overline{a_j} - \overline{a_k}}{2(n-1)} - \frac{1}{2} + \frac{1}{2} = \frac{\overline{a_i} - \overline{a_j}}{2(n-1)} + \frac{1}{2} = \bar{a}_{ij}$$

$$\bar{a}_{ij} + \bar{a}_{ji} = \frac{\overline{a_i} - \overline{a_j}}{2(n-1)} + \frac{1}{2} + \frac{\overline{a_j} - \overline{a_i}}{2(n-1)} + \frac{1}{2} = 1.$$

According to Definition 5, $\bar{A}$ is a consistent matrix.

4.3.3 Decision Variables

The decision variables are further divided into the input and output variables. Input variables are regarded as the cost associated with each decision-making unit (DMU). In this study, two cost indicators are considered: Computation Complexity, X_1 and Computation Intensity, X_2. On the other hand, output variables are regarded as the performance indicators of each DMU. This study evaluate the outcome of each DMU by examining how these DMU satisfied the four emergency response functions proposed by Lindell and Perry (1992). These four functions include (1) emergency assessment, (2) hazard operation, (3) population protection and (4) incident management. Table 4-2 depicts detail description of these four functions.

4.3.3.1 Input Variables

The first variable, computation complexity, describes the inherent difficulty associated with a computation task. This study majorly considers three core processing tasks (visualization, spatial analysis, advanced analysis) as illustrated in Table 4-1. It should be noted that computation complexity can be directly measured if the algorithm runtime complexity is explicitly given or the computation time for each processing tasks is benchmarked. In other situations, this study develops a complexity index scoring

system to estimate the computation complexity. In general, the complexity estimation system consists of three steps.

Step 1, split the processing task into sub-operations as summarized in the second column in Table 4-1.

Table 4-1 Typologies of core processing tasks

Core Processing Tasks	Sub-operations
Visualizing	Data Cleaning & Quality Control Low-Level Transformations Spatio-temporal Mapping & Registration
Spatial analysis	Data Cleaning & Quality Control Low-Level Transformations Spatio-temporal Mapping & Registration Data Subsetting, Filtering, Subsampling
Advanced Analysis	Data Cleaning & Quality Control Low-Level Transformations Spatio-temporal Mapping & Registration Data Subsetting, Filtering, Subsampling Object Segmentation Object Classification

Step 2, identify the complexity score for each sub-operations based on a scale from 1 to 9, where 1 indicates the low complexity and 9 indicates the high complexity.

Step 3, the overall complexity of the processing task is computed as the summation of all the complexity scores in the sub-operations.

Step 4, A scaling process $x' = \frac{x}{max(x)}$ is deployed to scale the range of all inputs to [0, 1].

Beside computational complexity, the other input variable is computational intensity. Compared to the computational complexity that reflects the difficulty associated with a processing algorithm, computational intensity measures the magnitude of a dataset. This study major considers LiDAR-derived applications. The total LiDAR

point count, denoted as $point\_count$, is selected as an indicator for the computational intensity. After the $point\_count$ of each DMUs are summarized, a similar scaling process as described in computational complexity step 4 is deployed to scale the range of all inputs to [0, 1]. It should be noted that under certain circumstances, the total point count is interchangeable with other indicators such as $File\_Szie$ or $Area$.

File_Szie: The $File\_Szie$ of a LiDAR data can be computed as k times the $point\_count$. k is the scaling coefficient indicating how many bytes each point occupied. According to ASPRS (2013), the value of k ranges from 27 to 35 $bytes/pts$ depending on the las format and the information it contains. Hence, for datasets that follow the same las format (same k), $File\_Szie$ of the data is equivalent to $point\_count$, or computational intensity.

Area: For the same dataset, the total point count can be estimated as the ratio between $area$ and $point\_density$. Supposed that for the same dataset, the point cloud follows a uniform distribution (with the similar $point\_density$)in different religions, then the computation intensity or $point\_count$ can be approximated based on the $Area$.

4.3.3.2 Output Variables

Sub-function and description of the output variables selected in this study are illustrated in Table 4-2. For the output variables; there are four main steps to determine the value of an output variable.

Table 4-2 Typologies of Emergency Response Functions.

	Sub-functions	Description
Emergency assessment	Threat detection, Damage assessment, Hazard/ environmental monitoring	consists of those diagnoses of past and present conditions and prognoses of future conditions that guide the emergency response
Hazard Operation	Hazard source control, Protection works, Contents protection	Refers to expedient hazard mitigation actions that emergency personnel take to limit the magnitude or duration of disaster impact (e.g., sandbagging a flooding river or patching a leaking railroad tank car).
Population protection	Search & rescue, Emergency medical care, Protective action section and implementation, Impact zone access control/security	Population protection refers to actions—such as sheltering-in-place, evacuation, and mass immunization—that protect people from hazard agents.
Incident Management	Logistics, Mobilization of emergency facilities/equipment, External coordination	Consists of the activities by which the human and physical resources used to respond to the emergency are mobilized and directed to accomplish the goals of the emergency response organization.
Note: the table is adopted from Lindell and Perry (1992)		

Step 1 Determine Information Demand level

The information demand level is determined by the vulnerability of a community. There are multiple ways to identify the vulnerability of a community, either from historical data or based on stakeholders' opinions.

One way to determine the information demand level is retrieving community vulnerability indicators from the historical database. In addressing to the four disaster functions proposed by Lindell and Perry (1992), different data can be retrieved from the historical database. For instance, the demand for population protection task can be derived from demographic geospatial data. For flooding hazard, the need for hazard operation can be estimated based on flood zone maps.

Under circumstances that the historical data are not available, the information demand level can be obtained directly from stakeholders' options. Assumed a group of q

stakeholders. For a given instance ξ, each of them has a linguistic judgment on the demand for information, denoted as $L = \{l_1, l_2, \ldots l_n\}$. Let S be a linguistic set and I be the intensity function. Then it can be easily obtained that $L \subseteq S$. Let ω be the conversion function from linguistic value to numerical value.

Then the total demand of a given instance ξ can be computed as $d = \sum_{i=1}^{n} \omega(l_i)$. In this case, we assumed that the linguistic value is uniformly distributed, and conversion function is given as $\omega(s_i) = \frac{I(s_i)}{T-1} = \frac{i}{T-1}$. Then the total demand for a given instance ξ is $d = \sum_{i=1}^{n} \frac{I(l_i)}{T-1}$.

Let the instance be evaluating the demand for disaster function F_j at location O_k. Then the total demand for a given instance $\xi(F_j, O_k)$, denoted as $d_k^j = \frac{\sum_q I(l_q^j)}{T-1}$.

Step 2 Identify Significance level of the information

Assumed that there is m processing task P that can satisfy n disaster functions F. Let $E = \{\varepsilon_i^j, i = 1,2,..m; j = 1,2,..n\}$ denoted a set of significance level of information processing task P_i in supporting disaster function F_j. Let S be a linguistic set and I be the intensity function. Then it can be easily obtained that $A \subseteq S$. Assumed the same uniform distribution for ε_i^j, then the significance level ε_i^j is computed as $\frac{I(\varepsilon_i^j)}{T-1}$

Step 3: Determine the value of output variables

Based on step 1 and step 2, then the outcomes of output variable at DMU (O_i, P_k) respect to disaster function, F_j can be computed as the product of the information level

at location O_k respect to function F_j and the significance level of the $Task_i$ respect to disaster function F_j, denoted as $\varepsilon(O_k, P_i) = d_k^j \cdot \varepsilon_i^j$.

For illustration, we considered a 7-tuple linguistic set given as follows:

$S = \{s_0 = None, s_1 = VeryLow, s_2 = Low, s_3 = Medium, s_4 = High,$

$s_5 = VeryHigh, s_6 = Perfect\}$

Then the linguistic conversion function $\omega(s_i) = \frac{I(s_i)}{6} = \frac{i}{6}$

Considering the situation that historical data are not available and the information demand is retrieved from stakeholders' judgment. Supposed there is a group of 7 decision makers, who evaluate the need for emergency assessment (F_3) respect to location O_1. Their linguistic judgments are given as: $L = \{s_0; s_2; s_4; s_2; s_3; s_3; s_1\}$, then the total demand for emergency assessment (F_3) respect to location O_1, is:

$d_1^3 = \frac{0+2+4+2+3+3+1}{6} = \frac{15}{6}$.

Let $\{s_1\}$ be the linguistic judgment on the significance level of information processing task P_4 in supporting disaster function F_3. Then $\varepsilon_4^3 = \frac{1}{6}$.

Then the outcomes of output variable $\varepsilon(O_1, P_4, F_3) = d_k^j \cdot \varepsilon_i^j = \frac{15}{36}$

Step 4: Data normalization

Because the DEA is a nonparametric method, the range of raw output data varies widely, after the outcomes of each output variable are determined, a data normalization

process is deployed to rescale the data to values between 0 and 1. The normalized value is computed as $y' = \frac{y - min(y)}{max(y) - min(y)}$.

4.3.4 Construction of preference cone constraint

The preference cone constraints for input and output variables are constructed based on experts' pairwise comparison of input and output variables based on the linguistic term, respectively. The relationship between inputs and outputs respectively forms a quadratic convex, and therefore refer to as cone constraint. The main purpose of the cone constraints is to prevent any input or output from being dominant in the DEA model. The sensitivity of cone constraints used in this study are depicted in the later section. The steps for constructing an input variables cone constraints is described as below:

Step 1, for each expert, he or she pairwise compares the input variables based on linguistic term, and the linguistic judgment preference matrix is described as R^k.

Step 2, the linguistic preference matrix R^k is converted to fuzzy preference judgment matrix, denoted as A^k, according to theorem 1.

Step 3, Based on thereon 2, the fuzzy preference judgment matrix of each individual decision-maker A^k is aggregated as a composite fuzzy preference judgment matrix $A = \sum_{k=1}^{d} w(k) A^k$. In this study, same weight is considered. Thus the composite matrix is simplified as $A = \frac{A^k}{d}$.

Step 4, the composite fuzzy preference judgment is converted to consistent matrix $\overline{A}$ based on thereon 3.

Step 5, Assuming $\overline{A} = (\overline{a}_{ij})_{n\times n}$ is final consistent group preference matrix with n decision variables. Let $A^* = \overline{A} - \lambda_A I_n$, where λ_A is the maximum eigenvalue of the matrix $\overline{A}$; I_n be the n-dimensional identity matrix. Then, group preference cone constraint is, and this cone is a closed convex cone.

For illustration step 5, assumed that we have a group preference matrix $\overline{A} = \begin{pmatrix} 0.5 & 0.2 \\ 0.8 & 0.5 \end{pmatrix}$, and maximum eigenvalue of the matrix $\overline{A}$ is $\lambda_A = 0.9$. Given that the 2-dimensional identify matrix $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, then $A^* = \begin{pmatrix} -0.4 & 0.2 \\ 0.8 & -0.4 \end{pmatrix}$.

4.3.5 Construction of linguistic group decision DEA model

A 2-tuple decision making unit (DMU), denoted as (O_i, P_j), is deployed in this study. It refers to execute the task P_j at location O_i. So the salience of each processing task is computed as the input-output relationship of each DMU (O_i, P_j). The classic CCR DEA model (Charnes et al. 1979) is integrated with a linguistic group decision preference cone constraint. The detail description of the proposed approach is illustrated in Figure 4-1. The model is designed to compute the information efficiency according the linguistic preference matrix by a group of decision-makers. The cone constraints for input and output decision variables, denoted as W and V respectively are constructed based on a group of decision-makers' pairwise comparison of the input and output variables based on the linguistic value. The detail model is described as below:

$$
\begin{aligned}
Max \quad & V^T Y_0 \\
s.t. \quad & W^T X_j - V^T Y_j \geq 0 \\
& W^T X_0 = 1 \\
& w \geq 0; v \geq 0 \\
& w \in W; v \in V \\
& W = \{w | A^* w \geq 0, w = (w_1, w_2, ...w_n)^T\} \\
& V = \{v | B^* v \geq 0, v = (v_1, v_2, ...v_n)^T\}
\end{aligned}
$$

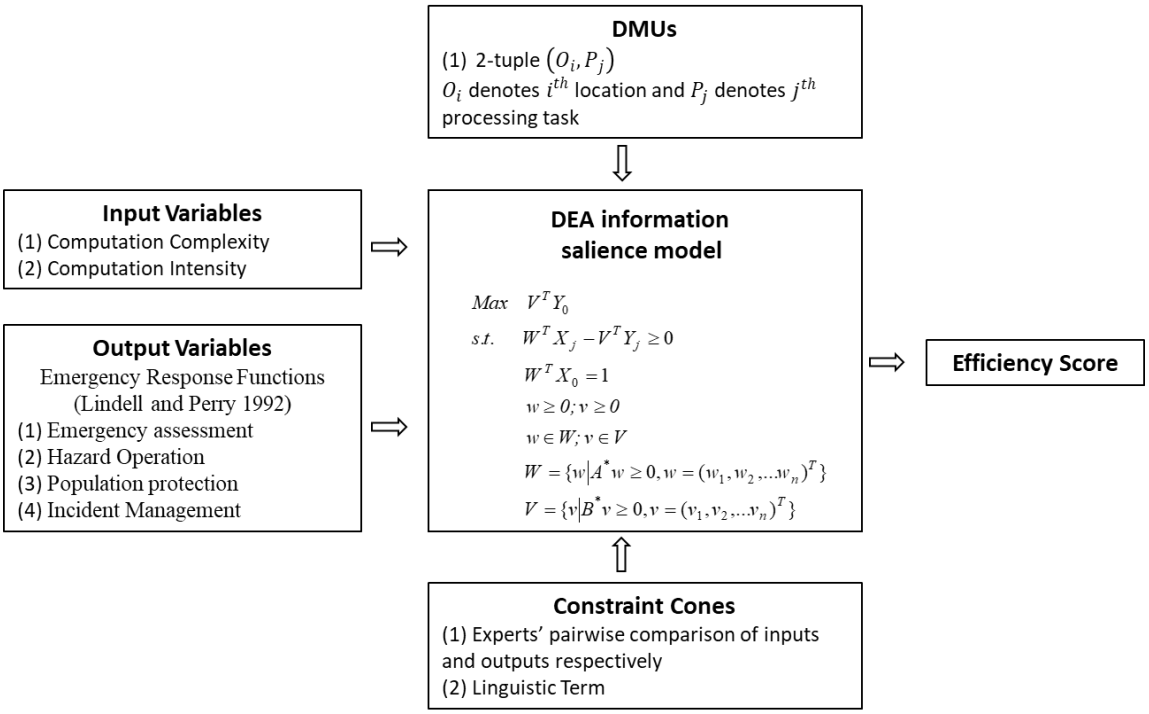


Figure 4-1 Description of proposed approach

Here, X_i and Y_i represent the input and output variables of DMU_i. The proposed DEA model is restrained by two weight relations, denoted as $w = (w_1, w_2, ...w_n)$ and $v = (v_1, v_2, ...v_n)$. These two weights represent the preference relation of the input and output decision variables respectively by a group of decision-makers. Then the dual model of the proposed model is described as follows:

Dual Problem

$$
\begin{aligned}
&Min \quad \theta \\
&s.t. \quad -X\lambda + \theta X_0 \in -U^*; \\
&\qquad Y\lambda - Y_0 \in -V^*; \\
&\qquad \lambda \geq 0.
\end{aligned}
$$

Here W^* and V^* represent the negative pole cone of the cone constraint W and V, respectively.

4.4. Case Study

4.4.1 Background

Hurricane Sandy (2012) was one of most deadliest and destructive hurricane in the Atlantic shoreline. It made landfall on southern New Jersey and took a major toll on the Ocean county of New Jersey. The shoreline area in Ocean County suffered the most server wind and surf from Hurricane Sandy. During the Hurricane, most of the public transformation was canceled. Many of the electric poles were subject to failure as a result of the strong wind. At the meantime, the water damage to critical electric equipment dramatically slowed the electric restoration. Besides, many of the gas stations were closed and the majority of the shoreline area were subject to fuel shortage. It is reported that over 5000 residential houses in the Ocean County were suffered from significant structural damaged during Hurricane Sandy (Sagara 2012). The Barnegat Peninsula, also known as Barnegat Bay Island and colloquially as "the barrier island", is a 20-mile long, narrow barrier peninsula located on the Jersey Shore in Ocean County, New Jersey. During Hurricane Sandy, the Barnegat Peninsula took a major hit and the entire island were sustained to significant damaged. On the ocean side, on the one hand, high velocity storm wave were striking the oceanfront dunes and house. On the bay side, on the other,

the house with relative low front door elevation were inundated by the flooding water. Moreover, the Barnegat Peninsula is isolated from the mainland only very limited accessibility: route 37 (Thomas A. Mathis and J. Stanley Tunney Bridges), route 528 (Mantoloking Bridge) and route 35. Therefore, the Barnegat Peninsula in Ocean County was selected as the study scope. Communities on the peninsula (as shown in Figure 4-2) include Point Pleasant Beach, Bay Head, Dover Beach, Mantoloking, Lavallette, Ortley Beach, Normandy Beach, Seaside Heights, and Seaside Park. In this section, 9 communities located Barnegat Peninsula in Ocean County with three different tasks are evaluated based on the proposed method.

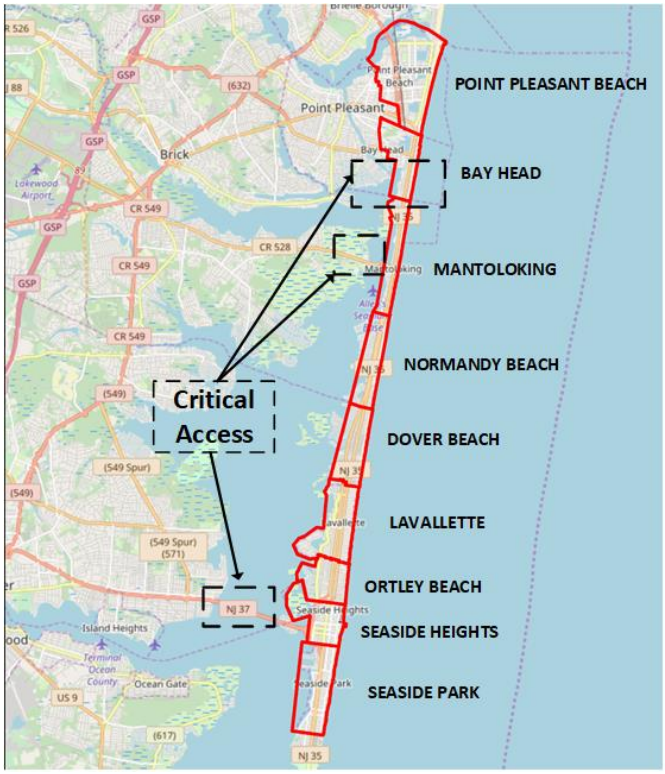


Figure 4-2 Map of Barnegat Peninsula in Ocean County

4.4.2 Data

Table 4-7 summarized the input-output used in this case study. The steps to derive input and output values are depicted in the previous section. Particular in this study, the

computation complexity is estimated based on the proposed scoring system. For simplification, this study selects the community area as the indicator for computation intensity. The area of each community is computed based on the 2010 census data shapefile in ArcMap. On the other hand, for the output variables, this case study adopted community vulnerability indicators directly from historical data based. For emergency assessment, this study address one major concern is safety evaluation of buildings after windstorms and floods by the ATC-45 manual (ATC 2004). Thus the building data is selected as indicators, which were retrieved from 2010 census data. For hazard operation, this hurricane Sandy based case study focuses on flood hazard. FEMA Flood Insurance Rate Map (FIRM), therefore, is chosen as the data resource. The value of this output variable is defined as percentage area of less than 100-return period flood zone. For the third output variable, population protection, the 2000 demographic data selected. It should be noted that the demographic information contained population breakdowns from different age groups, which could be of significant benefit in development evaluation or search and rescue operation strategies. However, for illustration purpose, only the total population was selected for this case study. The demographic data is also retrieved from 2000 Census Data. Finally, the value of incident management demand is estimated as the weighted sum of all critical infrastructures. The critical infrastructures considered in the case study includes of critical access (bridges, route), hospital, schools, gas stations. It is worthwhile mentioning that this study emphasizes the importance of critical access by assigning a weight of 10. All datasets used in this study are available in New Jersey Geographic Information Network (NJGIN). After the demand information were identified based on historical geospatial data based, experts from data processing team were invited

to determine the significance level of the data processing tasks in addressing to the four disaster functions Figure 4-3. In this study, three basic processing tasks were considered such as Visualizing (P_1), spatial analysis (P_2), and advanced analysis (P_3). The significance level of each processing tasks was summarized in Table 4-4 using a 7-tuple linguistic set described as below:

$$S = \{s_0 = None, s_1 = VeryLow, s_2 = Low, s_3 = Medium, s_4 = High,$$

$$s_5 = VeryHigh, s_6 = Perfect\}$$

Based on the information demand (Table 4-5), significance level of information (Table 4-4) and estimated complexity of each processing task (Table 4-6), the value of output variables was computed as shown in Table 4-7.

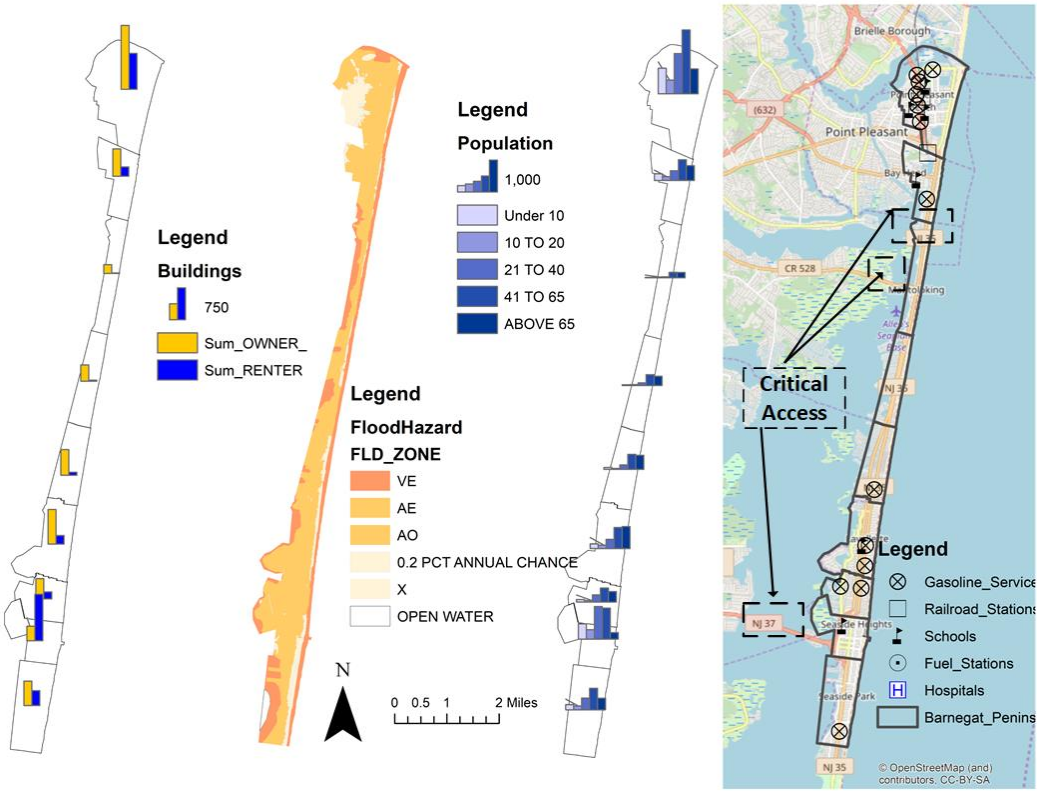


Figure 4-3 Indicators for disaster functions

Table 4-3 Area of each community

Community	Area (sq. mi)
O_1: POINT PLEASANT BEACH	1.87
O_2: BAY HEAD	0.75
O_3: MANTOLOKING	0.64
O_4: NORMANDY BEACH	0.68
O_5: DOVER BEACH	0.75
O_6: LAVALLETTE	1.05
O_7: ORTLEY BEACH	0.67
O_8: SEASIDE HEIGHTS	0.74
O_9: SEASIDE PARK	1.08

Table 4-4 Significance level of the information

	F_1: Emergency Assessment	F_2: Hazard Operation	F_3: Population protection	F_4: Incident Management
P_1: Visualizing	s_2	s_3	s_3	s_5
P_2: Spatial Analysis	s_4	s_5	s_5	s_1
P_3: Advanced Analysis	s_5	s_4	s_3	s_3

Table 4-5 Information demand

	F_1: Emergency Assessment	F_2: Hazard Operation	F_3: Population protection	F_4: Incident Management
	# of Buildings	1% Flood zone	Pop 2000	# Critical Infrastructure
O_1: POINT PLEASANT BEACH	2339	0.76	5388	10
O_2: BAY HEAD	856	0.93	1805	13
O_3: MANTOLOKING	215	0.97	400	13
O_4: NORMANDY BEACH	409	0.94	794	0
O_5: DOVER BEACH	676	0.88	1155	2
O_6: LAVALLETTE	1018	0.94	1937	3
O_7: ORTLEY BEACH	640	0.98	1136	2
O_8: SEASIDE HEIGHTS	1434	0.88	3040	11
O_9: SEASIDE PARK	930	0.78	1746	1

Table 4-6 Estimated Complexity of each processing task

Processing Task	Estimated Complexity
P_1: Visualizing	15
P_2: Spatial Analysis	27
P_3: Advanced Analysis	50

Table 4-7 Summary of Input-output data

	Inputs		Outputs			
DMU	X_1	X_2	F_1	F_2	F_3	F_4
(O_1, P_1)	0.30	1.00	0.09	0.05	0.15	0.18
(O_1, P_2)	0.54	1.00	0.18	0.08	0.26	0.04
(O_1, P_3)	1.00	1.00	0.23	0.06	0.21	0.11
(O_2, P_1)	0.30	0.40	0.03	0.06	0.05	0.06
(O_2, P_2)	0.54	0.40	0.07	0.10	0.09	0.01
(O_2, P_3)	1.00	0.40	0.08	0.08	0.07	0.03
(O_3, P_1)	0.30	0.34	0.01	0.06	0.01	0.24
(O_3, P_2)	0.54	0.34	0.02	0.10	0.02	0.05
(O_3, P_3)	1.00	0.34	0.02	0.08	0.02	0.14
(O_4, P_1)	0.30	0.36	0.02	0.06	0.02	0.00
(O_4, P_2)	0.54	0.36	0.03	0.10	0.04	0.00
(O_4, P_3)	1.00	0.36	0.04	0.08	0.03	0.00
(O_5, P_1)	0.30	0.40	0.03	0.05	0.03	0.04
(O_5, P_2)	0.54	0.40	0.05	0.09	0.06	0.01
(O_5, P_3)	1.00	0.40	0.07	0.07	0.04	0.02
(O_6, P_1)	0.30	0.56	0.04	0.06	0.06	0.06
(O_6, P_2)	0.54	0.56	0.08	0.10	0.09	0.01
(O_6, P_3)	1.00	0.56	0.10	0.08	0.07	0.03
(O_7, P_1)	0.30	0.36	0.02	0.06	0.03	0.04
(O_7, P_2)	0.54	0.36	0.05	0.10	0.05	0.01
(O_7, P_3)	1.00	0.36	0.06	0.08	0.04	0.02
(O_8, P_1)	0.30	0.40	0.06	0.05	0.09	0.20
(O_8, P_2)	0.54	0.40	0.11	0.09	0.15	0.04
(O_8, P_3)	1.00	0.40	0.14	0.07	0.12	0.12
(O_9, P_1)	0.30	0.58	0.04	0.05	0.05	0.02
(O_9, P_2)	0.54	0.58	0.07	0.08	0.08	0.00
(O_9, P_3)	1.00	0.58	0.09	0.06	0.07	0.01

4.4.3 DEA results

Scenario 0 takes into account the practical situation for disaster response, where the emphases were placed on both time and humanitarian relief. Cone constraints of both input variables and output variables were constructed to put emphases on the weight of computation complexity X_1 and population protectionF_3.The cone constraints used in scenario 0 are depicted as follows:

Scenario 0: Prioritization of the processing tasks, cone constraint in both input and output

Population protection (F_3), Less computation Complexity (X_2).

$$R_A = \begin{pmatrix} - & s_1 \\ s_5 & - \end{pmatrix}, A = \begin{pmatrix} 0.5 & 0.17 \\ 0.83 & 0.5 \end{pmatrix}, \overline{A} = \begin{pmatrix} 0.5 & 0.17 \\ 0.83 & 0.5 \end{pmatrix}, I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \lambda_A = 0.88 \text{; then}$$

$$A^* = \begin{pmatrix} -0.38 & 0.2 \\ 0.8 & -0.38 \end{pmatrix}$$

$$R_B = \begin{pmatrix} - & s_2 & s_0 & s_1 \\ s_4 & - & s_1 & s_2 \\ s_6 & s_5 & - & s_4 \\ s_5 & s_4 & s_2 & - \end{pmatrix} B = \begin{pmatrix} 0.5 & 0.33 & 0 & 0.17 \\ 0.67 & 0.5 & 0.17 & 0.33 \\ 1 & 0.83 & 0.5 & 0.67 \\ 0.83 & 0.67 & 0.33 & 0.5 \end{pmatrix}, \overline{B} = \begin{pmatrix} 0.5 & 0.39 & 0.17 & 0.28 \\ 0.61 & 0.5 & 0.28 & 0.39 \\ 0.83 & 0.72 & 0.5 & 0.61 \\ 0.72 & 0.61 & 0.39 & 0.5 \end{pmatrix},$$

$$\lambda_B = 1.87 \text{; then } B^* = \begin{pmatrix} -1.37 & 0.39 & 0.17 & 0.28 \\ 0.61 & -1.37 & 0.28 & 0.39 \\ 0.83 & 0.72 & -1.37 & 0.61 \\ 0.72 & 0.61 & 0.39 & -1.37 \end{pmatrix}$$

Based on the inputs and output and also the preference cone constraints determined by the group of decision-makers, the efficiency score, and rank of each individual DMU illustrated in Table 4-8. The first finding is that there is an apparent agreement between DEA efficiency scores and the information demand. The DEA result of scenario indicates that Seaside Heights (O_8) Point Pleasant Beach (O_1) and Bay Head (O_2) have the highest efficiency score. These three communities occupy nearly 60% of the population, and more than 50% of the houses in Barnegat Peninsula. Moreover, considering the critical infrastructure, route NJ37 in O_8 and route NJ35 in O_1 and O_2 are critical exits to the mainland area. Second, the priority of the processing task is constraint by the preference cone constraints. For instance, DMUs with P_3 are less preferred than

those without. This is because an input cone constraint that emphasis on less computation complexity task is added to the DEA model. Third, the selection of processing task is based on the information demand. For instance (O_3,P_1) ranks second among all DMUs. This is typically because there is a critical access (CR528) in Mantoloking O_3 that connecting the Barnegat Peninsula to the mainland. Stressing on the importance of critical infrastructure assessment, Visualization task P_1 is favourite.

Table 4-8 Scenario 0 result

Efficiency Rank	Efficiency score	DMU	Efficiency Rank	Efficiency score	DMU	Efficiency Rank	Efficiency score	DMU
1	1	(O_8, P_1)	10	0.42	(O_6, P_2)	19	0.35	(O_9, P_2)
2	0.89	(O_3, P_1)	11	0.42	(O_7, P_2)	20	0.33	(O_6, P_3)
3	0.76	(O_8, P_2)	12	0.41	(O_3, P_3)	21	0.32	(O_4, P_2)
4	0.65	(O_8, P_3)	13	0.39	(O_7, P_1)	22	0.3	(O_7, P_3)
5	0.58	(O_1, P_2)	14	0.38	(O_3, P_2)	23	0.28	(O_5, P_3)
6	0.57	(O_1, P_1)	15	0.38	(O_5, P_2)	24	0.27	(O_9, P_1)
7	0.52	(O_2, P_1)	16	0.38	(O_6, P_1)	25	0.26	(O_9, P_3)
8	0.49	(O_2, P_2)	17	0.36	(O_2, P_3)	26	0.23	(O_4, P_1)
9	0.47	(O_2, P_1)	18	0.35	(O_5, P_1)	27	0.20	(O_4, P_3)

4.4.4 Significance of the cone constraints

To verify the significance of the cone constraints in the DEA model, scenario 1 to 4 is performed. The detail settings of these four scenarios are described as below:

Scenario 1: No constraints

Scenario 2: Time-sensitive applications, Cone constraint in inputs

$$A^* = \begin{pmatrix} -0.38 & 0.2 \\ 0.8 & -0.38 \end{pmatrix}$$

Scenario 3: Prioritization of location, Cone constraint in outputs

$$B^* = \begin{pmatrix} -1.37 & 0.39 & 0.17 & 0.28 \\ 0.61 & -1.37 & 0.28 & 0.39 \\ 0.83 & 0.72 & -1.37 & 0.61 \\ 0.72 & 0.61 & 0.39 & -1.37 \end{pmatrix}$$

Scenario 4: Prioritization of the processing tasks, cone constraint in both input and output: Population protection (F_3), Less computation Complexity (X_2).

$$A^* = \begin{pmatrix} -0.38 & 0.2 \\ 0.8 & -0.38 \end{pmatrix} \text{ and } B^* = \begin{pmatrix} -1.37 & 0.39 & 0.17 & 0.28 \\ 0.61 & -1.37 & 0.28 & 0.39 \\ 0.83 & 0.72 & -1.37 & 0.61 \\ 0.72 & 0.61 & 0.39 & -1.37 \end{pmatrix}$$

For scenario 1, a normal CCR DEA model is applied to our case study. The estimation efficiency θ_1 is illustrated in the second column of Table 4-9. DEA is a nonparametric method, the calculation takes into consideration all four output variables and 2 inputs. Because there were no additional constraints on inputs and output variables, 12 DMUs were located in the efficiency frontier as depicted in Table 4-10. Each of the DMUs located on the efficiency frontier was evaluated based on different output variables. For instance, though, O_3 (Mantoloking) has only a few population and residential building, it has the maximum value in F_4 (incitement management.) Therefore, when considering the incident management F_4 tasks, O_3 should be top priorities. For scenario 2, the time-sensitive environment was simulated by applying a cone constraint to the input variables. The cone constraint A^* emphasized on the weight of computational complexity, which means that to achieve rapid computing, the preference was placed on the processing that requires less computation complexity (time). The efficiency result, denoted as θ_2 , is illustrated in Table 4-9. By comparing the result with scenario 1 and scenario 2, it can be obtained that high complexity task P_3

significantly reduced as a result of the input cone constraint. For example, take ($O_{3,}P_3$) as an example, the efficiency score decreased from 1.0 to 0.68 when the cone constraint in input variables is applied. Scenario 3 considered the situation of humanitarian relief. A cone constraint in output variables was applied to the DEA model. The cone constraint emphasize the importance of F_3 (Population protection) over other alternative functions. Therefore, it is expected that the efficiency will shift to locations that have larger populations. By comparing the scenario 1 and scenario 3, there witnesses an apparent reduced in the efficiency in locations with few populations. For example, O_3 (Mantoloking) has the fewest population among 9 communities. Therefore, when the preference for data processing was placed on population protection, the efficiency of processing information in O_3 dropped tremendously. On the other hand, in communities with largest population such as O_1 (Point Pleasant Beach), O_8 (Seaside Heights), the majority of the DMUs associated with these locations $((O_1, P_1), (O_1, P_2), (O_8, P_2), (O_8, P_3))$ remained on the new efficiency frontier.

Table 4-9 Efficiency results

DMU	θ_1	θ_2	θ_3	θ_4
	No Constraint	Constraint in input only	Constraint in output only	Constraints in both input and output
(O_1, P_1)	1.00	0.72	1.00	0.57
(O_1, P_2)	1.00	0.94	0.71	0.58
(O_1, P_3)	0.97	0.92	0.56	0.52
(O_2, P_1)	0.98	0.74	0.48	0.47
(O_2, P_2)	0.98	0.94	0.58	0.49
(O_2, P_3)	0.81	0.62	0.56	0.36
(O_3, P_1)	1.00	1.00	0.92	0.89
(O_3, P_2)	1.00	1.00	0.47	0.38
(O_3, P_3)	1.00	0.68	0.69	0.41
(O_4, P_1)	0.96	0.71	0.23	0.23
(O_4, P_2)	0.96	0.95	0.39	0.32
(O_4, P_3)	0.77	0.57	0.33	0.20
(O_5, P_1)	0.91	0.66	0.35	0.35
(O_5, P_2)	0.90	0.86	0.45	0.38
(O_5, P_3)	0.73	0.56	0.42	0.28
(O_6, P_1)	1.00	0.59	0.47	0.38
(O_6, P_2)	0.97	0.79	0.45	0.42
(O_6, P_3)	0.61	0.55	0.43	0.33
(O_7, P_1)	1.00	0.78	0.40	0.39
(O_7, P_2)	1.00	1.00	0.51	0.42
(O_7, P_3)	0.88	0.64	0.48	0.30
(O_8, P_1)	1.00	1.00	1.00	1.00
(O_8, P_2)	1.00	1.00	0.91	0.76
(O_8, P_3)	1.00	1.00	1.00	0.65
(O_9, P_1)	0.84	0.46	0.34	0.27
(O_9, P_2)	0.81	0.65	0.37	0.35
(O_9, P_3)	0.52	0.50	0.34	0.26

Table 4-10 Efficiency frontier

	θ_1	θ_2	θ_3	θ_4
Efficiency frontier	$(O_1, P_1)(O_1, P_2)(O_3, P_1)$ (O_3, P_2) (O_3, P_3) (O_6, P_1) $(O_7, P_1)(O_7, P_2)(O_8, P_1)$ $(O_8, P_2)(O_8, P_3)$	$(O_3, P_1)(O_3, P_2)$ $(O_8, P_1)(O_8, P_2)$ (O_8, P_3)	$(O_1, P_1)(O_8, P_1)$ (O_8, P_3)	(O_8, P_1)

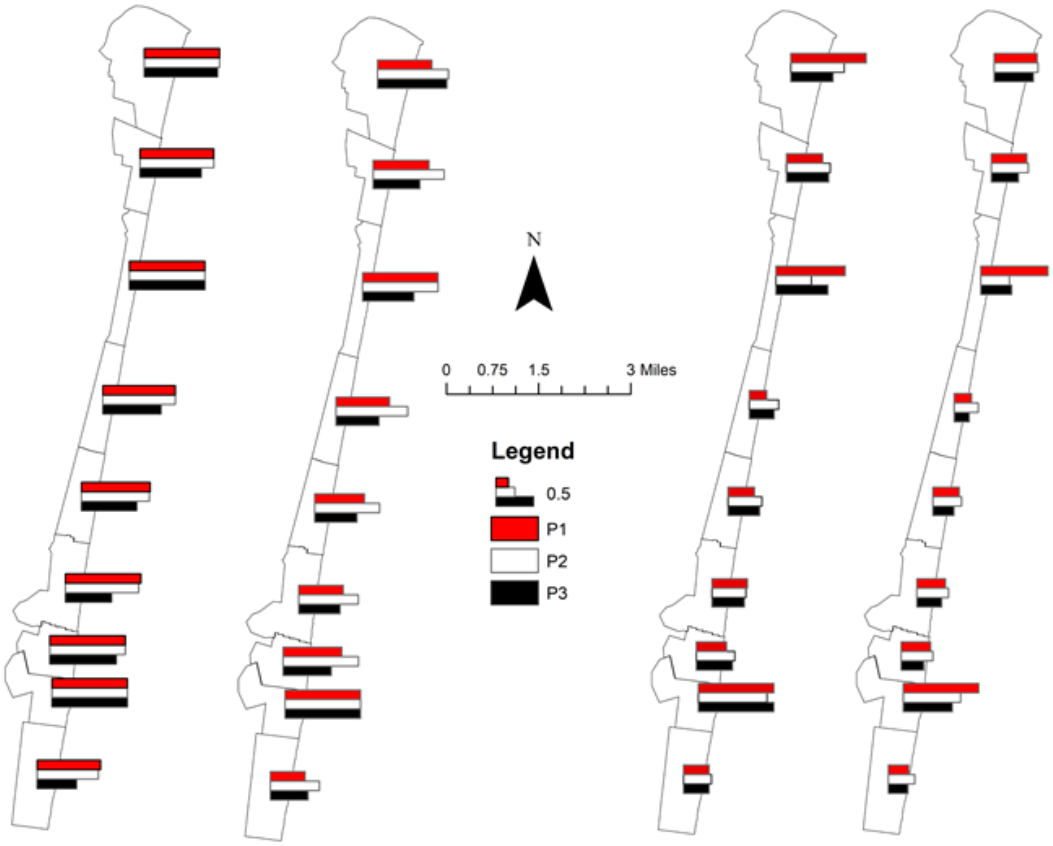


Figure 4-4 Efficiency Map

In sum, the integration of experts' preference into cone constraints can promote the understanding of information salience. For traditional DEA model (scenario 1) without cone constraints, it is likely to result in too much DMUs on the efficiency frontier. The balance between each input variables or output variables is not properly established. By applying the cone constraints, the efficiency scores of each DMUs can adjust according to the experts' preference such as less computation complexity jobs or population-preferred jobs.

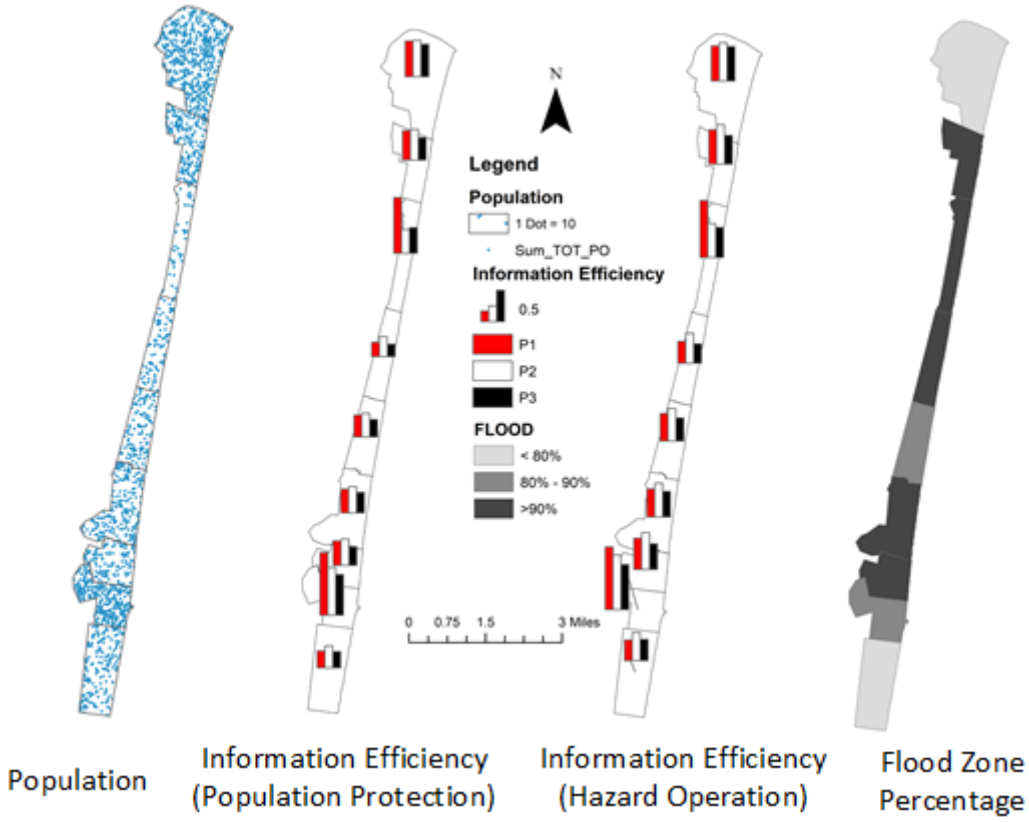


Figure 4-5 Comparison of information efficiency.

4.4.5 Sensitivity of the cone constraints

It is expected that the information efficiency scores or ranks will not be dominated by the experts' preference. In addressing this argument, a comparison study is performed based on scenario 5 and 6. The settings are illustrated as follows:

Scenario 5: Population protection (F_3) preferred, Less computation Complexity (X_2) Preferred.

$$A^* = \begin{pmatrix} -0.38 & 0.2 \\ 0.8 & -0.38 \end{pmatrix} \text{ and } B^* = \begin{pmatrix} -1.37 & 0.50 & 0.17 & 0.50 \\ 0.50 & -1.37 & 0.17 & 0.50 \\ 0.83 & 0.83 & -1.37 & 0.83 \\ 0.50 & 0..50 & 0.17 & -1.37 \end{pmatrix}$$

Scenario 6: Hazard Operation (F_2) preferred, Less computation Complexity (X_2)

Preferred. $A^* = \begin{pmatrix} -0.38 & 0.2 \\ 0.8 & -0.38 \end{pmatrix}$ and $B^* = \begin{pmatrix} -1.37 & 0.17 & 0.50 & 0.50 \\ 0.83 & -1.37 & 0.83 & 0.83 \\ 0.50 & 0.17 & -1.37 & 0.61 \\ 0.50 & 0.17 & 0.50 & -1.37 \end{pmatrix}$

In practical applications, the objective for data processing is often poorly defined and involves multiple criteria. It means that there exist no apparent discrepancies in each input or output variables. Under such circumstance, because of different backgrounds and different interests, experts might have a bias on the preference of data processing sequence. Therefore, it should be avoided that the information efficiency remains dominated by the preference of the experts, or in this case preference cone constraints. To verify whether the preference cone constraints are principal in determining the information efficiency, scenario 5 and 6 were simulated. Scenario 5 emphasis on the population protection tasks F_3 while scenario 6 emphasis on the importance of hazard operation F_2. The findings are described as below. From Figure 4-5, it can be obtained that, even the output cone constraint alters, the efficiency ranks maintains following similar patterns (Figure 4-6). It is typical because the information scores are majorly dominated by the input-output relationships rather than the cone constraints. For instance, considering the (O_3, P_1) in scenario 5, though there are only very few population in (O_3, P_1), it is still among the top efficiency tasks, because O_3 had the maximum number of critical infrastructure, which the visualization analysis (P_1) is in large demand.

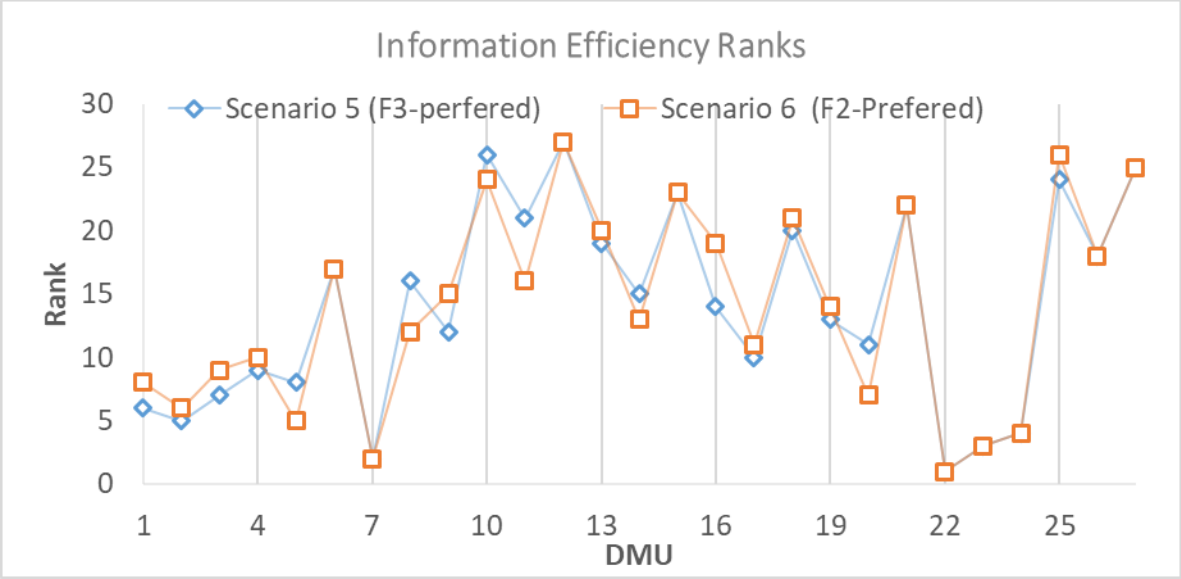


Figure 4-6 Efficiency Ranks of Scenario 5 and Scenario 6

Moreover, from the methodologic perspective, the output cone constraint in scenario 6 can be simplified as follows:

$$-1.37v_1 + 0.17v_2 + 0.50v_3 + 0.50v_4 \leq 0$$

$$0.83v_1 - 1.37v_2 + 0.83v_3 + 0.83v_4 \leq 0$$

$$0.50v_1 + 0.17v_2 - 1.37v_3 + 0.50v_4 \leq 0$$

$$0.50v_1 + 0.17v_2 + 0.50v_3 - 1.37v_4 \leq 0$$

Given that the weights for each output v_j is a non-negative upper bound and lower bound of the weight of hazard operation F_2 can be computed as below:

Upper bound: $v_2 \leq \frac{0.37}{0.51}(v_1 + v_3 + v_4) \leq \frac{0.37\times3}{0.51} max(v_1, v_3, v_4) = 2.17 max(v_1, v_3, v_4)$

Lower bound $v_2 \geq \frac{0.83}{1.37}(v_1 + v_3 + v_4) \geq \frac{0.83\times3}{1.37} min(v_1, v_3, v_4) = 1.81\, min(v_1, v_3, v_4)$

Therefore, the weight range of F_2 is $[1.81\, min(v_1, v_3, v_4)\,, 2.17 max(v_1, v_3, v_4)]$. In cases that there is an apparent discrepancy in the preference of the output or input

variables, the model can be modified by subtracting the abundant input or output variables.

4.5 Conclusions

In a time-sensitive environment such as disaster response decision-making, information salience is a noteworthy matter. Notably, with the growing amount of data, it is prone to result in "data-rich-but-information-poor" situation if the salience of information is not clearly specified (Timmerman et al. 2010). Disaster response often involves multiple criteria that the objective of information processing cannot be clearly defined. The present DEA model is a vigorous way to handle information salience with such a vague objective. In this research, the information salience or "information efficiency" is modeled using a DEA based model. The model integrates with two cone constraints on input and output variables, respectively. These cone constraints are constructed based on the linguistic preference from the decision makers. In general, the main contributions of this research are described as follows:

- An innovative perspective based on "information efficiency" is introduced for tasks prioritization during disaster response.
- Information needs preference was modeled as cone constraints using the linguistic term, which can more objectively reflect opinions from the decision makers.
- The group decision process is embedded in the traditional DEA model, which distinguishes the specific features of disasters.

The model is then validated based on a Hurricane Sandy based humanitarian relief response case study in Barnegat Peninsula. The need for time-sensitive humanitarian relief tasks are simulated as two emphases: (1) low computational complexity tasks; (2) Population protection. Based on these emphases, input and output cone constraints are constructed, respectively. Two input variable selected in this study are computational

complexity and computational intensity. Besides, the four disaster function proposed by (Lindell and Perry 1992) is selected as the output variables. From scenario 0, the result indicates that among the nine communities, Pleasant Beach (O_1), Bay Head (O_2) and Seaside Heights (O_8) are identified as communities with a top information need. Besides, processing task P_3 is less preferred because according to the input data, it has a higher computation complexity. In order to verify the importance of the cone constraints, four scenarios are compared. The comparison analysis illustrates that the integration of experts' preference into cone constraints can promote the understanding of information salience. Finally, simulations from scenario 5 and 6 clarify that information scores are majorly dominated by the input-output relationships rather than the cone constraints.

This study has two research tasks to be explored in future. One extension is the implementation of the present model for cloud computing purpose. This study considered majorly for the local computing situations. Thus the compound computational investment is simplified as two input variables: computation complexity and computation intensity. However, in cloud computing environment, the data processing have more options such as approximation computing, distributed computing. Therefore, the efforts in approximation, data transferring can be modeled as additional input variables. Another extension is validating the present method through real disaster decision making cases. The experts' preference matrix is simulated in this study. It is of profound meaning to test the method based on inputs from real decision-makers.

In sum, it can be seen that there are many reasons for identifying the information salience during disaster situation to avoid "data-rich-but-information-poor" situation. The value of the data is not achieved by collecting them or processing them, but by meeting

the decision's need. The comparison of information efficiency scores obtained from the present DEA model allows an information salience based evaluation. Given the importance of time-critical information during time-sensitive environment such as disaster response, it is hoped that this study contributes to promoting the effectiveness and efficiency of data processing.

Chapter 5 Accelerating Information Extraction from Spatial Disaster Data using Big Data Infrastructure

5.1 Introduction

During disasters, decision-makers seek information as a process of sense-making to form their point of view (Devin 1983). Information serves as evidences to assimilate into what is already known (e.g., experience, knowledge) (Kuhlthau 1991). The importance of information about disaster impacts can never be underestimated because disasters are often extreme events which can deviate significantly from past experiences. In reality, the value of data or information gathered during extreme events is often not fully exploited. For instance, during Hurricane Katrina, disaster response efforts were hindered by poor information flows, and inability to validate and process relevant information in a timely fashion (Thompson et al. 2006).

One reluctance in exploiting big spatial data comes from the doubt of whether such data can be processed in time. Disaster response involves making difficult decisions within a short time window. The time sensitivity associated with disaster response dictates the time sensitivity nature in data processing during natural disaster events. (Lippitt et al. 2014). The value of information in time-sensitive environments diminishes rapidly as time goes on (Hodgson et al. 2014). If the information arrives too late, it may have little if not any value, for decision makers during disaster response. To avoid this situation, data processing teams have to overcome the challenge of processing large data sets in a short amount of time. A typical spatial data application often includes a series of processing steps(e.g., data clean-up, data transformation, classification and feature extraction) to generate the final data products (e.g., visualization data, digital elevation model (DEM)). These processes are computation intensive as a result of the massive

amount of data (Li et al. 2018) as well as the spatial-temporal nature of these data sets. (Wang and Armstrong 2003).

The geospatial data processing tasks during disaster scenarios can be further complicated by the limited processing capability of the core as well as power outages or other types of cyber disruptions (Cao et al. 2015). Under normal circumstances, data processing may be performed in core. The computational capability of the core can be quickly overwhelmed during disastrous events due to the sudden ingestion of continuously updated large data sets collected by both centralized remote sensing systems and localized community sensor nodes such as social media. Moreover, it is reported that Hurricane Sandy (2012) caused more than 7 million customers without power. The core can be at the exceptional risk of being shut down as a result of power outages during disasters. To this end, for important processing tasks, it makes more sense not put all the eggs in one basket (core). Therefore, seeking alternative computational resources is essential during disaster environments. These computational resources can be (1) remotely High-Performance Computing (HPC) infrastructures that do not rely on the power in the affected area such as part of clusters (consisting of several personal computers), clouds (e.g., Amazon EC2, Microsoft AZURE), or (2) any other edge device that has backup power system as well as networking and computational capabilities (e.g., minivan-based mobile data collection system). On the one hand, High-Performance Computing (HPC) infrastructures are are now ubiquitous at a reasonable cost (Yang and Chen 2010), but reported to have the most potential in improving the performance and scalability of information processing (Han et al. 2009; Hegeman et al. 2014; Qiu et al. 2014). On the other hand, processing at edge device can couple data processing with data

collection, reducing the unnecessary data transferring to core. In addition, big data processing platforms, such as Hadoop, Apache Spark, Storm, have been developed to facilitate the use of a network of commodity computers for processing large data sets. In sum, these data processing infrastructures and platforms have great potentials in improving the data processing tasks during natural disasters.

To address the need of processing large spatial data sets, there are numerous attempts in deploying HPC infrastructures for processing big spatial data (Li et al. 2018). Much of the current efforts on disaster-related studies have emphasized on the custom rewriting of the spatial processing algorithms to enable distributed and parallel computing capability (Han et al. 2009; Huang et al. 2011; Wu et al. 2011). Much of the above-mentioned customization are designated to a designated application, which cannot extend to a more general purpose. Nevertheless, onsite rewriting and verifying all the required algorithms are laborious and time prohibited during disaster response time sensitive environment. In a more recent study, Li et al. (2018) proposed the concept of workflow-style architecture for large-scale LiDAR data processing. Compared to other applications (Han et al. 2009; Huang et al. 2011; Wu et al. 2011) that are packaged as a whole, the workflow-style processing depicts a computation task as a sequential of linking operations, where each operation is independent and executable. Such a workflow architecture offers the flexibility to insert and remove operations from the sequence to construct new applications. However, Li et al. (2018) only tested the workflow-style architecture in a relatively simple specified DEM application, the ability to chain a set of tools to support complex workflow-style processing is not fully implemented.

The overall goal of this study is to propose a workflow-style framework for spatial data processing to generate ready to use intelligence to support decision-making. By doing so, it is expected that the generic framework will enable complex spatial processing tasks to be constructed by chaining a set of elementary operations. Specifically, we first decompose complex LiDAR processing tasks into elementary operations. Then, the proposed solution uses LAStools, one of the most popular LiDAR processing tools, to realize these elementary operations. These elementary operations are reorganized as a topology-based flow on an Apache Strom based framework (Renart et al. (2017) to support stream processing on edge devices and HPC infrastructures. The prototype of this framework is validated on a Hurricane Sandy use case to evaluate its efficiency and effectiveness. The results suggested that the proposed approach can process data up to 69% faster than the typical processing approach using a standalone computer and reduced 89% of the unnecessary data transfer. The main contributions of this paper are described as follows:

- An innovative approach that explores the feasibility of deploying high-performance computing (HPC) to process time-sensitive disaster data
- A novel data processing paradigm that restructures complex processing tasks as workflow-style elementary operations that can support stream processing

The rest of the paper is structured as follows: Section 2 presents the related work. The proposed methodology is presented in section 3. Section 4 validates the approach using a Hurricane Sandy based use case. Finally, the last section provides the final conclusion and future work.

5.2 Related work

5.2.1 High-performance computing tools and strategies

In the case of processing an excessive amount of data in a short amount of time, High-Performance Computing (HPC) has repeatedly been highlighted as a solution (Alonso et al. 2004). There are numerous attempts in investigating the feasibility of distributing data processing workloads to more abundant computation nodes, such as remotely emulated edges (Han et al. 2009; Renart et al. 2017; Renart et al. 2017) or cloud (Hegeman et al. 2014) to speed-up the data processing. To achieve this goal, these studies either customize their own algorithms such that they can be ran in distributed manner (Awrangjeb et al. 2013; Wu et al. 2011) or extend the functions from existing platforms such as Apache Hadoop (Jian et al. 2015; Růžička et al. 2017), Spark (Brédif et al. 2015) and Apache Storm (Renart et al. 2017). Due to the exponential growth of computing power, data processing are undergoing major transformations. The first transformation is the shift from CPU computing to GPU computing (Liu 2013). For instance, Awrangjeb et al. (2013) used GPU for faster filtering of LiDAR point cloud in urban areas. Similar work in using GPU to accelerate data processing can be found in Lukač and Žalik (2013), Guan and Wu (2010), Sugumaran et al. (2011), and Meng et al. (2012). The second transformation is the shift from processing batch data to processing streaming data (or synonym for real-time and near real-time data) and supporting interactive analysis. Renart et al. (2017) proposed a content-driven online stream processing framework for handling LiDAR data. Kreylos et al. (2008) developed software for LiDAR data immersive visualization and interactive analysis for quality control and extraction of survey measurements. The third transformation is utilizing cluster, cloud, and edge computing. Han et al. (2009) implemented their LiDAR-to-DEM conversion approach

using a cluster of commodity computers and virtual grid. Hegeman et al. (2014) explored the use of an advanced and high-memory cloud-computing environment to process large-scale topographic LiDAR data. Renart et al. (2017) investigated the performance and scalability of using edge computing resource for LiDAR processing. Nevertheless, these LiDAR data related studies have limited their analytics applications to some specific tasks such as DEM generation and change detection. A generic workflow-style processing framework that can provide a mechanism to chain a set of tools to represent and realize complex processing tasks in time-sensitive environments is still largely missing. (Li et al. 2018).

5.2.2 Challenge in handling data processing problem in disaster response

Information for decision support during disasters requires handling an excessive amount of heterogeneous data. Because the increasing number of all types of data as well as the expanding strength of HPC and relative lower cost (Yang and Chen 2010), more and more agencies, began to implement HPC for their applications, especially in dealing with the massive amount of data in a time-sensitive environment (Yang et al. 2011). One growing area of applications for HPC is geospatial data processing during natural disasters. Yang et al. (2011) concluded the four challenges in handling geospatial data and these challenges are useful metrics to gauge the complexity of geospatial data processing tasks in an HPC environment.

- **Data intensity**: decision-makers are often overwhelmed by the data sets. The advancement of sensing technology (Yang et al. 2011) as well as the emerging of IoT (Renart et al. 2017) facilitate the rapid collection of the massive amount of disaster data. For example, regarding Hurricane Sandy, Rutgers mobile LiDAR system has collected over 575 GB of high-resolution and accurate 3D LiDAR data and images in

the aftermath of Hurricane Sandy (Gong 2013). Twitter released that people sent more than 20 million Tweets during the peak time (between Oct 27 and Nov 1) of Hurricane Sandy (2012). These data events are unprecedented in terms of data intensity, posing significant challenges to leverage these datasets for meaningful decision making. (Cui et al. 2010).

- **Computing Intensity:** The algorithms and models for deriving information from growing volumes of data for decision support are becoming more complicated with the increasing demands on more detail, relevant and accurate information. For instance, traditionally, the relatively low resolution airborne LiDAR-derived digital terrain models (DEMs) have been traditionally used for modeling large-scale terrains (Liu 2008). However, more complex algorithms have been recently developed to extend LiDAR-derived DEMs for more detail urban feature extraction (Priestnall et al. 2000) and pipeline risk assessment (Zhou et al. 2016).
- **Concurrent Intensity:** Decision-making during disaster response is a multi-level process, which involves different groups of decision-makers, experts as well as stakeholders. Moreover, recent development in IoT (Renart et al.) and distribuend geospatial information processing (Yang et al. 2008) have encouraged public participation in decision-making. Though it remains to debate whether the effort to encourage public or voluntary participation in disaster decision making is worthy or not (Irvin and Stansbury 2004). The study by Sakaki et al. (2010) has already shown that social sensors (information contributed by individuals such as VGI) could deliver faster earthquake notification than what have done by Japan Meteorological Agency. With the potential of more social sensors in the future, it is expected that the information support system in the future needs to be capable of handling intensive concurrent access.
- **Spatio-temporal intensity.** To better understand the past and predict the future, it is necessary to establish a time series of the geospatial data observations. The advancement of sensing technology enables revisiting the same area in a more timely fashion (Goodchild et al. 2007). For example, the Ocean County in New Jersey has been revisited three times by various governmental Airborne LiDAR survey teams to track the impacts of Hurricane Sandy (2012). One effective way to evaluate impacts

is change detection between multi-temporal data sets. Therefore, it is natural that the decision information support framework should be capable of not only processing newly collected data but also retrieving historical data for reference and comparison.

5.2.3 Existing studies for deploying HPC in disasters

While the HPC tools are widely available, the remaining question is how these tools fit into the disaster context. In literature, existing studies for deploying HPC in disasters can be summarized into three categories: (1) using HPC strategies for disaster forecasting or prediction; (2) customizing functions to support HPC; and (3) developing big spatial data management systems to support efficient query and retrieval of data sets. For the first category, there were extensive research efforts in investigating the feasibility of adopting HPC strategies in disasters such as earthquakes (Meng et al. 2012; Vugteveen et al. 2014), hurricanes (Allen 2007; Westerink et al. 2004) and dust storms (Xie et al. 2010). For instance, Meng et al. (2012) implemented GPU-based parallel computing to detect the earthquake. In another study, Allen (2007) reviewed the ongoing development of a Dynamic Data Driven Application System in forecasting hurricane events. However, much of these works were emphasizing disaster forecasting or predicting, which tend to be less time-sensitive than it is in the disaster response scenarios. Other works have focused on finding customized solutions for specific algorithms such as LiDAR-to-DEM conversion (Guan and Wu 2010; Han et al. 2009; Huang et al. 2011; Jian et al. 2015; Li et al. 2018). For example, Han et al. (2009) proposed a parallel approach using a PC cluster and virtual grid for LiDAR-to-DEM conversion. Huang et al. (2011) customized the IDW interpolation algorithm to support parallelism. Guan and Wu (2010) exploited the thread level parallelism on multi-core platforms. Similar works have been conducted on rewriting building feature extraction (Lee 2012), hydrological feature extraction

(Tesfa et al. 2011), vertex decimation for data reduction (Sugumaran et al. 2011) algorithms to support parallelism. Though all these studies indicated that the HPC strategies would be of remarkable benefits in accelerating the processing speed, these studies are geared towards specific analytics applications and are difficult to be extended to more general applications. The third research focus is on developing distributed data organization systems. Aji et al. (2013) presented a Hadoop-GIS system to support spatial data management, in particular for query of massive spatial data. Coddington et al. (1999) introduced a Web-based distributed Geographic Information System (GIS) for decision support. Qiu et al. (2014) presented a framework to enable cloud computing in emergency management systems. Li et al. (2018) proposed a tile-based system for LiDAR management and parallel processing. A common limitation of these studies is that the methods proposed in these studies require further development and validation before the system or the framework can be applied in real disaster decision-support scenarios. There are a wealthy portfolio of existing LiDAR processing tools, such as LASTools, Terrasolid, CloudCompare, PointTool, and ArcGIS LiDAR Analyst, etc. Only very few studies (Kersting and Kersting 2005; Li et al. 2018) exploited the capability of these tools in a HPC environment. To this end, the scoping of this study is to develop a general framework for LiDAR stream processing by coupling one of the most popular LiDAR processing tool, LAStool, with Apache Storm platform. The focus is on how to design and provide a framework that can allow users to model the LiDAR data processing workflow in an existing data analytics environment, transform the workflow into a computational graph that can be executed by a cluster of computers, and optimize the data movement and processing speed.

5.3 Methodology

A streaming processing framework is proposed in this study for information extraction from large quantities of LiDAR data (Figure 0-1). First, a tile-based system is used to manage the LiDAR data. The LiDAR data can be data sets collected from disaster scenes using different sensing methods such as airborne LiDAR or mobile LiDAR. . These datasets will be uploaded to the cloud. Second, the complex processing tasks are decomposed into elementary operations, which can be parallelized on the available computation nodes in the cluster or cloud. Last but not least, the processing workflow is converted to a Storm-based topology for stream processing in the cloud.

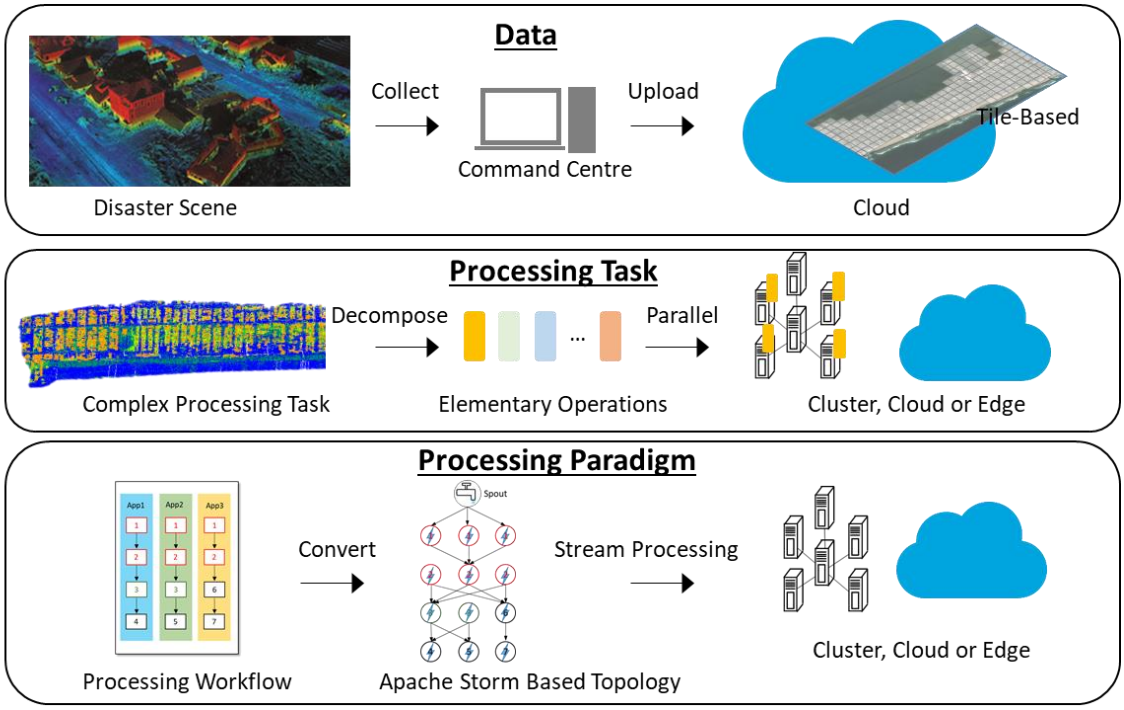


Figure 5-1 Methodology

5.3.1 Tile-based LiDAR management system

All the LiDAR data are managed using a tile-based structure. A multi-scale tile system is utilized to store and manage the LiDAR data. In the tile-based structure, the entire data are segmented into rectangular shapes called tiles. Figure 5-2 depicts the tiles

for airborne LiDAR data and mobile LiDAR data respectively. Figure 5-2a demonstrates the tiles for USGS EAARL-B Post-Sandy LiDAR. Figure 5-2b displays the tiles of the archived Rutgers mobile LiDAR survey team data in Seaside Park, New Jersey. The size of a tile is determined by the resolution of the data as well as the purpose of the applications. For instance, the tile size for airborne LIDAR data is often much larger than the tile size for the mobile LiDAR due to the relatively low resolution. A typical tile size for airborne LiDAR is 1000 meter by 1000 meter, which supports community-wide applications. A typical tile size for mobile LiDAR data is 25 meter by 25 meter, which supports much localized assessment such as individual building analysis or debris impact analysis.

In the tile-based system, each of the tiles contains two attributes, $Tile\_Name$ and $bounding\ box$. The naming rule $Tile\_Name$ is described as follows. Basic information is required to include the $Tile\_Name$. This information includes but not limited to (1) type of the data, (2) data collection system, (3) collection date; (4) data format. For example, $Mobile\_Rutgers\_20160912.las$ means the a mobile LiDAR data tile collected using Rutgers Mobile LiDAR system in Sep 12 2016 as .las format. The second important attribute is $bounding\ box$. A $bounding\ box$ is an object determined by four parameters: upper latitude, lower latitude, left longitude and right longitude, denoted as $Upper\_Lat, Lower\_Lat,\ Left\_Long,$ and Right_Long. Then the rectangular shape of a $bounding\ box$ can be represented as the four corner points as depicted in Figure 5-2c. This tile-based LiDAR management system enables indexing mechanism that will support a key function, namely spatial query. As long as the specific location (x, y) is given, the corresponding LiDAR data set can be identified rapidly by matching (x, y)

with the list of $bounding\ box$ such that $Upper\_Lat < y < Lower\_Lat$ and $Left\_Long < x < Right\_Long$.

Figure 5-2 Tile-based LiDAR management system

(a) Tiles for USGS EAARL-B Post-Sandy Airborne LiDAR;
(b) Tiles of the archived Rutgers mobile LiDAR;
(c) Details of a bounding box

In the proposed system, all the data are uploaded to the cloud for storage. Cloud storage is a service where data is remotely maintained, managed, and backed up. The service allows the users to store files online so that they can access them from any location via the Internet. Comparing to local storage, cloud storage has three advantages. First, the cloud storage paradigm helps to undertake information gathering. Disaster response often requires fusing information from different sensors as well as combining data from different physically detached data centers. The cloud storage paradigm enables information gathering that would otherwise be extremely hard to accomplish if not outrightly impossible without the benefit of cloud storage. Second, cloud storage can, to a certain degree, reduce the unnecessary computation or bandwidth overload in the

response center. Instead of direct file transferring, remote end-users can retrieve the data in the cloud via a web link. Last but not least, cloud storage can also reduce the risk of data loss due to power outages in disasters or physical damage to the data storage medium.

5.3.2 Decomposition of a processing task into elementary operations

LiDAR analytics are inherently time-consuming due to complex geospatial algorithms (Wang and Armstrong 2003) as well as diverse selection of processing parameters (Guan et al. 2013). However, HPC that performs a single complex processing task does not take advantage of its parallel processing capabilities. Therefore, the complex processing tasks need to be divided into elementary operations in a way that "the same operations" can be computed in several nodes simultaneously. To this end, seven core operation categories based on the prevailing LiDAR applications can be summarized in Table 5-1. Table 5-2 lists examples of application-specific operations as mapped to the core operation categories defined in Table 5-1. All three applications have similar operations, although they use spatiotemporal datasets for different purposes or may handle different data types. Thus, we argue that an efficient framework that can support the parallelization of these core operations can be flexible and scalable to a diverse set of analytics applications. As a matter of fact, many processing algorithms in existing LiDAR processing software packages can be mapped to these core operation categories. Table 5-3 provides a sample mapping of these tools to the core operation categories.

In this study, LAStool was selected as the tool to devise analytics for various disaster assessment tasks. LAStools are widely known for their blazing speeds and high

productivity (Hug et al. 2004; Li et al. 2018). They provide over 40 functions that cover a variety of applications (e.g., rasterizing, triangulating, classification) for handling and processing of LiDAR data (Hug et al. 2004). More importantly, compared to other software (e.g., CloudCompare, PointTools, Terrasolid), LAStools support command line that can be easily extended to big data analytics platforms such as Apache Storm or Apache Spark. For instance, in this study, we adopted a third party software WINE to execute windows .exe files on a Linux system. Figure 5-3 depicts a workflow for a typical building classification task using LAStools. Description and command line for each elementary operation are also listed in the figure.

The composition of the analytic applications encapsulates several application-level data processing structures. First, original datasets can often be partitioned into tiles, and several categories of operations listed in Table 5-1 can be executed on each tile independently. Spilting large datasets into tiles leads to a bag-of-tasks processing pattern. Second, processing of a single tile or a group of tiles can be expressed as a hierarchical coarse-grain data flow pattern (Beynon et al. 2001; Plale and Schwan 2000; Tan et al. 2010). For example, transformation, filtering, mapping, and segmentation operations can be composed as a workflow. The segmentation operation itself may consist of a pipeline of lower level operations as well. Third, several types of operations such as aggregation and classification can be represented as MapReduce style (Dean and Ghemawat 2008; Dean and Ghemawat 2010) computations.

LAStools Function	Description	Command Line
LasNoise	Remove Noise From Original Data	C:\lasnoise -i " C:\Air_Eaarl-Pre_20121026.laz" -step 1 -isolated 2 -remove_noise -odir " C:\" -o “Air_Eaarl-Pre_20121026.las"
Las2Las	Project the data to New Jersey State Plane	C:\las2las -i " C:\ Air_Eaarl-Pre_20121026.las " -target_sp83 NJ -odir “C:\" -o "Air_Eaarl-Pre_20121026_18N.las"
LasClip	Clip the LiDAR data to the Barnegat Peninsula, NJ	C:\lasclip -i C:\Air_Eaarl-Pre_20121026_18N.las -poly BarnegatPeninsula.shp -o C:\Air_Pre_BarnegatPeninsula.las -v
LasGround	Classify Ground Points	C:\lasground_new -i " C:\Air_Pre_BarnegatPeninsula.las " -town -ultra_fine -odir " C:\Air_Pre_BarnegatPeninsula_Ground.las"
LasHeight	Computes the height of each point above the ground	C:\lasheight -i "C:\Air_Pre_BarnegatPeninsula_Ground.las" -odir "C:\" -o "Air_Pre_BarnegatPeninsula_Height.las"
LasClassify	Classify buildings and high vegetation	C:\lasclassify -i “C:\Air_Pre_BarnegatPeninsula_Height" -odir “C:\" -o "Air_Pre_BarnegatPeninsula_Class“ -verbose -planar 0.2

Figure 5-3 A typical building classification task breakdown

Table 5-1 Core operation categories in spatial disaster data processing

Core Operation Categories	Example Operations	Data Assess pattern	Computation Complexity
Data Cleaning & Quality Control	Transformations to reduce effects of sensor/measurement artifacts. Transform sensor acquired measurements to domain specific variables.	Mixture of local and global pattern	Moderate computational complexity.
Low-Level Transformations	Transformations of a dataset to another format. E.g., coordinate transformation (such as UTM to GCS), value conversion (such as. RGB to grayscale conversion), or as geometry transformation (3D to 2D projection).	Mainly local pattern	Low to moderate, mainly data-intensive computations
Data Subsetting, Filtering, Subsampling	Select portions of a dataset corresponding to regions in the atlas and/or time intervals. Select portions of a dataset based on value ranges. Subsample data to reduce resolution and data size.	Local as well as indexed pattern	Low to moderate, mainly data-intensive computations
Spatio-temporal Mapping & Registration	Create composite dataset from multiple spatially co-incident datasets. Create derived dataset from spatially co-incident datasets obtained at different times.	Irregular local and global data pattern	Moderate to high computational complexity.
Object Segmentation	Segment “base level” objects such as ground, road, dune, vegetation, and buildings. Extract features from “base level” objects.	Irregular, but primarily local data pattern	High computational complexity.
Object Classification	Classify “base level” individual objects at finer details such as utility poles, building types, and transportation assets through a possibly iterative combination of clustering, machine learning and human input (active learning).	Irregular local and global data patterns	High computational complexity.
Change Detection, Comparison, and Quantification	Quantify changes over time in domain-specific low-level variables, base level objects, and high-level objects. Construct “change objects” to describe changes in low-level domain specific variables, base level, and high-level objects. Spatial queries for selecting and comparing segmented regions and objects.	Mixture of local and global data patterns as well as indexed	High Complexity and data-intensive computations.

Table 5-2 Example application scenarios mapped to the core operation categories

Operation Category	Weather Prediction	Monitoring and Change Analysis	Pathology Image Analysis
Data Cleaning and Low-Level Transformations	Remove anomalous measurements from MODIS and convert spectral intensities to the value of interest.	Remove unusual readings. Convert signal intensities to color and other values of interest.	Color normalization. Thresholding of pixel and regional grayscale values.
Data Cleaning and Low-Level Transformations	Spatial selection/crossmatch to find the portion of a dataset that is corresponding to a given geographic region.	Spatial selection/crossmatch to find portion of a dataset corresponding to a given geographic region	Selection of regions within an image. Thresholding of pixel values.
Data Cleaning and Low-Level Transformations	Mapping tiles to map projection. Generation of a mosaic of tiles to get complete coverage.	Registering low and high-resolution images corresponding to same regions.	Deformable registration of images to an anatomical atlas.
Object Segmentation	Segmentation of regions with similar land surface temperature.	Segmentation of buildings, trees, plants, etc.	Segmentation of nuclei and cells. Compute texture and shape features
Object Classification	Classification of segmented regions.	Classification of buildings, trees, plants.	K-means clustering of nuclei into categories.
Spatio-temporal Aggregation	Time-series calculations on changing land and air conditions.	Aggregation of labeled buildings, trees, plants into residential, industrial, vegetation areas.	Aggregation of object features for per image features.
Change Detection, Comparison, and Quantification	Spatial and temporal queries on classified regions and aggregation to look for changing weather patterns.	Characterize vegetation changes over time and are	Spatial queries to compare segmented nuclei and features.

Table 5-3 Corresponding tools for the seven core operation categories

Core Operation Categories	Tools in Lastool	Tools in Cloud Compare	Tools in Terrasolid
Data Cleaning & Quality Control	LasControl, LasDuplicate, LasInfo, LasNoise, LasPrecisiom, LasReturn, LasThin, LasValidate, LasView,	Noise Filter, SOR (Statistical Outlier Removal) filter, Remove Duplicate Points, Hidden Points Removal	TerraMatch: Calibration and Strip Adjustment, Tie Lines tools, Match tools (e.g., apply correction, find intensity correction,), etc.;
Low-Level Transformations	Blast2Dem, Blast2Iso, Las2Dem, Las2Iso, Las2Las, Las2Shp, Las2tin, las2Txt, Las2Zip, Shp2Las, LasPublish,	Fit Tool (plane, sphere, 2D polygon, 2.5D quadric), Unroll, Rasterize and Contour Plot, Contour Plot to Mesh	Projection tool (Coordinate Transformations, Geoid adjustment), Convert Storage Format (to kmz, dgn, etc.)
Data Subsetting, Filtering, Subsampling	Las2Las, LasCanopy, LasClip, LasGrid, LasIndex, LasCoverage, LaSsort, LasSpilt,	Subsampling Tool (by random, space, octree)	Point Filtering Tools (by classification, intensity)
Spatio-temporal Mapping & Registration	LasColor, LasTrack, LasPlane	Align (point pairs picking), Match Boundary Box Centers, Match Scales, Fine Registration	TerraPhoto: Camera Calibration Tool, Color Correction Tool, Improving Image Positioning Tool, Color Points, and Selection Shapes Tools, Manage Trajectories Tool.
Object Segmentation	Las2Boundry, LasClassify, LasHeight, LasGround,	Label Connect Component, Cross Section/ Unfold, Section, Facet Detection, RANSAC shape detection	TerraScan: Macro Classification tool (Classify / By intensity; Classify / Surface Points, Classify Using Brush, etc.), Power Lines using Least Squares Fitting, TerraModel (Surface Modeling)
Object Classification	lasclassify, lasheight, lasground,	CANUPO Classification, Cloth Simulation Filter (CSF)	TerraScan: Macro Classification tool (Classify / By intensity; Classify / Surface points, Classify Using Brush and etc.),
Change Detection, Comparison, and Quantification	-	Compute 2.5D Volume	TerraScan Change Detection Tool

5.3.3 Constructing Apache Strom Based Topology

This study is built on the Apache Storm based stream processing platform. In Apache Storm, a processing workflow is referred to a topology. A topology is a graph stream transformations, consisting of nodes which are either a spout or a bolt. The detail steps for are described in the following sections..

5.3.3.1 Converting application process to spouts and bolts

One fundamental difference between a processing workflow and a topology is that a processing workflow ends eventually, whereas a topology runs forever until being killed. A topology is a network of nodes. In a topology, a stream of data flows from spouts to bolts or from one bolt to another bolt. A spout is the entry point, and it is the source of streams in the storm topology. The spout is connecting to a queue system that will pull messages from the message emitters. If the message is non-empty, the spout will check whether the profile specified in the message can match with the existing topology. The spout then emits a new message and pass it to all bolts subscribed in the stream. Bolts are the actual processing logic. It takes tuples as input and converts them to executable command. In the previous section, complex applications are broken up into smaller elementary operations. Then, each operation is compiled as an independently executable file using LAStools. Apache storm is flexible with the programming language. This study adopted a third party software WINE to execute .exe files in LAStools on a Linux system. Compared to the application based workflow, the topology based workflow can improve the robustness of the computation framework. For instance, there are three Bolt1s in the topology as shown in Figure 5-4b. These three bolts are identical but located in three independently processors. In another word, if anyone or two bolts

fail, the topology can maintain functioning. Also, compared to the application based data processing, the topology based computing can reduce the waste of computation resource. In application-based data processing (Figure 5-4a), functions 2 are running three times in all three applications. On the other hand, in the topology-based streaming processing, only the Bolt2 in the middle receive a message from the previous Bolt1, and therefore, the function 2 is triggered once.

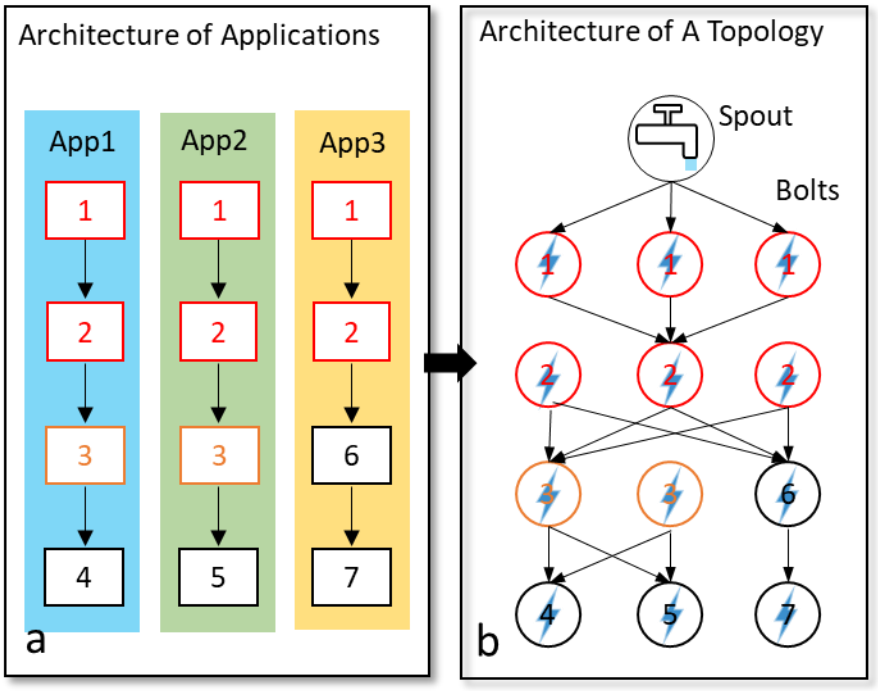


Figure 5-4 Details steps for application-topology conversion

5.3.3.2: Spouts and bolts communication

The communication strategy between two nodes of the cluster is called stream grouping. A stream grouping defines how the messages are distributed from one set of upstream parallel processing task to another set of downstream parallel processing tasks. A stream grouping tells a topology how to send tuples between nodes and controls how the messages are routed in the topology. Commonly used groupings include shuffle grouping, fields grouping, global grouping, and all grouping. For instance, in Global

Grouping, messages are sent from upstream tasks to a single idle downstream task in a round-robin fashion, thereby achieving load balancing for distributed tasks. In this research, to reduce the computation of duplicate functions, global grouping is selected as the primary means of message exchange among bolts. Figure 5-5 depicts the pseudocode for grouping corresponding to the topology in Figure 5-4b. Besides grouping, the default implementation in Apache Strom also allows tasks within the nodes to communicate using TCP, which can be efficient because of the loopback adapter but can be further improved using shared memory based communication (Kamburugamuve et al. 2016).

Grouping:
SetBolt(Bolt1).allGrouping(Spout);
SetBolt(Bolt2).globalGrouping(Bolt1);
SetBolt(Bolt3).globalGrouping(Bolt2);
SetBolt(Bolt6).globalGrouping(Bolt2);
SetBolt(Bolt4).globalGrouping(Bolt3);
SetBolt(Bolt5).globalGrouping(Bolt3);
SetBolt(Bolt7).globalGrouping(Bolt6);

Figure 5-5 Pseudocode for Grouping

5.3.3.3 *Stream processing and data movement*

What transport through a topology are messages. A message is defined as the quintuplet: (Header, Action, Data, Location, Topology). The data, location, and topology fields may be empty, or they may contain a message payload with the location of the user or the topology to be uploaded. The header includes a semantic profile in addition to the credentials of the sender. The action field of a message defines its reactive behavior when matching occurs. Data will be routed based on the location specified in the profile. Besides, the location and topology are specified in the Location and Topology field respectively. A tuple is the most basic data structure in Apache Storm, which contains

processing logic, and links between nodes indicate how data should be passed around between nodes. When a spout or bolt emits a tuple to a stream, it sends the tuple to every bolt that subscribed to that stream. Global grouping sends tuples generated by all instances of the source to a single target instance

The topology-based data processing is implemented to support stream processing. In batch processing model, a set of data is ingested before the data analytics. However, in the streaming model, data is fed into analytics tools piece-by-piece. Computation machines, either core or edge, will process the incoming data as soon as it arrives. Such a process will decrease the latency of ingesting all the data to meet the real-time or near real-time processing need. These mechanisms are achieved by exploiting the message system as depicted in the pseudocode in Figure 5-6. In the proposed work, a topology runs continuously over a stream of incoming data unless being terminated. Instead of specifying the exact name of the data, the header of quintuplet message includes the semantic profile of the analytics data. If and only if there exists $Data$ that matches with the $SematicProfile$ $(header)$ in the designated$Location$, the data analytics $(Action)$ will be executed.

Pseudocode

```
BEGIN
Get Tuple(Header, Action, Data, Location, Topology)
if Tuple.Data in Tuple.Location match Tuple.Header.profile
   Execute Tuple.Action
END
```

Figure 5-6 Pseudocode for stream processing mechanism

The framework is designed to minimize unnecessary data movement. The need for transferring data to the core can be determined based on the practical decision-making

needs. Without losing generality, we propose three data movement rules in the case of cloud edge computing in a disaster response scenario. First, data collected by edge devices (such as minivan based mobile data collection system) are prioritized to be processed locally in the edges. This means only the required portion of the analytics products are transferred to the core for further processing. Second, the visualization data are preferred to host as a WebGL based renderer (Schütz 2016) at the edge. Third, any other data that are less time sensitive are suggested to store at the edge for future usage. To do so, a new bolt is implemented following the similar mechanism as the one displayed in Figure 5-6. The only difference is that the $Action$ is referred to specifying data transferring tasks.

5.3.4 Major advantages of the stream processing approach

The proposed Apache Storm based stream processing will have three significant advantages over current computation frameworks. Figure 5-7 depicts a detailed comparison of the prevailing and proposed data processing framework.

Task Parallelism. Most of the existing LiDAR processing algorithms are built to execute on a standalone machine. They often do not support the flexibility of automated parallelism. In cases where timely processing is required, complex manual interactions are required to manage the job assigning and resource allocation. The fact that computation resources are not always located at the same place (e.g., edges, cloud) further complicates such a process. Data transfer and workload balancing among different edges will consume an additional amount of time. The proposed topology based processing framework is particularly suitable for parallel computing. The task queues are

pushed to the master computing node, and the master node would automatically level the available computing resource in the network.

Dynamically resource allocation. In conventional processing algorithms, tasks are running on a fixed number of executors. A Storm Cluster can be configured to make use of “dynamic resource allocation” which allows altering of the number of executors to be allocated to a task over time. For instance, if a process needs more executors, the number of resources deployed for data analytics can be increased to match to the actual workload. On the other hand, in a multiprocessor system, the proposed framework can also realize the task parallelism by executing a different process on the same or different data.

Inter-task Communication. In most current computing frameworks, all tasks are running independently. There is no communication between tasks. In the proposed framework, inter-task communication can be achieved by merging task operations into storm topologies. A topology-based processing framework enables (1) data and information sharing; (2) mechanism to inform task that data is available to read or write; and (3) mechanism to prevent tasks from interfering with each other. Inter-task communication can help to coordinate the data processing tasks better, balance the processing workloads, reduce the execution of duplication functions, and as a result achieve better utilization of computation resource while decreasing the data processing time.

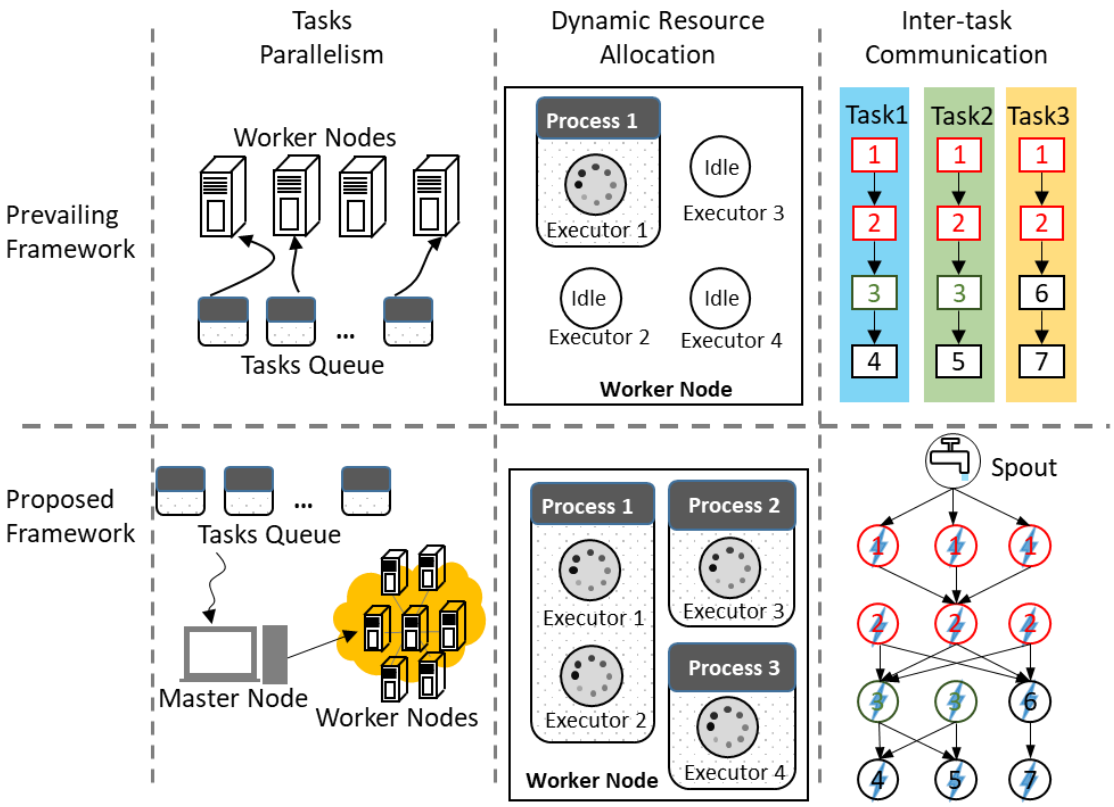


Figure 5-7 Comparison of prevailing and the proposed processing framework

5.4 Disaster response example

5.4.1 Background

A Hurricane Sandy use case is selected in this study for evaluating the proposed data processing framework. Hurricane Sandy (2012) is one of the most destructive hurricanes in the Atlantic shoreline area. It made landfall in southern New Jersey and took a major toll on the Atlantic shoreline area. It toppled many major transportation infrastructures in the state of New Jersey and New York city and put millions of people in the dark. The data used in this study are airborne LiDAR data collected using the second-generation Experimental Advanced Airborne Research Lidar (EAARL-B) system. The purpose of the data acquisition project was to produce highly detailed and accurate digital elevation maps of the New Jersey coastline. The proposed framework is designed

to be used in the initial state of Hurricane Sandy to meet the four emergency response functions proposed by Lindell and Perry (1992). The detail steps for determining the prioritization of the data processing applications are introduced in Chapter 4. As shown in Chapter 4, terrain analysis, interactive 3D analysis, and change detection are among the highly ranked data processing applications. In this chapter, we will demonstrate how to accelerate these data processing applications using a distributed data processing framework.

5.4.2 Topology Construction

Once the priorities of data analytics are identified, these applications are decomposed into a series of elementary operations. Figure 5-8 lists all the elementary operations and their workflow in three applications. LAStools are used as the exemplary LiDAR processing tools in this study. According to the method described in the previous section, the three applications are broken down into elementary operations and then converted to an apache storm topology as shown in Figure 5-9.

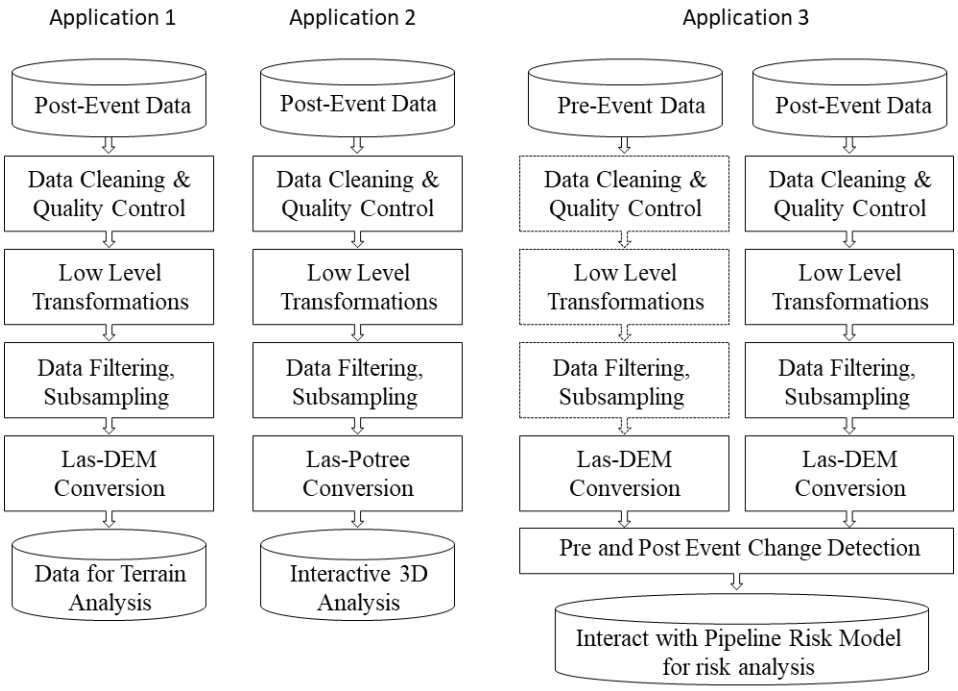


Figure 5-8 Detail workflow of the desired applications

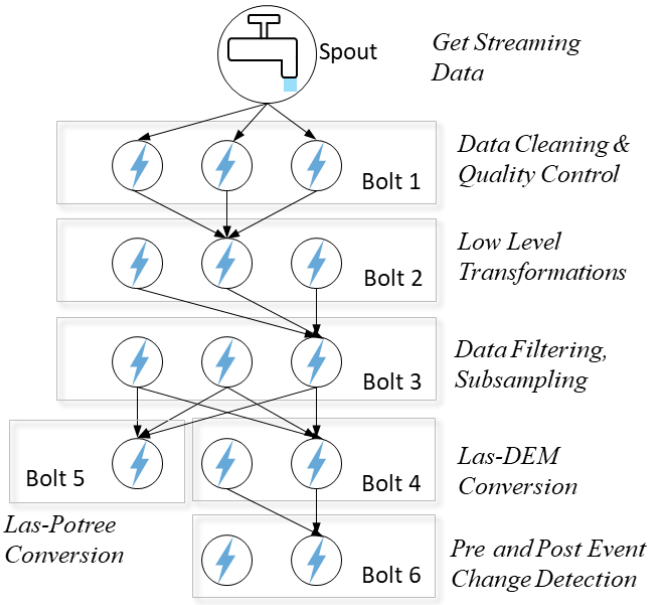


Figure 5-9 A topology of the desired applications

This use case was implemented in the Amazon EC2 which allows the creation and management of the virtual computing environment. A cluster was set up, and its structure is shown in Figure 5-10. The cluster consists of five nodes: one Nimbus, one Zookeeper,

and three Supervisor nodes. A nimbus node can be considered as the master node located in the core, which assigns computation tasks to workers (supervisors) node. A zookeeper node is a coordinator that monitors the performance of supervisors, which is also considered to implement in core. A supervisor node or work node is the machine that is actually doing the data processing. Three supervisor nodes were implemented in the Amazon EC2 to emulate three edge devices. Each of the five nodes is simulated using a t2.micro instance in Amazon EC2, with 1 GB memory and 8 GB storage. In each of the rendezvous point, a streaming engine based on Storm version 1.1.1 was installed

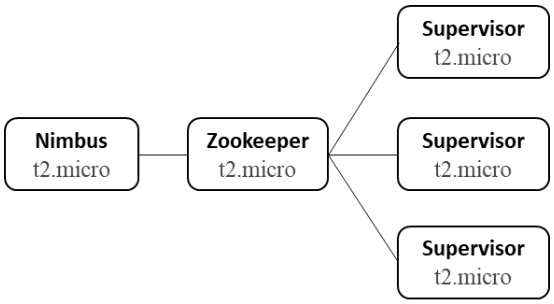


Figure 5-10 Architecture of the cluster

5.4.3 Experiment evaluation

In this section, the performance and scalability of the data processing approach and their realization as an EC2 cluster were evaluated on the 2012 EAARL-B pre and 2012 EAARL-B post LiDAR datasets. The analysis area includes the entire New Jersey Coastline. The analysis area was further divided into 27 locations. At each location, the pre- and post-disaster LiDAR datasets were used in the analysis. For each location, the size of pre- and post-event datasets ranges from 500 KB to 14.5 MB.

In the first experiment, the performance of the topology-based stream processing is evaluated against a traditional single computer processing approach. Based on the Figure 5-10, a virtual environment was set up in the cloud (Amazon EC2). Both scenarios

are executed in the same worker (supervisor) node. In the proposed work, the number of worker nodes is set to be one using $conf.setNumWorkers(1)$. In the reference analysis, the typical analysis mode was executed in the same t2.micro instance directly. It should be pointed out that the LAStools have the multi-core LiDAR processing capability. In this experiment, such capability was not enabled in both scenarios. The evaluation was repeated using different size of the data sets and the corresponding elapsed time is shown in Figure 5-11. The time breakdowns of each individual sub-operations were also reported. The first column of each group in Figure 5-11 displays the total elapsed time using a typical standalone processing mode. The total elapsed time can be estimated as the summation of the elapsed time of each application. The second column of each group in the figure shows the total elapsed time using our stream processing approach. The total elapsed time consisted of two parts: execution time of the topology (three applications) and the latency. The latency was calculated as the sum of all latency in each Spout or Bolt respectively. The experiment was repeated with data sets of varying size ranging from 73 thousand points to 3232 thousand points.

The first finding that can be noticed in Figure 5-11 is that the total elapsed time for both methods increase when the data size increases. Second, compared to the conventional approach, the stream processing approach consumed 42% less computation time. For the t2.micro instance used in the experiment, there were four processors in each node. Processing in the reference analysis did not offer the flexibility of dynamically executor allocation. As a result, the potential computational capability of the four executors was not fully exploited. On the other hand, the stream processing approach supported the scaling of processors. Third, from Figure 5-11, it can be noted that the total

latency in the stream processing approach increased as the data size increased. It incurred an average of additional 25% of the processing time. In this experiment, it can be summarized that our proposed approach outperformed the conventional processing that built on a single executor regarding the processing time. The saving in total elapsed time is achieved by enabling allocation more available processors to the processing jobs. It should be pointed that the t2.micro instance that we used in the experiment have four processors. It is expected to have a more significant time saving if we deployed CPUs with more processors (e.g., t2.2xlarge. m5.4xlarge) for the data processing.

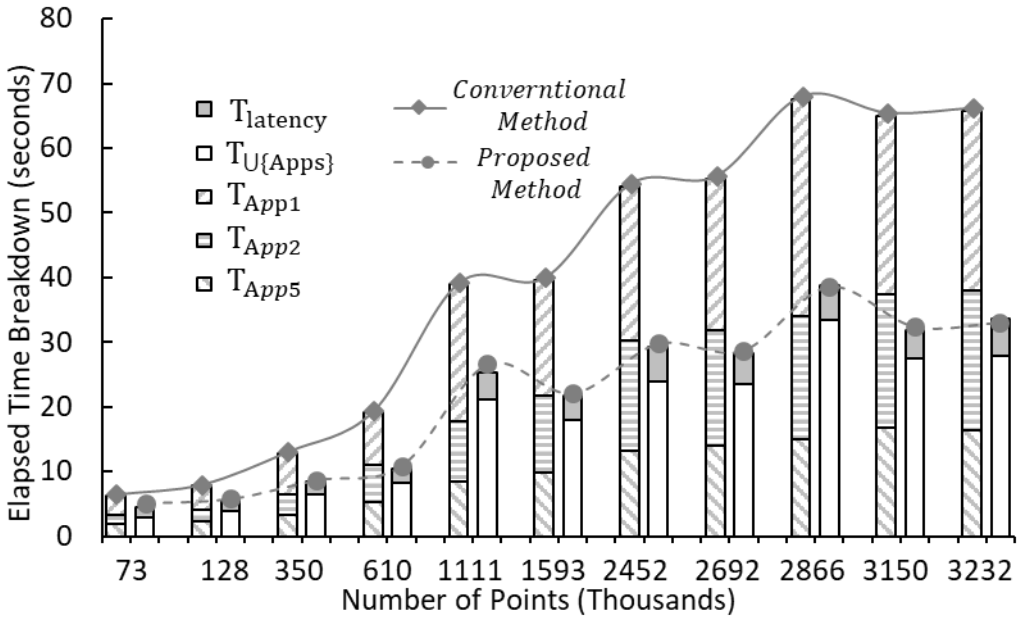


Figure 5-11 Performance Results of a single pair of datasets

In the second experiment, we emulated a real disaster response scenario by deploying our work on a queue of 27 pairs of LiDAR datasets. The data processing was deployed in the same virtual environment in the Amazon EC2 with three worker nodes. To prove the gain in data processing speed, we set up four analysis scenarios. In the scenario 1, a reference group test was performed using the conventional processing in

standalone mode. In the scenario 2, one supervisor node was activated in the cluster using ($conf.setNumWorkers(1)$). In the scenario 3 and 4, two ($conf.setNumWorkers(2)$) and three ($conf.setNumWorkers(3)$) supervisor nodes were deployed in the cluster respectively. In the experiment, scenarios 1 and 2 were compared to evaluate the performance of the proposed stream processing approach over the conventional approach. Scenarios 2, 3 and 4 were formed to assess the scalability of the stream processing approach. It can be noted that all four scenarios were simulated in the cloud (Amazon EC2). All the processing were executed using the same kind of t2.micro instances with 1vGPU and with 1 GB memory and 8 GB storage.

Figure 5-12 displays the elapsed time of all the worker nodes in the tested scenarios. The first finding is that, compared to the conventional processing, there were significant time savings when our proposed work is deployed to process the same data set queue. Notably, when a single worker node was deployed in the processing task, our proposed work consumed 28.4% of less time than that of the conventional method. Second, our approach can achieve automatic workload balancing. In the two worker node scenario, the nimbus node assigned tasks to the two worker nodes based on their availability. The occupation time for the two nodes is 7.50 and 8.12 minutes, respectively. Besides, the number of tasks finished by these two worker nodes are 16 and 11 respectively. Moreover, it can be obtained that in the three worker node scenario, the workloads were more evenly distributed. Three worker nodes finished ten, ten, and seven tasks in 5.66, 5.62, and 5.56 minutes respectively. Last but not least, the scalability of deploying more worker nodes in the cluster for processing is non-linear. In the experiment, it is expected that the two and three worker nodes cluster would consume

50% and 67% less time than the single worker node cluster. In our analysis, however, the system elapsed time is only 37.1% and 56.2% less. One main reason behind this phenomena is that additional latency was introduced for tasks distribution and workload balancing. Nevertheless, the proposed approach enables workload sharing by activating more worker nodes in the cluster. Based on these findings, we can argue that our approach can exploit the availability of the potential computation resources.

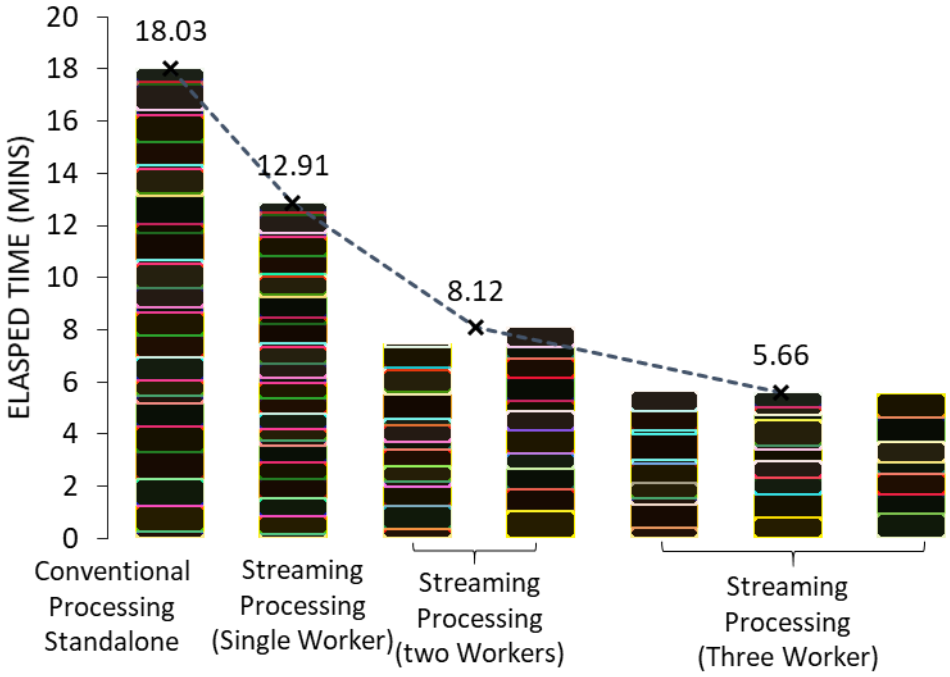


Figure 5-12 Performance of streaming processing on datasets queue

To further investigate the uncertainty in the latency of deploying more worker nodes, the four scenarios were repeated ten times. Figure 5-13 summarized the mean and standard deviation of the system elapsed time. The system elapsed time was computed as the maximum occupation of all the nodes deployed in the system. According to the experiment results, the average elapsed time for the standalone conventional processing, one worker, two workers, and three worker streaming processing were 19.45, 12.80, 8.07, and 5.96 minutes respectively. Compared to the conventional processing in standalone

mode, stream processing had 34%, 59% and 69% of time-saving when one worker node, two worker nodes, and three worker nodes were deployed in the cluster. Another finding that is worthy of mentioning is that the uncertainty (standard deviation in Figure 5-13) of the processing time tends to increase when more worker nodes were activated in the cluster. This uncertainty is largely due to the overhead in job scheduling in the Nimbus node.

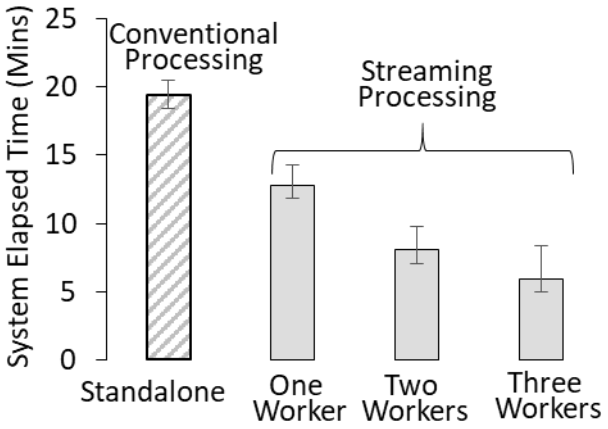


Figure 5-13 System Elapsed Time comparison

The last part of this study concerns the analysis of the accessibility and storage. As LiDAR data often consume a lot of storage space, movement and storage of LiDAR datasets cause significant burdens, in particular in post-disaster scenarios where computational resources and data communication capacities are often limited. However, not all data sets are necessarily required to transfer to the core immediately. The burden of core storage and the pressure on the network can be relieved if the edges only submit the required part of the data and transfer them to the core. This work is implemented to follow three rules to reduce the transferring of an excessive amount of data to the core. First, data collected by edge devices (such as minivan based mobile data collection system) are prioritized to be processed locally in the edges, only the required portion of

the analytics products are transferred to the core for further processing. Second, the visualization data are preferred to host as WebGL based renderer (Schütz 2016) at the edge. Third, any other data that are less time sensitive are suggested to store at the edge for future usage. Based on these rules, only the result from Application 1 and 3 were required to transfer to the core to couple with other information such history data for further analysis. Visualization data generated in Application 2 were hosted locally at the edge as WebGL based renderer, which supports remote access and interaction. Other data are considered as less time sensitive and stored locally at the edge. Figure 5-14 demonstrates the breakdown of the data that require storage, host at the edge as well as transfer to the core in the second experiment. Among a total size of 2092 MB of data generated, only a small portion (235 MB) was critical and required immediately transfer to the core for advanced processing. Visualization data (372 MB) generated from App2 can be hosted locally as a web-based platform. This web-based platform can put visualization LiDAR data into the browser of a remote client without incurring the overhead of transferring the massive amount of data. Besides, the rest 71% data were either raw or abundant data that can store in the cloud or locally for future usage.

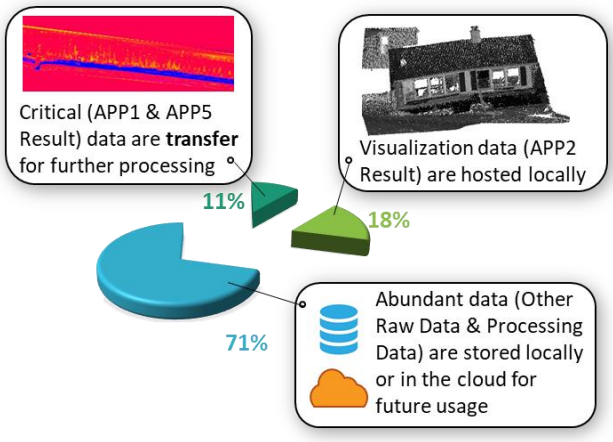


Figure 5-14 Data accessibility breakdown

5.5 Conclusion and future work

The scoping of this study is to propose a stream processing framework that exploits the HPC to process LiDAR data sets for time-sensitive applications. The feasibility of the work is demonstrated using a Hurricane Sandy based use case. The framework consists of three key components: (1) a tile-based data management system in the cloud, (2) formalism for decomposition of the complex processing tasks into elementary operations; (3) Apache-storm based stream processing.

Based on the use case, the results indicated a 69% computation time saving when a three-supervisor cluster instead of a conventional standalone machine is used to process the data. It is also expected that the overall processing time can be shortened if more computation resources are available to use. Moreover, our approach also contributes to the relieving the burden of data movement and storage in the core, therefore reducing the demand on the network. According to our use case, only a minority (11%) of the data were critical and required immediate transfer to the core for further processing.

Furthermore, this work explored the feasibility of restructuring complex processing tasks into workflow-style elementary operations that can support stream processing in HPC infrastructures. It is expected that the proposed work can contribute to the better delivery of on-demand information in disaster response more efficiently and effectively. One limitation of this work is that the uncertainty in latency in stream processing is not fully explored. According to the case study, the latency and its uncertainty increase when more supervisor nodes are deployed. To determine the optimized processing strategy, this needs to be further investigated in future research.

Chapter 6 Conclusions and Future Research

Severe weather events such as hurricanes, ice storms, surge, and flooding have been occurring across the U.S and around the world, threatening places where economic and industrial activities are heavily concentrated. These extreme events are now increasing observed and monitored with a loosely coupled network of geospatial sensors. Analysis of these datasets offers tremendous opportunities in improving the resilience and adaptability of coastal communities in the face of future natural disasters. Despite the high values in these data sets, the vast size and complex processing requirements of these new data sets make it challenging to effectively use them in coastal community management applications, in particular emergencies. Yet, unprocessed data are intangible and non-consumable, which is often resulting in 'data-rich-but-information-poor" situation.

The overarching goal of this research is to research, develop, and evaluate a data processing framework that is capable of efficiently processing the emerging large geospatial data sets and extract crucial information to enhance disaster management during large-scale extreme events. This research systematically studied the fundamental aspects of big spatial disaster data including the anatomy of big spatial disaster data, data processing patterns, data quality issues, uncertainty propagation along the analytics pipeline, and adaptive processing in time-sensitive environments. More specifically, this dissertation addresses the following research questions.

1. What is the basic anatomy of big spatial disaster data?
2. What are the core operation categories and processing patterns with big spatial disaster data?

3. How does the uncertainty associated with spatial disaster data sets propagate through a given processing pipeline?
4. How to adequately represent users' dynamic and complex information needs and processing requirement during coastal resilience investigations in a unified framework?
5. How to dynamically adapt 3D disaster data analytics given user information needs and processing requirements and algorithm and dataset descriptions?

In Chapter 2, I characterized the basic anatomy of big spatial disaster data to highlight the challenges and opportunities in using these emerging data sets in coastal community management applications during extreme events. I also characterized data processing patterns associated with the emerging big spatial disaster data sets and abstracted these patterns into core operation categories. These work laid the foundation for realizing cloud-based computing of these data sets for disaster response applications.

In Chapter 3, I used a case study based approach to demonstrate approaches for quantifying uncertainty propagation in processing geospatial data sets. More specifically, I proposed a method to identify the optimal strategy for approximation parameter selection in interpolating Light Detection and Ranging (LiDAR) data into Digital Elevation Models (DEMs). The method is developed to address the need to model accuracy loss in rapid generation of DEMs, which are essential pieces of information used in disaster response and flooding simulation.

In Chapter 3, I proposed a DEA based information salience model to prioritize the sequence of the information processing tasks. The model provides a unified way of representing user information needs and balancing these needs to realize optimized data

processing sequences. More specifically, this model integrates the DEA efficiency score with linguistic group decision process. The proposed model is tested against a hurricane sandy based case study in the Barnegat Peninsula, New Jersey. The results indicate that the proposed model prototyped a framework for information articulation between decision-makers and the data processing team. The proposed model will help to accelerate the data-information transliteration and reduce the possible ‘data-rich-but-information-poor” situation

Based on Chapter 3, I proposed in Chapter 4 a stream data processing approach that realized accelerated information extraction from large quantities of geospatial data given various user information needs. The approach is capable of representing complex spatial data analytics into a workflow centric data analysis representation and levering the flexible computing resources in the cloud and at the edge to improve information extraction from these large data sets.

Throughout this dissertation research, I used extensively Hurricane Sandy related data sets as use cases to evaluate the proposed approaches. The results demonstrated the proposed approaches provide a scalable approach for information extraction from spatial disaster data within a realistic time bound. It is important to recognize that this research does not focus on developing algorithms for data processing tasks such as segmentation and object recognition. Instead, it focuses on formulating mechanisms to integrate existing spatial data analytics into the emerging big data processing frameworks and to address the particular challenges in using the big spatial disaster data for coastal resilience decision support. In terms of future research, it is beneficial to investigate the

development of dedicated disaster data processing algorithms and integrate them into the framework developed in this research.

The short-time disasters, as well as the long-term environmental changes, are prone to pose a pressing challenge to coastal communities. To improve the resilience (both capability and adaptability) of the community, precise and thorough information on the formation, development and the consequence of the disasters as well as the impacts to the communities is essential. The immerging sensing technology, as well as the flooding data, will, on the one hand, improve information gathering, and on the other hand, compound the difficult in information processing for decision making during disaster time sensitive environment. Amongst many research needs, this study proposes a prototype study to address the sensitivity issue in disaster response by closing the gap between information processing and decision-making. It is hoped that the methods presented in this dissertation will motivate innovative research in closing the loop among information gathering, data processing and decision-making.

Bibliography

Adger, W. N. (2006). "Vulnerability." *Global environmental change*, 16, 268–281.

Aguilar, F. J., Mills, J. P., Delgado, J., Aguilar, M. A., Negreiros, J. G., and Pérez, J. L. (2010). "Modelling vertical error in LiDAR-derived digital elevation models." *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, 103–110.

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., and Saltz, J. (2013). "Hadoop gis: a high performance spatial data warehousing system over mapreduce." *Proceedings of the VLDB Endowment*, 6(11), 1009-1020.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., and Dustdar, S. (2013). "Quality control in crowdsourcing systems: Issues and directions." *IEEE Internet Computing*, 17, 76–81.

Allen, G. (2007). "Building a dynamic data driven application system for hurricane forecasting." *Computational Science–ICCS 2007*, 1034-1041.

Alonso, S., Chiclana, F., Herrera, F., and Herrera-Viedma, E. (2004). "A learning procedure to estimate missing values in fuzzy preference relations based on additive consistency." *Lecture notes in computer science*, 3131, 227-238.

Arun, P. V. (2013). "A comparative analysis of different DEM interpolation methods." *The Egyptian Journal of Remote Sensing and Space Science*, 16(2), 133-139.

ASPRS (2004). "ASPRS guidelines vertical accuracy reporting for lidar data (version 1.0).", American Society of Photogrammetry and Remote Sensing.

ASPRS (2013). "LAS SPECIFICATION, VERSION 1.4 – R13 ", The American Society for Photogrammetry & Remote Sensing

ATC (2004). "ATC-45 field manual: safety evaluation of
buildings after wind storms and floods." Applied Technology Council, Redwood City.

Awrangjeb, M., Zhang, C., and Fraser, C. S. (2013). "Automatic extraction of building roofs using LIDAR data and multispectral imagery." *ISPRS Journal of Photogrammetry and Remote Sensing*, 83, 1-18.

Banker, R. D., and Natarajan, R. (2008). "Evaluating contextual variables affecting productivity using data envelopment analysis." *Operations research*, 56(1), 48-58.

Barbarosoğlu, G., Özdamar, L., and Cevik, A. (2002). "An interactive approach for hierarchical analysis of helicopter logistics in disaster relief operations." *European Journal of Operational Research*, 140(1), 118-133.

Barbier, G., Zafarani, R., Gao, H., Fung, G., and Liu, H. (2012). "Maximizing benefits from crowdsourced data." *Computational and Mathematical Organization Theory*, 18(3), 257-279.

Bater, C. W., and Coops, N. C. (2009). "Evaluating error associated with lidar-derived DEM interpolation." *Computers & Geosciences*, 35(2), 289-300.

Beall, C., Lawrence, B. J., Ila, V., and Dellaert, F. "3D reconstruction of underwater structures." IEEE, 4418–4423.

Beamon, B. M., and Balcik, B. (2008). "Performance measurement in humanitarian relief chains." *International Journal of Public Sector Management*, 21, 4–25.

Beard, M. K., Buttenfield, B. P., and Clapham, S. B. (1991). *NCGIA Research Initiative 7: Visualization of Spatial Data Quality: Scientific Report for the Specialist Meeting 8-12 June 1991, Castine, Maine*, National Center for Geographic Information and Analysis.

Becker, J., Manville, V., Leonard, G., and Saunders, W. (2008). "Managing lahars the New Zealand way: a case study from Mount Ruapehu volcano." *Natural Hazards Observer*, 32, 4–6.

Beynon, M. D., Kurc, T., Catalyurek, U., Chang, C., Sussman, A., and Saltz, J. (2001). "Distributed processing of very large datasets with DataCutter." *Parallel Computing*, 27(11), 1457-1478.

Bharosa, N., Lee, J., and Janssen, M. (2010). "Challenges and obstacles in sharing and coordinating information during multi-agency disaster response: Propositions from field exercises." *Information Systems Frontiers*, 12, 49–65.

Bhatla, A., Choe, S. Y., Fierro, O., and Leite, F. (2012). "Evaluation of accuracy of as-built 3D modeling from photos taken by handheld digital cameras." *Automation in construction*, 28, 116–127.

Brandon, C. M., Woodruff, J. D., Donnelly, J. P., and Sullivan, R. M. (2014). "How unique was Hurricane Sandy? Sedimentary reconstructions of extreme flooding from New York Harbor." *Scientific reports*, 4, 7366.

Brédif, M., Vallet, B., and Ferrand, B. (2015). "Distributed Dimensonality-Based Rendering of LIDAR Point Clouds." *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(3), 559.

Briese, C., and Pfeifer, N. "Airborne laser scanning and derivation of digital terrain models." *Proc., Fifth Conference on Optical*.

Brilakis, I., Fathi, H., and Rashidi, A. (2011). "Progressive 3D reconstruction of infrastructure with videogrammetry." *Automation in Construction*, 20, 884–895.

Bruinsma, G. W. J. (2010). *Adaptive workflow simulation of emergency response*, University of Twente.

Bruneau, M., Chang, S. E., Eguchi, R. T., Lee, G. C., O'Rourke, T. D., Reinhorn, A. M., Shinozuka, M., Tierney, K., Wallace, W. A., and von Winterfeldt, D. (2003). "A framework to quantitatively assess and enhance the seismic resilience of communities." *Earthquake spectra*, 19, 733–752.

Callison-Burch, C., and Dredze, M. "Creating speech and language data with Amazon's Mechanical Turk." *Proc., Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics, 1-12.

Cao, V.-H., Chu, K., Le-Khac, N.-A., Kechadi, M. T., Laefer, D., and Truong-Hong, L. "Toward a new approach for massive LiDAR data processing." *Proc., Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015 2nd IEEE International Conference on*, IEEE, 135-140.

Cash, D., Clark, W. C., Alcock, F., Dickson, N. M., Eckley, N., and Jäger, J. (2002). "Salience, credibility, legitimacy and boundaries: linking research, assessment and decision making."

Chapin, S. J., Katramatos, D., Karpovich, J., Grimshaw, A. S., and others "The legion resource management system." Springer, 162–178.

Charaniya, A. P., Manduchi, R., and Lodha, S. K. "Supervised parametric classification of aerial lidar data." *Proc., Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, IEEE, 30-30.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). "Measuring the efficiency of decision making units." *European journal of operational research*, 2(6), 429-444.

Charnes, A., Cooper, W. W., and Rhodes, E. (1979). "Measuring the efficiency of decision-making units." *European journal of operational research*, 3(4), 339.

Chen, F., Zhai, Z., and Madey, G. "Dynamic adaptive disaster simulation: developing a predictive model of emergency behavior using cell phone and GIS data." Society for Computer Simulation International, 5–12.

Coddington, P. D., Hawick, K. A., and James, H. A. "Web-based access to distributed high-performance geographic information systems for decision support." *Proc., Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*, IEEE, 12 pp.

Comfort, L. K., Ko, K., and Zagorecki, A. (2004). "Coordination in rapidly evolving disaster response systems: The role of information." *American Behavioral Scientist*, 48, 295–313.

Cooper, H. M., Fletcher, C. H., Chen, Q., and Barbee, M. M. (2013). "Sea-level rise vulnerability mapping for adaptation decisions using LiDAR DEMs." *Progress in Physical Geography*, 37(6), 745-766.

Csanyi, M. N., and Toth, C. K. (2007). "Point positioning accuracy of airborne LiDAR systems: A rigorous analysis."

Csanyi, N., and Toth, C. K. (2007). "Improvement of lidar data accuracy using lidar-specific ground targets." *Photogrammetric Engineering & Remote Sensing*, 73(4), 385-396.

Cui, D., Wu, Y., and Zhang, Q. "Massive spatial data processing model based on cloud computing model." *Proc., Computational Science and Optimization (CSO), 2010 Third International Joint Conference on*, IEEE, 347-350.

Cutter, S. L., Barnes, L., Berry, M., Burton, C., Evans, E., Tate, E., and Webb, J. (2008). "A place-based model for understanding community resilience to natural disasters." *Global environmental change*, 18, 598–606.

Dai, F., and Lu, M. (2010). "Assessing the accuracy of applying photogrammetry to take geometric measurements on building products." *Journal of construction engineering and management*, 136, 242–250.

Dean, J., and Ghemawat, S. (2008). "MapReduce: simplified data processing on large clusters." *Communications of the ACM*, 51(1), 107-113.

Dean, J., and Ghemawat, S. (2010). "MapReduce: a flexible data processing tool." *Communications of the ACM*, 53(1), 72-77.

Delgado, M., Herrera, F., Herrera-Viedma, E., and Martinez, L. (1998). "Combining numerical and linguistic information in group decision making." *Information Sciences*, 107(1-4), 177-194.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. "Imagenet: A large-scale hierarchical image database." *Proc., Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, 248-255.

Dervin, B., and Nilan, M. (1986). "Information needs and uses." *Annual review of information science and technology*, 21, 3-33.

Dong, L., and Shan, J. (2013). "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques." *ISPRS Journal of Photogrammetry and Remote Sensing*, 84, 85-99.

Dorband, J., Palencia, J., and Ranawake, U. (2003). "Commodity computing clusters at goddard space flight center." *Journal of Space Communication*, 1, 113–123.

Earle, P. S., Bowden, D. C., and Guy, M. (2012). "Twitter earthquake detection: earthquake monitoring in a social world." *Annals of Geophysics*, 54.

Eilat, H., Golany, B., and Shtub, A. (2008). "R&D project evaluation: An integrated DEA and balanced scorecard approach." *Omega*, 36(5), 895-912.

Erskine, R. H., Green, T. R., Ramirez, J. A., and MacDonald, L. H. (2007). "Digital elevation accuracy and grid cell size: effects on estimated terrain attributes." *Soil Science Society of America Journal*, 71(4), 1371-1380.

Ezequiel, C. A. F., Cua, M., Libatique, N. C., Tangonan, G. L., Alampay, R., Labuguen, R. T., Favila, C. M., Honrado, J. L. E., Canos, V., Devaney, C., and others "UAV aerial imaging applications for post-disaster assessment, environmental management and infrastructure development." IEEE, 274–283.

Fan, Z., Qiu, F., Kaufman, A., and Yoakum-Stover, S. "GPU cluster for high performance computing." IEEE, 47–47.

Fernández-Lozano, J., Gutiérrez-Alonso, G., and Fernández-Morán, M. Á. (2015). "Using airborne LiDAR sensing technology and aerial orthoimages to unravel roman water supply systems and gold works in NW Spain (Eria valley, León)." *Journal of Archaeological Science*, 53, 356-373.

Fisher, P. F. (1999). "Models of uncertainty in spatial data." *Geographical information systems*, 1, 191–205.

Gahegan, M., and Ehlers, M. (2000). "A framework for the modelling of uncertainty between remote sensing and geographic information systems." *ISPRS Journal of Photogrammetry and Remote Sensing*, 55, 176–188.

Gao, J. (1997). "Resolution and accuracy of terrain representation by grid DEMs at a micro-scale." *International Journal of Geographical Information Science*, 11(2), 199-212.

Garson, G. D., and Biggs, R. S. (1992). *Analytic mapping and geographic databases*, Sage.

Gillespie, T., Chu, J., Frankenberg, E., and Thomas, D. (2007). "Assessment and Prediction of Natural Hazards from Satellite Imagery." *Progress in Physical Geography*, 31(5), 459-470.

Gleason, C. J., and Im, J. (2012). "Forest biomass estimation from airborne LiDAR data using machine learning approaches." *Remote Sensing of Environment*, 125, 80-91.

Gong, J. (2013). "Mobile lidar data collection and analysis for post-sandy disaster recovery." *Computing in Civil Engineering (2013)*, 677-684.

Gong, J., Zhou, H., Gordon, C., and Jalayer, M. (2012). "Mobile terrestrial laser scanning for highway inventory data collection." *Computing in Civil Engineering (2012)*, 545–552.

Goodchild, M. F. (1994). "Integrating GIS and remote sensing for vegetation analysis and modeling: methodological issues." *Journal of Vegetation Science*, 5(5), 615-626.

Goodchild, M. F., and Glennon, J. A. (2010). "Crowdsourcing geographic information for disaster response: a research frontier." *International Journal of Digital Earth*, 3, 231–241.

Goodchild, M. F., Yuan, M., and Cova, T. J. (2007). "Towards a general theory of geographic representation in GIS." *International journal of geographical information science*, 21(3), 239-260.

Green, R. H., Doyle, J. R., and Cook, W. D. (1996). "Preference voting and project ranking using DEA and cross-evaluation." *European Journal of Operational Research*, 90(3), 461-472.

Guan, H., Li, J., Zhong, L., Yongtao, Y., and Chapman, M. (2013). "Process virtualization of large-scale lidar data in a cloud computing environment." *Computers & geosciences*, 60, 109-116.

Guan, X., and Wu, H. (2010). "Leveraging the power of multi-core platforms for large-scale geospatial data processing: Exemplified by generating DEM from massive LiDAR point clouds." *Computers & Geosciences*, 36(10), 1276-1282.

Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy." ACM, 729–736.

Gupta, S., Weinacker, H., and Koch, B. (2010). "Comparative analysis of clustering-based approaches for 3-D single tree detection using airborne fullwave lidar data." *Remote Sensing*, 2(4), 968-989.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K., and Pollefeys, M. (2017). "Semantic3D. net: A new Large-scale Point Cloud Classification Benchmark." *arXiv preprint arXiv:1704.03847*.

Haile, A. T., and Rientjes, T. (2005). "Effects of LiDAR DEM resolution in flood modelling: a model sensitivity study for the city of Tegucigalpa, Honduras." *Isprs wg iii/3, iii/4*, 3, 12-14.

Han, S. H., Heo, J., Sohn, H. G., and Yu, K. (2009). "Parallel processing method for airborne laser scanning data using a PC cluster and a virtual grid." *Sensors*, 9(4), 2555-2573.

Han, Y., Sheth, A., and Bussler, C. "A taxonomy of adaptive workflow management."

Hasselman, T., Wathugala, G., Urbina, A., and Paez, T. L. "Top-Down vs. Bottom-Up Uncertainty Quantification for Validation of a Mechanical Joint Model." *Proc., 23rd International Modal Analysis Conference, Orlando, FL*.

Hatzikyriakou, A., Lin, N., Gong, J., Xian, S., Hu, X., and Kennedy, A. (2015). "Component-Based Vulnerability Analysis for Residential Structures Subjected to Storm Surge Impact from Hurricane Sandy." *Natural Hazards Review*, 05015005.

Hegeman, J. W., Sardeshmukh, V. B., Sugumaran, R., and Armstrong, M. P. (2014). "Distributed LiDAR data processing in a high-memory cloud-computing environment." *Annals of GIS*, 20(4), 255-264.

Henderson, T. C., and Bhanu, B. "Three point seed method for the extraction of planar faces from range data." *Proc., Proceedings of the Workshop on Industrial Applications of Machine Vision*, 181-186.

Hengl, T. (2006). "Finding the right pixel size." *Computers & Geosciences*, 32(9), 1283-1298.

Henn, A., Gröger, G., Stroh, V., and Plümer, L. (2013). "Model driven reconstruction of roofs from sparse LIDAR point clouds." *ISPRS Journal of photogrammetry and remote sensing*, 76, 17-29.

Herrera, F., Herrera-Viedma, E., and Verdegay, J. (1996). "Direct approach processes in group decision making using linguistic OWA operators." *Fuzzy Sets and systems*, 79(2), 175-190.

Hirokawa, R., Kubo, D., Suzuki, S., Meguro, J.-i., and Suzuki, T. "A small UAV for immediate hazard map generation." 2725.

Hodgson, M. E., Battersby, S. E., Davis, B. A., Liu, S., and Sulewski, L. (2014). "Geospatial data collection/use in disaster response: a united states nationwide survey of state agencies." *Cartography from pole to pole*, Springer, 407–419.

Hodgson, M. E., and Bresnahan, P. (2004). "Accuracy of airborne lidar-derived elevation." *Photogrammetric Engineering & Remote Sensing*, 70(3), 331-339.

Honig, M. I., and Coburn, C. (2008). "Evidence-based decision making in school district central offices: Toward a policy and research agenda." *Educational Policy*, 22(4), 578-608.

Horan, T. A., and Schooley, B. L. (2007). "Time-critical information services." *Communications of the ACM*, 50(3), 73-78.

Hsiao, K., Liu, J., Yu, M., and Tseng, Y. "Change detection of landslide terrains using ground-based LiDAR data." *Proc., XXth ISPRS Congress, Istanbul, Turkey, Commission VII, WG*, 5.

Huang, F., Liu, D., Tan, X., Wang, J., Chen, Y., and He, B. (2011). "Explorations of the implementation of a parallel IDW interpolation algorithm in a Linux cluster-based parallel GIS." *Computers & Geosciences*, 37(4), 426-434.

Huang, Y. (2000). "Evaluation of information loss in digital elevation models with digital photogrammetric systems." *The Photogrammetric Record*, 16(95), 781-791.

Hug, C., Krzystek, P., and Fuchs, W. "Advanced lidar data processing with LasTools." *Proc., XXth ISPRS Congress*, 12-23.

Hwang, C.-L., Lai, Y.-J., and Liu, T.-Y. (1993). "A new approach for multiple objective decision making." *Computers & operations research*, 20(8), 889-899.

Irvin, R. A., and Stansbury, J. (2004). "Citizen participation in decision making: Is it worth the effort?" *Public administration review*, 64(1), 55-65.

ISO (2002). "Guidelines for quality and/or environmental management systems auditing." The International Organization for Standardization.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and others "KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera." ACM, 559–568.

Jian, X., Xiao, X., Chengfang, H., Zhizhong, Z., Zhaohui, W., and Dengzhong, Z. (2015). "a Hadoop-Based Algorithm of Generating dem Grid from Point Cloud Data." *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(7), 1209.

Joyce, K. E., Wright, K. C., Samsonov, S. V., and Ambrosia, V. G. (2009). *Remote sensing and the disaster management cycle*, INTECH Open Access Publisher.

Jwa, Y., Sohn, G., and Kim, H. (2009). "Automatic 3d powerline reconstruction using airborne lidar data." *Int. Arch. Photogramm. Remote Sens*, 38(Part 3), W8.

Kalal, Z., Mikolajczyk, K., and Matas, J. "Forward-backward error: Automatic detection of tracking failures." *Proc., Pattern recognition (ICPR), 2010 20th international conference on*, IEEE, 2756-2759.

Kamburugamuve, S., Ekanayake, S., Pathirage, M., and Fox, G. "Towards high performance processing of streaming data in large data centers." *Proc., Parallel and Distributed Processing Symposium Workshops, 2016 IEEE International*, IEEE, 1637-1644.

Kersting, J., and Kersting, A. P. B. "LIDAR DATA POINTS FILTERING USING ARCGIS'3D AND SPATIAL ANALYST." *Proc., Proceedings of the 25th ESRI User Conference, San Diego, California, USA*, Citeseer.

Kervyn, F. (2001). "Modelling topography with SAR interferometry: illustrations of a favourable and less favourable environment." *Computers & geosciences*, 27(9), 1039-1050.

Khazaei, H., Misic, J., and Misic, V. B. (2012). "Performance analysis of cloud computing centers using m/g/m/m+ r queuing systems." *IEEE Transactions on parallel and distributed systems*, 23(5), 936-943.

Kienzle, S. (2004). "The effect of DEM raster resolution on first order, second order and compound terrain derivatives." *Transactions in GIS*, 8(1), 83-111.

Kobler, A., Pfeifer, N., Ogrinc, P., Todorovski, L., Oštir, K., and Džeroski, S. (2007). "Repetitive interpolation: A robust algorithm for DTM generation from Aerial Laser Scanner Data in forested terrain." *Remote sensing of environment*, 108(1), 9-23.

Korah, T., Medasani, S., and Owechko, Y. "Strip histogram grid for efficient lidar segmentation from urban environments." *Proc., Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, IEEE, 74-81.

Kreylos, O., Bawden, G. W., and Kellogg, L. H. "Immersive visualization and analysis of LiDAR data." *Proc., International Symposium on Visual Computing*, Springer, 846-855.

Kuhlthau, C. C. (1991). "Inside the search process: Information seeking from the user's perspective." *Journal of the American society for information science*, 42(5), 361.

Kwan, M.-P., and Lee, J. (2005). "Emergency response after 9/11: the potential of real-time 3D GIS for quick emergency response in micro-spatial environments." *Computers, Environment and Urban Systems*, 29(2), 93-113.

Labiak, R. C., Van Aardt, J. A., Bespalov, D., Eychner, D., Wirch, E., and Bischof, H.-P. "Automated method for detection and quantification of building damage and debris using post-disaster LiDAR data." *Proc., SPIE Defense, Security, and Sensing*, International Society for Optics and Photonics, 80370F-80370F-80378.

Lee, C. A., Gasster, S. D., Plaza, A., Chang, C.-I., and Huang, B. (2011). "Recent developments in high performance computing for remote sensing: A review." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4, 508–527.

Lee, H. J. "Parallel Algorithm for Building Extraction from LiDAR Data." *Proc., Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1.

Lefsky, M. A., Cohen, W. B., Parker, G. G., and Harding, D. J. (2002). "Lidar remote sensing for ecosystem studies: Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists." *BioScience*, 52, 19–30.

Leslar, M., Hu, B., and Wang, J. (2014). "Error analysis of a mobile terrestrial LiDAR system." *Geomatica*, 68(3), 183-194.

Li, M., Cheng, L., Gong, J., Liu, Y., Chen, Z., Li, F., Chen, G., Chen, D., and Song, X. (2008). "Post-earthquake assessment of building damage degree using LiDAR data and imagery." *Science in China Series E: Technological Sciences*, 51, 133-143.

Li, Z., Hodgson, M. E., and Li, W. (2018). "A general-purpose framework for parallel processing of large-scale LiDAR data." *International Journal of Digital Earth*, 11(1), 26-47.

Lim, E. H., and Suter, D. "Multi-scale conditional random fields for over-segmented irregular 3D point clouds classification." *Proc., Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, IEEE, 1-7.

Lin, N., Emanuel, K., Oppenheimer, M., and Vanmarcke, E. (2012). "Physically based assessment of hurricane surge threat under climate change." *Nature Climate Change*, 2, 462–467.

Lindell, M. K., and Perry, R. W. (1992). *Behavioral foundations of community emergency planning*, Hemisphere Publishing Corp.

Lippitt, C. D., Stow, D. A., and Clarke, K. C. (2014). "On the nature of models for time-sensitive remote sensing." *International Journal of Remote Sensing*, 35, 6815–6841.

Liu, L. (2013). "Computing infrastructure for big data processing." *Frontiers of Computer Science*, 7, 165–170.

Liu, X. (2008). "Airborne LiDAR for DEM generation: some critical issues." *Progress in Physical Geography*, 32(1), 31-49.

Liu, X., Zhang, Z., Peterson, J., and Chandra, S. "The effect of LiDAR data density on DEM accuracy." *Proc., Proceedings of the International Congress on Modelling and Simulation (MODSIM07), Christchurch, New Zealand*.

Longstaff, P. H., Armstrong, N. J., Perrin, K., Parker, W. M., and Hidek, M. A. (2010). "Building resilient communities: A preliminary framework for assessment." *Homeland Security Affairs*, 6.

Lukač, N., and Žalik, B. (2013). "GPU-based roofs' solar potential estimation using LiDAR data." *Computers & Geosciences*, 52, 34-41.

Lwin, K. K., and Murayama, Y. (2011). "Estimation of building population from LIDAR derived digital volume model." *Spatial analysis and modeling in geographical transformation process*, Springer, 87-98.

Ma, B. N., and Mark, J. W. (1995). "Approximation of the mean queue length of an M/G/c queueing system." *Operations Research*, 43(1), 158-165.

Marjanović, M., Kovačević, M., Bajat, B., and Voženílek, V. (2011). "Landslide susceptibility assessment using SVM machine learning algorithm." *Engineering Geology*, 123(3), 225-234.

May¹, N. C., and Toth, C. K. (2007). "Point positioning accuracy of airborne LiDAR systems: A rigorous analysis."

Mayer, H. (2008). "Object extraction in photogrammetric computer vision." *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(2), 213-222.

McCarthy, T., Fotheringham, S., Charlton, M., Winstanley, A. C., and O'Malley, V. (2007). "Integration of LIDAR and stereoscopic imagery for route corridor surveying." *Mobile mapping technology*, 37, 1125-1130.

McLaughlin, R. A. (2006). "Extracting transmission lines from airborne LIDAR data." *IEEE Geoscience and Remote Sensing Letters*, 3(2), 222-226.

McNie, E. C. (2007). "Reconciling the supply of scientific information with user demands: an analysis of the problem and review of the literature." *Environmental science & policy*, 10(1), 17-38.

Meng, X., Currit, N., and Zhao, K. (2010). "Ground filtering algorithms for airborne LiDAR data: A review of critical issues." *Remote Sensing*, 2(3), 833-860.

Meng, X., Wang, L., Silván-Cárdenas, J. L., and Currit, N. (2009). "A multi-directional ground filtering algorithm for airborne LIDAR." *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(1), 117-124.

Meng, X., Yu, X., Peng, Z., and Hong, B. (2012). "Detecting earthquakes around Salton Sea following the 2010 Mw7. 2 El Mayor-Cucapah earthquake using GPU parallel computing." *Procedia Computer Science*, 9, 937-946.

Mezian, M., Vallet, B., Soheilian, B., and Paparoditis, N. (2016). "Uncertainty Propagation for Terrestrial Mobile Laser Scanner." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 41.

Miyazaki, H., Nagai, M., and Shibasaki, R. (2015). "Reviews of Geospatial Information Technology and Collaborative Data Delivery for Disaster Risk Management." *ISPRS International Journal of Geo-Information*, 4, 1936–1964.

Mongus, D., Lukač, N., and Žalik, B. (2014). "Ground and building extraction from LiDAR data based on differential morphological profiles and locally fitted surfaces." *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 145-156.

Moore, I. D., Gessler, P., Nielsen, G., and Peterson, G. (1993). "Soil attribute prediction using terrain analysis." *Soil Science Society of America Journal*, 57(2), 443-452.

Morsdorf, F., Meier, E., Kötz, B., Itten, K. I., Dobbertin, M., and Allgöwer, B. (2004). "LIDAR-based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management." *Remote Sensing of Environment*, 92, 353–362.

Mountrakis, G., Im, J., and Ogole, C. (2011). "Support vector machines in remote sensing: A review." *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247-259.

Mukherjee, S., Joshi, P., Mukherjee, S., Ghosh, A., Garg, R., and Mukhopadhyay, A. (2013). "Evaluation of vertical accuracy of open source Digital Elevation Model (DEM)." *International Journal of Applied Earth Observation and Geoinformation*, 21, 205-217.

NSF (2015). "New U.S.-Japan collaborations bring Big Data approaches to disaster response." <https://www.nsf.gov/news/news_summ.jsp?cntn_id=134609>.

Olsen, M. J. (2013). *Guidelines for the use of mobile LIDAR in transportation applications*, Transportation Research Board.

Olsen, M. J., Cheung, K. F., Yamazaki, Y., Butcher, S., Garlock, M., Yim, S., McGarity, S., Robertson, I., Burgos, L., and Young, Y. L. (2012). "Damage assessment of the 2010 Chile earthquake and tsunami using terrestrial laser scanning." *Earthquake Spectra*, 28, S179–S197.

Orlovsky, S. (1978). "Decision-making with a fuzzy preference relation." *Fuzzy sets and systems*, 1(3), 155-167.

Ouyang, M. (2014). "Review on modeling and simulation of interdependent critical infrastructure systems." *Reliability engineering & System safety*, 121, 43-60.

Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., and Phillips, J. C. (2008). "GPU computing." *Proceedings of the IEEE*, 96(5), 879-899.

Pandey, P., and Pompili, D. "MobiDiC: Exploiting the untapped potential of mobile distributed computing via approximation." *Proc., Pervasive Computing and Communications (PerCom), 2016 IEEE International Conference on*, IEEE, 1-9.

Passalacqua, P., Do Trung, T., Foufoula‐Georgiou, E., Sapiro, G., and Dietrich, W. E. (2010). "A geometric framework for channel network extraction from lidar: Nonlinear diffusion and geodesic paths." *Journal of Geophysical Research: Earth Surface*, 115(F1).

Pederson, P., Dudenhoeffer, D., Hartley, S., and Permann, M. (2006). "Critical infrastructure interdependency modeling: a survey of US and international research." *Idaho National Laboratory*, 25, 27.

Peterson, K., Ziglar, J., and Rybski, P. E. "Fast feature detection and stochastic parameter estimation of road shape using multiple LIDAR." *Proc., Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, IEEE, 612-619.

Plale, B., and Schwan, K. "dQCOB: managing large data flows using dynamic embedded queries." *Proc., High-Performance Distributed Computing, 2000. Proceedings. The Ninth International Symposium on*, IEEE, 263-270.

Plaza, A. J., and Chang, C.-I. (2007). *High performance computing in remote sensing*, CRC Press.

Polat, N., Uysal, M., and Toprak, A. S. (2015). "An investigation of DEM generation process based on LiDAR data filtering, decimation, and interpolation methods for an urban area." *Measurement*, 75, 50-56.

Poulter, B., and Halpin, P. N. (2008). "Raster modelling of coastal flooding from sea-level rise." *International Journal of Geographical Information Science*, 22, 167–182.

Priestnall, G., Jaafar, J., and Duncan, A. (2000). "Extracting urban features from LiDAR digital surface models." *Computers, Environment and Urban Systems*, 24(2), 65-78.

Qi, H., and Altinakar, M. S. (2011). "A GIS-based decision support system for integrated flood management under uncertainty with two dimensional numerical simulations." *Environmental Modelling & Software*, 26(6), 817-821.

Qiu, M., Ming, Z., Wang, J., Yang, L. T., and Xiang, Y. (2014). "Enabling cloud computing in emergency management systems." *IEEE Cloud Computing*, 1(4), 60-67.

Rabbani, T., Van Den Heuvel, F., and Vosselmann, G. (2006). "Segmentation of point clouds using smoothness constraint." *International archives of photogrammetry, remote sensing and spatial information sciences*, 36(5), 248-253.

Renart, E., Balouek-Thomert, D., Hu, X., Gong, J., and Parashar, M. "Online Decision-Making Using Edge Resources for Content-Driven Stream Processing." *Proc., e-*

Science (e-Science), 2017 IEEE 13th International Conference on, IEEE, 384-392.

Renart, E. G., Diaz-Montes, J., and Parashar, M. "Data-driven stream processing at the edge." *Proc., Fog and Edge Computing (ICFEC), 2017 IEEE 1st International Conference on*, IEEE, 31-40.

Renschler, C. S., Frazier, A., Arendt, L., Cimellaro, G.-P., Reinhorn, A. M., and Bruneau, M. (2010). "A framework for defining and measuring resilience at the community scale: The PEOPLES resilience framework." *US Department of Commerce National Institute of Standards and Technology, Office of Applied Economics Engineering Laboratory NIST GCR*, 10–930.

Repo, A. J. (1986). "The dual approach to the value of information: an appraisal of use and exchange values." *Information processing & management*, 22(5), 373-383.

Reutebuch, S. E., McGaughey, R. J., Andersen, H.-E., and Carson, W. W. (2003). "Accuracy of a high-resolution lidar terrain model under a conifer forest canopy." *Canadian journal of remote sensing*, 29(5), 527-535.

Romanowicz, R. J., Young, P. C., Beven, K. J., and Pappenberger, F. (2008). "A data based mechanistic approach to nonlinear flood routing and adaptive flood level forecasting." *Advances in Water Resources*, 31(8), 1048-1056.

Rottensteiner, F., and Briese, C. (2002). "A new method for building extraction in urban areas from high-resolution LIDAR data." *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 34, 295–301.

Rottensteiner, F., Trinder, J., Clode, S., and Kubik, K. (2005). *Automated delineation of roof planes from lidar data*, na.

Růžička, J., Orčík, L., Růžičková, K., and Kisztner, J. (2017). "Processing LIDAR Data with Apache Hadoop." *The Rise of Big Spatial Data*, Springer, 351-358.

Sabou, M., Bontcheva, K., and Scharl, A. "Crowdsourcing research opportunities: lessons from natural language processing." *Proc., Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, ACM, 17.

Sagara, E. (2012). "Hurricane Sandy's destruction: Aerial assessment shows nearly 72K buildings damaged in NJ." *The Star Ledger. November*, 18.

Sakaki, T., Okazaki, M., and Matsuo, Y. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proc., Proceedings of the 19th international conference on World wide web*, ACM, 851-860.

Sakaki, T., Okazaki, M., and Matsuo, Y. "Earthquake shakes Twitter users: real-time event detection by social sensors." ACM, 851–860.

Sanders, B. F. (2007). "Evaluation of on-line DEMs for flood inundation modeling." *Advances in Water Resources*, 30, 1831–1843.

Scheidl, C., Rickenmann, D., and Chiari, M. (2008). "The use of airborne LiDAR data for the analysis of debris flow events in Switzerland." *Natural Hazards and Earth System Sciences*, 8(5), 1113-1127.

Schnabel, R., Wahl, R., and Klein, R. "Efficient RANSAC for point‐cloud shape detection." *Proc., Computer graphics forum*, Wiley Online Library, 214-226.

Schön, B., Bertolotto, M., Laefer, D. F., and Morrish, S. "Storage, manipulation, and visualization of LiDAR data." International Society of Photogrammetry and Remote Sensing.

Schütz, M. (2016). "Potree: Rendering large point clouds in web browsers." *Technische Universität Wien, Wiedeń*.

Sherman, H. D., and Zhu, J. (2006). *Service productivity management: Improving service performance using data envelopment analysis (DEA)*, Springer Science & Business Media.

Shi, J., and Malik, J. (2000). "Normalized cuts and image segmentation." *IEEE Transactions on pattern analysis and machine intelligence*, 22(8), 888-905.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.

Singh, A. (1989). "Review article digital change detection techniques using remotely-sensed data." *International journal of remote sensing*, 10(6), 989-1003.

Son, H., Kim, C., and Kim, C. (2014). "Fully automated as-built 3D pipeline extraction method from laser-scanned data based on curvature computation." *Journal of Computing in Civil Engineering*, 29(4), B4014003.

Soudarissanane, S., Lindenbergh, R., and Gorte, B. "Reducing the error in terrestrial laser scanning by optimizing the measurement set-up." International Society for Photogrammetry and Remote Sensing.

Starcke, K., and Brand, M. (2012). "Decision making under stress: a selective review." *Neuroscience & Biobehavioral Reviews*, 36(4), 1228-1248.

Sugumaran, R., Oryspayev, D., and Gray, P. "GPU-based cloud performance for LiDAR data processing." *Proc., Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications*, ACM, 48.

Sutherland, W. J., Fleishman, E., Mascia, M. B., Pretty, J., and Rudd, M. A. (2011). "Methods for collaboratively identifying research priorities and emerging issues in science and policy." *Methods in Ecology and Evolution*, 2(3), 238-247.

Suveg, I., and Vosselman, G. (2004). "Reconstruction of 3D building models from aerial images and maps." *ISPRS Journal of Photogrammetry and remote sensing*, 58, 202–224.

Tan, W., Missier, P., Foster, I., Madduri, R., De Roure, D., and Goble, C. (2010). "A comparison of using Taverna and BPEL in building scientific workflows: the case of caGrid." *Concurrency and Computation: Practice and Experience*, 22(9), 1098-1117.

Tesfa, T. K., Tarboton, D. G., Watson, D. W., Schreuders, K. A., Baker, M. E., and Wallace, R. M. (2011). "Extraction of hydrological proximity measures from DEMs using parallel processing." *Environmental Modelling & Software*, 26(12), 1696-1709.

Thompson, S., Altay, N., Green III, W. G., and Lapetina, J. (2006). "Improving disaster response efforts with decision support systems." *International Journal of Emergency Management*, 3(4), 250-263.

Thomson, A. (1998). "Supervised versus unsupervised methods for classification of coasts and river corridors from airborne remote sensing." *International Journal of Remote Sensing*, 19(17), 3423-3431.

Timmerman, J. G., Beinat, E., Termeer, K., and Cofino, W. (2010). "Analyzing the data-rich-but-information-poor syndrome in Dutch water management in historical perspective." *Environmental management*, 45(5), 1231-1242.

Timmerman, J. G., Boer, J., Hisschemöller, M., and Mulder, W. (2001). "Specifying information needs: improving the working methodology." *Regional Environmental Change*, 2(2), 77-84.

Timmerman, J. G., Ottens, J. J., and Ward, R. C. (2000). "The information cycle as a framework for defining information goals for water-quality monitoring." *Environmental management*, 25(3), 229-239.

Todd, P., and Benbasat, I. (1992). "The use of information in decision making: an experimental investigation of the impact of computer-based decision aids." *Mis Quarterly*, 373-393.

Tolone, W. J., Wilson, D., Raja, A., Xiang, W.-n., Hao, H., Phelps, S., and Johnson, E. W. (2004). "Critical infrastructure integration modeling and simulation." *ISI*, 3073, 214-225.

Trinder, J., and Salah, M. (2012). "Aerial images and LiDAR data fusion for disaster change detection." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1(4), 227-232.

Vaughan, H. H., Waide, R. B., Maass, J. M., and Ezcurra, E. (2007). "Developing and delivering scientific information in response to emerging needs." *Frontiers in Ecology and the Environment*, 5(4), 1-4.

Verma, V., Kumar, R., and Hsu, S. "3d building detection and modeling from aerial lidar data." *Proc., Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, IEEE, 2213-2220.

Vilaplana, J., Solsona, F., Teixidó, I., Mateo, J., Abella, F., and Rius, J. (2014). "A queuing theory model for cloud computing." *The Journal of Supercomputing*, 69(1), 492-507.

Vitner, G., Rozenes, S., and Spraggett, S. (2006). "Using data envelope analysis to compare project efficiency in a multi-project environment." *International Journal of project management*, 24(4), 323-329.

Voigt, S., Kemper, T., Riedlinger, T., Kiefl, R., Scholte, K., and Mehl, H. (2007). "Satellite image analysis for disaster and crisis-management support." *IEEE transactions on geoscience and remote sensing*, 45, 1520–1528.

Vos, P., Meelis, E., and Ter Keurs, W. (2000). "A framework for the design of ecological monitoring programs as a tool for environmental and nature management." *Environmental monitoring and assessment*, 61(3), 317-344.

Vugteveen, P., van Katwijk, M. M., Rouwette, E., and Hanssen, L. (2014). "How to structure and prioritize information needs in support of monitoring design for Integrated Coastal Management." *Journal of sea research*, 86, 23-33.

Wähner, K. (2014). "Real-time stream processing as game changer in a big data world with hadoop and data warehouse." *InfoQ (September 10, 2014)*.

Walker, N. D. (1996). "Satellite assessment of Mississippi River plume variability: causes and predictability." *Remote sensing of environment*, 58, 21–35.

Wang, S., and Armstrong, M. P. (2003). "A quadtree approach to domain decomposition for spatial interpolation in grid computing environments." *Parallel Computing*, 29(10), 1481-1504.

Ward, R. C., Loftis, J. C., and McBride, G. B. (1986). "The “data-rich but information-poor” syndrome in water quality monitoring." *Environmental management*, 10(3), 291-297.

Webster, T. L., Forbes, D. L., Dickie, S., and Shreenan, R. (2004). "Using topographic lidar to map flood risk from storm-surge events for Charlottetown, Prince Edward Island, Canada." *Canadian Journal of Remote Sensing*, 30(1), 64-76.

Wei, G., Vasilakos, A. V., Zheng, Y., and Xiong, N. (2010). "A game-theoretic method of fair resource allocation for cloud computing services." *The journal of supercomputing*, 54(2), 252-269.

Westerink, J. J., Feyen, J. C., Atkinson, J. H., Luettich, R. A., Dawson, C. N., Powell, M. D., Dunion, J. P., Roberts, H. J., Kubatko, E. J., and Pourtaheri, H. (2004). "A new generation hurricane storm surge model for southern Louisiana." *Bulletin of the American Meteorological Society, in review*.

Wilson, J. P., and Gallant, J. C. (2000). "Secondary topographic attributes." *Terrain analysis: Principles and applications*, 87-131.

Wolock, D. M., and Price, C. V. (1994). "Effects of digital elevation model map scale and data resolution on a topography‐based watershed model." *Water Resources Research*, 30(11), 3041-3052.

Wright, L. D., Nichols, C. R., Cosby, A. G., Danchuk, S., D'Elia, C. F., and Mendez, G. R. (2015). "Trans-disciplinary Collaboration to Enhance Coastal Resilience: Envisioning a National Community Modeling Initiative."

Wu, H., Guan, X., and Gong, J. (2011). "ParaStream: a parallel streaming Delaunay triangulation algorithm for LiDAR points on multicore architectures." *Computers & geosciences*, 37(9), 1355-1363.

Xiao, Z., Song, W., and Chen, Q. (2013). "Dynamic resource allocation using virtual machines for cloud computing environment." *IEEE transactions on parallel and distributed systems*, 24(6), 1107-1117.

Xie, J., Yang, C., Zhou, B., and Huang, Q. (2010). "High-performance computing for the simulation of dust storms." *Computers, Environment and Urban Systems*, 34(4), 278-290.

Xiong, K., and Perros, H. "Service performance and analysis in cloud computing." *Proc., Services-I, 2009 World Conference on*, IEEE, 693-700.

Yang, B., Dong, Z., Zhao, G., and Dai, W. (2015). "Hierarchical extraction of urban objects from mobile laser scanning data." *ISPRS Journal of Photogrammetry and Remote Sensing*, 99, 45-57.

Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., Bambacus, M., and Fay, D. (2011). "Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?" *International Journal of Digital Earth*, 4(4), 305-329.

Yang, C., Li, W., Xie, J., and Zhou, B. (2008). "Distributed geospatial information processing: sharing distributed geospatial resources to support Digital Earth." *International Journal of Digital Earth*, 1(3), 259-278.

Yang, J., and Chen, Z. "Cloud computing research and security issues." *Proc., Computational intelligence and software engineering (CiSE), 2010 international conference on*, IEEE, 1-3.

Yu, Y., Li, J., Guan, H., Wang, C., and Yu, J. (2015). "Semiautomated extraction of street light poles from mobile LiDAR point-clouds." *IEEE Transactions on Geoscience and Remote Sensing*, 53(3), 1374-1386.

Yuan, C. (2012). *High performance computing for massive LiDAR data processing with optimized GPU parallel programming*, The University of Texas at Dallas.

Zhang, K., Yan, J., and Chen, S.-C. (2006). "Automatic construction of building footprints from airborne LIDAR data." *IEEE Transactions on Geoscience and Remote Sensing*, 44, 2523–2533.

Zhou, Z., Gong, J., and Guo, M. (2015). "Image-based 3D reconstruction for posthurricane residential building damage assessment." *Journal of Computing in Civil Engineering*, 30, 04015015.

Zhou, Z., Gong, J., Roda, A., and Farrag, K. (2016). "Multiresolution Change Analysis Framework for Postdisaster Assessment of Natural Gas Pipeline Risk." *Transportation Research Record: Journal of the Transportation Research Board*(2595), 29-39.

Zhu, Z., and Brilakis, I. (2007). "Comparison of civil infrastructure optical-based spatial data acquisition techniques." *Computing in Civil Engineering (2007)*, 737–744.

Ziadat, F. M. (2007). "Effect of contour intervals and grid cell size on the accuracy of DEMs and slope derivatives." *Transactions in GIS*, 11(1), 67-81.