8-19-2016 12:00 AM

Using Physical and Social Sensors in Real-Time Data Streaming for Natural Hazard Monitoring and Response

Yelena Kropivnitskaya, *The University of Western Ontario*

Supervisor: Dr. Kristy Tiampo, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Geophysics

Abstract

Technological breakthroughs in computing over the last few decades have resulted in important advances in natural hazards analysis. In particular, integration of a wide variety of information sources, including observations from spatially-referenced physical sensors and new social media sources, enables better estimates of real-time hazard. The main goal of this work is to utilize innovative streaming algorithms for improved real-time seismic hazard analysis by integrating different data sources and processing tools into cloud applications. In streaming algorithms, a sequence of items from physical and social sensors can be processed in as little as one pass with no need to store the data locally. Massive data volumes can be analyzed in near-real time with reasonable limits on storage space, an important advantage for natural hazard analysis.

Seismic hazard maps are used by policymakers to set earthquake resistant construction standards, by insurance companies to set insurance rates and by civil engineers to estimate stability and damage potential. This research first focuses on improving probabilistic seismic hazard map production. The result is a series of maps for different frequency bands at significantly increased resolution with much lower latency time that includes a range of high-resolution sensitivity tests.

Second, a method is developed for real-time earthquake intensity estimation using joint streaming analysis from physical and social sensors. Automatically calculated intensity estimates from physical sensors such as seismometers use empirical relationships between ground motion and intensity, while those from social sensors employ questionnaires that evaluate ground shaking levels based on personal observations. Neither is always sufficiently precise and/or timely. Results demonstrate that joint processing can significantly reduce the response time to a damaging earthquake and estimate preliminary intensity levels during the first ten minutes after an event. The combination of social media and network sensor data, in conjunction with innovative computing algorithms, provides a new paradigm for real-time earthquake detection, facilitating rapid and inexpensive risk reduction. In particular, streaming algorithms are an efficient method that addresses three major problems in hazard estimation by improving resolution, decreasing processing latency to near real-time standards

and providing more accurate results through the integration of multiple data sets.

Keywords

Pipelining, stream computing, high performance computing, parallel computing, physical sensors, social sensors, hazard estimators

Co-Authorship Statement

This thesis is prepared in integrated-article format and the following manuscripts were written by Yelena Kropivnitskaya:

Kropivnitskaya, Y., Qin, J., Tiampo, K., & Bauer, M. (2015). Pipelining Implementation for High Resolution Seismic Hazard Maps Production. *International Conference On Computational Science, ICCS 2015 – Computational Science at the Gates of Nature* (pp. 1473-1482). Reykjavik: Procedia Computer Science.

Kropivnitskaya, Y., Tiampo, K., Qin, J., & Bauer, M. (2016). Impact of the Ground Motion Prediction Equation Changes on Eastern Canada Hazard Maps. *2016 CSCE Annual Conference, 51*. London, Canada.

Kropivnitskaya, Y., Tiampo, K., Qin, J., & Bauer, M. (2016). Real-Time Earthquake Intensity Estimation Using Streaming Data Analysis of Social and Physical Sensors. *Pure and Applied Geophysics*, accepted with minor revisions.

Kropivnitskaya, Y., Tiampo, K., Qin, J., & Bauer, M. (2016). The Predictive Relationship between Earthquake Intensity and Tweets Rate for Real-Time Ground Motion Estimation. *Seismological Research Letter*, submitted on 07/19/2016.

The work for these projects was completed under supervision and financial support of Dr. Kristy Tiampo. The research was funded by a MITACS Accelerate Grant, an NSERC Discovery Grant, and the NSERC and ICLR CRD.

Acknowledgments

The results presented in this thesis are the outcomes of four years journey, where I have been accompanied and supported by many people. Here I would like to express my gratitude for all of them.

First and foremost, I express my gratitude to my supervisor Dr. Kristy Tiampo for her support, patience and guidance, for her ability to be a scientist, a supervisor, a partner and good human being at the same time. I admire her professionalism, she taught me a lot of things from academia to life scale. This work would not be possible without her participation and support.

I thank the members of my research committee Dr. Gail Atkinson and Dr. Robert Scherbakov for their teaching and discussions about this work and their valuable comments that improved the final outcome.

Dr. Jinhui Qin and Dr. Michael Bauer played an important role in the realization of all projects covered in this thesis. Their help, constructive criticism, fresh ideas, professionalism, and openness to new projects improved the quality of this work. I thank SOSCIP and SHARCNET for providing computing resources and technical support during my work.

Thank you to Dr. Gero Michel and Tahinde Frederick, who opened for me the door to the catastrophe modelling field and gave me new experiences, skills and knowledge. All together that gave new perspective to this work.

Also I am grateful to the people from the Department of Earth Sciences of Western University for the favorable working environment, for their support and readiness to help. Special thanks to Claire Mortera, Marie Schell, Kevin Jordan, Miyako Maekawa, Jen Heidenheim, Bernie Dunn, Gerhard Pratt and to all the graduate students, postdocs and labmates from the Computational Laboratory for Fault System Modeling, Analysis, and Data Assimilation and from the Engineering Seismology Laboratory.

Thanks to all my friends who helped me not to dwell on the work: Colin Sproat, Beth Hooper, Hadis Samadi Alinia, Xiaoming Zhang, Ivan Kobyzev, Slava Kohut,Alexsandr

Larychev, Sergey Vassilyev, Tatyana Katsaga, Leonid Nesteruk and Viktoriya Karpenko. I appreciate our friendship and your support.

Thanks to my parents Nataliya and Yevgeniy, to my sister Yulia and my better half Konstantin for their love, patience and support during my good and bad times, without them this journey would have not been possible.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 General Introduction

1.1 Objectives

Today near-real time big data processing approaches strengthen a community's resilience to natural hazard throughout the broad implementation of natural hazards applications, exposure estimation techniques, disaster impact monitoring, recovery effort coordination and vulnerability mitigation methods.

The overall volume of available data in the natural hazards field has been increasing exponentially over the last decade (IDC, 2012) with the appearance of new physical and social sensors. In general, this phenomenon is often described as a complex feature resulting from the growth of big data. Initially, big data was defined as relatively massive datasets that cannot be captured, stored, managed, and analyzed by standard database software tools (Manyika et al. 2011). This definition has been changed over time and today big data is not described in terms of large volumes alone. In reality these datasets include other characteristics defined as the 'V's' of big data, such as volume, velocity, variety, veracity, variability and validity (Tiampo et al., 2016). Data velocity decribes the rate of data production or usage (for instance, the rate of readings taken from a sensor operating in real-time). Data variety is a property refers to structural complexity of big data sets. Also, big data can be uncertain, imprecise and incomplete, and this is defined by the property of veracity. Variability refers to data whose meaning is constantly changing in time. Finally, validity is related to the data accuracy and correctness in the context of usage (Tiampo et al., 2016). Because there is some overlap between these terms, the following discussion will focus on volume, velocity, variety and variability.

The growth in both data volume and variability has had a significant demand on the capacity improvement of hazard monitoring tools. Response methodologies can be improved by adding new, complementary data to existing analysis techniques. In practice, these advances are complicated by the fact that data often is not unified and needs to be processed using various technologies and methods that are all factors studied

under the complex field of big data today (Hu et al., 2014). For example, data integration becomes important in cases where new data sources can significantly speed up the calculations that are critical for natural hazard monitoring applications. Joint streaming processing of data from physical and social sensors for natural hazards monitoring and mitigation can significantly improve not only temporal but also spatial resolution of existing analysis approaches such as hazard intensity mapping. The explosion of big data has made existing data management methods inefficient and new approaches have been proposed to deal with the various aspects associated with their incorporation and analysis. One potential solution includes analyzing large volumes of data from multiple sources on the fly without storing it, often called "data in motion" or streaming analysis. Streaming analysis focuses on rapid processing of continuous unstructured data. As a result, data is continuously analyzed and transformed in memory. As a result, the streaming approach addresses problem arising on the overlap between big data characteristics described above, such as volume, velocity, variability and variety and affects storage, transmission, and access of that data (Tiampo et. al., 2016). The processing of data streams works by processing "time windows" of data in memory across a cluster of servers. Streaming analysis can be applied to data sources such as seismic monitoring stations, deep ocean bottom pressure sensors in tsunami stations, sea level stations, water level and flow sensors in river gauging stations, satellite and airborne altimetry systems and other physical sensors. In addition, big streams of social sensors create data packets with high frequency in real time that can be imported into a streaming environment as well.

Here we use the IBM InfoSphere Streams application for the development of near real-time hazards assessment and response applications. The goal of this work was to develop an efficient application, aimed at better characterizing the exposure and hazard associated with large earthquakes. The work has been done in collaboration with the consortium of Canadian academic institutions, the high-performance computing network SHARCNET, in order to use modern computer science advantages such as advanced cloud and streaming computing to improve existing solutions and instruments used in hazard and risk assessment. First, the resolution of seismic hazard maps was significantly improved through optimization of those algorithms that are used to estimate particular values along with a pronounced reduction in their execution time.

In the second phase of this work, an approach to near real-time earthquake intensity estimation using streaming data analysis from social (Twitter data) and physical (seismographs and accelerometers) sensors is developed. It was found that seismic intensity levels correlate well with tweet rates in the ten minutes following an earthquake in regions with a significant number of Twitter users, such as California, Chile, and Japan. Four models describing that correlation are proposed: linear, two-segment linear, three-segment linear and exponential. The relationships were validated and calibrated based on a historical data set of intensity records and tweets archives. These relationships can be used for near real-time intensity estimation techniques where Twitter data is employed either as a complimentary or sole data source. In this case, "near real-time" is related to the time delay between the occurrence of a hazard event and the release of the processing results to government, emergency response and public users. Taken together, these results demonstrate the significance of streaming computing and joint processing techniques in the natural hazards field.

1.2 Background

Natural hazards are naturally occurring events that often cause large amounts of economic and human loss and that cannot be controlled or eliminated. Over the last several decades, climate change has been altering the natural hazard patterns by increasing their frequency and intensity (Houghton, 2001). At the same time, constantly increasing population density, particularly along the coasts, adds complexity to the problem (Hinrichsen, 1998). Despite the increase in number and intensify of natural disasters (Houghton, 2001), the effect of natural hazards on people and the environment can be mitigated, wherein the level of ability to mitigate and the level of mitigation resources directly correlates with general economic situation in the certain country (Alexander, 1997). Natural events are more risky in developing countries where the economic situation forces human priorities to change and communities to accept a lower standard of living and insufficient hazard mitigation actions (UN/ISDR, 2006).

Proper mitigation of natural hazards, such as forecasting, early warning systems, monitoring and disaster response, requires the ability to process large amounts of data at high speeds. Since the end of the 2000s, this requirement was one of the reasons for the

rapid growth of the big data field. In the context of emergency response, big data is comprised of three major components: spatially distributed data sources; infrastructure capable of analyzing large amount of data; and people managing analysis of data and delivery of results (Cimellaro, 2010).

Big data is one of the most promising areas for natural disasters risk mitigation. As the amount of wireless telephone and internet users grows in developing countries over the next decade (Poushter, 2016), the ability to use new information sources for effective mitigation techniques has a great potential. Cell phones can provide detailed information on population density and movement and social networks such as Twitter or Facebook can assist with monitoring and response to natural hazards as well as provide data suitable for analysis during and after an event.

The exploitation of new data sources also brings new concerns related to the amount of data being collected, the ability to access that data and the accompanying security and privacy (Herold, 2016). Also, while available methods of gathering information work well in developed countries, they are still less successful in places with a smaller user base. As a result, existing uncertainty and multiple requirements make the assessment of new methods difficult.

Big data is not only addressing as the volume of data, its variety and variability (the number of data types and how it is changing with time) and velocity (the speed of data processing) (Laney, 2001). There also are issues related to data complexity, or the lack of a trivial description of data interactions, and recency, defined as the ability of a social sensor to react in real-time. Therefore, big data should be seen as an ecosystem that is a function of the new kinds of information that can be obtained and the necessary additional computational resources and analytical tools required by various new players in the field (Cimellaro, 2010). Big data is a complex social system defined by the three Cs (Letouze, 2013):

- Digital bread crumbs (Cimellaro, 2010). This is the data created by digitizing human actions and interactions. The majority is passively generated by multiple digital devices and services, primarily as a side effect of their operation. It can

be divided into three categories: digital content, exhaust data and sensing data (Costello et al., 2009).

- Capacities and analytics. These define the infrastructure, methods, instruments and skills required to process new information and include visualization techniques, algorithms and machine learning approaches (Letouze, 2013).

- Communities. These describe the players in the ecosystem such as data generators, processors and end users. Potentially this includes the entire population (Letouze, 2013).

To date, there are a number of big data solutions proposed and developed in the natural hazards field. These methods can be divided by the application purpose into three categories: disaster monitoring, hazard mapping and disaster response applications. Note that the research of this thesis falls into each of these categories. Below the overview of existing solutions in each category is provided.

1.2.1 Disaster monitoring

Proper monitoring is vital for disaster mitigation and big data solutions are already being used for that purpose in certain areas. For instance, remote sensing techniques include monitoring satellites with panchromatic, multispectral, infrared and thermal sensors for the near real-time smoke and fire detection (Nirupama and Simonovic, 2002). Further advancements in monitoring techniques can be accomplished by improving the existing systems and mastering new data sources.

One of the biggest avenues for improvement of existing monitoring solutions is risk measurement systems. New satellite data can provide more detailed information for hazard detection and mapping. For example, the National Aeronautics and Space Administration (NASA) Gravity Recovery and Climate Experiment (GRACE) satellites launched in 2002 are used to improve monitoring of groundwater depletion (Tapley et al., 2004) and the NASA Soil Moisture Active Passive mission (SMAP) (Entekhabi et al., 2010) will be able to provide the more accurate data related to soil moisture. These two

sources, when combined together, can help with detection of arid lands and identification of potentially food-insecure areas.

Another area where existing monitoring can be improved is early warning systems. For example, tsunami risk can be assessed by analyzing wave surface information obtained from satellite altimeters together with GPS data on ionospheric perturbations (Occhipinti et al., 2008). However, the associated requirement for real time ionosphere monitoring for anomaly detection make this solution hard to use in reality. Therefore, new big data solutions will be necessary in order to provide a more realistic approach to real-time tsunami mapping.

Early warning systems are also an area where new data sources can be used to improve monitoring techniques. For example, accelerometers build into cell phones can be used to detect earthquakes (Kong et al., 2016) while hand water pump sensors can provide hydrological data (Taylor and Alley, 2001).

New data sources combined with existing networks of sensors can offer great opportunities for improving accuracy and coverage of monitoring systems. For instance, the United States Geological Survey (USGS) uses social media data in conjunction with seismometers to detect earthquakes. USGS presented and evaluated an earthquake detection procedure that relies on Twitter data. Tweet-frequency time series were extracted from positive tweets and clearly presents large spikes correlated with the origin times of widely felt earthquakes (Earle et al., 2012). A recent study by Jongman et al. (2015) describes how social networks can be utilized to obtain images and descriptions of ongoing events in real time. Another example is the Global Flood Detection System, implemented in the Philippines and Pakistan, has been combined with Floodtags, an analytics platform based on Twitter posts. This combination proved to work better in highly populated areas during unexpected flood events, such as breaches of flood defense (Jongman et al., 2015).

Despite the fact that the examples above demonstrate that incorporation of new sensor networks is a promising approach to improve existing hazard event detection solutions, there are still many challenges. The variety of geographical areas will require different

thresholds and new ways of sampling the data in order to ensure success. Certain communities may require additional awareness about the usefulness of their social network posts or cell phone data disaster detection and response.

1.2.2 Hazard Mapping with Big Data

Correct assessment of vulnerabilities and hazard exposure is crucial for risk estimation and mitigation. Big data solutions are able to access some exposure characteristics such as information about population, infrastructure and economic situation and identify the vulnerability of certain geographical areas and countries. Also, while traditional hazard mapping systems rely on historical and statistical data available, multiple big data methods developed during recent years work well for environments where data is less available data or there is a complete absence of historical records. Musaev et al. (2014) developed the landslide detection and mapping system LITMUS by integrating multiple data sources from social sensors (Twitter, Instagram, and YouTube) and physical sensors (USGS seismometers and Tropical Rainfall Measuring Mission (TRMM) satellite). The system demonstrated better performance than traditional techniques employed by USGS for real-time hazard mapping.

Satellite images that are available today at relatively low cost have resulted in the rapid development of algorithms designed for their mapping to the existing landscape. Earlier approaches based on expert models do not allow scaling to global coverage, more frequent sampling or the use of data from multiple sources (Pesaresi and Freire, 2014). Better satellite images have driven the creation of new systems designed to process them. For example, GlobCover38 and the moderate-resolution imaging spectroradiometer (MODIS) Land Cover Type39, developed over the last decade, can classify urban areas with accuracy close to 96% (Klotz et al., 2014). Both systems, combined with global population information, are capable of performing risk analysis for countries lacking data suitable for other existing solutions.

Global Human Settlement Layer (GHSL), developed by the European Union's Joint Research Centre, and the Global Urban Footprint (GUF), created by the German Aerospace Center, is designed to perform global mapping at the best existing spatial

details. Both systems are capable of detecting small settlements based on the images received from the satellites. GHSL is expected to be more flexible as it is capable of quickly retrieving and integrating large amounts of image data as well as providing detailed information on current development such as buildings and other exposure characteristics (Pesaresi et al., 2013.). Maps produced by GHSL already show how vulnerable existing building are because of the way they were built or their age. Both solutions are currently being tested but preliminary results show them to be reliable (Klotz et al., 2014).

Another relatively new hazard mapping method is crowd sourcing, also called participatory mapping. Most of the existing projects involve volunteers with local knowledge to map images and social media information such as videos and photos. One of the projects is OpenStreetMap (OSM), started in 2004. Today OSM is an organization with more than 1.5 million users from over 80 countries who work together on creating open digital map of the entire world (Palen et al., 2014). Users are organized into groups by geographical area. There is a single database with information about existing building, roads etc. edited by project users remotely. Currently OSM does not cover all territories evenly but the data available in the database is of high quality when compared to other sources (Haklay, 2010). The Humanitarian OSM Team (HOT) is a part of OSM focused on disaster applications. HOT has proven its importance by providing detailed maps after earthquake in Haiti (2010), Typhoon Yolanda (2014) and the earthquake in Nepal (2015). HOT's Missing Maps Project is focused on mapping world's most vulnerable places, mostly in developing countries, to improve disaster response in future (Missing Map, 2015). Other HOT projects are aimed to map specific vulnerabilities. For example, one is working on infrastructure vulnerable to floods in Dar es Salaam, Tanzania (HotOSM, 2016) while another is collecting exposure data for risk modelling software in Indonesia (HotOSM, 2016). At the same time there is not enough transparency about how the coordination of efforts works in OSM. It is not always clear who the users are and if there is a potential way to improve the consistency and accuracy of the data (Palen et al., 2014). More detailed evaluation may be required to identify potential areas of improvement and application of OSM and similar projects.

Another type of data that has been used in big data hazard mitigation projects is wireless details. Wireless call records combined with phone metadata can be used to assess the level of disaster preparedness and identify exposures. Call detail records (CDRs) are metadata related to wireless calls with details about the number of calls performed between towers and the amount of airtime used. Collected in developing countries, where cell phone penetration rate is over 90% (Parkes, 2013), such data can provide detailed and dynamic information related to population. Several researches have been conducted where CDRs were used to obtain information about the migration of people (Bengtsson et al., 2011), population density and the dynamics in Europe, Haiti and New York (Deville et al., 2014).

Data extracted from phone records can be useful not just for building exposure maps or designing evacuation routes but also for tasks like evaluating the risk of disease after a disaster (Wesolowski et al., 2012), adjusting exposure maps according to the patterns of seasonal migration (Pentland, 2012), determining the economic status of certain communities to identify groups of people missed by official surveys. Projects have been conducted in Rwanda (Blumenstock, 2014), Latin America (Frias-Martinez et. al., 2012) and UK (Eagle, 2010)).

Research using wireless data is still an emerging area and at this point some applications have been well tested while others are less so. One of the promising applications is the ability to use CDRs to assess social network features (Aldrich et al., 2014). For example, during the earthquake in Haiti in 2010, phone records combined with social networks were used to connect refuges with their family members in other areas of the country and provide assistance exactly where it was needed (Eagle et al. 2009, Onnela et al. 2007). Phone records are a promising area for research, but compared to social media data they do contain less of the information required for vulnerability management. For example, they do not routinely incorporate education, age, health or access to water or other infrastructure sources. As a result, they have a lower potential for practical applications than other social sensors.

1.2.3 Disaster response

As a part of International Decade for Natural Disaster Reduction (1990–1999) and additional efforts, several organizations have been created to collect, analyze and share data used for effective natural disaster response. While these organizations are proven to be effective (UN/ISDR, 2015), several challenges remain to be addressed. Warning delays need to be reduced, especially during floods and storms. In addition, telecommunication systems in developing countries require an upgrade.

The success of big data applications to the field of natural hazard response depends on its ability to improve situational awareness and short-term impact assessment (Letouze et al., 2013).

1.2.3.1 Situational awareness

Situational awareness is defined as "all knowledge that is accessible and can be integrated into a coherent picture, when required, to assess and cope with a situation" (Sarter and Woods, 1991). As mentioned above, disaster events tend to correlate with activity in social media and as a result can contribute to the enhancement of situational awareness. In one example, the USGS uses Twitter not only for earthquake detection worldwide, but also for alert posts (Smyrl et al. 2011). The Billion Prices Project developed by the Massachusetts Institute of Technology (MIT) collects information about prices posted online to monitor food security and detect and post inflation patterns (BBT, 2016). The public sector uses disease information obtained from digital sources to create alerts about possible outbreaks (WHO, 2016). To take advantage of existing platforms during emergency more of the following is required: capacity to be able to create tools and platforms, proper information delivery mechanisms, and adequate guidelines and procedures to provide required information to all parties.

1.2.3.2 Short-term effect assessment

Full and timely assessment of the immediate impact of a disaster heavily relies on multiple data sources available for analysis. Remote sensing is one of the main sources today for this kind of assessment. Its main functions are better situational awareness and

proper damage mapping with large coverage, damage assessment for the most critical properties, such as office buildings and homes, and the evaluation of impact on critical infrastructure such as water pipes, energy grids, and roads (Stow et al., 2015).

One of the projects that has shown promising results was the damage assessment with laser scanning devices capable of creating 3D images of an area (Qiu et al., 2014). Satellite images are being used more now for evaluating damage caused by hurricanes in tropical forests (Rossi et al., 2009) and tsunamis (Cervone et al., 2013). They provide high accuracy data for damage assessment after earthquakes, such as the ones that occurred in Japan on 11 March 2011 (Yamazaki et al., 2013) and in Haiti on 12 January 2010 (Pham et al., 2015). Also, new automated image processing solutions are being developed (Polli et al., 2013; Nex et al., 2014). At the same time, social sensors such are being used increasingly to improve the accuracy and details of the disaster impact. Crowd sourcing platforms in social media c allow for the implementation of algorithms to speed up the processing of images received from satellites and drones by breaking it into smaller chunks and involving multiple contributors for simultaneous processing (Meier, 2014).

During the wildfires in Russia (2010), digital activists used social networks to coordinate the efforts of volunteer firefighters (Asmolov, 2014). Big data, together with social media, was able to direct the efforts of decentralized groups and improve interactions between them. Such technologies can help to synchronize the efforts of organizations involved in disaster response. Thus health professionals providing aid during the 2010 earthquake in Haiti successfully coordinated their efforts using Twitter. The examples above demonstrate the potential applications and usefulness for similar platforms and technologies. An overview of existing Big Data analysis approaches is provided in the following section.

1.3 Big Data analysis approaches

Big data involves combining data crumbs with advanced data processing tools to perform the following types of analysis:

- Descriptive analysis – qualitative description of changes or concerns such as hurricane damage assessment or flood detection on early stages.

- Predictive analysis – statements and conclusions about non-observable or hard-to-measure entities. For example, information about migration and interaction patterns between people during disasters can be obtained from the analysis of wireless calls, cell phone relocation and recharge patterns. This type of analysis is more focused on what will happen in the future and can be used for tasks such as more detailed weather forecasts or the prediction of sudden disasters.

- Prescriptive analysis – predictive analysis considers possible future situations by examining the most probable pathways. Once the list of scenarios has been identified each can be assessed further with predictive analysis.

- Discursive analysis – disaster mitigation with the help of the third C (above), communities. This is achieved by creating additional awareness and preparedness while providing early warning and real-time feedback (Data-Pop Alliance, 2015).

For all of the types of analysis listed above, stream computing techniques can be used to decrease processing velocities without relying on the need to increase storage volume. Below, the theoretical overview of the streaming processing concept is discussed.

1.3.1 Big Data Streaming Processing Concept Overview

Streaming computing is a relatively new approach to manage the volume, velocity, variety and variability of big data. A data stream is defined as a sequence of values that are normally both infinite and change with time. Simple examples are heart rate measurements or stock market tickers. More complex events or patterns also can be represented as a single or multiple data streams. The processing of data streams involves querying, filtering and transforming of values arriving from the producers, and feeding the relevant data, formatted properly, to end users. The high-performance clusters can be used to process high-dimensional streams from multiple sensors. In many cases, sequential implementations of even single-pass algorithms cannot keep up with the speed

of real-world streams due to the high cost of similarity computation. In this cases real-time stream clustering implemented through parallelization (Cugola and Margara, 2011).

The main goal of data stream processing is to observe the evolution of the data and analysis products with time. In many cases, more recent data is of higher importance than the older one. Streaming applications generally focus on live data while historical values are often aggregated and archived in case they are needed later for additional analysis. Thus, time becomes a key factor in data analysis from streams. Clearly, this requirement makes utilization of traditional database systems quite difficult as they deal mostly with static data, and are designed to query and retrieve the values as they were at a particular point of time.

Another common requirement of data streams processing is real-time or near real-time evaluation of fresh data and uninterrupted delivery of results. Data is evaluated continuously, a principle which again differs with well-known models of relational databases that store static data (Babcock et al.,2002).

At the same time, quite often, future data rates of new values being produced by the stream sources are unknown and, in most cases, the processing application does not control the arrival of data. Thus, the processor needs to constantly monitor the arrival of new values and must be ready to compare them to the data acquired and processed earlier (Babu and Widom, 2001).

The infinite nature of streams and the requirement for continuous and often real-time evaluation of fresh data, as well as the other requirements discussed earlier, demand new approaches to data processing. These approaches are expected to be different from the models used in conventional data management systems and have resulted in the new field of streaming data research. Below is a brief overview of the key aspects of the existing stream processing models.

Earlier models proposed for stream processing emerged from the field of relational databases, and based on relations between old and new values. In some cases, these models were limited to append-only updates, often compared to Tapestry, an

experimental mail system developed at the Xerox Palo Alto Research Center (Terry et al., 1992). Data streaming was seen as a sequence of data tuples continuously arriving at a given order. At a certain point, the Tapestry model was a convenient approach as it could be considered more as an extension to existing systems, and it allowed the field to maintain its existing approaches, query operators and languages intact.

One of the popular models views a stream as a continuously arriving unbounded sequence of tuples, being appended to existing data with a timestamp or another time-based index (Golab and Özsu, 2003). The timestamp indicates the time of arrival or observation and implies the time when the tuple was produced as well as its sequential order. Each tuple may include additional attributes according to various requirements and defined data structure.

This model allows more than one tuple for each given timestamp. Examples of the systems based on this approach are STREAM (Arasu et al., 2007), TelegraphCQ (Chandrasekaran et al., 2003) and Aurora (Abadi et al., 2003). The STREAM system provided the formal definition of the model (Arasu et al.,2007), where a stream is a set of tuple and timestamp pairs. Tuples can be simple, with only one data type object, or more complex. In that case Data Definition Language (DDL) (Sullivan and Heybey, 1998) or COUGAR, with more broad data types (Bonnet et al., 2001), may be used to define a tuple. Some implementations use XML to represent each tuple instead of storing them in the form of relational table (Chen et al., 2000). In the general case, every data stream format is independent of the existing data model.

The main goal of the timestamp is to establish the sequence of tuples. Their order allows the determination of which tuples have already been processed. Thus, the timestamp does not always represent a measured time and can be a simple integer value that is increased for each new tuple. At the same time, tuples are not guaranteed to arrive in correct order, even with timestamps (Golab and Özsu, 2003) because these are being applied after the arrival. In such cases, post-processing may be required without strictly following the existing timestamps.

One of the products developed for streaming computations is InfoSphere Streams announced by IBM in April 2009. This product is designed for streamed data analysis and is capable of processing large amounts of data at high speed. It is able to perform advanced analysis and, as a result, can be used to aid in decision making processes related to the actual acquisition of the data itself or the allocation of critical resources. As a result, as more data is collected better intelligence is provided in real-time or near real-time.

1.3.2 InfoSphere Streams Platform

Streams is a development platform and a runtime environment which can be used to create applications capable of processing various multiple streams, filter the data and correlate large volumes of continuous values according to predefined rules that have been programmed in advanced. In addition, alerts also can be defined to take actions within a defined time frame.

InfoSphere Streams can process data arriving from a variety of sources such as video cameras, sensors, stock tickers or other data feeds, including traditional databases. Each input stream must be defined within the application. Expected input may be text, number or different non-relational data such as video, audio, radar or sonar inputs. Subsequently, input format operators can be defined which will perform the required actions on the data (such as filtering, various transformations and applying functions).

Developers can create applications in the InfoSphere Streams Studio using IBM Streams Processing Language (earlier known as Streams Processing Application Declarative Engine), a declarative (non-procedural) language that was customized for the stream processing application. Programs ready for execution must be deployed to the Streams Runtime environment. Streams Live Graph then can be used to monitor cluster performance, including each virtual machine and communications between them (Ebbers et al., 2013).

Streams Processing Language (SPL) used in InfoSphere Streams, is a distributed data-flow composition programming language. It is a full-featured language similar to C++ or

Java™. SPL is extensible, meaning that it is possible to add new keywords, concepts, and structures to the source language, and supports user-defined data types. Developers can create custom functions in SPL or embed a code in C++ or Java, including user-defined operators programmed earlier and elsewhere.

Operators in InfoSphere Streams can be represented as vertices of a directed graph. They interact with each other and perform actions on one or multiple data streams which can be external or internal, produced by other parts of the application. Each SPL program consists of the following basic building blocks:

- Stream — a sequence of tuples, usually structured and infinite. In most cases it is an input to operators and can be consumed through the definition of a window or sequentially tuple by tuple.
- Tuple — a set of values together with their type definition. Tuple type is usually defined by its stream structure.
- Stream type — name and data type of each value in a tuple.
- Window — a group of tuples limited by specified count number, time frame or a set of attributes.
- Operator — a function that takes stream data, processes it and creates an output stream. It's the main building block of SPL.
- Processing element (PE) — a single execution block of an SPL program. It usually includes one or more operators.
- Job — Streams application which can be executed. It usually consists of one or more PEs.

When a program written on SPL is compiled, an Application Description Language (ADL) file is generated. This file contains a description of the application structure, including details about each PE, such as binary files that must be executed, data formats and execution restrictions, among others. There also are several pre-defined operators

defined in SPL. Each one has its own logic that requires a set of parameters and a list of input and output streams. Thus the input and output streams definition is a part of each operator invocation. Below is the list of operators available in InfoSphere Streams:

- Source — consumes streamed data as an input.
- Sink — sends output data as a stream to external storage or other consumers.
- Functor — changes arriving stream data by filtering it, performing various transformations and applying functions.
- Sort — arranges arriving stream data according to defined conditions.
- Split — transforms input stream into two or more output streams.
- Join — joins streams according to defined conditions.
- Aggregate — aggregates streams data according to defined conditions.
- Barrier — combines and coordinates data arriving in a stream.
- Delay — pauses data flow.
- Punctor — used to specify data values that should be processed together.

A port is a point where an operator is applied to a stream. Different operators have different numbers of input and output ports. A functor, for example, has one input port and one output port. A source has no input ports and a sink has no output ports. A split and a join usually use multiple input and output ports (Sakr, 2013).

Research under this thesis employs Streams for rapid, high-resolution seismic hazard estimation (Chapter 2), sensitivity analysis of hazard maps (Chapter 3) and joint data processing of physical and social sensors (Chapter 4, 5). In this work seismometers placed in areas likely to be shaken (strong motion recorders) are refered as physical sensors. Strong motion instruments record those earthquake ground motions capable of damaging buildings or of weakening soils. In order to better understand the associated

research, the next section contains a brief explanation of the strong motion data processing workflow.

1.4 Strong Motion Processing Steps

The steps below have originally been developed as a part of the Caltech EERL project and improved by Trifunac and Lee (1978). These steps were described in greater detail by Shakal et al. (2003).

1. Baseline Correction. The goal of this steps is to estimate the raw data zero-mean. For that purpose existing data points are interpolated to equal-interval sampling. When handling large numbers of records, the most appropriate approach is simple baseline correction using a constant or linear trend. In certain cases, additional corrections may be performed on some records. If a complex baseline correction is required it can be performed with long-period filtering during Step 4, below. The output of Step 1 is called Volume (or Phase) and is still considered to be raw data.

2. Instrument Correction. The values adjusted to baseline in the previous step are corrected for instrument response using a simple finite-difference operator. In the case of frequency-domain processing the process of dividing existing data spectrum by the instrument response spectrum replaces the finite-difference operator.

3. High Frequency Filtering. High frequency noise is removed by applying a specific type of filter, depending on the processing approach. Caltech/USC uses an Ormsby filter with corner frequency of 23 Hz and a termination frequency of 25 Hz. CSMIP uses a Butterworth filter with corner frequency around 80% of the final sampling rate and a 3rd or 4th order decay. Then the resulting data is decimated to adjust the sample rate to the distribution frequency. Again, Caltech/USC uses 50 points/second while CSMIP is at 100 points/second.

   To improve the high-frequency accuracy in the case of time-domain processing, the instrument correction is performed prior to this decimation, rather than after (Shakal and Ragsdale, 1984).

4. Initial Long Period Filtering. Values with periods longer than 15 seconds are removed from decimated data. Recent research shows that keeping longer periods for larger earthquakes may be appropriate and Caltech/USC traditionally uses higher thresholds. Velocity and displacement are obtained by integrating acceleration values and filtering them with the same low-frequency filter.

5. Maximum-Bandwidth Response Spectrum estimation. Pseudo-velocity (PSV) response spectra for 0, 2, 5, 10, and 20% of critical value are calculated using the method of Nigam and Jennings (1969). The resulting spectra are plotted for 0.04 to 15 seconds (the full bandwidth) so that the best filter can be chosen based upon comparison.

6. Long-Period Filter Selection. This is the most difficult processing step. The maximum period with useful data is determined as the intersection of the maximum-bandwidth spectrum and the average noise spectrum determined for the system. The long period minimum value was obtained in Step 5. A number of long-period filters with values between minimum and maximum long period are applied to the data from Step 3. Filter periods that give signal-to-noise ratio (SNR) higher than 2 are the most effective in reducing noise. The final value for the filter is chosen after comparing displacement plots to each other as well as to data from nearby stations. In case of CSMIP, the choice is made by a team of experts after analysis of available information and discussion.

7. Finally acceleration, velocity and displacement values from the previous step are saved as files for distribution. The values may be plotted with tripartite logarithmic scaling or the linear scaling depending on the requirements and application.

1.5 Outline

This chapter provides a brief overview of existing big data applications in the natural hazard field such as hazard monitoring, mapping and disaster response, and provides definitions for the streaming paradigm necessary for understanding the primary concepts of the work.

In the second chapter a detailed explanation of the pipelining implementation of the existing EqHaz program suite (Assatourians & Atkinson, 2013) for production of high resolution seismic hazard maps is provided based on IBM InfoSphere Streams. The results of this work have been published in the Procedia Computer Science at the International Conference on Computational Science – Computational Science at the Gates of Nature. During the hazard mapping pipelining implementation, two main processing operators, the catalog generation operator using Monte Carlo simulations and the map generation operator, were identified as bottlenecks. As a result, the workload of the hazard mapping workflow has been decomposed and these two operators have been split into multiple pipelines for parallel execution. Implementation on the experimental environment of a cluster of four machines, each with dual Xeon quad-core 2.4GHz CPU, 16GB RAM and running Linux, resulted in mean hazard calculations and mapping procedures with a resolution up to 2,500,000 points and a near-real-time processing time of approximately 5-6 minutes. The proposed approach has a significant potential for both the applications of emergency hazard policy makers and providers, but also for the scientific community as a tool for hazard maps sensitivity testing, because it can simplify the model selection process and provide insights as to which input factors and parameters significantly affect the results and are the main sources of uncertainty.

In the third chapter, several sensitivity tests were performed for the ground motion prediction models proposed by Atkinson and Adams (2013) and used as input to the most recent generation of Canadian seismic hazard maps released in 2015 and the basis for updated seismic provisions of National Building Code of Canada (NBCC) 2015. The results are published in the conference proceedings of 2016 Canadian Society for Civil Engineering (CSCE) Annual Conference. The primary outcomes are a series of mean hazard maps at predicted pseudo acceleration (PSA) for a 2475 year return period at three different periods of 0.2, 1.0 and 2.0 seconds. These are calculated for three sensitivity tests for both the absolute difference and amplification ratio for comparison to each base model.

Chapter 4 presents an approach for the estimation of real-time earthquake intensities using streaming data analysis from social and physical sensors. The results are accepted,

with minor revisions, to the journal of Pure and Applied Geophysics. In this work, Twitter is defined as potentially significant data source for post-disaster response systems. Ground motion and Twitter datasets for South Napa earthquake used in this study to select and validate the predictive relationships between Modified Mercalli Intensity (MMI) and tweets rate in the ten minutes following an earthquake. Regression analysis of intensity versus the logarithmic mean of the number of tweets per minute identifies four potential predictive equations (linear, two-segment linear, three-segment linear and exponential) within a legitimate range of values for each model. In addition, the implementation of two streaming applications for processing of data from physical and social sensors are presented. Moreover, both applications were linked into one application for joint processing from both types of sources. Intensity maps for the Napa earthquake are obtained from each application and evaluated for one minute intervals after the earthquake.

In Chapter 5, the validation and calibration of the predictive relationship obtained in the previous chapter is performed for California, Japan and Chile. These results have been submitted to Seismological Research Letters. Data sets used in the validation and calibration processes for each region and challenges associated with social sensors data processing are presented in the paper. Results from the statistical validation of models with and an analysis of the residuals are provided, including whisker diagrams for the residuals for every model and region and residuals plots versus existing and potential predictors. In addition, the calibration procedure is explained as it relates to the most important output of this work – spatially calibrated relationships between intensity and tweet rates for the California, Japan and Chile regions. The results clearly show the importance of complementary data from social sensors and techniques for real-time processing for emergency management and response applications.

1.6 References

Abadi, D. J., Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S., . . . Zdonik, S. (2003). Aurora: a new model and architecture for data stream management. *The VLDB*, 12(2):120–139.

Aldrich, D., & Meyer, M. (2014). Social capital and community resilience. *American Behavioral Scientist.*

Alexander, D. (1997). The Study of Natural Disasters 1977–1997: Some Reflections on a Changing Field of Knowledge. In Disasters (pp. 284-304). *Blackwell Publishers Ltd.*

Arasu, A., Babcock, B., Babu, S., Cieslewicz, J., Datar, M., Ito, K., . . . Widom, J. (2007). STREAM: The Stanford data stream management system. (M. Garofalakis, J. Gehrke, & R. Rastogi, Eds.) *Data Stream Management: Processing High-Speed Data Streams.*

Asmolov, G. (2014). Crowdsourcing as an Activity System: Online Platforms as Mediating Artifacts A Conceptual Framework for the Comparative Analysis of Crowdsourcing in Emergencies. *CEUR Workshop Proceedings*, 1148.

Assatourians, K. & Atkinson, G. M., 2013. EqHaz: An open-source probabilistic seismic-hazard code based on the Monte Carlo simulation approach.. *Seismol. Res. Lett*., 84(3), pp. 516-524.

Atkinson, G. M. & Adams, J., 2013. Ground motion prediction equations for application to the 2015 national seismic hazard maps of Canada. *Can. J. Civil Eng*., Volume 40, pp. 988-998.

Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and issues in data stream. *21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*., 1–16.

Babu, S., & Widom, J. (2001). Continuous queries over data streams. *ACM SIGMOD Record*, 30(3):109–120.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & von Schreeb, J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. *PLoS Med*, 8.

Blumenstock, J. (2014). Calling for better measurement: Estimating an individual's wealth and well-being from mobile phone transaction records. *Proceedings of Knowledge Discovery in Data.*

Bonnet, P., Gehrke, J., & Seshadri, P. (2001). Towards sensor database systems. *2nd International Conference on Mobile Data Management MDM* 2001, 3–14.

Cervone, G., & Manca, G. (2011). Damage Assessment of the 2011 Japanese Tsunami Using High-Resolution Satellite Data. *The International Journal for Geographic Information and Geovisualization*, 200-203.

Chandrasekaran, S., Cooper, O., Deshpande, A., Franklin, M. J., Hellerstein, J. M., Hong, W., . . . Shah, M. A. (2003). TelegraphCQ: continuous dataflow processing. *ACM SIGMOD International Conference on Management of Data*, 668–668.

Chen, J., DeWitt, D. J., Tian, F., & Wang, Y. (2000). NiagaraCQ: a scalable continuous query system for. *ACM SIGMOD Record*, 29(2):379–390.

Cimellaro, G. P. (2010). Framework for analytical quantification of disaster resilience. *Engineering Structures*, 3639-3649.

Costello, A., Abbas , M., Allen , A., Ball , S., Bell, S., Bellamy, R., . . . Patterson, C. (2009). Managing the health effects of climate change: *Lancet and University College London Institute for Global Health Commission. Lancet.*

Cugola, G., & Margara, A. (2011). Processing flows of information: From data stream to complex event. *ACM Computing Surveys*, 44(3):15:1–15:62.

Data-Pop Alliance. (2015). Big Data for Resilience: Realising the Benefits for Developing Countries. *Paris: Data-Pop Alliance.*

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., . . . Tatem, A. J. (2014, Nov 11). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45), 15888–15893.

Eagle, N., Macy, M., & Claxton, R. (2010). Network Diversity and Economic Development. *Science*.

Eagle, N., Pentland, A., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *PNAS*, 15274–15278.

Earle, P., Bowden, D., & Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of GeoPhysics*, 708–715.

Ebbers, M., Ghisloti de Souza, R., Lima, M. C., McCullagh, P., Nobles, M., VanStee, D., & Waters, B. (2013). Implementing IBM InfoSphere BigInsights on IBM System X. Second Edition. *International Business Machines Corporation*.

Entekhabi, D. N. (2010). The soil moisture active passive (SMAP) mission. *P. IEEE*, 704–716.

Frias-Martinez, Vanessa, & Virsesa. (2012). On the relationship between socioeconomic factors and cell phone usage. *International Conference on Information and Communication Technologies and Development*.

Golab, L., & Özsu, M. T. (2003). Processing sliding window multi-joins in continuous queries over data. 29th international conference on Very large data bases VLDB 2003, *VLDB '03*, 500–511.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning. B, *Planning & design*, 682.

Herold, R. (2016). 10 Big Data Analytics Privacy Problems. Retrieved from Secureworld: https://www.secureworldexpo.com/10-big-data-analytics-privacy-problems

Hinrichsen, Don. Coastal Waters of the World: Trends, Threats, and Strategies. *Washington D.C. Island Press*, 1998.

Houghton, J. Y. (2001). Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge, United Kingdom and New York, NY: Cambridge University Press.*

Hu, H., Wen, Y., Chua, T.S. and Li, X. (2014) Toward Scalable Systems for Big Data Analytics*: A Technology Tutorial. IEEE Access*, 2, 652-687.

IDNDR Secretariat. (1999). Proceedings of the International Decade for Natural Disaster Reduction Forum. *Programme Forum 1999.* Geneva: IDNDR.

Jongman, B., Winsemius, H., & Jeroen, C. (n.d.). Declining vulnerability to river floods and the global benefits of adaptation. *Proceedings of the National Academy of Sciences.*

Klotz, A. C., Hmieleski, K. M., Bradley, B. H., & Busenitz, L. W. (2014). New venture teams a review of the literature and roadmap for future research. *Journal of Management*, 40(1), 226-255.

Kong, Q. A.-W. (2016). MyShake: A smartphone seismic network for earthquake early warning and beyond. *Science Advances*, 1-8.

Laney, D. 3D data management controlling data volume, velocity and variety. 3D data management controlling data volume, velocity and variety, *Application Delivery Strategies*, File 949, 2001.

Letouzé, E., Meier, P., & Vinck, P. (2013). Big data for conflict prevention: New oil and old fires. New York: *Technology and the Prevention of Violence and Conflict, International Peace Institute.*

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H. Big data: The Next Frontier for Innovation, Competition, and Productivity. *San Francisco, CA, USA: McKinsey Global Institute*, 1-137, 2011.

Meier. (2014). Crowd Computing Satellite & Aerial Imagery. *Digital Humanitarians*.

Missing Maps. (2015). Retrieved from Missing Maps project website: http://www.missingmaps.org/

Musaev, A., Wang, D., & Pu, C. (2014). LITMUS: Landslide detection by integrating multiple sources. *The 11th International Conference on Information Systems for Crisis Response and Management.*

Nex, F. E. (2014). Automated processing of high resolution airborne images for earthquake damage assessment. *SPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 315-321.

Nigam, N. C., & Jennings, P. C. (1969). Calculation of Response Spectra from Strong-Motion Earthquake Records. *Bull. Seism. Soc. Amer*.(59), 909-922.

Nirupama, & Simonovic, S. (2002). Role of Remote Sensing in Disaster Management. *London, Ontario, Canada: Water Resources Research Report.*

Occhipinti, G. A. (2008). Tsunami detection by GPS: how ionospheric observation might improve the Global Warning System. *GPS World*, 50–56.

Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., . . . Barabási , A. L. (2007, January 27). Structure and tie strengths in mobile communication networks. (H. E. Stanley, Ed.) *Proceedings of the National Academy of Sciences of the United States of America*, 104(18), 7332–7336.

Parkes, S. (2013). ITU releases latest global technology development figures. Retrieved from ITU: http://www.itu.int/net/pressoffice/press_releases/2013/05.aspx#.V42D76K6O7K

Pesaresi, M. H. (2013). A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results. Selected Topics in Applied Earth Observations and Remote Sensing, *IEEE Journal* of 6(5), 2102–2131.

Pesaresi, M., & Freire, S. (2014). Producing a Global Reference Layer of Built-Up by Integrating Population and Remote Sensing Data. *Ispra, Italy: European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen.*

Pham, T.-T.-H., Apparicio, P., Gomez, C., Weber, C., & Mathon, D. (2014). Towards a rapid automatic detection of building damage using remote sensing for disaster management: The 2010 Haiti earthquake. *Disaster prevention and management*, 53-66.

Polli, D., Dell'Acqua, F., & Candela, L. (2013). Mapping Earthquake Damage from Post-Event only VHR SAR Texture Maps: Zooming into Poor Estimation Cases. *Remote Sensing and Geoinformation not only for Scientific Cooperation*.

Poushter J., S. R. (2016). Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies. *Pew Research Center*.

Tiampo, K.F., McGinnis, S., Kropivnitskaya, Y., Qin, J., Bauer, M.A. Big data challenges and hazards modeling, invited chapter, *Insurance Catastrophe Risk Modeling*, ed. G. Michel, submitted with minor revisions, 2016

Qiu, S., Cao, C., Zhang, B., Xi , S., Wang , X., Yan , B., & Chuanrong , L. (2014). Feasibility study of remote sensing using structured light for 3D damage assessments after natural disasters. *I. S. Photonics, Ed. SPIE Asia Pacific*, 92632R-92632R.

Rossi, E., Rogan, J., & Schneider, L. (2012). Mapping forest damage in northern Nicaragua after Hurricane Felix (2007) using MODIS enhanced vegetation index data. *GIScience & Remote Sensing*, 385-399.

Sakr, S. (2013). An introduction to InfoSphere Streams. Retrieved from www.ibm.com: http://www.ibm.com/developerworks/library/bd-streamsintro/bd-streamsintro-pdf.pdf

Sarter, N. B., & Woods, D. D. (1991). Situation Awareness: A critical but ill-defined phenomenon. *International Journal of Aviation Psychology*(1), 45-57.

Shakal, A. F., & Ragsdale, J. T. (1984). Acceleration, Velocity and Displacement Noise Analysis of the CSMIP Accelerogram Digitization System. *8th World Conference Earthquake Engineering*, 2, 111-118.

Shakal, A. F., Huang, M. J., & Graizer, V. (2003). Strong-Motion Data Processing (Vol. B). (W. H. Lee, H. Kanamori, P. C. Jennings, & C. Kisslinger, Eds.) *Amsterdam: Academic Press*.

Smyrl, L., Kern, T., & Allen, J. (2011). USGS Twitter Earthquake Dispatch. *@USGSted*, 1.

Soden, R., & Palen, L. (2014, May 27-30). From Crowdsourced Mapping to Community Mapping: The Post-Earthquake Work of OpenStreetMap Haiti. *COOP 2014-Proceedings of the 11th International Conference on the Design of Cooperative Systems*, 311-326.

Stow, D., Lippitt, C., & Lloyd , L. (2015). Time-Sensitive Remote Sensing Systems for Post-Hazard Damage Assessment. *Time-Sensitive Remote Sensing*, 13-28.

Sullivan, M., & Heybey, A. (1998). Tribeca: a system for managing large databases of network traffic. *USENIX Annual Technical Conference ATEC'98*, 2–2.

Tapley, B., Bettadpur, M., Watkins, C., & Reigberg, C. (2004). The Gravity Recovery and Climate Experiment: Mission overview and early results. *Geophys. Res. Lett.*

Taylor, C. J. (2001). Groundwater level monitoring and the importance of long-term water-level data. *Denver: US Geological Survey*.

Terry, D., Goldberg, D., Nichols, D., & Oki, B. (1992). Continuous queries over append-only databases. *SIGMOD '92,* 321–330.

The Billion Prices Project. (n.d.). Retrieved from The Billion Prices Project: http://bpp.mit.edu/

Tiampo, K.F., McGinnis, S., Kropivnitskaya, Y., Qin, J., Bauer, M.A. Big data challenges and hazards modeling, invited chapter, *Insurance Catastrophe Risk Modeling, ed. G. Michel*, submitted with minor revisions, 2016

Trifunac, M. D., & Lee, V. (1978). Uniformly Processed Strong Earthquake Ground Accelerations in the Western United States of America for the Period from 1933 to 1971: Corrected Acceleration, Velocity and Displacement Curves. *Report CE* 78-01.

UGMHOT, H. O. (2012). Evaluation of OpenstreetMap Data in Indonesia - A Final *Report. Departmentof Geodetic & Geomatics Engineering, Faculty of Engineering.*

United Nations/International Strategy for Disaster. (2004). *Living with Risk: A Global Review of Disaster Reduction Initiatives. Geneva: UN/ISDR*.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. J., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science* (338), 267-270.

World Health Organization. (2010, November 22). Global infectious disease surveillance. Retrieved from http://www.who.int: http://www.who.int/mediacentre/factsheets/fs200/en/

Yamazaki, F., Daiki, H., & Suzuki, K. (2013). Use of airborne optical and thermal imagery for the detection of building damage due to the 2012 Tsukuba tornado. *10th International Conference on Urban Earthquake Engineering*. Tokyo.

Chapter 2

2 A Pipelining Implementation for High Resolution Seismic Hazard Maps

The results covered in this chapter are published in conference proceedings of International Conference on Computational Science, ICCS 2015 – Computational Science at the Gates of Nature.

Seismic hazard maps are a significant input into emergency hazard management that play an important role in saving human lives and reducing the economic effects after earthquakes. Despite the fact that a number of software tools have been developed (McGuire, 1976, 1978; Bender & Perkins, 1982, 1987; Robinson, et al., 2005, 2006; Field, et al., 2003), map resolution is generally low, potentially leading to uncertainty in calculations of ground motion level and underestimation of the seismic hazard in a region. In order to generate higher resolution maps, the biggest challenge is to handle the significantly increased data processing workload.

In this study, a method for improving seismic hazard map resolution is presented that employs a pipelining implementation of the existing EqHaz program suite (Assatourians & Atkinson, 2013) based on IBM InfoSphere Streams – an advanced stream computing platform. Its architecture is specifically configured for continuous analysis of massive volumes of data at high speeds and low latency. Specifically, it treats processing workload as data streams. Processing procedures are implemented as operators that are connected to form processing pipelines. To handle large processing workload, these pipelines are flexible and scalable to be deployed and run in parallel on large-scale HPC clusters to meet application performance requirements. As a result, mean hazard calculations are possible for maps with resolution up to 2,500,000 points with near-real-time processing time of approximately 5-6 minutes.

2.1 Introduction

Today, probabilistic seismic hazard analysis (PSHA), based on the total probability theorem (Cornell, 1968; McGuire, 2004), is the most widely used method for estimating seismic hazard analysis. This is particularly true for hazard map calculations, where the probability of a ground motion acceleration value occurring at any given site is estimated by integrating conditional probabilities over all possible distance and magnitude values. There are a number of methodological considerations in PSHA mapping, and hazard map resolution is one of the most important of those technical issues today (Musson & Henni, 2001). In this context, resolution is defined as the spatial grid resolution that is used to produce map or grid density. A coarse grid results in loss of details for smaller source zones and, as a result, could underestimate the maximum ground motion level value (Musson & Henni, 2001). However, hazard computation with a finer grid requires considerably more computational resources, especially with a complex model. For instance, calculation for more than 6.5 trillion locations is necessary in order to obtain a hazard map for Eastern Canada with a resolution of one dot per meter. Additionally, if a Monte Carlo simulation approach is used for the PSHA calculation (Musson, 1999, 2000), where resolving power improves with the number of simulations used to generate synthetic data, then the processing workload is even heavier and also requires additional computational power. For example, 100,000 simulations of 100 years duration is equivalent to 10,000,000 years of seismicity data. However, significant technological breakthroughs in high performance computing (HPC) have taken place over the last few decades. For example, a cluster of computers can be connected together with advanced networks to provide increased computational power on demand. Complex problems can be partitioned into smaller tasks and programmed into pipelines and deployed on these computer clusters. Heavy workload can be split and distributed onto those pipelines and processed in parallel to meet application performance needs.

IBM InfoSphere Streams (Streams), an advanced stream computing platform, is designed specifically for parallelism and deployment across computing clusters for continuous analysis of massive volumes of data with high speeds and low latency. In particular, Streams treats processing workload as data streams. Processing procedures are

implemented as operators that are connected together to form processing pipelines. To handle large processing workload, these pipelines are flexible and scalable. Here, PSHA mapping programs are designed to be deployed and run in parallel on large-scale HPC clusters. The result is more efficient real-time seismic hazard analysis through distributed computing networks.

The main goal of this work is to utilize innovative computational algorithms for PSHA mapping by integrating different input data sources and existing processing tools into a streaming and pipelined computing application. A set of high-resolution maps for different frequencies calculated in terms of the probability of exceeding a certain ground motion level across Eastern Canada are obtained. The motivation for this study is not the estimation of seismic hazard in Eastern Canada, but is to demonstrate that the use of the pipelining and streaming techniques provided by Streams that makes possible the production of high-resolution hazard maps in near-real time with no limitations on resolution. This approach could be used in any region in the world where seismic sources and ground motion characteristics are available. The results can also be used as input for sensitivity analysis of hazard maps for any input models, something that has not been done before, largely because of difficulties with computational implementation.

2.2 Monte Carlo Approach for PSHA Mapping

One of the straightforward and flexible PSHA methodologies involves the use of Monte Carlo simulations (Musson, 1999). Here seismic source models representing the spatial and temporal historical earthquakes occurrence in the region are used as the first stage in Monte Carlo simulations for the generation of a synthetic earthquake catalogue. As a result, the synthetic catalogue shows a set of probable earthquakes in the region over a long time period, on the order of 100 years. In the next stage, this synthetic catalogue is used to estimate a distribution of ground motions at a number of sites using selected ground motion prediction equations (GMPEs) and/or attenuation parameters (Musson, 1999). Finally, probabilities of exceeding a certain ground motion level at every point across a region are calculated. The maps take into account uncertainties in the earthquake location, size and the resulting ground motions that can affect region.

Today there is a number of both free and commercial software tools available to perform PSHA, including those based on the Monte Carlo simulation approach (EQRISK (McGuire, 1976); FRISK (McGuire, 1978); SeisRisk (Bender & Perkins, 1982, 1987); Fortran codes produced by National Seismic Hazard Mapping Project (NHSMP) from the U.S. Geological Survey (USGS); CRISIS (Ordaz, et al., 2013); EQRM (Robinson, et al., 2005, 2006); OpenSHA (Field, et al., 2003)). The Monte Carlo PSHA tool implemented here is the EqHaz software suite of open-source FORTRAN programs developed by Assatourians and Atkinson (2013). This suite consists of three programs. EqHaz1 creates the synthetic earthquake catalogues generated by the user-specified seismicity parameters. EqHaz2 produces the ground motion catalogues at a site, mean hazard probability curves and mean hazard motions at specified return periods calculated for a site or grid of points. EqHaz3 de-aggregates the hazard, producing the relative contributions of the different earthquake sources as a function of distance and magnitude. The EqHaz suite has a number of limitations which mean that it cannot be applied to hazard calculations with more than 1,000,000 records in each synthetic catalog and for more than 100,000 grid points. These limits are barriers on the real-time production of high-resolution hazard maps.

2.3 Pipelining Implementation in Streams

IBM has been actively involved in streaming concept development, resulting in the IBM InfoSphere Streams product (Streams) (IBM, 2014). Streams is a software package that provides a runtime platform, programming model, and toolkits. Stream Processing Language (SPL) is the programming language used in Streams for building applications in stream processing – a data-centric programming model.

The building blocks of a SPL application are data streams and operators. Data streams consist of data packets (i.e. tuples) structured for user defined schema. Operators define the operations on data streams. Operators and streams are assembled together for each application as a data-flow graph which defines the connections among data sources (inputs), operators, and data sinks (outputs).

The deployment of a Streams application is supported by the underlying runtime system. The Streams runtime model consists of distributed processes. Single or multiple operators form a processing element (PE). The Streams compiler and runtime services make it easy determine where to best deploy those PEs, either on a single machine or across a cluster in order to meet the resource requirements (Bauer, et al., 2010).

The advantage of using these techniques is that it allows for the use of multiple computational units without the need to manage allocation, synchronization or communication among them. In addition, Streams Studio is an integrated development environment based on Eclipse with a Streams plug-in. The Streams programming interface has a set of toolkits of built-in operators (IBM, 2014) that can be used directly to speed up the development work. A subset of Streams built-in operators used in the current implementation is listed in Table 1**Error! Reference source not found.**, in order to better understand the implementation procedure.

In the implementation for PSHA mapping in Streams, the EqHaz1 and EqHaz2 Fortran source codes have been modified slightly and compiled into shared system object libraries and linked into Streams applications. Streams allows for the creation of new primitive operators in a more traditional native language, such as Java™ or C++ (IBM, 2014), that can directly call those PSHA procedures from the shared libraries in the streaming environment. In this application, two primitive operators have been implemented. One is the Catalog operator, which encapsulates EqHaz1 procedures for the catalogue generation. The other is the Map operator, which encapsulates EqHaz2 procedures for the map generation. The pipelined implementation is showed on Figure 1 captured from Streams Studio. In order to clearly display the entire pipeline it is divided into two stages in the figure. The connection between these stages is shown by circles filled with the same color. The first stage shows the process of input parameters gathering. The second one contains PEs performing PSHA procedures and outputting results. Algorithm 1 in Appendix A presents all the associated steps.

All models considered as input streams are located on the Internet at the site www.seismotoolbox.ca. In Algorithm 1, the developed application waits for a new user-

request for map generation or for an indication of any updates in the models such that the maps must be updated (Step 1).

Table 1: Built-in stream operators used in the implementation (IBM, 2014)

Name	Base explanation	Purpose in the implementation
InetSource	Periodically checks listed remote (Internet or intranet network) data sources, acquires data and generates a stream from them	To gain all input model parameters needed for analysis, including seismic zones model, ground motion prediction equations (GMPEs) and attenuation parameters
Custom	Receive and send any number of streams to perform user-specified analysis on stream	To normalize all input seismic sources and GMPE models from different sources to the same formats
Functor	Transforms input tuples into output ones and filter if needed	To filter user-defined parameters for two different stages: catalog generation and ground motion estimation
Barrier	Synchronizes tuples from several streams	To synchronize all input parameters before analysis begins
Split	Under a user specified condition splits a stream into several output streams	To partition the workload of Monte Carlo simulations of large synthetic catalogs and ground motion estimation on many sites
TCPSink	Writes data tuples to a TCP socket	Configured as a TCP server to sink outputs in JSON format

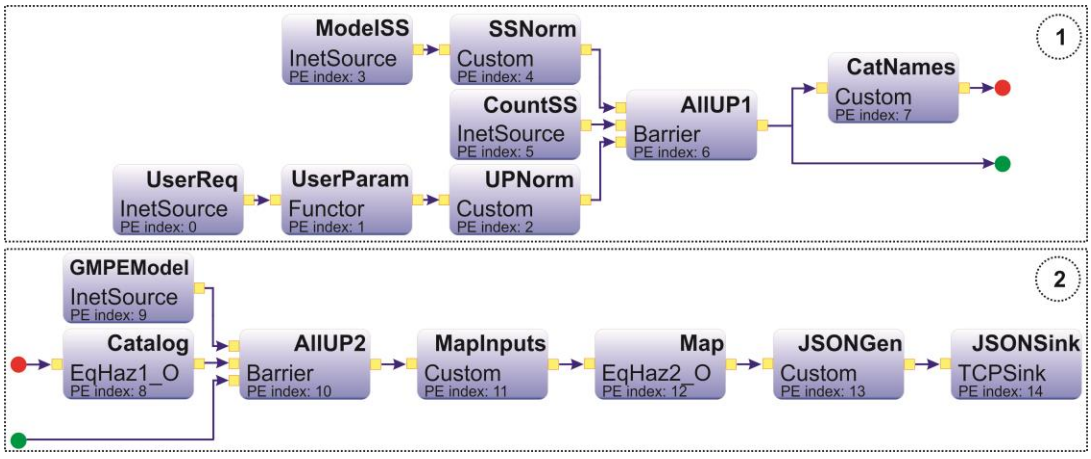

Figure 1: PSHA mapping pipelined application graph

As soon as this occurs, it generates an input stream for the synthetic catalog generation procedure (Step 2). Afterwards, the stream goes to the catalogue generation primitive operator developed from the EqHaz1 program (Step 3). Next, the generated synthetic catalog, GMPEs from F4 and other user parameters from F3 are combined to form an input stream for the maps generation module, based on the EqHaz2 program (Steps 4-6). Subsequently, the output stream is converted to JavaScript Object Notation (JSON) format, which is an open standard used to transmit data objects consisting of attribute-value pairs (Step 7). The JSON map file contains grid point coordinates represented by two attributes: "lat" for latitude and "lon" for longitude. The calculated value is represented by spectral acceleration (in our case pseudo acceleration (PSA)) for a given period of time at some level of probability (in this example, 2% in a 50 year return period) at each point of the grid. Spectral acceleration is a common measure of ground motion intensity in building engineering. It represents the maximum acceleration that a ground motion will cause in a linear oscillator for specified natural period and damping level. This is denoted as the "count" attribute. The total output with the set of map coordinates and values sinks to a TCP port (Step 8), where it is taken by a specifically designed JAVA script to be displayed on the webpage. The dynamic gmaps-heatmap.js JavaScript library is used for visualization purposes (Wied, 2014).

As introduced previously, the catalog generation operator has the greatest processing workload because it contains Monte Carlo simulations. The map generation operator also involves a heavy workload if the calculations are required on a dense grid. Therefore, either or both are the bottlenecks of the entire processing scheme for generating high resolution PSHA maps. To overcome EqHaz limitations and to reduce the application execution time, the workload has been decomposed and these two components have been split into multiple pipelines and the executions are parallelized. The complete application graph of the implementation is shown on Figure 2 and Figure 3.

The pipelined procedure is shown in Algorithm 2 in the Appendix A. As a result, each single operator executes on a single core. The whole process is divided into four stages. The first stage is exactly the same as that shown in Figure 1, because the process of getting input parameters remains the same as before. The second stage (Figure 2) shows

the decomposition and parallel implementation of Monte Carlo simulations for the synthetic catalogue. The third stage (Figure 3) presents the PEs performing spatial grid decomposition. Stage 4 of Figure 3 shows the final output. The connections between stages are shown with solid circles of the same color.

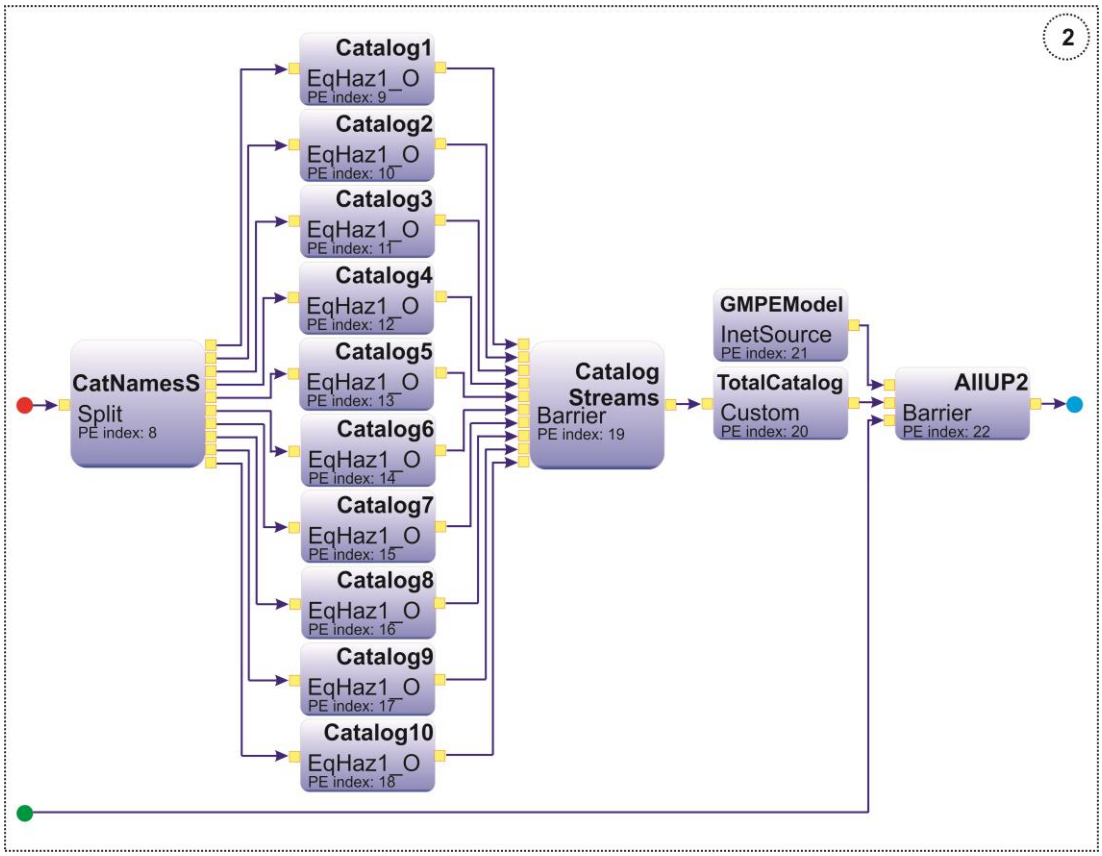

Figure 2: PSHA mapping pipelined with the workload splitting application graph (Stage 2)

That implementation provides an increase in the number of records in the synthetic catalog to as many as 10,000,000 records and the number of sites for each hazard map is stepped-up to 2,500,000 sites, allowing for the production of high resolution maps and significantly decreasing the execution time, as detailed below.

In Algorithm 2, the input stream for synthetic catalogue generation operator is divided into several streams in order to account for the EqHaz1 limit - the requested number of records in each catalogue has to be under 1,000,000 (Step 3). After the workload split is

performed in (Step 3), each feeds into the catalogue generators to be processed in parallel (Steps 4-13) in order to optimize execution.

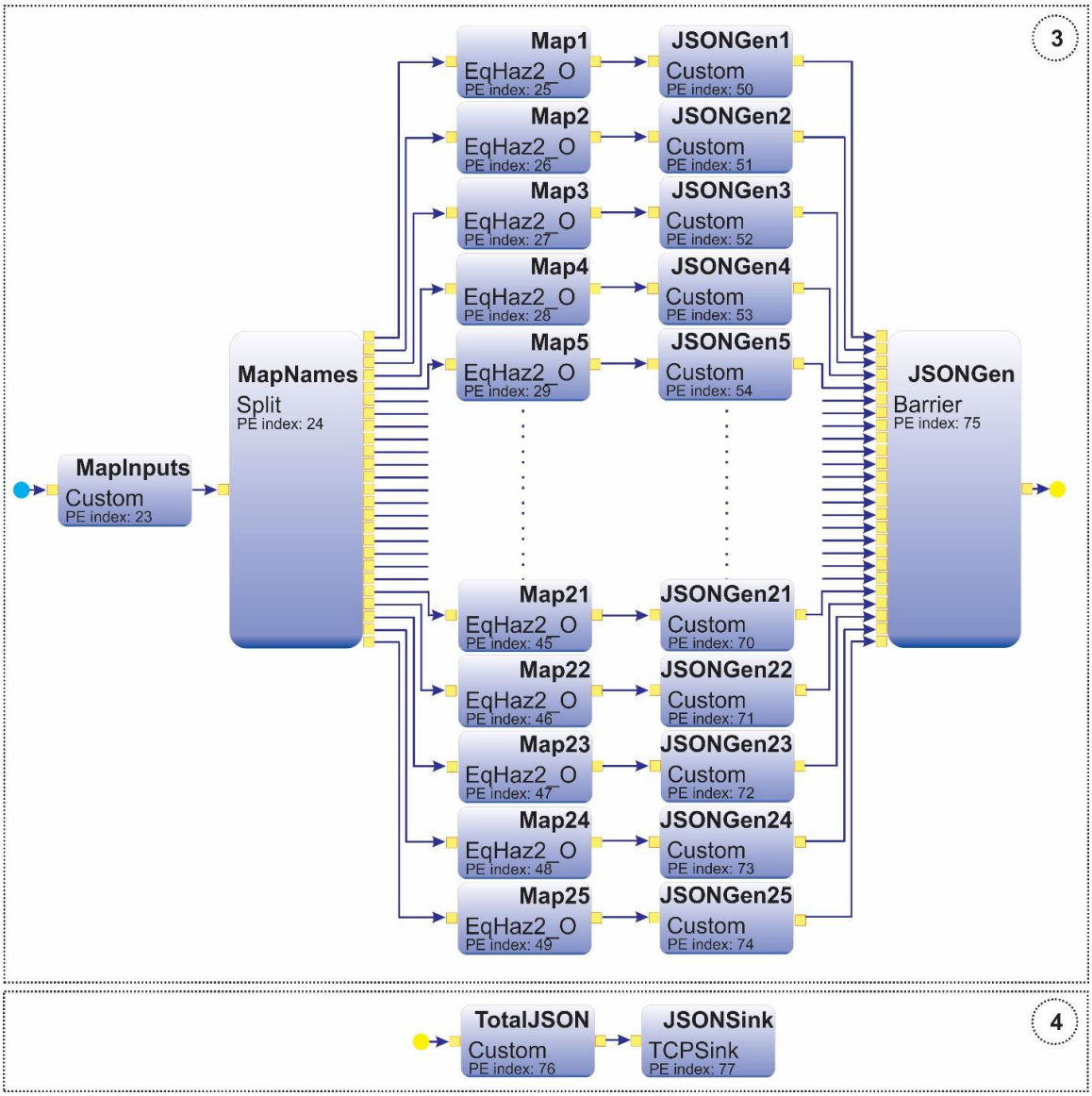

Figure 3: PSHA mapping pipelined with the workload splitting application graph (Stages 3 and 4)

To satisfy the EqHaz1 limitations, the total number of records in the catalogue requested by the user should be less than the maximum number of records in each subcatalog multiplied by the number of parallel pipelines (10 in this example). When all subcatalogs are generated, the total synthetic catalog will be formed (Steps 14-15).

In order to overcome the EqHaz2 limit for the number of grid points, the input stream to the Map operator is split so that there are no more than 100,000 grid points for each Map operator (Step 18). In our case the entire grid is divided into 25 subgrids, where every subgrid contains less than 100,000 points. The EqHaz2 calculations for the ground motion acceleration parameters are performed on (Steps 19-43) and each output stream is converted to JSON format (Steps 44-69). Subsequently, results of all streams are merged together and sent to output port to be visualized.

2.4 Experimentation

The experimental environment in this work consists of a cluster of four machines, each with dual Xeon quad-core 2.4GHz CPU, 16GB RAM and running Linux. The computation power of the cluster is provided by the total of 32 cores. From the application point of view, at least 32 processes can be run in parallel on the cluster for handling heavy workload. InfoSphere Streams Version 3.2 has been configured and installed on the cluster. Files with input models (F1 and F2 contain GSC 2011 composite model of 39 Eastern Canada zones, F4 contain ground motion databases for Eastern Canada (Atkinson & Adams, 2013)) are available online from the official website of the engineering seismotoolbox (www.seismotoolbox.ca). To demonstrate the performance improvement by running multiple pipelines in parallel for PSHA mapping, the execution time has been measured for both implementations with the same processing workload. Figure 1 is for sequential processing (Seq.) without the workload splitting, while Figure 2 and Figure 3 are for running multiple pipelines in parallel (Paral.) on the cluster with the workload splitting. For the parallel experiment, 10 pipelines were employed at the catalogues generation stage and compiled into 10 PEs, and 25 pipelines were implemented at the map generation stage and compiled into 25 PEs. For the processing workload used in the experiment, the number of records in every synthetic catalog is 1,000,000 and the number of grid points is 100,000, because these are maximum limits for the sequential programs. To measure performance time, two additional operators were written to measure the execution time and output the results into a file: the first operator estimates the execution time of total synthetic catalog generation (Catalog); the second assesses the ground motion calculation performance time (Map). Timing results for

Catalog and Map, in addition to the total execution time and the speedup (SF) factor, are summarized in the Table 2.

Table 2: Timing and speed results

Catalog (number of records = 1,000,000)			Map (number of grid points = 100,000)			Total		
Seq. (s)	Para. (s)	SF (Seq./Para.)	Seq. (s)	Para. (s)	SF (Seq./Para.)	Seq. (s)	Para. (s)	SF (Seq./Para.)
5.68	0.43	13.2	3780.3	138.2	27.4	3785.98	138.63	27.3

The total speedup factor obtained is 27.3 and is the performance improvement with the pipelining implementation and paralleled executions. It should be noted that the performance increase is obtained not by the improvement of every primitive operator execution, but simply by subdivision of input streams and primitive operators pipelining.

An example of a final output map is shown in Figure 4. Red lines demonstrate a spatial decomposition of map grid. Each red rectangle contains 100,000 points and the total number of points is 2,500,000 for a map resolution equal to one value per kilometer.

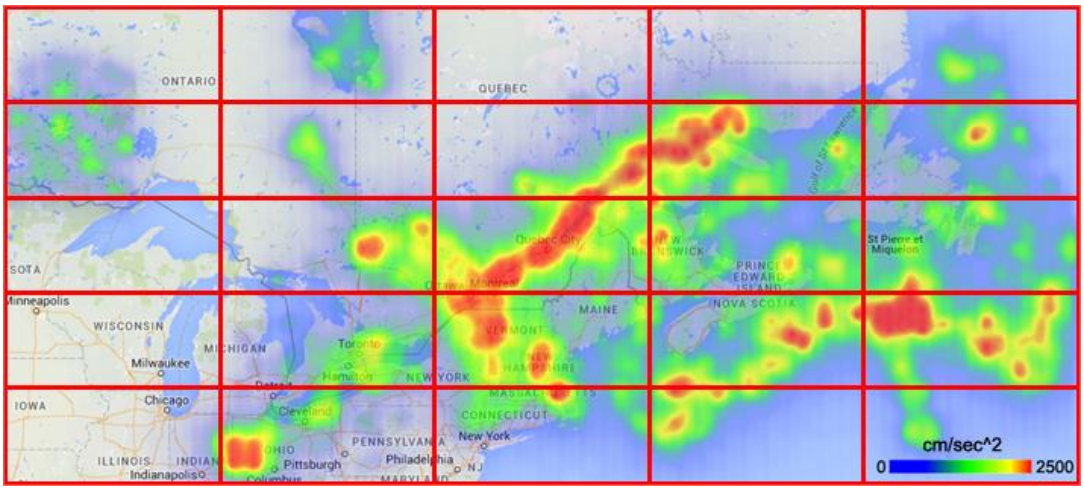

Figure 4: Mean hazard map for a 2475 year return period for pseudo acceleration at a T=0.1 sec period

The total procedure execution time is 5.6 minutes. This example demonstrates a significant improvement in PSHA mapping procedure. For comparison, Assatourians and Atkinson (2013) produced hazard maps with a resolution of 3326 points created by the

EqHaz suite. The PSHA procedure execution time for that case was approximately 90 minutes (Assatourians & Atkinson, 2013) with the experimental environment consists of one machine with dual quad-core 2.4 GHz CPU, 8 Gb RAM and running Windows 7 Professional 64-bit (personal communications, K. Assatourians).

2.5 Conclusions

Seismic hazard maps have many practical and important applications and are widely used by policymakers to assign earthquake-resistant construction standards, by insurance companies to set insurance rates, by civil engineers to estimate structural stability and by emergency hazard agencies to estimate hazard potential and the associated response. Therefore, the results of this work are important for emergency hazard providers, demonstrating that the use of the Streams platform make it possible to produce high resolution hazard maps in near-real time. Here we have provided an example for Eastern Canada, but the method can be applied to any region where PSHA input parameters are available. The pipelining implementation is flexible and scalable for extension and deployment onto a larger cluster. As a result, it is possible to obtain even higher resolution with better performance time.

In the future, this work could be extended to achieve increased performance from additional decomposition on two levels: synthetic catalogs and map generation stages, with even greater cluster support. Additional performance tests should be done to determine the optimum number of parallel pipelines. In the long term, this work could be used as a basis for a universal PSHA streams toolkit development.

2.6 References

Assatourians, K. & Atkinson, G. M., 2013. EqHaz: An open-source probabilistic seismic-hazard code based on the Monte Carlo simulation approach.. *Seismol. Res. Lett.*, 84(3), pp. 516-524.

Atkinson, G. M. & Adams, J., 2013. Ground motion prediction equations for application to the 2015 national seismic hazard maps of Canada. *Can. J. Civil Eng.*, Volume 40, pp. 988-998.

Bauer, M. A., Biem, A., McIntyre, S. & Xie, Y., 2010. A Pipelining Implementation for Parsing X-ray Diffraction Source Data and Removing the Background Noise. *s.l., J. Phys.: Conf. Ser.*.

Bender, B. & Perkins, D. M., 1982. SEISRISK II: A computer program for seismic hazard estimation. *Open-File Report*, pp. 82-293.

Bender, B. & Perkins, D. M., 1987. SEISRISK III: A computer program for seismic hazard estimation. *U.S. Geol. Surv. Bull.* , Volume 1772.

Cornell, C. A., 1968. Engineering seismic risk analysis. *Bull. Seismol. Soc. Am.*, 58(5), pp. 1583-1606.

Field, N., Jordan, T. A. & Cornell, C. A., 2003. OpenSHA: A developing community-modeling environment for seismic hazard analysis. *Seismol. Res. Lett.*, 74(4), pp. 406-419.

IBM, 2014. IBM Knowledge Center. [Online] Available at: http://www-01.ibm.com/support/knowledgecenter/

McGuire, R., 1976. Fortran program for seismic risk analysis. U.*S. Geol. Surv. Open-File Rept*., pp. 76-67.

McGuire, R., 1978. FRISK: Computer program for seismic risk analysis using faults as earthquake sources. *U.S. Geol. Surv. Open-File Rept.,* pp. 78-1007.

McGuire, R., 2004. Seismic Hazard and Risk Analysis.. *Oakland, California.: Earthquake Engineering Research Institute.*

Musson, R., 1999. Determination of design earthquakes in seismic hazard analysis through Monte Carlo simulation. J.Earthquake Eng., Issue 3, pp. 463-474.

Musson, R., 2000. The use of Monte Carlo simulations for seismic hazard assessment in the U.K.. *Ann.Geofis*., Issue 43, pp. 1-9.

Musson, R. & Henni, P., 2001. Methodological considerations of probabilistic seismic hazard mapping. Soil Dynamics and Earthquake Engineering, Issue 21, pp. 385-403.

Ordaz, M., Martinelli, F., D'Amico, V. & Meletti, C., 2013. CRISIS2008: A flexible tool to perform probabilistic seismic hazard assessment. *Seismol. Res. Lett.,* 84(3), pp. 495-504.

Robinson, D., Dhu, T. & Schneider, J., 2006. Practical probabilistic seismic risk analysis: A demonstration of capability. *Seismol. Res. Lett.,* 77(4), pp. 453-459.

Robinson, D., Fulford, G. & Dhu, T., 2005. EQRM: Geoscience Australia's Earthquake Risk Model: *Technical Manual: Version 3.0. GA Record 2005/01, Geoscience Australia, Canberra*, p. 148.

Wied, P., 2014. Dynamic Heatmaps for the Web. [Online] Available at: http://www.patrick-wied.at/static/heatmapjs/

Chapter 3

3 Impact of the Ground Motion Prediction Equation Changes on Eastern Canada Hazard Maps

The material covered in this chapter has been published in the conference proceeding of 2016 Canadian Society for Civil Engineering Annual Conference, London, Ontario, Canada.

Accurate seismic hazard assessment is one of the most important steps on the way to reduce seismic risk. Probabilistic Seismic Hazard Assessment (PSHA) (Cornell, 1968; McGuire, 2004, 2008) is most common method used today in national seismic codes, including the National Building Code of Canada (NBCC) and National Building Code of United States of America. The new generation of Canadian seismic hazard maps was released in 2015 as a basis for updated seismic provisions of NBCC 2015. Its application implemented a new set of representative ground-motion prediction equations proposed by Atkinson and Adams, 2013, in which a set of three alternative weighted ground motion prediction equations (GMPE) are used to describe epistemic uncertainty using a logic trees approach to quantify the distribution of all outcomes. A logic tree aproach weights each potential outcome in order to incorporate the epistemic uncertainty related to the inputs of PSHA and thus enable estimation of the resulting hazard uncertainty. Here sensitivity tests have been performed with the aim to compare the proposed GMPE model with different weights in the probabilistic logic tree (±25-75% weight change applied) and to examine the total amplitude range of ground motion data represented on hazard maps at 0.5, 1.0 and 5.0 Hz frequency. The results of this study show the importance of a correct epistemic uncertainty estimation in PSHA in general and for the Eastern Canada region in particular. Weight changes in each node of the GMPE logic tree for Eastern Canada leads to the significant changes in ground motion values on the hazard maps. In particular, at 5.0 Hz frequency, the difference from the initial base model varies between 14.73% to 64.6% and 19.44% to 100% at 1.0 Hz frequency. At 0.5 Hz frequency the results are even more sensitive and show a 25.71% to 142.85% difference in ground motion estimation.

3.1 Introduction

Seismic hazard maps are a series of maps in different frequency bands calculated in terms of probability of exceeding a certain ground motion level at many points across a region. The maps are probabilistic because they take into consideration the uncertainties in the earthquake magnitude, location and the resulting ground motions that can affect a set of sites. There are many practical applications of hazard maps, the most important among them are the production of earthquake resistant construction standards, insurance rate calculation, structural stability estimation, potential hazard and the associated response assessment. The national seismic hazard maps for Canada are under continuous improvement by the Geological Survey of Canada (GSC). As a result, several seismic hazard models have been proposed and five generations of seismic hazard maps have been created (NRCC, 1953, 1970, 1985, 2005, 2010).

Canada has a vast territory and a variety of tectonic regimes. As a result, the production of seismic hazard maps takes into account the country divided into eastern, western and stable (central) regions (Adams and Atkinson, 2003). Each region has its own particular qualities, principles and approaches of hazard mapping. This study describes models, approaches and results only for Eastern Canada, which is known as a region of moderate seismicity. Despite the fact that detailed source characteristics and wave propagation properties of particular active fault sources are generally unavailable in this region, several generations of the source zone models, earthquake occurrence patterns and ground motion relations have been developed to quantify seismic hazard characteristics in the region. The variety of proposed models increases the resulting level of uncertainty and results in potential error in the decision making process, related policies and standards. As a result, there is a strong interest in the sensitivity analysis of hazard maps as they relate to the main input model parameters.

Aside from the site effects (soil and amplification characteristics) the most important input models for hazard map production are the seismic source zone model and the GMPE model. The seismic source zone model contains detailed seismicity specifications for a region, allowing for uncertainty, while a GMPE model is a statistical model developed for different tectonics regions to predict ground shaking at any site in a region.

Previously, Atkinson and Goda (2011) investigated the effect of seismicity models and a new GMPE on seismic hazard assessment for four Canadian cities (Ottawa, Montreal, Quebec City, Vancouver) and concluded that the selection of correct GMPE model has an important effect on the resulting hazard assessment. Atkinson and Adams (2013) proposed a new set of GMPE for the 2015 National Seismic Hazard Maps. The main difference between the newly proposed GMPE model and other traditional models is in the means of quantifying epistemic uncertainty. In the past, the existing peer-reviewed GMPE models used different assigned weights representing that uncertainty and the relative confidence in each model (Atkinson, 1995). The updated method proposes using of existing peer-reviewed GMPEs to construct a set of three weighted equations (central, upper and lower) for each region and event type. The central GMPE is a representative GMPE and the upper and lower bounding equations quantify the uncertainty in the central GMPE. All three sets of GMPEs have different weights assigned to the logic tree used in PSHA. Each of these approaches has advantages and disadvantages (Atkinson, 2011). The first approach is more traditional and widely used than the representative model. However, in Canada the representative approach has been approved to the 2015 seismic hazard map production process (Atkinson and Adams, 2013).

Our motivation in this study is to expand the boundaries of the sensitivity analysis of hazard maps to the new GMPE model. The sensitivity tests performed in this study compared the complex effect of changes to GMPE models not for one particular site, as has been done before, but on a number of sites (hazard map). For high-resolution hazard map production calculations must be carried out on a large number of sites, and this requires powerful computational resources. Kropivnitskaya et al. (2015) demonstrated that these steps in the hazard mapping process, such as Monte Carlo simulations and resolution mapping, can be split and distributed into pipelines in parallel. Near real-time and low-latency parallel computing of hazard maps have been achieved by using streaming and pipelining computing paradigms through IBM InfoSphere Streams platform (IBM, 2015) which reads input data and models from external sources (in our case www.seismotoolbox.ca), executes hazard calculations in parallel on a number of sites and visualize resulting hazard map (Kropivnitskaya et al., 2015).

3.2 Hazard Maps Production Methodology

Seismic hazard analysis provides an estimate of the effects from natural earthquakes on the man-made structures at a given site of interest. There are two widely used seismic hazard analysis approaches: deterministic and probabilistic. In the deterministic methodology, hazard is estimated for a specific magnitude at a fixed source-to-site distance (Reiter, 1990; Anderson, 1997; Krinitzsky, 2002). The probabilistic approach quantifies the probability of exceeding various ground-motion levels at a site or a map of sites given by all possible earthquakes in a region (Cornell, 1968; McGuire, 1977). PSHA is a powerful tool for seismic hazard map production (Adams and Atkinson, 2003; Trifunac, 1990). The classic probabilistic hazard calculation is relatively simple when all input parameters and models are specified. Given that the seismic hazard source model and attenuation relationships are known or assumed, PSHA is performed in the following steps:

1. Determine the seismic hazard source model that provides N earthquake scenarios E_n for magnitude (m_n), location (L_n), and rate (r_n);

2. Determine the distance D_n to the site of interest;

3. Calculate the distribution of possible ground-motion levels for the scenario (1):

$$p_n(\ln PGA) = \frac{1}{\sigma_n \sqrt{2\pi}} e^{-\frac{(\ln PGA - g(m_n, D_n))^2}{2\sigma_n}}, \qquad (1)$$

where

ln*PGA* is a natural logarithm of peak-ground acceleration;

$g(m_n,D_n)$ is the mean of ln*PGA* given by attenuation relationship;

σ_n is the standard deviation of ln*PGA* given by the attenuation relationship.

4. Find the probability of exceeding each ln*PGA* by integration (2):

$$P_n(>\ln PGA) = \frac{1}{\sigma_n\sqrt{2\pi}} \int_{\ln PGA}^{\infty} e^{-\frac{(\ln PGA - g(m_n, D_n))^2}{2\sigma_n}} d\ln PGA; \qquad (2)$$

5.Obtain the annual rate at which each ln*PGA* is exceeded Rn due to the scenario (3):

$$R_n(>\ln PGA) = r_n P_n(>\ln PGA); \qquad (3)$$

6.Calculate the total annual rate of exceeding each ln*PGA* (4):

$$R_{total}(>\ln PGA) = \sum_{n=1}^{N} R_n(>\ln PGA) = \sum_{n=1}^{N} r_n P_n(>\ln PGA); \qquad (4)$$

7.Compute the probability of exceeding each ground-motion level in the next *T* years of this annual rate using the Poisson distribution (5):

$$P_{Poissonian}(>\ln PGA, T) = 1 - e^{-R_{total}T}. \qquad (5)$$

An alternative procedure to perform PSHA proposed by Musson (1999) uses Monte Carlo simulations, where in the first stage a long-time synthetic earthquake occurrence catalogue is simulated. This synthetic catalogue is used to estimate a distribution of ground motions at any site from selected GMPE. Today there are a number of free and commercial software tools available to perform PSHA including Monte Carlo simulation approach (EQRISK, FRISK (McGuire, 1978), SeisRisk (Bender and Perkins, 1987), Fortran codes produced by National Seismic-Hazard Mapping Project by the USGS, CRISIS (Ordaz et al., 2013), EQRM (Robinson et al., 2005, 2006), OpenSHA (Field et al., 2003). One such tool is the EqHaz software suite of open-source FORTRAN programs developed by Assatourians and Atkinson (2013). This suite consists of three programs. EqHaz1 creates the synthetic earthquake catalogues generated by the user-specified seismicity parameters. EqHaz2 produces the ground motion catalogues and mean hazard probability curves at a site, as well as mean hazard motions at specified return periods calculated for a grid of points. These modules have some calculation limits that do not allow their use these modules for PSHA mapping purposes. In particular, the number of records in the each synthetic catalog is limited to 1,000,000 and hazard map produced by EqHaz2 could have no more than 100,000 points that for Eastern Canada

means the maximum hazard map resolution that can be obtained is about one value per 25 kilometers. EqHaz3 de-aggregates the hazard, estimating the relative contributions of the earthquake sources in terms of distance and magnitude. Kropivnitskaya et al. (2015) took the EqHaz1 and EqHaz2 Fortran source code, compiled them into system object libraries and implemented the base PSHA procedures and functions in the streaming environment IBM InfoSphere Streams. As a result, pipelining and parallel execution on the experimental environment of a cluster of four machines, each with dual Xeon quad-core 2.4 GHz CPU, 16 GB RAM and running Linux overcame the EqHaz limitations mentioned above. In particular, the number of sites for hazard maps are stepped-up to 2,500,000 sites that gives a hazard map with resolution about one value per kilometer. The number of records in the synthetic catalog has been increased in 10 times up to 10,000,000 records (Kropivnitskaya et al., 2015). These achievements allowed the creation of hazard maps for a sensitivity testing in near-real time for the entire territory of Eastern Canada with relatively high resolution.

3.2.1 Seismic Source Zone Model

Eastern Canada has a relatively low rate of earthquake activity, the result of its location in a stable continental region within the North American Plate. Approximately 450 earthquakes occur in this region every year. On average, four of those are $M > 4$ and thirty of $M > 3$ per year, three events of $M > 5$ every decade. Approximately twenty-five are reported as felt events. In most cases an $M > 3$ event is strong enough to be felt while an $M > 5$ event is the threshold for damage in this region. The GSC manages a seismograph network that can detect all events in Eastern Canada of $M > 3$ as well as all events with $M > 2.5$ or more in areas with a high density population (Halchuk, 2000). As of today, there is no solid understanding of what causes earthquakes in Eastern Canada. This region is the part of the stable interior of the North American Plate which extends across the western Atlantic Ocean to the mid-ocean ridge. Seismic activity in these areas is related to regional stress fields, with earthquakes concentrated in regions of crustal weakness. Therefore, there is no correlation between plate interaction and the seismic activity rate and magnitude in this region (Halchuk, 2000).

The seismicity zoning and occurrence model used in this work as input for synthetic data generation are a composite model of thirty-nine Eastern Canada zones provided by GSC (Adams an Halchuk, 2003, 2004; Halchuk and Adams, 2008) shown in Figure 5. The model includes regional geological and seismological features and consists of two alternative source zone models: the H model, which is based on historical seismicity, and the R model, which is based on regional tectonic structure.

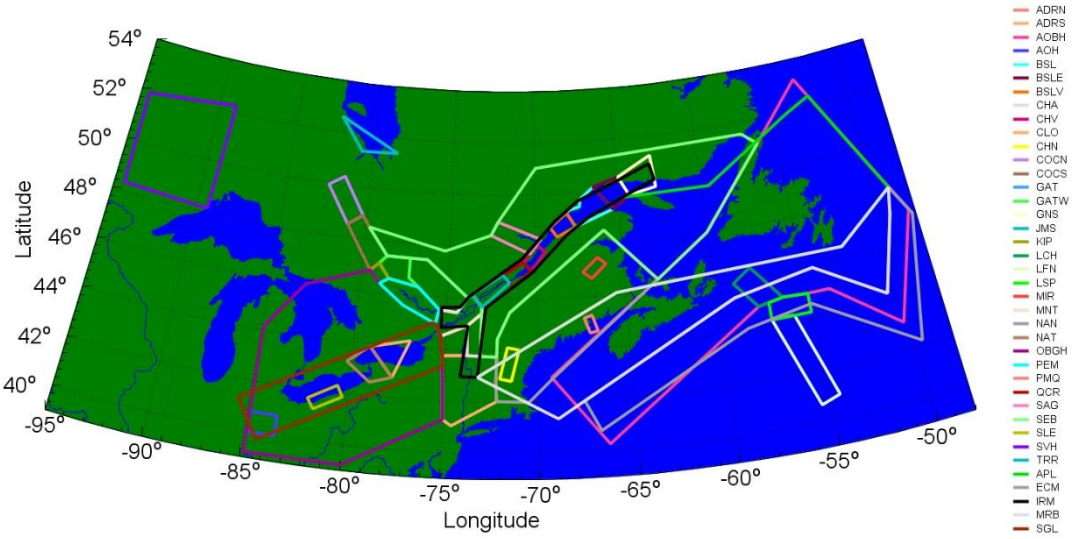

Figure 5: Eastern Canada Seismic Zones Composite Model: small clusters of seismicity are represented by H model; large zones compose R model

3.2.2 GMPE Model

A GMPE represents the shaking amplitude as a function of earthquake magnitude, distance from the source, site conditions, and other variables and are calculated for peak ground acceleration and velocity. The selection of the correct GMPE model is a critical step in PSHA containing two types of uncertainty. The first is aleatory uncertainty, random variability of amplitudes about a median prediction equation. This type of GMPE uncertainty is handled by integrating over the distribution of ground-motion amplitudes about the median. The second type is epistemic uncertainty, which affects the correct value of the median, and in most cases is handled by considering alternative GMPEs in a logic tree format (e.g., Bommer and Scherbaum, 2008). The logic tree model is based on branches with alternative models and then weights assigned to each of those. As a result,

the branches and their weights are intended to represent the underlying continuous distribution of possible ground motions. There are two alternative approaches for epistemic uncertainty estimation in GMPEs by using logic trees. One of them uses multiple existing peer-reviewed GMPEs, with weights assigned to each GMPE based on the judgment of the analyst concerning their relative merits or applicability. Another method uses existing peer-reviewed GMPEs, data analysis, and judgment to define a representative suite of models to capture the uncertainty, including one or more central model along with high and low alternatives. Atkinson (2011) argued that the development of a representative suite is a superior approach to building ground-motion characterization logic trees, in comparison with the more widely practiced use of multiple GMPEs drawn from the literature.

Three sets of representative GMPE (central, low and high) have been developed in Eastern Canada based on five peer-reviewed equations Pezeshk, Zhandieh and Tavakoli (2011), Atkinson and Boore (2006), Atkinson (2008), Silva, Gredor and Daragh (2002) with single corner (variable stress), Gredor and Daragh, (2002) with double corner (with saturation). In Figure 6 examples of several GMPE equations are shown with pseudo spectral acceleration at different period of time for event with different magnitude. Spectral acceleration is the most commonly used measure of ground motion intensity in building engineering practice. For specified natural period and damping level, spectral acceleration represents the maximum acceleration that a ground motion will cause in a linear oscillator. In other words, pseudo spectral acceleration (PSA) is equal to spectral displacement times the square of the natural frequency (Baker et. al, 2006). The representative central equation developed shows the geometric mean ground motions of the five alternatives. It can be used as input for ground motion estimation for all site classes B, C conditions (the average shear-wave velocity in the top 30 m Vs30=760 m/s) in Canada. Atkinson (2011) demonstrated that the representative equation method is comparable to the alternative equations approach, but the weights of each GMPE in the logic tree for PSHA calculations must be selected in a consistent manner.

For Eastern Canada the proper logic tree weights proposed are 0.5, 0.25, and 0.25 for the central, low and high equations weight, respectively. This weights schema is used and

designated as the base model in this study. To obtain the relative performance of the models and to study the representation of epistemic uncertainty, three sensitivity exercises were performed here. In particular, a set of hazard maps were created considering different weights in the PSHA logic tree for central, low and high equations in the representative set. In each case the weight difference in the each node of the logic tree comparing with basic model equals ±25%-75%. During the first test the weights of low and central equations in the logic tree have been decreased up to 25% and 50% relatively to teach 0 and high equation weight has been increased by 75% and reached the maximum weight equals 1. During the second test the maximum weight has been assigned to the central equation when at the same time for low and high equations 0 weight has been assigned. The third test represent the case in which the low equation received the maximum weight and high and central equation have not been taken into account. Logic tree weights used in this study are the end members for GMPE logic tree sensitivity analysis and future work should include looking at the whole range of weight values from 0 to 1 with some certain (for example, 0.05) step.

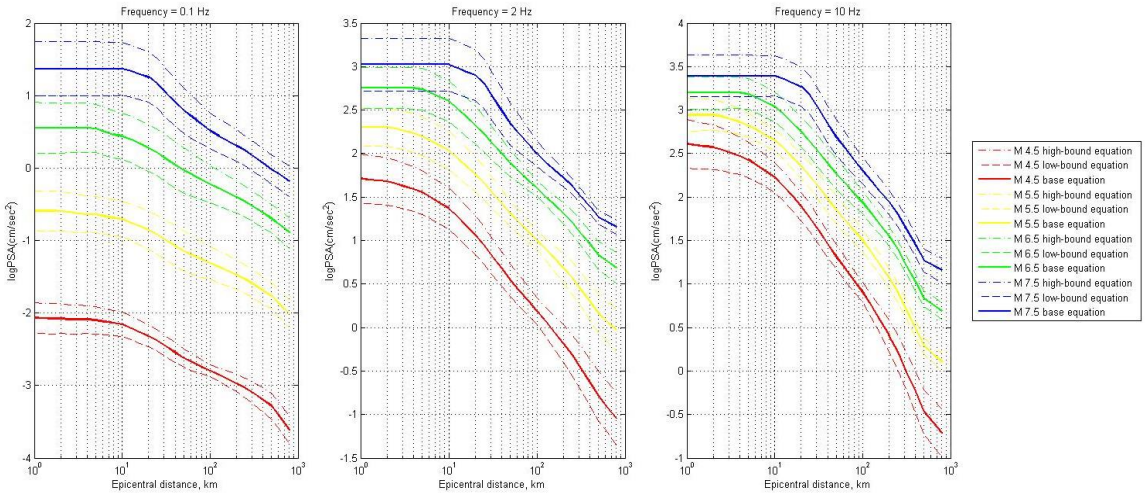

Figure 6: PSA values at 0.1Hz, 2Hz and10 Hz for Eastern North America GMPE versus epicentral distance

3.3 Results

The mean hazard maps of predicted PSA with 2475 year return period were calculated based on the input parameters explained above. 2475 year return period represents 2% exceedance probability in 50 years. This return period is commonly used in building design because of its ability to capture the effects of rare but large earthquakes. The hazard maps produced with input base model are shown on Figure 7, representing PSA at three different periods equal 0.2 sec, 1.0 sec and 2.0 sec.

Three sensitivity tests have been performed for different GMPE models. The effects of modification in the GMPE model are significant at every vibration period in all three sensitivity tests. Test 1, presented in Figure 8, compared predicted spectral acceleration calculated for the base model with predicted spectral acceleration calculated for the model with maximum weight assigned to the high-bound GMPE of the representative set. As expected, the difference of the estimated level of ground motion with the high-bound model is generally positive in all cases and varies in the range from 1.81 to 58.47 cm/sec2 at 0.5 Hz frequency, from 3.6 to 128.53 cm/sec2 at 1.0 Hz frequency, from 4.76 to 398.89 cm/sec2 at 5.0 Hz frequency.

During the second test (Figure 9), maximum weight was assigned to the central GMPE of the representative set. The results obtained in this case are also significantly different when compared to the base model results. The estimated difference varies from 0.00063 to 17.92 cm/sec2 at 0.5 Hz frequency, from – 0.00059 to 36.37 cm/sec2 at 1.0 Hz frequency, and from 0.0011 to 142.95 cm/sec2 at 5.0 Hz frequency.

In the third test (Figure 10) the maximum weight has been assigned to the low-bound GMPE and results have been compared with the results of the base model. This test results in a negative difference that is in the range of -49.3 to -1.79 cm/sec2 at 0.5 Hz frequency, from -100.06 to -4.08 cm/sec2 at 1.0 Hz frequency, from -756.84 to -11.67 cm/sec2 at 5 Hz frequency.

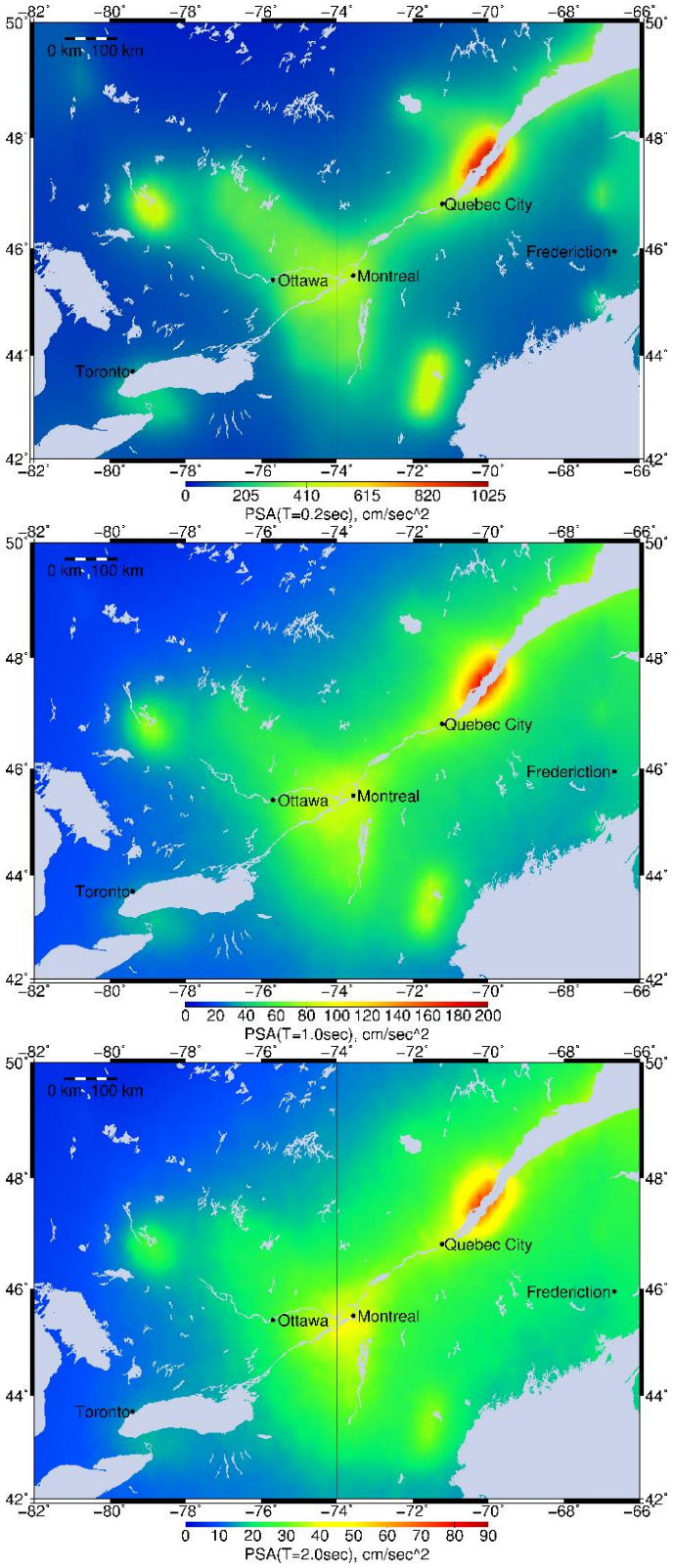

Figure 7: 2475 year return period mean hazard maps (with base GMPE model)

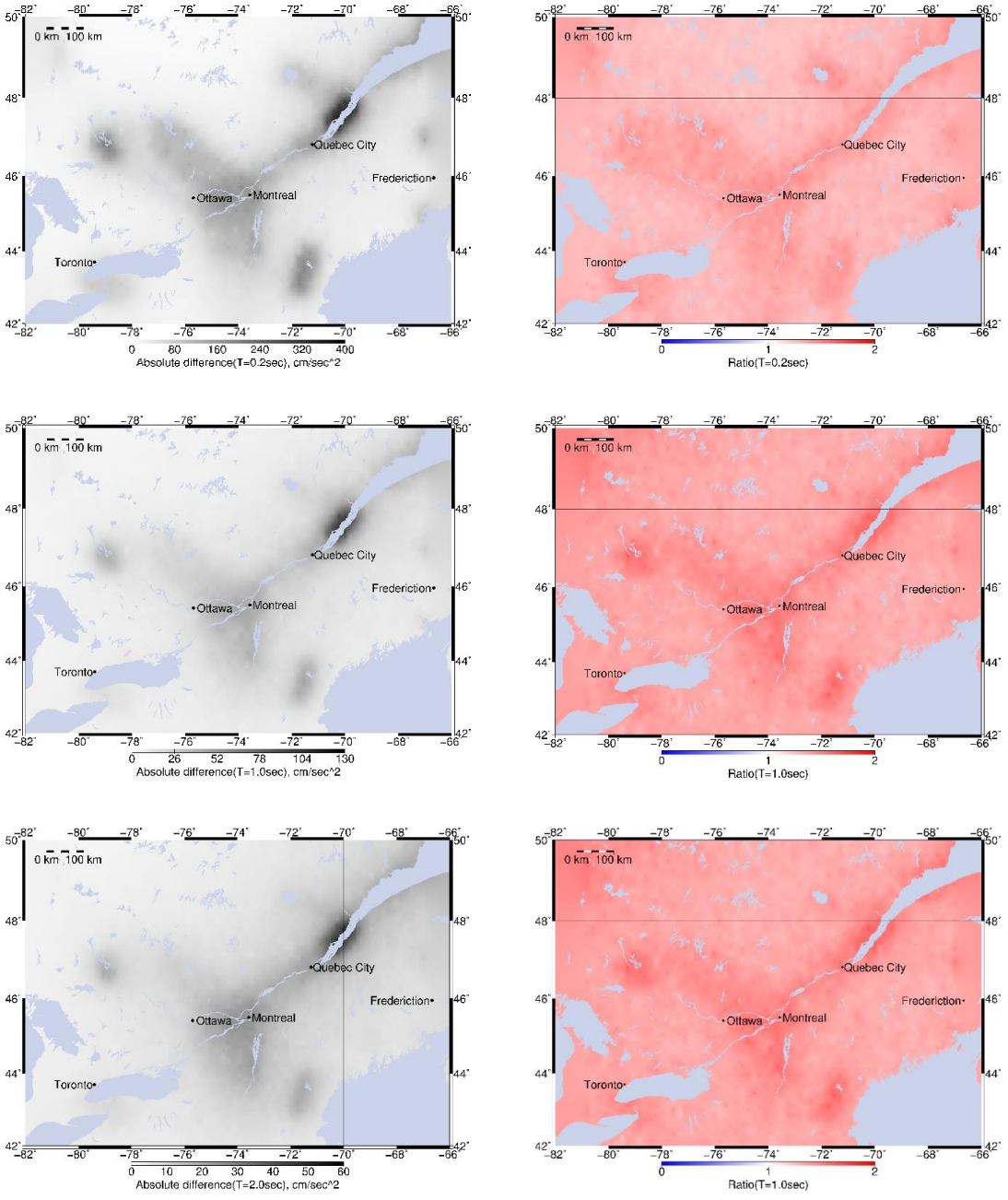

Figure 8: Sensitivity test of 2475 year return period mean hazard maps with high-bound GMPE

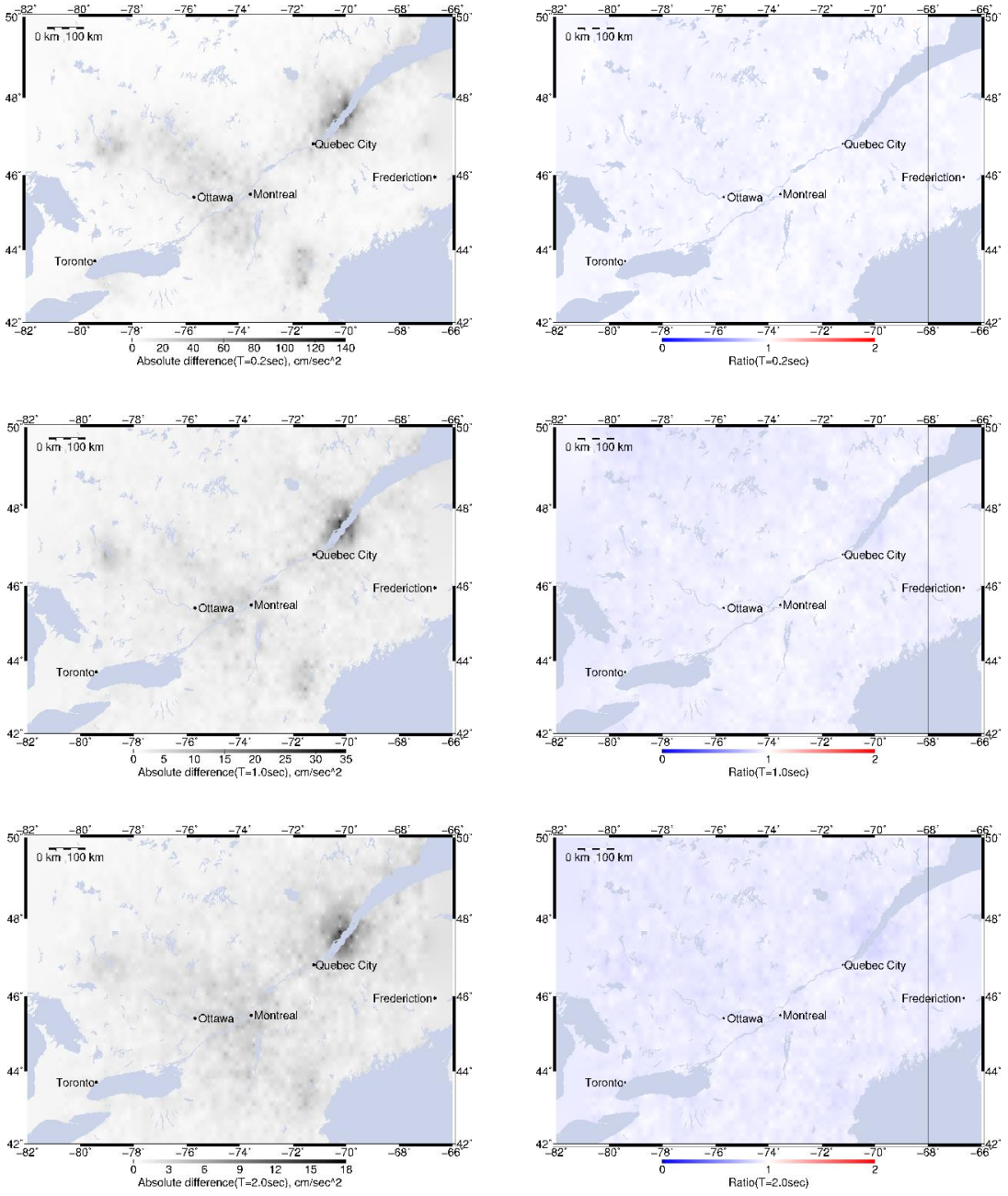

Figure 9: Sensitivity test of 2475 year return period mean hazard maps with medium GMPE

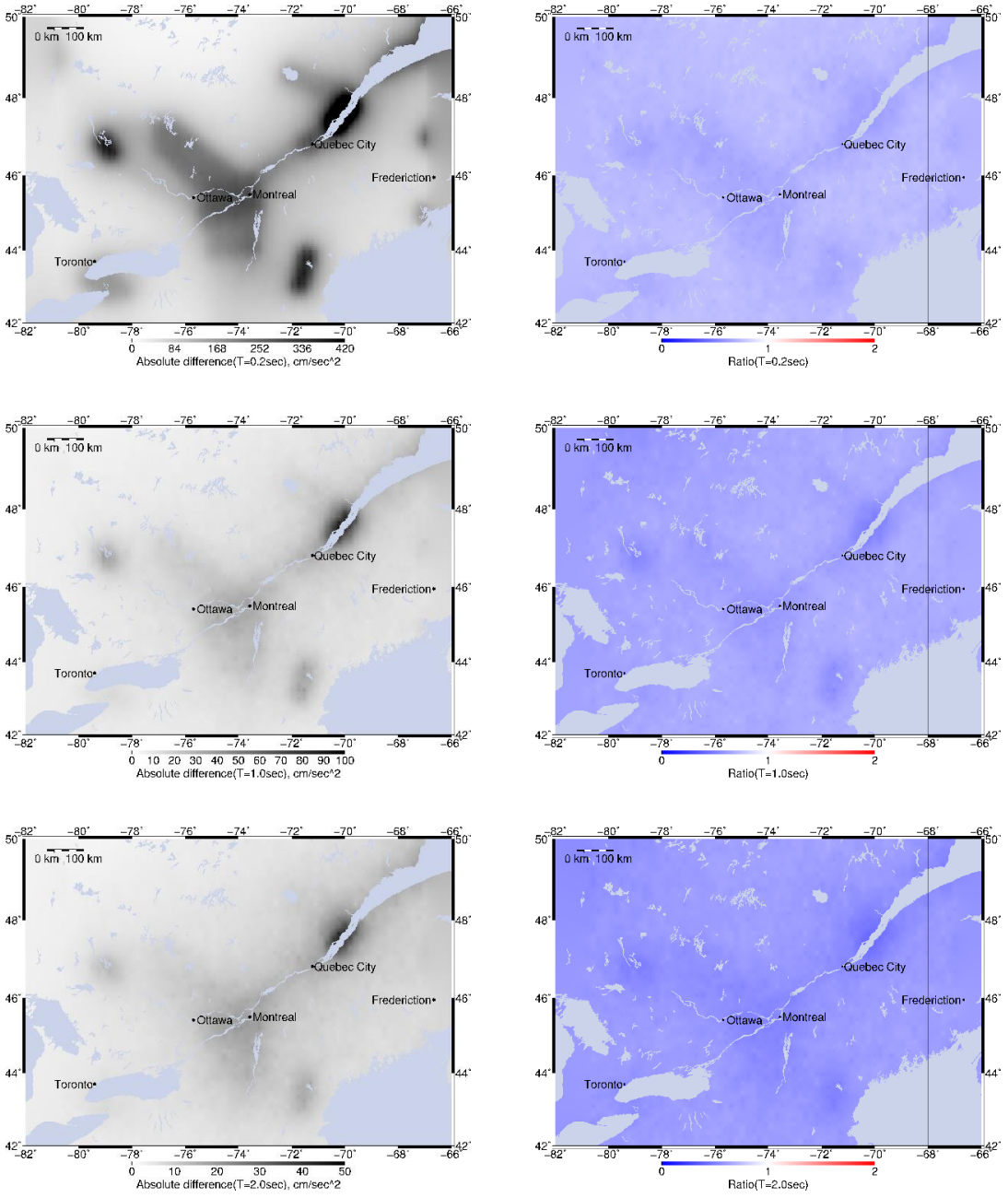

Figure 10: Sensitivity test of 2475 year return period mean hazard maps with low-bound GMPE

The mean changes on every vibration period for each test performed also were analyzed. The maximum change occurs in Test 1, in which 100% weight in the GMPE logic tree has been assigned to the high-bound equation of the representative GMPE set. The minimum mean difference appeared in the Test 2 case in which the medium equation of

the representative GMPE set was applied. In general, the predicted spectral acceleration is more sensitive to the applied changes at low frequencies and close distances than at high frequencies and far distances. That means the areas close to the seismic zones with higher hazard are more sensitive to the changes in probabilistic logic tree.

3.4 Conclusions

Hazard maps have important significance in the seismic risk mitigation programs. Both overestimation and underestimation of the true seismic hazard and risk may significantly increase economic and human losses. Attention should be paid to the development and sensitivity analysis of key input models for PSHA. GMPE models play an important role in ground-motion characterization and seismic hazard map production. In particular, the correct representation of epistemic uncertainty through weights assigned in the representative set of GMPE logic tree has a significant impact on the seismic hazard assessment and, as a result, is one of the most important factors in PSHA and should be handled carefully according to the experts' recommendations.

3.5 References

Adams J., and Atkinson, G. 2003. Development of seismic hazard maps for proposed 2005 National Building Code of Canada. *Canadian Journal of Civil Engineering*, 30 (2): 255-271. doi: 10.1139/102-070.

Adams, J., Halchuk S., 2003. Fourth generation seismic hazard maps of Canada: Values for over 650 Canadian localities intended for the 2005 National Building Code of Canada. *Geological Survey of Canada Open File* 4459: 1-155.

Adams, J., Halchuk S., 2004. Fourth-generation seismic hazard maps for the 2005 national building code of Canada. *Proceedings of the 13th World Conference on Earthquake Engineering*, Vancouver, Canada. Paper 2502 on CD-ROM.

Anderson, J.G. (1997). "Benefits of Scenario Ground Motion Maps", *Engineering Geology*, Vol. 48, No. 1, pp. 43–57.

Assatourians, K., and G. M. Atkinson (2013). EqHaz: An open-source probabilistic seismic-hazard code based on the Monte Carlo simulation approach. *Seismol. Res. Lett.* 84, no. 3, 516–524.

Atkinson, G. (1995). Ground motion relations for use in eastern hazard analysis. I*n Proceedings, 7th 650 Canadian Conference on Earthquake Engineering, Montreal, June 1995*, p. 1001-1008.

Atkinson, G. (2008). Ground-motion prediction equations for eastern North America from a referenced empirical approach: implications for epistemic uncertainty, *Bull. Seism. Soc. Am*.98, 1304-1318

Atkinson G. 2011. An empirical perspective on uncertainty in earthquake ground motions. *Canadian Journal of Civil Engineering* 38(9): 1002-1015

Atkinson, G. and J. Adams (2013). Ground motion prediction equations for application to the 2015 national seismic hazard maps of Canada. *Canadian Journal of Civil Engineering*, 40, 988-998.

Atkinson, G. M., and D. M. Boore (2006). Earthquake ground-motion prediction equations for eastern North America, *Bull. Seism. Soc. Am.*96, 2181-2205.

Atkinson, G. and K. Goda (2011). Effects of seismicity models and new ground motion prediction equations on seismic hazard assessment for four Canadian cities. *Bull. Seism. Soc. Am.*, 101, 176-189.

Bender, B., and D. M. Perkins (1987). SEISRISK III: A computer program for seismic hazard estimation. *U.S. Geol. Surv. Bull*. 1772.

Bommer, J. and F. Scherbaum (2008). The use and misuse of logic trees in probabilistic seismic hazard analysis. *Earthquake Spectra*, 24, 997-1009.

Cornell, C. A. (1968). Engineering seismic risk analysis. *Bull. Seismol. Soc. Am*. 58, no. 5, 1583–1606.

Field, N., T. A. Jordan, and C. A. Cornell (2003). OpenSHA: A developing community-modeling environment for seismic hazard analysis. *Seismol. Res. Lett*. 74, no. 4, 406–419.

Gregor, N. J., W. J. Silva, I. G. Wong, and R. R. Youngs (2002). Ground-motion attenuation relationships for Cascadia subduction zone megathrust earthquakes based on a stochastic finite-fault model, *Bull. Seism. Soc. Am*.92, 1923-1932.

Halchuk, S., (2000) Earthquake zones in Eastern Canada http://www.earthquakescanada.nrcan.gc.ca/ zones/eastcan-eng.php

Halchuk, S., and J. Adams (2008). Fourth generation seismic hazard maps of Canada: maps and grid values to be used with the 2005 National Building Code of Canada, *Geological Survey of Canada, Open File* 5813.

IBM (2015) IBM Knowledge Center http://www-01.ibm.com/support/knowledgecenter/

Krinitzsky, E.L. (2002). "How to Obtain Earthquake Ground Motions for Engineering Design", *Engineering Geology*, Vol. 65, No. 1, pp. 1–16.

Kropivnitskaya Y., Qin J., Tiampo K., Bauer M. (2015). A Pipelining Implementation for High Resolution Seismic Hazard Maps Production. International Conference On Computational Science, *ICCS 2015 — Computational Science at the Gates of Nature.* Volume 51:1473–1482

McGuire, R. (1978). FRISK: Computer program for seismic risk analysis using faults as earthquake sources. *U.S. Geol. Surv. Open-File Rept*. 78-1007.

McGuire, R. (2004). Seismic Hazard and Risk Analysis. *Earthquake Engineering Research Institute, Oakland, California*.

McGuire, R. (2008). Probabilistic seismic hazard analysis: Early history. *Earthq. Eng. Struct. Dyn.* 37, 329–338.

Musson, R.M.W (1999). Determination of design earthquakes in seismic hazard analysis through Monte Carlo simulation. *J.Earthquake Eng.*, 3, 463-474.

NRCC. 1953. 1960. 1965. 1970. 1975. 1977. 1980. 1985. 1990. 1995. 2005. 2010. National Building Code of Canada, *Associate Committee on the National Building Code, National Research Council of Canada,* Ottawa, ON.

Ordaz, M., F. Martinelli, V. D'Amico, and C. Meletti (2013). CRISIS2008: A flexible tool to perform probabilistic seismic hazard assessment. *Seismol. Res. Lett.* 84, no. 3, 495–504, doi: 10.1785/0220120067.

Pezeshk, S., A. Zandieh and B. Tavakoli (2011). Ground-motion prediction equations for eastern North America from a hybrid empirical method. *Bull. Seism.Soc. Am*.

Reiter, L. (1990). Earthquake Hazard Analysis: Issues and Insights, *Columbia University Press*, New York, U.S.A.

Robinson, D., T. Dhu, and J. Schneider (2006). Practical probabilistic seismic risk analysis: A demonstration of capability. *Seismol. Res. Lett*. 77, no. 4, 453–459.

Robinson, D., G. Fulford, and T. Dhu (2005). EQRM: Geoscience Australia's Earthquake Risk Model: Technical Manual: Version 3.0, GA Record 2005/01, *Geoscience Australia*, Canberra, 148 pp.

Silva, W. J., N. J. Gregor, and R. Darragh (2002). Development of regional hard rock attenuation relations for central and eastern North America, *Technical Report, Pacific Engineering and Analysis*, El Cerrito, CA.

Chapter 4

4 Real-Time Earthquake Intensity Estimation Using Streaming Data Analysis of Social and Physical Sensors

The material covered in this chapter is accepted for publication with minor revisions to the journal of Pure and Applied Geophysics on 06/07/2016.

Earthquake intensity is one of the key components of the decision-making process for disaster response and emergency services. Accurate and rapid intensity calculations can help to reduce total loss and the number of casualties after an earthquake. Modern intensity assessment procedures handle a variety of information sources, which can be divided into two main categories. The first type of data is that derived from physical sensors, such as seismographs and accelerometers, while the second type consists of data obtained from social sensors, such as witness observations of the consequences of the earthquake itself. Estimation approaches using additional data sources or that combine sources from both data types tend to increase intensity uncertainty due to human factors and inadequate procedures for temporal and spatial estimation, resulting in precision errors in both time and space. Here we present a processing approach for the real-time analysis of streams of data from both source types. The physical sensor data is acquired from the U.S. Geological Survey (USGS) seismic network in California and the social sensor data is based on Twitter user observations. First, empirical relationships between tweet rate and observed Modified Mercalli Intensity (MMI) are developed by using data from the M6.0 South Napa, CA earthquake that occurred on August 24, 2014. Second, the streams of both data types are analyzed together in simulated real-time to produce one intensity map. The second implementation is based on IBM InfoSphere Streams, a cloud platform for real-time analytics of big data. To handle large processing workloads for data from various sources, it is deployed and run on a cloud-based cluster of virtual machines. We compare the quality and evolution of intensity maps from different data sources over 10-minute time intervals immediately following the earthquake. Results

from the joint analysis show that it provides more complete coverage, with better accuracy and higher resolution over a larger area than either data source alone.

4.1 Introduction

Earthquakes are a natural phenomenon that regularly produces significant damage and loss of life. According to the USGS (2015a), every year several million earthquakes occur worldwide. Most are not dangerous due to either their small magnitude or remote location. But others can cause significant economic loss and casualties. For example, the M7.8 Nepal earthquake that occurred on April 25, 2015 killed more than 9,000 people, injured more than 23,000 people and destroyed 436,344 houses (NDRRP, 2015). Even in well-studied regions with modern building codes such as California, estimated rates of potentially dangerous seismic activity in the north San Francisco Bay area have changed with time. While during the second half of the 19th century, occurrence rates were estimated at one M≥6.0 earthquake every four years, after the 1906 earthquake, the seismic activity rates significantly decreased until the M6.9 Loma Prieta earthquake in 1989. Today, scientists expect larger and more frequent earthquakes on the basis of increasing regional stresses (USGS, 2015c).

Despite the fact that the joint efforts of the scientific and engineering communities have significantly reduced the impact of earthquakes as a result of practical actions and procedures that help to minimize losses after an earthquake, their causes, properties and impacts on human society are still an important topic of scientific research. While engineers are able to build earthquake-resistant buildings, bridges and other infrastructure elements and emergency response services continuously improve population preparedness and risk mitigation, the scientific community continues its studies and simulations of the earthquake source, including the size and extent of the resulting ground motion. Today, continuous analysis of massive volumes of real-time seismic data can be streamed and processed at high speeds and low latency. This technological advantage can be incorporated into post-disaster emergency response systems.

One of the most important measures of damage in emergency response systems is the earthquake intensity. Earthquake intensity quantifies the severity of ground shaking at a

given distance from the epicentre and provides a direct measure of the likely damage. Intensity maps are the most common method for spatial representation of intensity levels across a given region. Public and private organizations use intensity maps for both disaster planning and post-earthquake response (Wald et al., 2003). The MMI scale is the primary intensity measure used in North America. MMI is a twelve step scale, numbered from I to XII. The numbers represent shaking levels from slight shaking to total destruction (Wood and Neumann, 1931; Richter, 1958).

There are two general approaches for the estimation of intensity levels. The first approach is designated instrumental intensity level and is based on a regression of kinematic parameters recorded at individual seismic stations. During an earthquake, energy travels in the form of waves from the epicentre and causes ground movement. The two most common measures of ground motion are peak ground acceleration (PGA) and peak ground velocity (PGV), measured in the east-west (EW) direction, north-south (NS) and vertical (UP) directions. In this case, the intensity level is most accurate for locations adjacent to the seismic stations and less accurate at those location where the ground shaking is obtained through interpolation (Wald et al., 2006b).

The second approach for intensity calculations is based on the shared information from people who experience an earthquake. The current state-of-art is the “Did You Feel It?” program for Community Internet Intensity Maps (CIIM) developed by the USGS. Here, individuals fill in a simple online questionnaire in order to share information about their earthquake experiences and observations. A CIIM processes the completed questionnaires and assigns average intensity level to each ZIP code. The CIIM is updated in time as additional data are received (Wald et al., 2006a).

The fact that people share their experience and observations in social networks can be extended to improve the CIIM results. In particular, any related observations posted online that can be linked to their geographical location can be incorporated into CIIM analyses. The micro-blogging service Twitter is one potentially significant data source, as it has 255 million active users around the world and connects them to the Internet through both their phones and computers (Campagne et al., 2012). In addition, one of the most

important features of Twitter messages (tweets) that makes them useful for improving earthquake response is their real-time nature (Sakaki, 2010). Sakaki (2010) first proposed using Twitter users as earthquake sensors and designated them 'social sensors'. To date, several studies have demonstrated the application of Twitter data in post-disaster response systems (Earle et al., 2010, 2011). Earle et al., (2010, 2011) demonstrated how instrument-based event detection and estimation of earthquake location and magnitude could be supplemented by Twitter data. Sakaki (2010, 2013) constructed earthquake monitoring, detection and early warning system in Japan based on tweet data. That system sends emails to registered users based on seismic events detected with 96% probability, using a special tweets classifier based on the tweet keywords, context and the number of words. System notification is delivered much faster when compared to the warnings issued by the Japan Meteorological Agency. Crooks (2012) analyzed the spatial and temporal characteristics of Twitter activity during the M5.8 earthquake that occurred on the east coast of the United States, August 23, 2011 and concluded that Twitter data can complement other sources of data and enhance situational awareness among people. Burks et al. (2014) created a regression model using tweets in the 10 minutes following an earthquake and integrated it with earthquake characteristics such as moment magnitude, source-to-site distance and shear-wave velocity in the top thirty meters. The main contribution of that work was the demonstration of Tweeter potential as a near-real-time complementary data source for earthquake intensity estimation.

In this paper, we take another step forward and present a streaming implementation of an approach for earthquake intensity estimation that integrates data from both social and physical sensors. Results are shown for the M6.0 South Napa, CA earthquake that occurred on August 24, 2014 at 3:20 a.m. local time. The North San Francisco Bay Area portion of the San Andreas Fault system is a complex network of primarily right-lateral strike-slip faults accommodating motion between the North American and Pacific plates. This network of faults is approximately 80 kilometers wide and trends north-northwest in the area of the West Napa Fault. The West Napa Fault transfers slip between a group of related faults (the Contra Costa Shear Zone) which has a maximum slip rate of one millimeter per year.

The epicenter of the South Napa earthquake was located to the south of the city of Napa and to the northwest of American Canyon on the West Napa Fault (Figure 11). Fifteen thousand people experienced severe shaking, 106,000 people felt very strong shaking, 176,000 felt strong shaking, and 738,000 felt moderate shaking. The duration of shaking lasted from 10 to 20 seconds, depending on location. One person was killed, approximately 200 people were injured and over $400 million in damage occurred as a result of this event (USGS, 2014).

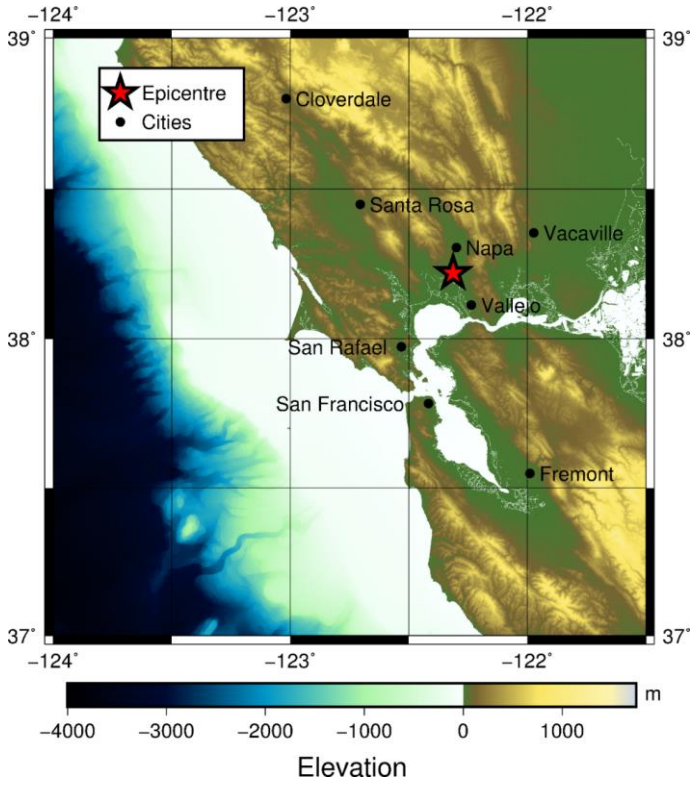

Figure 11: Topography map of the region (based on ETOPO1 dataset (Amante and Eakins, 2009)).

Here we assume that the rate of tweets that directly reference earthquake shaking is correlated with the intensity of shaking. The goal is to calculate the number of tweets per minute that directly identify each particular earthquake intensity level, based on a set of terms typically used in MMI calculations (Wald et al., 2003).

4.2 Selection and Validation of Predictive Relationship between MMI and Tweets Rate

In order to develop an empirical relationship between seismic intensity and tweets rate we used two data sets. The first database is provided by the Northern California Earthquake Data Center (NCEDC) and consists of all aggregated reported intensity estimations from instrumental ground motion recordings and represents the most accurate intensity levels after the earthquake. Each value reflects the average estimation of the ground shaking experienced by the public or an assessment of damage level. Figure 12 shows the MMI map with main cities in the region, epicenter of earthquake and seismic stations in the region (NCEDC, 2014).

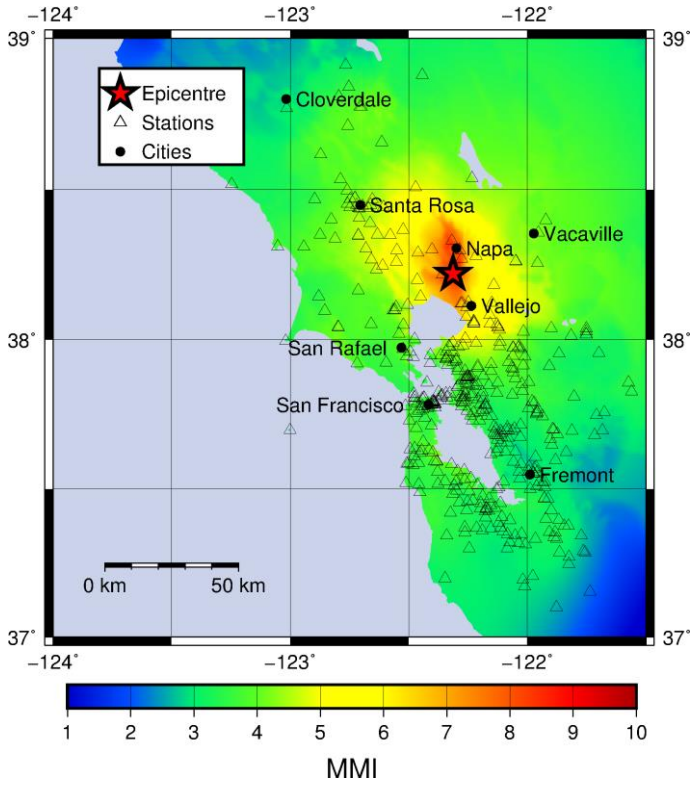

Figure 12: USGS Intensity Map (created from NCEDC data (2014)).

The second dataset is the archive of Twitter records from August 24, 2014 (downloaded from https://archive.org/details/twitterstream). Tweets related to the earthquake (positive tweets) are identified using keywords listed in Table 3.

Table 3: Keywords used for positive tweets filtering

Word	Language
Σεισμός	Greek
地震	Chinese
زلزال, هزة أرضية	Arabic
terremoto, seísmo, sismo, temblor, temblor de tierra	Spanish
Землетрясение	Russian
Aardbeving	Dutch
tremor de terra, terremoto	Portugeese
Deprem	Turkish
Terremoto	Italian
tremblement de terre, séisme	French
Erdbeben	German
地震	Japanese
רעידת אדמה	Hebrew
지진	Korean
Jordbävning	Swedish
earthquake, quake	English

There are several well-known challenges associated with using Twitter data for analytical purposes. One of them is the precise geolocation detection of tweets. Some tweets are geotagged, meaning that they contain the current user location indicator at the time of tweeting. However, the geotagging feature is rarely used by users. For instance, Graham et.al (2014) observed only 0.7% geotagged tweets among 19.6 million tweets. For tweets containing specific cities, the percentage of geotagging was between 2% and 5% (Severo et. al, 2015). The location of a Twitter user can also be obtained from a field in the user account description. Of user accounts with tweets containing some location information, 7.5% contained latitude and longitude values, 57% included a named location, 20.4% referenced information that helped to identify a country, while 15.1% provided humorous or non-spatial information (Takhteyev et. al, 2012). In this study we began with geotagged tweets and for tweets with a location assigned in the user profile we assigned that as the location of the current tweet. Finally, for tweets with no georeference parameters we used a text-based geolocation algorithm. That algorithm analyzes the tweets' content and assigns location coordinates according to a coordinates list of major California cities, towns, their shortcuts and keywords. For example, we assigned "SF Bay

Area", "SF" and "Bay Area" to San Francisco city coordinates and "South Napa", "SW Napa" and "napa" as city of Napa coordinates.

The Twitter dataset used here was gathered on August 24, 2014, with a total volume of 1.8 gigabytes. The streaming approach provides low-latency high-volume access to tweets. The stream of positive tweets on August 24, 2014 is shown in Figure 13.

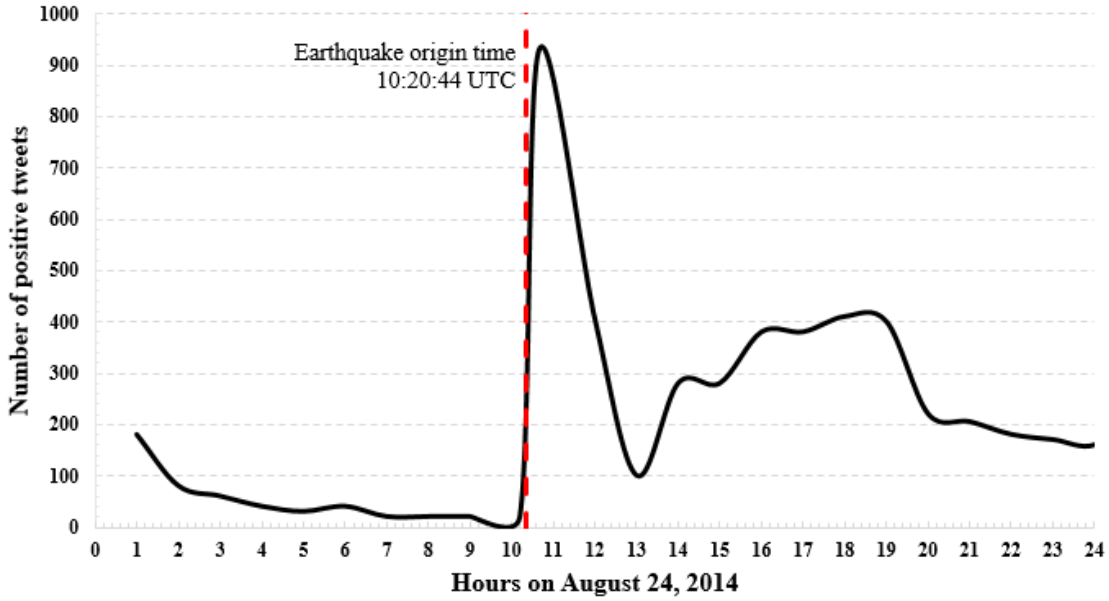

Figure 13: Number of positive tweets in the Napa region on the day of the earthquake.

We limited our analysis to the ten minute period following the earthquake. There were 747 positive tweets during that time interval: 348 were geotagged or contained a user location, 399 were assigned a geolocation based on the above algorithm. Figure 14 shows a map of the spatial distribution of the number of tweets in this analysis. Figure 15 presents a map of population density for northern California. We anticipated that the number of tweets would correlate with population density or, more specifically, the amount of twitter users should correlate with population density. However, comparison of Figures 14 and 15 suggests that population density in this particular case is not as closely correlated with the spatial pattern of the number of tweets in the first ten minutes after an event.

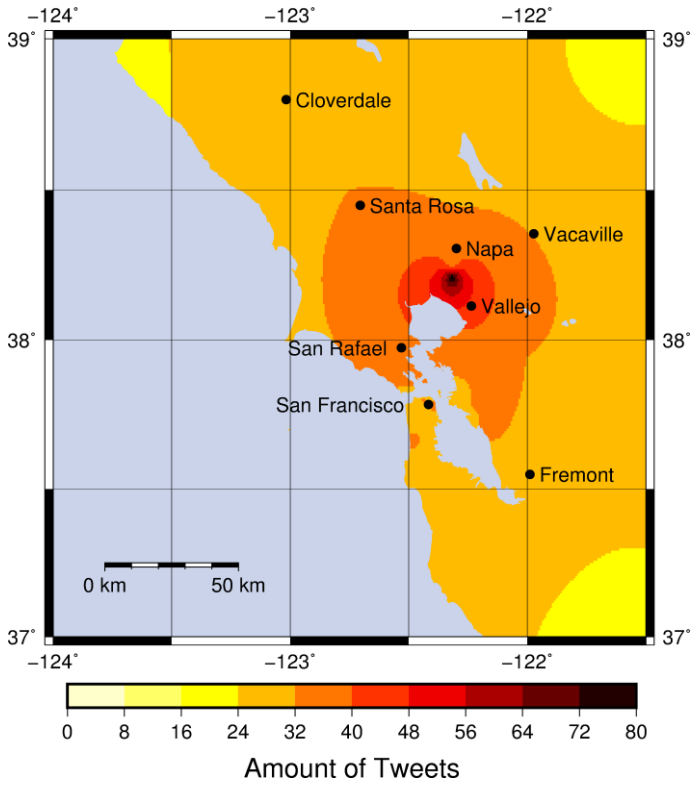

Figure 14: Number of positive tweets ten minutes after the earthquake.

There are three likely explanations for this phenomena. The first reason is related to the local time of the event. The earthquake occurred at 03:20:44 Pacific Daylight Time, when the majority of the population was likely asleep and not actively use cell phones or laptops to the extent that they would during the daytime. The second reason is associated with the time interval of interest. As observed in Figure 14, people tweeted more immediately after the earthquake in those regions close to the epicentre and with high intensity levels. In those areas with high population density that are further from the event we did not observe high tweet numbers during the first ten minutes. The third reason may be associated with the geolocation technique. The majority of non-geotagged tweets mention the word “Napa” and their location had been assigned to the Napa city coordinates in our analysis. However, at some time after earthquake occurrence, many observers already know what happened and may mention Napa Valley in their tweets, even if they were not actually in the city of Napa. Clearly, with these challenges in

current geolocation algorithms there is a significant level of uncertainty in an analysis using data only from social sensors.

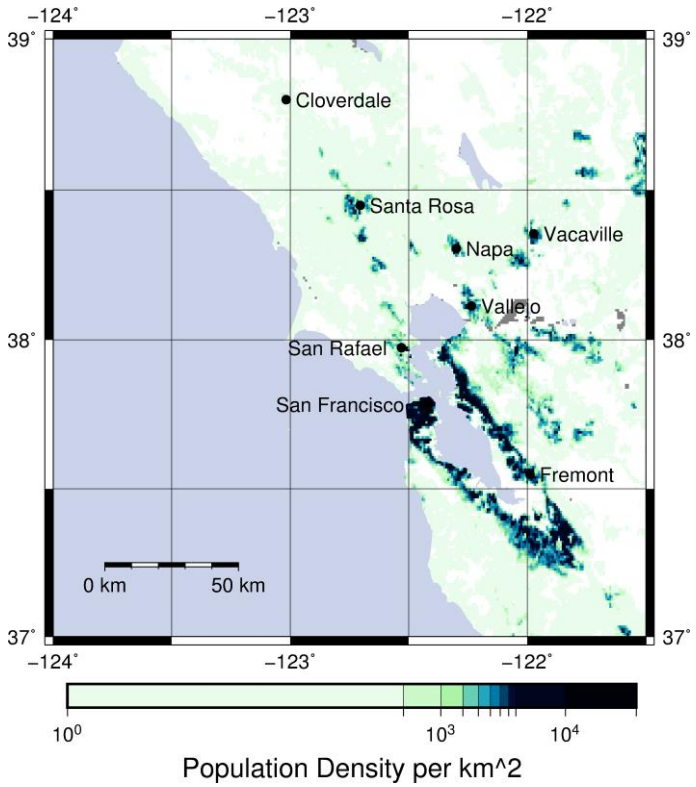

Figure 15: Population density in the region (data from GPWv3 (CIESIN, 2005)).

In order to employ tweet rate as an additional earthquake sensor, i.e. to predict MMI of an actual earthquake, we generated a tweet-frequency time series using a sliding time window of one minute over time steps of five seconds, normalized to the number of tweets per minute. A logarithmic transformation is used here in order to reduce the positive skewness in the data. Figure 16 presents the relationship between MMI values and the logarithmic rate of positive tweets obtained in this study. For each intensity point corresponding to the same latitude and longitude location from the first data set we automatically assigned a value of observed tweet rate, calculated as tweets per minute at the tenth minute after the earthquake. The selection of a ten minutes time interval is based on the analysis that showed that the number of positive tweets is constantly increasing over a longer period of time and reaches its maximum value approximately eleven minutes after the earthquake. In addition, because we are interested in emergency

applications of intensity maps, here we consider only a ten minute interval following the earthquake. We regress MMI against the logarithmic mean of the number of tweets per minute to obtain predictive equations within a legitimate range of values for each model (see Table 4). We regress the average ground-motion values for specified MMI levels in order to approximately follow the appropriate trend instead of producing a relationship overly influenced by the greater statistical volume of data at lower intensities. We applied a least squares solution with 95% confidence bounds for four models: linear, exponential, two-segment linear, three-segment linear. The selection of appropriate model is based on the fact that a relatively precise and unbiased model is a simplest model that produces random residuals. The three-segment model demonstrates the best fit to the data with the highest coefficient of determination (R-squared) of 0.53 and the lowest root-mean-square (RMS) error of 0.65 MMI units. However, because R-squared and RMS error cannot determine whether the model estimates and predictions are biased, we also assessed the residual plots. The residuals between predicted and observed data shown in Figure 17 for each model demonstrate normality in every case.

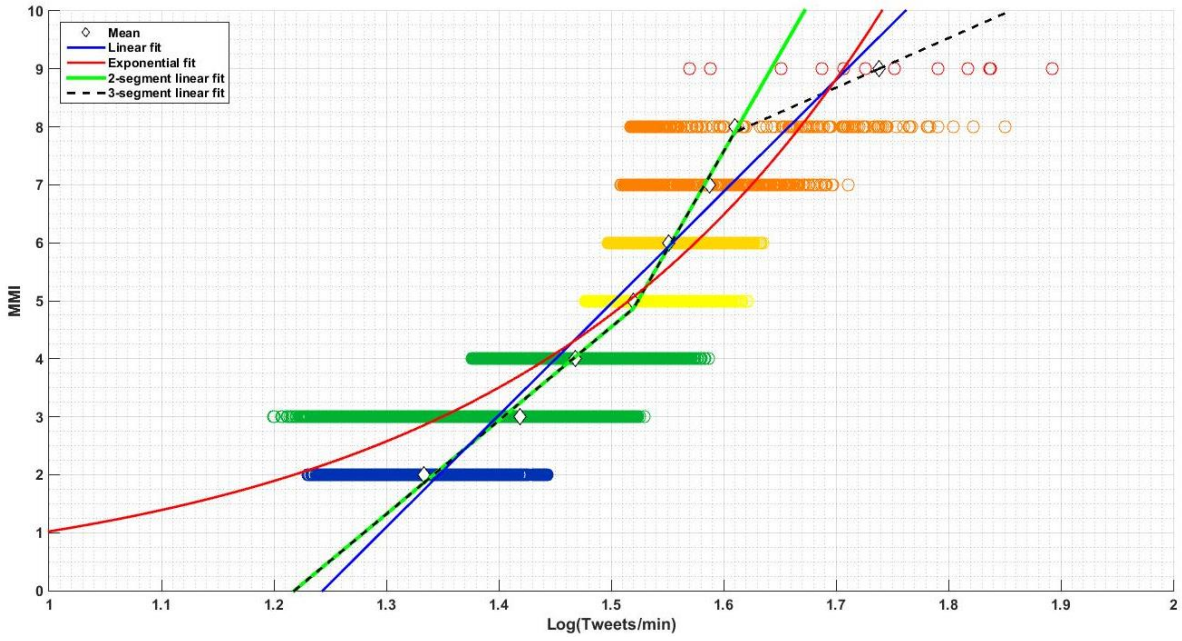

Figure 16: Combined tweet rate dataset (i.e. colored circles at each MMI level) used to derive average log(Tweets/min) (diamonds) for each MMI level (II - blue circles, III – light green circles, IV – green circles, V –light yellow circles, VI – yellow

circles, VII - light orange circles, VIII – orange circles, IX – red circles). The lines show different regression results.

Table 4: Equations to predict MMI from tweets rate

Model Name	Equation	Coefficient of determination, R^2	RMS error	Valid range of values Ntweets/min
Linear	$MMI = 11.73\,log(N_{tweets/min}) - 13.47$	0.32	0.79	[14.08; 148.4]
Exponential	$MMI = 0.0168 * e^{3.67*log(N_{tweets/min})}$	0.26	0.82	(0; 61.7]
Two-segment linear	$MMI = 16.13\log(N_{tweets/min}) - 19.65$ $\log(N_{tweets/min}) < 1.52$ $MMI = 33.75\log(N_{tweets/min}) - 46.42$ $1.52 < \log(N_{tweets/min}) < 1.61$	0.49	0.69	[16.53; 53.8]
Three-segment linear	$MMI = 16.13\log(N_{tweets/min}) - 19.65$ $\log(N_{tweets/min}) < 1.52$ $MMI = 33.75\log(N_{tweets/min}) - 46.42$ $1.52 < \log(N_{tweets/min}) < 1.61$ $MMI = 8.55\log(N_{tweets/min}) - 5.86$ $1.61 < \log(N_{tweets/min}) < 1.74$	0.53	0.65	[16.53; 122.7]

Several important observations can be made from Figure 16. First, both two and three-segment models show the positive tweets rate increases slowly up to MMI less than V. The three-segment model also has a greater slope as MMI increases to VIII. At low MMI levels shaking is light and often goes unnoticed (USGS, 2015b). Starting at MMI V it becomes moderate and is felt by nearly everyone, explaining the increase in slope after MMI V. At severe levels of shaking (greater than VIII), even specially designed structures are slightly damaged, potentially affecting communication infrastructure and decreasing the number of devices with internet access and decreasing the rate of positive tweets.

Twitter data is a simple proxy for MMI level estimation. However, due to the high level of over and underestimation resulting from the use of Twitter data alone, we propose that

it should be implemented jointly with instrumental intensity. In that case, two completely different data sources are analyzed in real-time using a streaming environment and methodology.

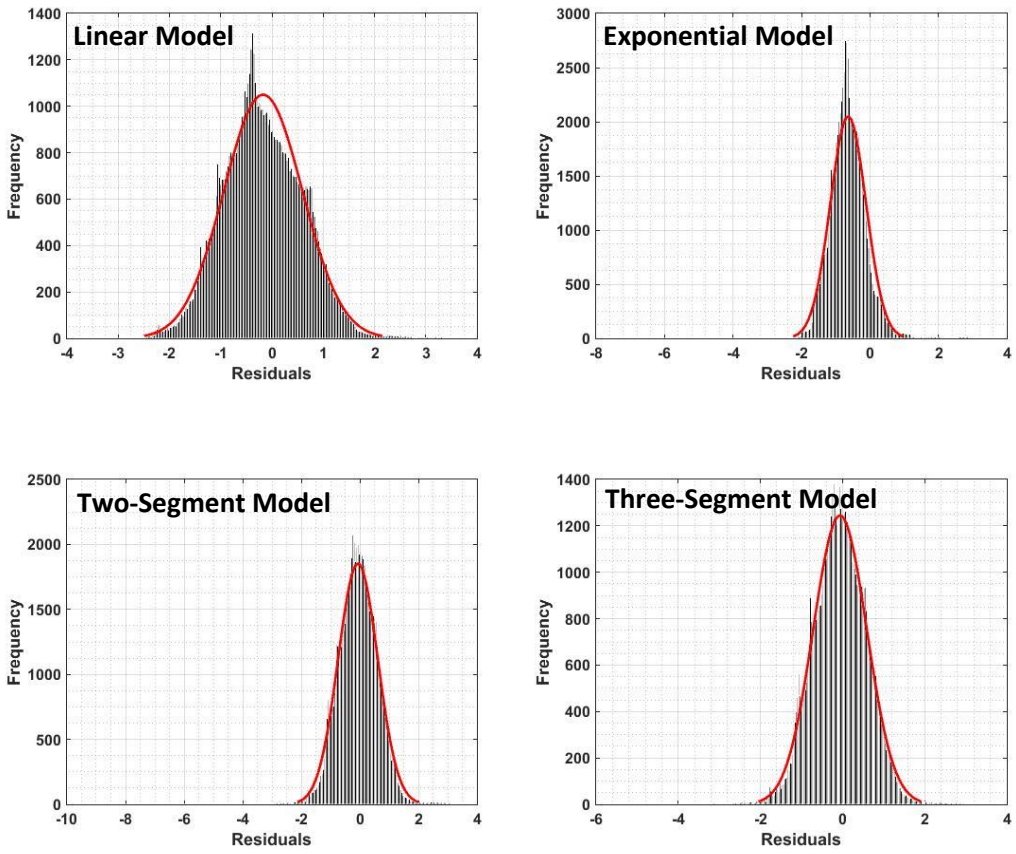

Figure 17: Residuals between predicted and observed data for each model.

4.3 Streaming Methodology and Environment

Both input data types, whether from seismic stations or from Twitter, have a time-dependent nature and can be classified as data streams. A data stream is a sequence of tuples (data packets) received at a sequence of positive real time intervals. In this case it is not possible to process the arriving data as a traditional database. Traditional databases are not designed to be used for continuous data loading and continuous queries (Terry et al., 1992). A Twitter data stream represents a massive volume of data with an average of 5,700 tweets per second and requires streaming approaches to be processed with low latency.

IBM created InfoSphere Streams (Streams), a product which provides a runtime platform, a datacentric programming model and a Stream Processing Language (SPL) specifically for complex streaming data analysis. Although there are other software packages for streaming data applications, we chose Streams for this application because it is flexible in accommodating different data sources. In Streams, an application is scalable for deployment on a larger HPC cluster in order to meet the application needs, and it is easy to implement with the support from various specialized toolkits.

For a Streams application, each processing procedure generally is implemented as an operator, and these operators subsequently are connected to form processing pipelines. The processing performance is the critical component of the streaming data application in terms of meeting certain real-time or near real-time requirements important to this work. In order to promptly and efficiently handle processing workload by making use of the computational resources of a HPC cluster, Streams provides toolkit operators that can easily split an operator's workload into multiple data streams. In the meantime, a processing pipeline also can be duplicated into multiple pipelines in order to digest those data streams, which makes an application scalable. The Streams runtime platform makes it easy to deploy those pipelines on a large-scale HPC cluster without the need to manually manage allocation, synchronization or communication among the operators. The Streams runtime model consists of distributed processes. Single or multiple operators form a processing element (PE). The Streams compiler and runtime services determine where best to deploy those PEs, either on a single machine or across a cluster, in order to meet the resource requirements and application performance needs (IBM, 2014).

Streams come with a standard toolkit for application development. For example, the toolkit provides different types of data source adapters that can be used to monitor different data sources and pull data from multiple sources at the same time. Various data sources could be a local or remote file directory, a TCP or UDP port, or any web URI. Because data from different sources can be of different types and arrive at a different pace, the toolkit also provides utility operators for synchronizing and/or merging multiple data streams. There also are several specialized toolkits available to speed-up the development work for various applications. For example, here we process Twitter data

which is in JSON format (detailed in Section 4), and a specialized JSON toolkit for Streams is used to parse and filter out the information needed from those tweet messages. Detailed implementation of this work is presented in the next section.

The experimental environment in this work consists of a cluster of four virtual machines (VMs), each configured with 8-core 2.4GHz CPU, 16GB RAM. InfoSphere Streams Version 3.2 has been configured and installed on the VM cluster. The cluster was available through the Southern Ontario Smart Computing Innovation Platform (SOSCIP) cloud, the first research-dedicated analytics cloud in Canada. The SOSCIP cloud also provides access to a broad range of software tools for application development and data analytics, which can be combined with user-specific software configurations to create customized VMs to meet project demands.

4.4 Implementation and Results

Two separate applications were developed for implementation of intensity level mapping in Streams. The first is for streaming and processing of data from physical sensors and the second from social sensors. Subsequently, both applications were linked into one application for joint processing from both sources.

The pipelined implementation to analyze datasets from seismic stations is shown on Figure 18 (details are provided in Algorithm 1, Appendix B).

Here, the input data are strong motion records from seismic network of USGS/NSMP (USGS National Strong-Motion Project) obtained from the Center for Engineering Strong Motion Data (CESMD, 2014), which provides raw and processed data for earthquake engineering applications in cooperation with the USGS and the California Geological Survey (CGS). The USGS/NSMP network contains 33 stations in the area of interest (mapped as triangles on Figure 19).

The duration of strong shaking during the earthquake was generally 10 to 15 seconds or less, recorded in time steps of 20 milliseconds. The minimum duration of the strong motion seismic time series was approximately 20 seconds while the maximum duration was approximately 80 seconds.

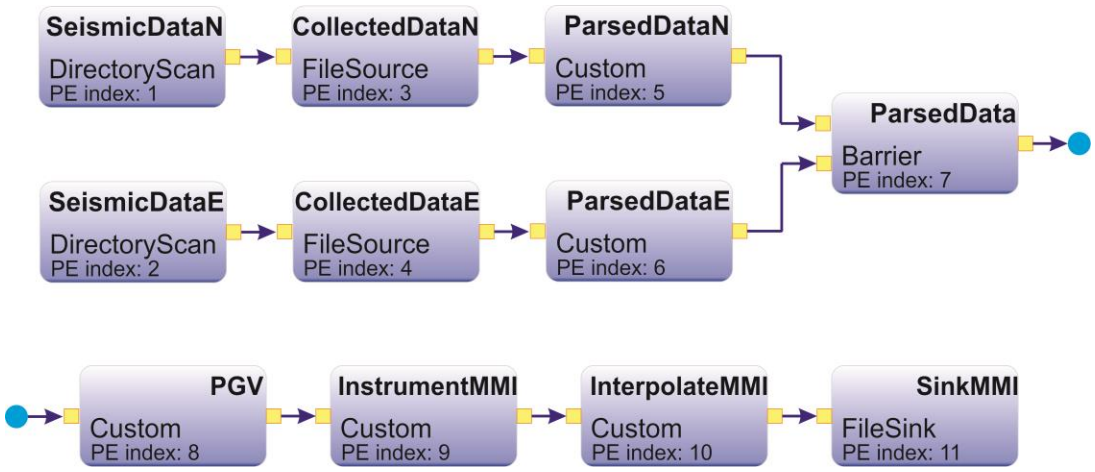

Figure 18: Pipelined application graph for intensity mapping based on seismic data (captured from Streams Studio).

The strong-motion velocity time series are processed and reviewed by USGS/NSMP (noise filtering, baseline or sensor offset corrections performed). However, the corresponding raw, unprocessed data also is available and could be processed here in a preprocessing step using the streaming paradigm. In order to simulate the real-time nature of the data processing as it would occur in an actual earthquake, we arranged the data input so that the start of each time series corresponded to the actual time recording began after the earthquake occurrence. The initiation of each time series was at different times, although they did overlap. The total duration of shaking in the streaming input was approximately four minutes.

The seismic data is in SMC format, which uses ASCII character codes and consists of text headers, integer headers, real headers, and time-series coordinates and values. The header includes information about the earthquake and the recording physical sensor. Each file also contains either a single time series of acceleration, velocity or displacement or a set of response spectra or Fourier amplitude spectra of corrected acceleration calculated from a single time series. Analog strong-motion records contain traces corresponding to three orthogonal components of motion that are located in three separate SMC-format files to represent the record (USGS, 2010).

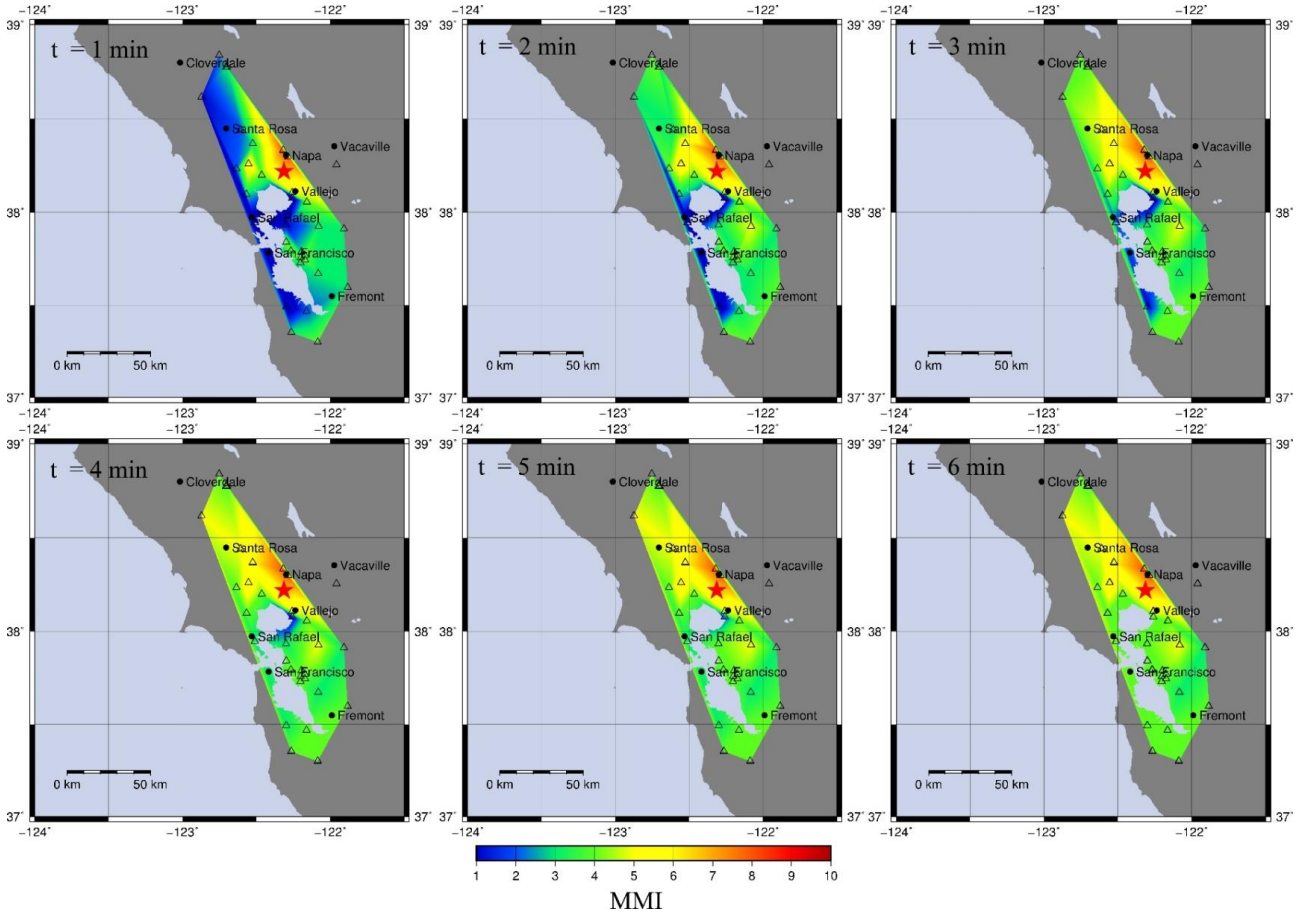

Figure 19: MMI based on physical sensors data (calculated using PGV values and empirical relationship from Atkinson and Kaka, 2007) estimated at one minute intervals after the Napa Valley earthquake. Red star represents the epicenter location. Triangles represent the location of seismic stations

Here the horizontal components (EW, NS) of velocity time series are used in the standard empirical relationship for PGV and MMI. From Atkinson and Kaka, 2007, the relationship between PGV and observed MMI can be represented by equations (1) and (2), as follows:

$$MMI = 4.37 + 1.32*(logPGV),\ logPGV \leq 0.48 \quad (1)$$

$$MMI = 3.54 + 3.03*(logPGV),\ logPGV \geq 0.48 \quad (2)$$

In our case these files have the following naming convention:

{StID.NComp.NP.--_v}.smc

where StID is the station identificator and Ncomp is the component name, which has one of three values: HNN is the north direction component, HNE is the east direction component, HNZ is the vertical component of record.

Two DirectoryScan operators control the streaming of the input data source into the horizontal components of the velocity time series, shown as PE 1 and PE 2 on Figure 18. In real situations, the implementation of these Streams source adaptors could be deployed to directly monitor the data collected at each seismic station, instead of pulling data from a data center, for real-time streaming processing. The first operator monitors the NS component files and the second monitors the EW component files. Subsequently, data from the input files is transformed to two streams by FileSource operators, shown as PE 3 and PE 4 on Figure 18. The next two operators, PE 5 and PE 6, are responsible for input stream parsing and extraction of station coordinates, time of records and north and east components of the velocity values. The PE 7 is used to synchronize the two streams coming from PE 5 and PE 6. In PE 8, PGV is estimated as the maximum horizontal component over a five seconds time interval. Next, the intensity calculation custom operator PE 9 calculates the MMI intensity level based on equations (1) and (2).

After the intensity level is derived at station locations, it is interpolated over a grid with spacing of three seconds in the latitude and longitude direction over the entire area, as limited by station endpoints (Figure 19). The interpolation operator PE 10 performs Delaunay triangulation for a set of points on a plane (Delaunay, 1934). This interpolation approach is chosen here in order to decrease the roughness of interpolated surface due to the very small number of data points. Finally, the sink operator PE 11 writes the results to the output file in csv format. The output file is refreshed as more data arrives, which can be plugged into a visualization tool to dynamically display MMIs in the area on-the-fly.

Figure 19 shows the output from Streams as the evolution of intensity estimates based on the seismic data, starting one minute after the Napa Valley earthquake and at one minute time intervals. However, as expected from the limited amount of data points available from the seismic stations (33 stations), the accuracy is low. Approximately one quarter

of the total area is covered by instrumental intensity results and the interpolated values are not smooth, with discontinuities not typically found in final intensity maps.

The pipelined implementation of application of Twitter data processing is shown on Figure 20 (details are provided in Algorithm 2, Appendix B).

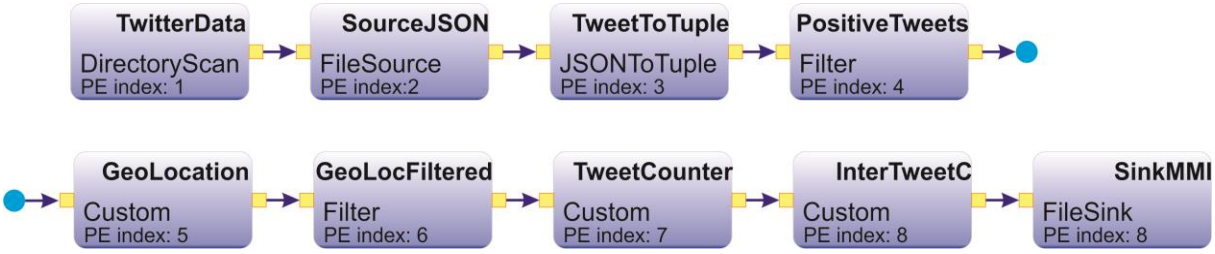

Figure 20: Pipelined application graph of intensity mapping based on Twitter data (captured from Streams Studio).

Twitter data has four main objects: Tweets, Users, Entities, and Places. The anatomy of these objects is complicated and has a number of attributes. The attributes needed for our implementation and for Twitter input data schema are:

a) coordinates – Tweet object attribute in geoJSON format that contains the geographic location of this Tweet reported by the user;

b) location – User object attribute, a string data type that contains the location for the account's user-defined profile;

c) created_at - Tweet object attribute, a string format that contains the UTC creation for each tweet;

d) text – Tweet object string attribute, which is the actual UTF-8 text of the status update (Twitter, 2015).

Each tweet in the archive has a creation time attribute with one second accuracy, allowing for the creation of a time series with one second time steps. This time series was fed into Streams, starting at the time of earthquake occurrence, again in order to simulate real-time processing of the information.

Input Twitter data coming from the source is monitored by DirectoryScan operator PE 1 (Figure 20) and initially read in JavaScript Object Notation (JSON) format PE 2, an open standard format used to transmit data objects consisting of attribute–value pairs. The data is transformed from JSON to defined Streams tuples in PE 3. The streaming tuples are filtered for positive tweets when passing through operator PE 4. Not all of the positive tweets have a coordinates attribute. For those positive tweets without a specified location, the location parsing operator PE 5 assigns coordinates according to the geolocation algorithm presented in Section 2. In this case, the tweet coordinates obtained are approximate. Subsequently, these positive tweets are filtered PE 6 again in order to exclude tweets from other regions. The logarithmic number of positive tweets per minute is calculated every five seconds PE 7, which is achieved in Streams by applying a "sliding time window" to data streams when passing through an operator, i.e. the time window is in size of 1 minute and the sliding step is 5 seconds. In step PE 8, the logarithmic number of tweets at every location is interpolated over a spatial grid of ten minutes in the latitude and longitude direction for the entire area, using the gridding algorithm with continuous splines in tension of Smith and Wessel (1990). Finally, sink operator PE 9 outputs the results in csv format, which can be plugged into a visualization tool to display.

Figure 21 shows the results obtained from Streams as the logarithmic number of tweets per minute starting from one minute after the earthquake at one minute intervals. For instrumental intensity (Figure 19), the maximum MMI for the entire area occurs at the seventh minute. Figure 21 shows that the logarithmic number of tweets per minute increased by a factor of four over the ten minute time interval of interest. Also, unlike instrumental intensity, the values are more completely distributed in space (comparing Figures 19 and 21).

The pipelined implementation of joint processing is shown on Figure 22 (details are provided in Algorithm 3, Appendix B). Both data sets and all the operators from the above two applications explained above are employed in this analysis. However, because that data is arriving continuously from both sources at the same time, calculations occur at smaller spatial and temporal discretization. Again, the logarithm of the number of

tweets per minute are assumed to be correlated with MMI. The correlation analysis PE 17 was performed using the three-segment linear prediction equation (Table 4), the predictive model that resulted from the study presented in Section 2. MMI was estimated at every location in the joint streaming analysis of PE 17. In step PE 18, the resulting values of MMI are interpolated on a grid spacing of three seconds in the latitude and longitude directions (Smith and Wessel, 1990).

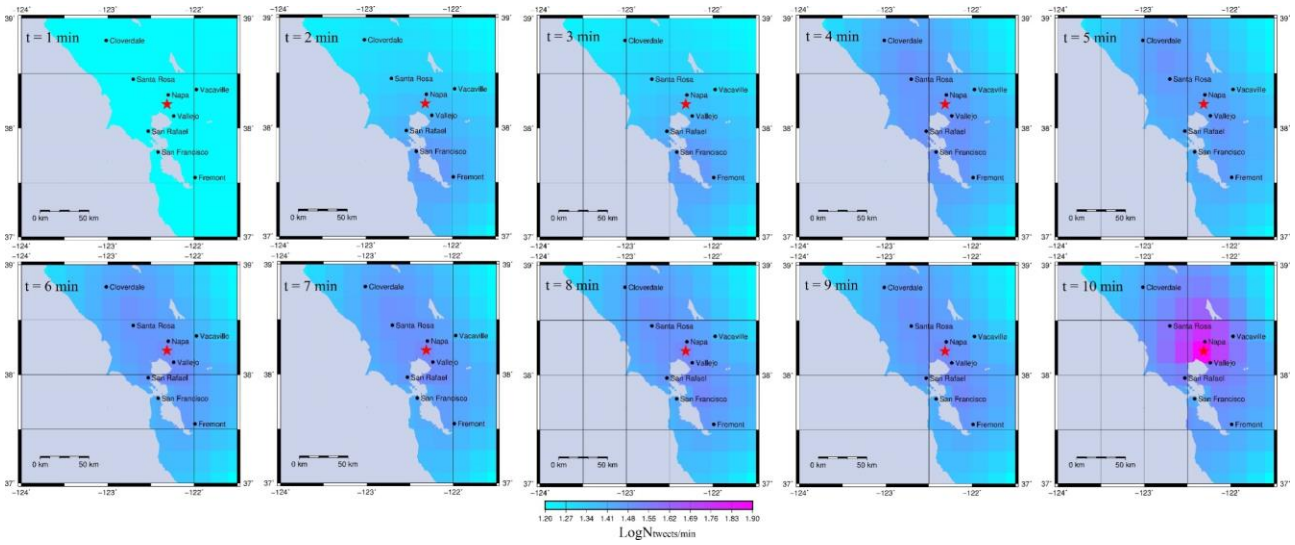

Figure 21: Logarithmical number of tweets at one minute intervals after the Napa Valley earthquake. The red star represents the epicenter location.

The maps represent the MMI level starting from one minute after earthquake up to ten minutes, at one minute intervals. To quantify the improvement between maps generated based on data just from physical sensors (Figure 19) and those obtained from joint processing from physical and social sensors (Figure 23) we calculated the RMS error between those maps obtained at the final minute for physical sensors data (six minutes after the earthquake) and the USGS instrumental intensity map provided in Figure 12. Note that the RMS error calculated here is different from that used to select the best model (Table 4).

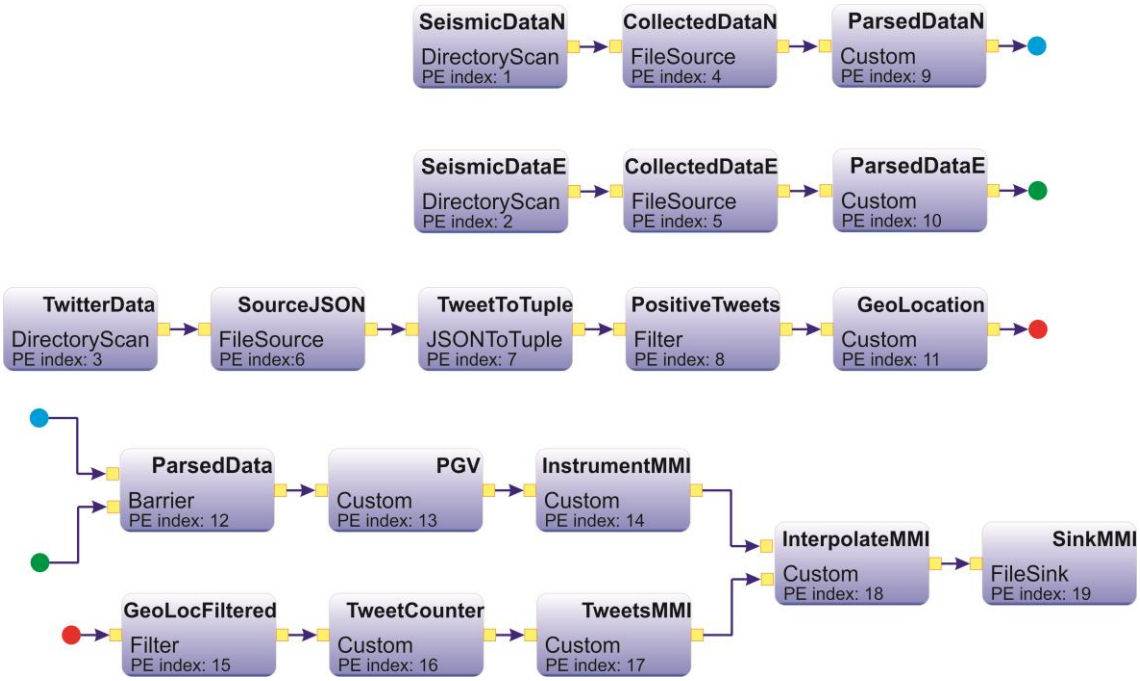

Figure 22: Pipelined application graph of intensity mapping based on joint data (captured from Streams Studio).

The RMS error for the maps using only instrumental intensity is 2.8 MMI units. In contrast, for the RMS for the jointly processed intensity map is 0.58 MMI units. The lower RMS, and the improvement in resolution, for the joint processing case is a result of the increased number of data points used on the final interpolation step compared to those of the maps shown in Figure 19. Results from the joint intensity calculation for this case study are shown in Figure 23.

To measure performance characteristics we used the maximum values of the InfoSphere Streams built-in metric for resource utilization monitoring for every host node on the cluster (see Table 5).

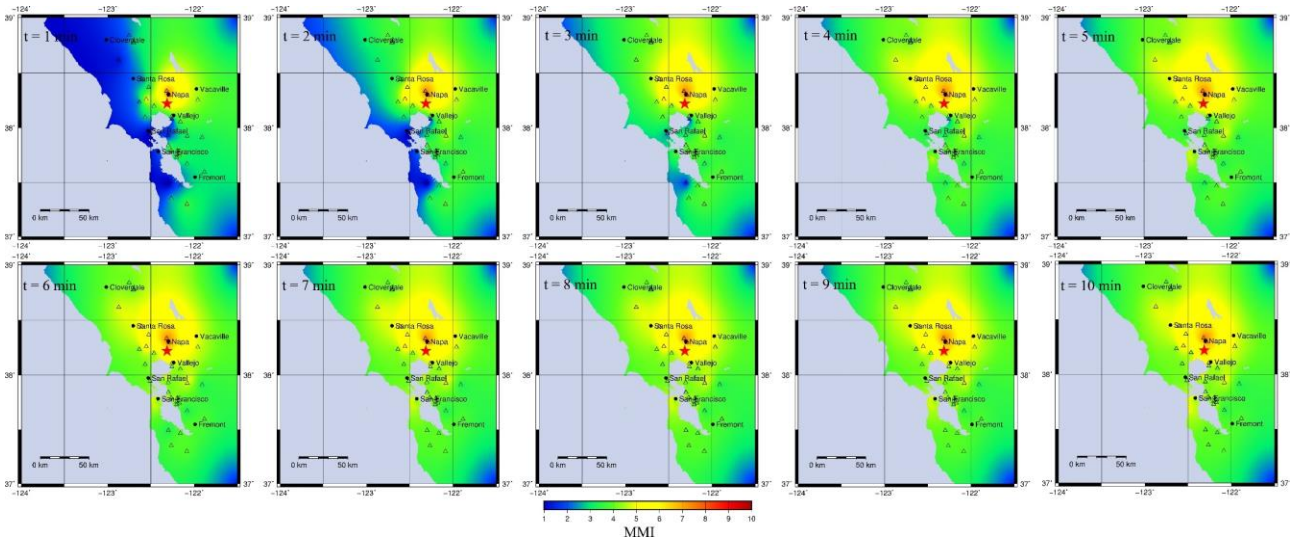

Figure 23: MMI after joint data processing from physical and social sensors at one minute intervals after the Napa Valley earthquake. Red star represents the epicenter location. Triangles represent the seismic station locations.

CPU in milliseconds (ms) represents the time that was used by all PEs on a node. The memory consumption metric shows the amount used by the all PEs on a node, in kilobytes (kB). Load average is a common metric on Linux systems that measures the average CPU load over a period of time. A higher metric represents a system that is increasingly overloaded. InfoSphere Streams multiplies the raw load average from Linux by 100, and normalizes it by the number of processors on each cluster in that instant. In our case all four hosts were loaded to a level that was less than moderate, providing evidence that the processing workload involved in this particular case could be handled by fewer computational resources for the same performance level.

Table 5: Performance metrics over 10 minute time interval

Node name on the cluster	CPU (ms)	Memory consumption (kB)	Load average
Cluster-parent	9,542	1,682,004	4
Cluster-child1	10,880	2,654,176	1
Cluster-child2	9,530	1,197,148	26
Cluster-child3	7,800	1,196,550	1

The purpose of this Streams implementation and the execution experiment here is to demonstrate its viability and potential real-time performance. In addition, it allows data to

be processed continuously on-the-fly while it being collected from multiple sources, which is different from the traditional data processing implementation.

4.5 Conclusion

This work presents the successful application and potential for emergency management purposes using the analytics cloud. The streaming concept is applied using the IBM InfoSphere Streams product using multiple data sources from an actual earthquake in order to demonstrate its application in real-time hazard mapping.

Joint streaming processing using tweets and seismic records that were recorded within the ten minutes following the Napa Valley earthquake (2014) was used to estimate MMI at particular sites in California. We demonstrate that the logarithmic number of tweets can be used as a proxy for shaking intensity and can be used as a supplementary data source, in conjunction with existing networks of physical sensors, such as seismic stations, in order to improve intensity estimates in real-time. Results from the joint analysis show that it provides more complete coverage, with better accuracy and higher resolution over a larger area than either data source alone. Future studies will examine the extent to which the twitter regression relation is applicable in other seismic areas, both in California and worldwide, and how potential errors in geolocation affect the accuracy in space and time.

In many areas, the importance of this additional data source could be very significant, due to the complete or partial lack of traditional data sources as a result of the high cost of their installation and ongoing operation. This work demonstrates that Twitter data could be used as independent data source of MMI estimation. In particular, it has significant potential for regions that may not have an extensive seismic network. Finally, implementation of this approach is not event or location dependent, so this approach could be applied to other regions and types of hazard. For example, it could be employed in areas where access is restricted due to flood inundation. Incorporation of social sensor data with traditional data sources using advanced computational processing methods can provide more complete and accurate coverage for rapid hazard response.

4.6 References

Amante, C. and B.W. Eakins, 2009. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. *NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA*. doi:10.7289/V5C8276M [Accessed 10/10/2015].

Atkinson, G. & Kaka, S., 2007. Relationships between felt intensity and instrumental ground motions for earthquakes in the central United States and California. *Bull. Seism. Soc. Am.*, Issue 97, pp. 497-510.

Burks, L., Miller, M. & Zadeh, R., 2014. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. s.l., *Tenth U.S. National Conference on Earthquake Engineering Frontiers of Earthquake Engineering*.

Campagne, J., Dux, J., Guyot, P. & Julien, D., 2012. Twitter reaches half a billion accounts - More than 140 million in the U.S.. [Online] Available at: http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

Center for International Earth Science Information Network - CIESIN - Columbia University, and Centro Internacional de Agricultura Tropical - CIAT. 2005. Gridded Population of the World, Version 3 (GPWv3): Population Density Grid. Palisades, NY: *NASA Socioeconomic Data and Applications Center (SEDAC).* http://dx.doi.org/10.7927/H4XK8CG2. Accessed 10/10/2015

Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J., 2012. Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS* , 17(1), pp. 124-147.

Delaunay, B., 1934. Sur la sphère vide. A la mémoire de Georges Voronoï. Bulletin de l'Académie des Sciences de l'URSS, *Classe des sciences mathématiques et naturelles*, Issue 6, pp. 793-800.

Earle, P., Bowden, D. & Guy, M., 2011. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), pp. 708-715.

Earle, P. et al., 2010. OMG Earthquake! Can Twitter Improve Earthquake Response?. *Seismological Research Letters*, 81(2), pp. 246-251.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographe*r, 66(4), 568–578. http://doi.org/10.1080/00330124.2014.907699

IBM, 2014. IBM Knowledge Center. [Online] Available at: http://www-01.ibm.com/support/knowledgecenter/[Accessed 2015].

Northern California Earthquake Data Center, 2014. *UC Berkeley Seismological Laboratory.* Dataset. doi:10.7932/NCEDC.

Nepal Disaster Risk Reduction Portal, 2015. Incident Report of Earthquake 2015. [Online] Available at: drrportal.gov.np

Richter, C., 1958. Elementary Seismology. *San Francisco: Freeman.*

Sakaki, T., Okazaki, M. & Matsuo, Y., 2010. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. Raleigh, NC, *World Wide Web Conference (WWW)*.

Sakaki, T., Okazaki, M. & Matsuo, Y., 2013. Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development, *IEEE Trans. Knowledge and Data Eng,* vol. 25, no. 4, pp. 919-931

Severo, M., Giraud, T., & Pecout, H. (2015). Twitter data for urban policy making: an analysis on four European cities. In C. Levallois (Ed.), *Handbook of Twitter for Research*. EMLYON

Smith, W. H. & Wessel, P., 1990. Gridding with continuous curvature splines in tension. *Geophysics*, Issue 55, pp. 293-305.

Takhteyev Y., Wellman B., Gruzd A., 2012. Geography of Twitter Networks, *Social Networks*, Volume 34, Issue 1, January 2012, pages 73–81

Terry, D., Goldberg, D., Nichols, D. & Andoki, B., 1992. Continuous queries over append-only databases. *SIGMOD*, pp. 321-330.

The Center for Engineering Strong Motion Data, 2014. *CESMD Internet Data Report.* [Online] Available at: http://www.strongmotioncenter.org/cgi-bin/CESMD/archive.pl

Twitter, 2015. The Twitter Platform Documentation. [Online] Available at: https://dev.twitter.com/overview/documentation [Accessed 2015].

United States Geological Survey, 2010. National Strong Motion Project. [Online] Available at: http://escweb.wr.usgs.gov/nsmp-data/smcfmt.html [Accessed 2015].

United States Geological Survey, 2014. M6.0 - 6km NW of American Canyon, California. [Online] Available at: http://earthquake.usgs.gov/earthquakes/eventpage/nc72282711#general_summary

United States Geological Survey, 2015a. Earthquake Facts and Statistics. [Online] Available at: http://earthquake.usgs.gov/earthquakes/eqarchives/year/eqstats.php [Accessed 2015a].

United States Geological Survey, n.d. The Modified Mercalli Intensity Scale. [Online] Available at: http://earthquake.usgs.gov/learn/topics/mercalli.php [Accessed 2015b].

United States Geological Survey, n.d. The San Andreas and Other Bay Area Faults. [Online] Available at: http://earthquake.usgs.gov/regional/nca/virtualtour/bayarea.php [Accessed 2015c].

Wald, D., Quitoriano, V. & Dewey, J., 2006a. USGS "Did you feel it?" community internet intensity maps: macroseismic data collection via the internet. Geneva, Switzerland, *First European Conference on Earthquake Engineering and Seismology*.

Wald, D., Worden, B., Quitoriano, V. & Pankow, K., 2006b. ShakeMap Manual: Technical Manual, Users Guide, and Software Guide, *Boulder: United States Geological Survey.*

Watson, D. F., 1981. Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *Comput. J.*, 24(2), p. 167–172.

Wood, H. & Neumann, F., 1931. Modified Mercalli Intensity Scale of 1931. *Seismological Society of America Bulletin*, 21(4), pp. 277-283.

Chapter 5

5 The Predictive Relationship between Earthquake Intensity and Tweets Rate for Real-Time Ground Motion Estimation

The material covered in this chapter is submitted to the Seismological Research Letters journal on July 19th 2016.

The standard measure for evaluation of immediate effect of an earthquake on Earth's surface, people, and man-made structures is intensity. Intensity estimates are widely used for emergency response, loss estimation and distribution of public information after earthquake occurrence. Intensity measures are designed to standardize the measurements of seismic effect and their subsequent evaluation and response (Wald et al., 2003). Modern intensity assessment procedures process a variety of information sources. Those sources are primarily from two main categories: physical sensors (seismographs and accelerometers) and social sensors (witness observations of the earthquake consequences). Acquiring new data sources in the second category can help to speed up the existing procedures of the intensity calculation and improve the accuracy of those assessments in a more timely fashion. One potentially important data source in this category is the widespread micro-blogging platform Twitter, ranked number nine worldwide as of January 2016 by number of active users, ~320 million (Twitter, 2016). In our previous studies, empirical relationships between positive tweets rate and observed Modified Mercalli Intensity (MMI) were developed using data from the M6.0 South Napa, CA earthquake (Napa earthquake) that occurred on August 24, 2014 (Kropivnitskaya et al., 2016). These relationships allow us to stream data from social sensors, supplementing data from physical sensors in order to produce real-time intensity maps. The streaming application implementation is based on IBM InfoSphere Streams, a cloud platform for real-time analytics on big data. These relationships could potentially decrease latency in the intensity calculations for future earthquakes in California and in other places around the world. However, there is a strong need for their validation and calibration in regions other than California. In this study, we validate empirical

relationships between tweets rate and observed MMI using new data set from earthquakes that occurred in California, Japan and Chile during the period March-April 2014. The statistical complexity of the validation test and calibration process is complicated by the fact that the Twitter data stream is limited for open public access, reducing the number of available tweets. In addition, in this analysis only spatially limited positive tweets (marked as a tweet about earthquake) are incorporated into the analysis, further limiting the data set and restricting our study to a historical data set. In this work, the predictive relationships for California is recalibrated slightly and a new set of relationships is estimated for Japan and Chile.

5.1 Introduction

Earthquake intensity is a location-specific characteristic of the post-seismic damage effect that depends not only on the magnitude of the earthquake, but also on the distance from the earthquake epicenter to the site of interest and the geological features of the surrounding area (USGS, 2015). The traditional understanding the earthquake intensity is based on estimates of the infrastructure damage at the specific site of interest estimated from the subjective perception of professional observers and/or volunteers who witnessed the earthquake and its consequences (USGS, 2013). For this purpose, specially designed questionnaire are used to be filled out and summarized at certain locations. A successful example of the modern implementation of this approach is the Did You Feel It? program created by United States Geological Survey (USGS) that collects information from people who experience an earthquake and volunteer to share their observations online to create Community Internet Intensity Maps (CIIM) with observations and extent of damage (Wald et al., 2006).

In addition to the intensity evaluation method based on human observations, another approach is to determine intensity from the peak ground motion, either velocities or accelerations, at a station nearby the site of interest. Empirical relationships are then employed to calculate intensity level (Wald et al., 1999). Results can be obtained much faster by comparing the observation-based analysis and the instrumental intensity level. A successful realization of this method is the ShakeMap application, also developed by the USGS (Wald et al., 1999).

The intensity evaluation approaches can be divided into two main categories, according to the input information source type, above. The first evaluation category employs social sensor data from people who witness the consequences of the earthquake. The second utilizes data from physical sensors, such as seismometers and accelerometers to estimate intensity. Although the first method historically estimated intensity at a much slower rate than that of physical sensors, today electronic questionnaires and observers' reports can be supplemented with auxiliary online data sources from social networks, where people also share their observations after the earthquakes. In our case, we access data from the online social networking service Twitter. Twitter enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets while unregistered users can only read those tweets. Users can access Twitter through the website interface, SMS or a mobile device application that is the true source of the real-time nature of tweets (Twitter, 2016). Twitter data is immediately available in a data stream, which can be mined using stream mining techniques (Schonfeld, 2009). In this study we work with data following the stream model. In this model, data arrive at high speed and data mining algorithms must be able to predict intensity level in real-time and under strict constraints of space and time (Bifet et al., 2010).

Previously, studies have shown Twitter data potential for earthquake detection (Earle et al., 2010, 2011; Sakaki, 2010; Crooks, 2012; Burks et al., 2014) and intensity estimation (Kropivnitskaya et al., 2016). Kropivnitskaya et al. (2016) created four empirical predictive relationships (linear, two-segment linear, three-segment linear, exponential) that link the positive tweet rates in the first ten minutes following the earthquake with the instrumental intensity level in MMI scale units using regression analysis of data from physical and social sensors during the Napa earthquake. Figure 24 shows the ratio between the combined data (instrumental and Twitter data) and the MMI intensity values records reported by the USGS intensity levels in the Napa region. Despite the fact that the ratio between estimated and actual intensity is relatively low for this particular earthquake and that this joint processing of social and physical data demonstrates significant high potential of proposed models forin the near real-time predictive streaming applications, the empirical relationships between earthquake intensity and

tweet rates models still needs to be validated for all of California and other seismically active regions of the world in the world and spatially calibrated if necessary.

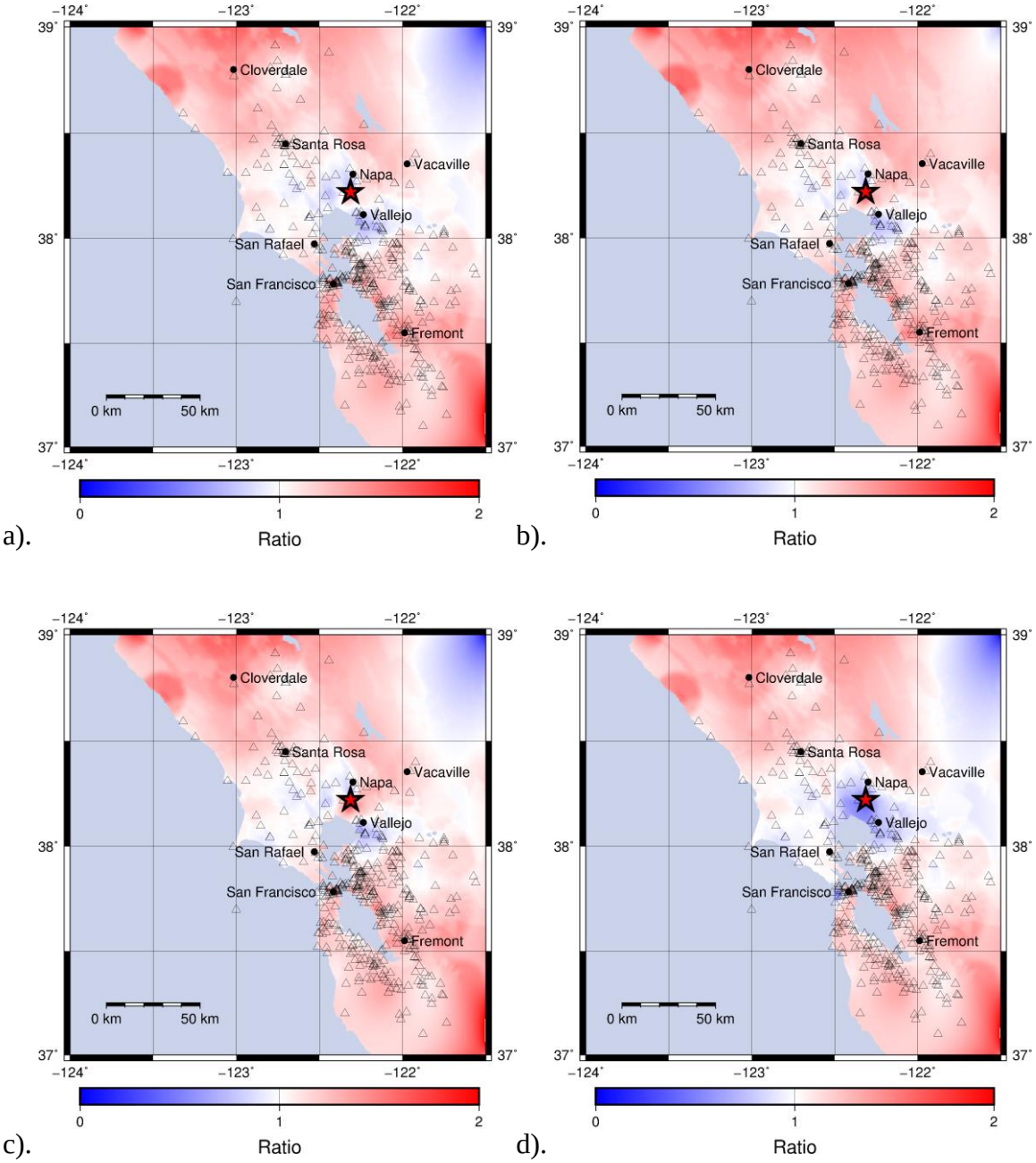

Figure 24: Ratio between combined intensity level (from physical and social sensors) and instrumental intensity level (triangles – seismic stations) after the Napa earthquake (red star - epicenter). a) Linear model, b) exponential model, c) two-segment model and d) three-segment model.

The statistical complexity of validating and calibrating the model is complicated by the fact that Twitter data stream is limited for open public access. The basic levels allow only up to 1% of the total tweets volume to be streamed (Twitter, 2016). For our purposes only spatially limited positive tweets (tweets about earthquake) are used, so the rate limitations are critical and historical data are used in our application instead of a real data stream. The relationships are validated and calibrated for three regions: California, Chile and Japan. The degree of social media engagement across these countries is relatively high. The proportion of each country's population that has a Twitter account is 36% in the USA, 24% in Japan and 33% in Chile (Dawson, 2012). Therefore, potential for the success of Twitter streaming applications in these regions is high.

5.2 Data preparation

In order to validate whether the numerical results quantifying hypothesized relationships between logarithmic tweet rates following ten minutes after an earthquake and earthquake intensity on MMI scale, obtained from Kropivnitskaya et al. (2016), can be used in other regions and for other earthquakes in California, we selected independent events for testing purposes (Table 6).

Table 6: List of earthquakes used in validation and calibration processes

Date and Time	Magnitude	Depth (km)	Epicenter Location
2014-03-02 20:11	6.5	119	111km NNW of Nago, Japan
2014-03-10 5:18	6.8	16.6	78km WNW of Ferndale, California
2014-03-13 17:06	6.3	79	15km NNE of Kunisaki-shi, Japan
2014-03-16 21:16	6.7	20	64km WNW of Iquique, Chile
2014-03-17 5:11	6.4	21	80km WNW of Iquique, Chile
2014-03-22 12:59	6.2	20	91km WNW of Iquique, Chile
2014-03-29 4:09	5.1	5.1	2km E of La Habra, California
2014-04-01 23:46	8.2	25	94km NW of Iquique, Chile

Note that the magnitude scale used in this study is moment magnitude based on the seismic moment of the earthquake, which is equal to the rigidity of the Earth multiplied by the average amount of slip on the fault and the size of the area that slipped (USGS, 2016). Those events listed in the Table 6 are shown in Figure 25 (California), Chile

(Figure 26) (Chile) and Japan (Figure 27) (Japan). Note that none of these events were included in the original analysis (Kropivnitskaya et al., 2016).

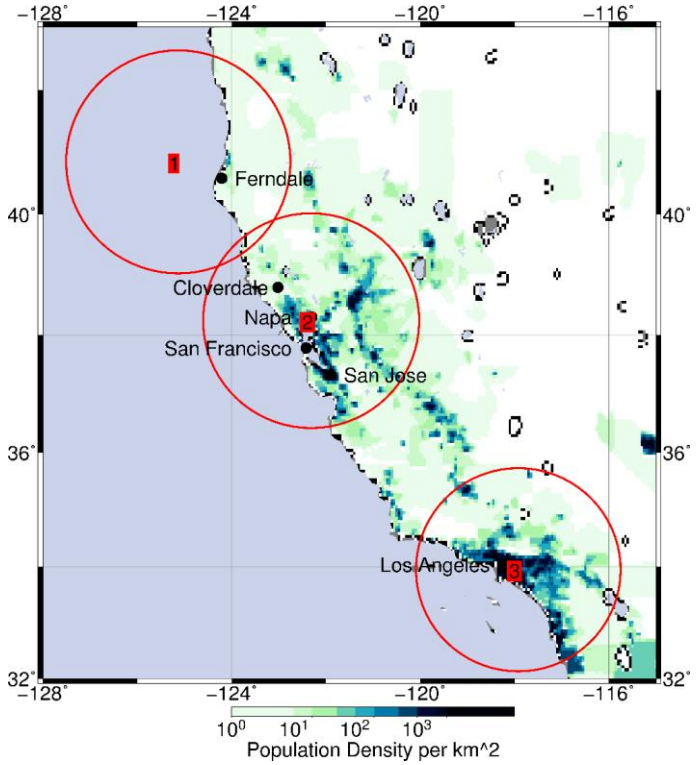

Figure 25: California population density map with earthquake epicenters used in the validation and calibration process (1 – Ferndale earthquake, 2 – Napa Valley earthquake, 3 – La Habra earthquake) and areas covered for analysis (circles).

The selection of seismic events used in validation is dictated by two constraints. The first is to the result of the limited number of freely available, Twitter data sets of unlimited volume. Gathering information from social media feeds is, in essence, a web-mining process (Kosala and Blockeel 2000, Sakaki et al., 2010, Russell 2011). It entails three operations: extracting data from the data providers (in this case Twitter) via application programming interfaces (APIs); parsing, integrating, and storing these data in a resident database; and then analyzing these data to extract information of interest. However, currently available Twitter tools offer limited capabilities for information gathering procedures (Twitter, 2016). As a result we used archived historical dataset of Twitter

records from March-April 2014 (downloaded from https://archive.org/details/twitterstream).

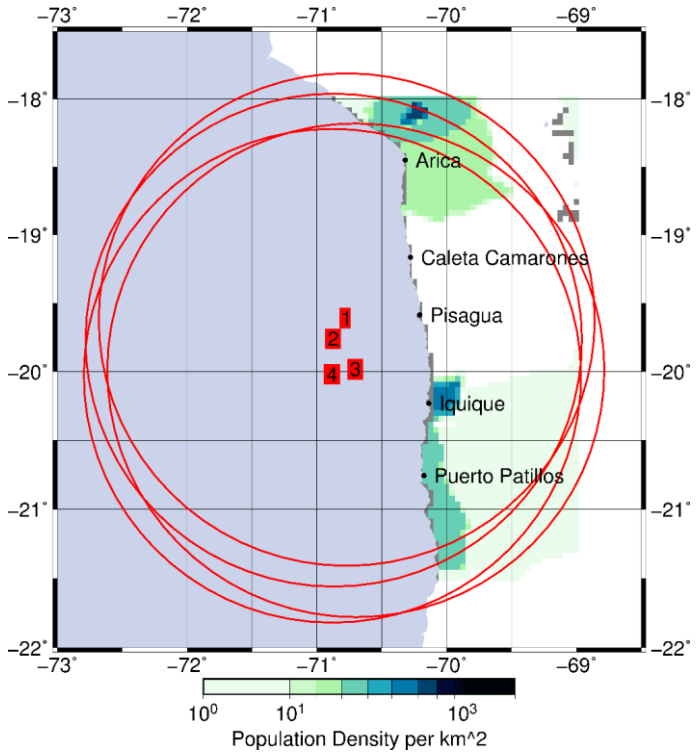

Figure 26: Chile population density map with earthquake epicenters used in the validation and calibration process (1 - 64km WNW of Iquique, 2 - 80km WNW of Iquique, 3 - 91km WNW of Iquique, 4 - 94km NW of Iquique) and areas covered for analysis (circles).

The second limitation is related to the lack of geotagged data within available Twitter data sets. We selected seismic events that both showed a spike in the amount of positive tweets within ten minutes following the earthquake and where at least fourteen of those tweets were geotagged (the minimum of valid range of values in the models used (Kropivnitskaya et al., 2016)). The geotagged tweets contain the current user location indicator at the time of tweeting and can be directly used in the location-specific intensity estimation algorithms. However, Graham et al. (2014) showed that only 0.7% tweets are geotagged among 19.6 million tweets. In this case, other tweets that do not have a direct specific location reference, but contain a link to specific cities can be used.

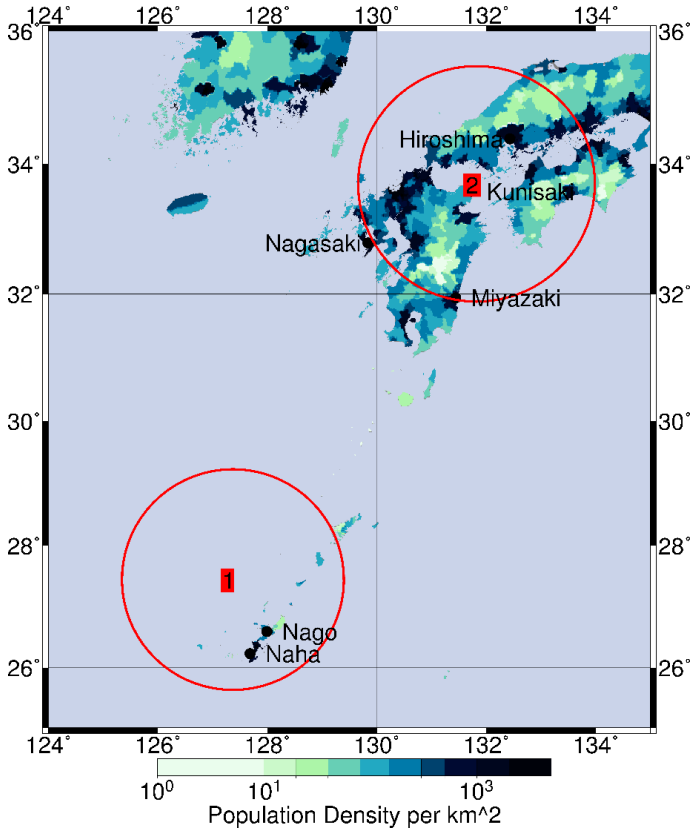

Figure 27: Japan population density map with earthquake epicenters used in the validation and calibration process (1 – Nago earthquake, 2 - Kunisaki-shi) and areas covered for analysis (circles).

The percentage of geotagging in this case is between 2% and 5% (Severo et al., 2015). The geotag also can be obtained from a field in the user account description (7.5% of profiles contain latitude and longitude values, 57% include a named location, 20.4% referenced information that can be used to identify a country, while 15.1% provided humorous or non-spatial information (Takhteyev et al., 2012)). In the approach used by Kropivnitskaya et al. (2016), all three types of geotagged technics were used and the same logic has been implemented here. A text-based geolocation algorithm has been improved by employing a location database extension for California, Japan and Chile. In location database updating, the complete, shortened or abbreviated names are included for any settlements with a population of more than five thousand people in a 200 km radius of the epicenter of the earthquakes listed in Table 6 (see Figures 25 through 27).

After building a tweets data set for each event that is limited in time and space, we generated a tweet-frequency time series with one second windows time bins and normalized to number of tweets per minute. For each point from the Twitter data set we automatically assigned an MMI intensity value from the instrumental intensity database of the USGS National Earthquake Information Center corresponding to the same (closest) latitude-longitude location. For California we selected the M6.8 Ferndale, CA earthquake (March 10, 2014) and the M5.1 La Habra, CA earthquake (March 29, 2014) (Figure 25). The four prediction models (Kropivnitskaya et al., 2016) shown on Figure 28.

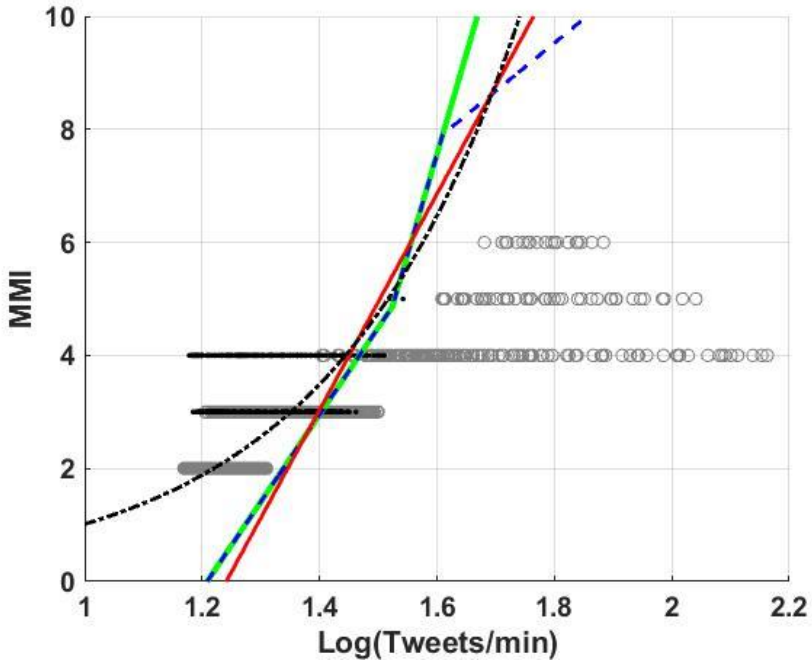

Figure 28: Observed California earthquake data (black dots – Ferndale earthquake, grey circles – La Habra earthquake) with prediction models obtained for Napa earthquake (red – linear, black - exponential, green – two-segment linear, blue – three-segment).

5.3 Validation of Relationship

The validation process analyzes the goodness of fit of the regression for Napa earthquake, determining whether the regression residuals are random, and checking whether the model's predictive performance deteriorates substantially when applied to data for the

Ferndale and La Habra earthquakes not used in the original model estimation. Both events are not large enough to validate empirical relationships at intensity levels higher than VI. The Ferndale earthquake is an offshore event that occurred near areas of relatively low population. The La Habra event had a relatively small magnitude and occurred in areas of generally well-mitigated urban seismic risk. As a result, the three-segment model is excluded from the validation test.

An important observation can be made from Figure 28. Both earthquakes' datasets show that the positive tweets rate increases slower than predicted by the empirical relationships for the Napa earthquake derived in Kropivnitskaya et al. (2016).

The residuals for both events are shown with whisker diagrams for different models in Figures 29 and 30, and display the distribution of the residuals based on the five number summary: minimum, first quartile, median, third quartile, and maximum.

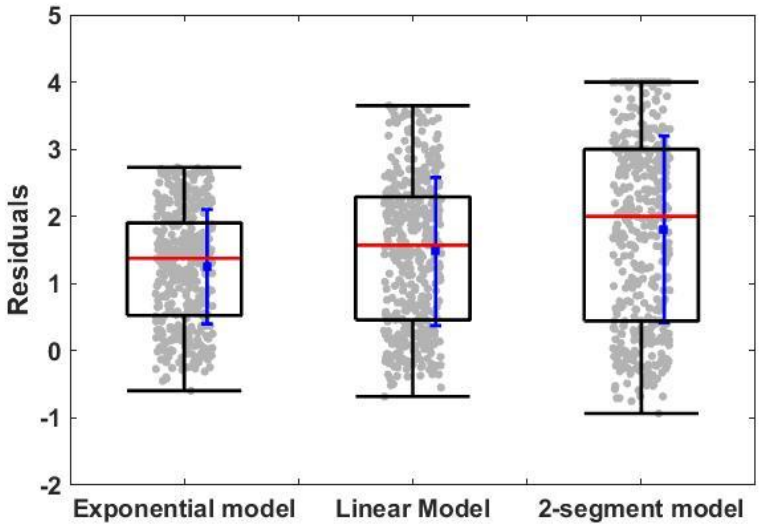

Figure 29: Whisker diagram for the Ferndale earthquake residuals (red – median, blue square – mean, error bars – standard deviation).

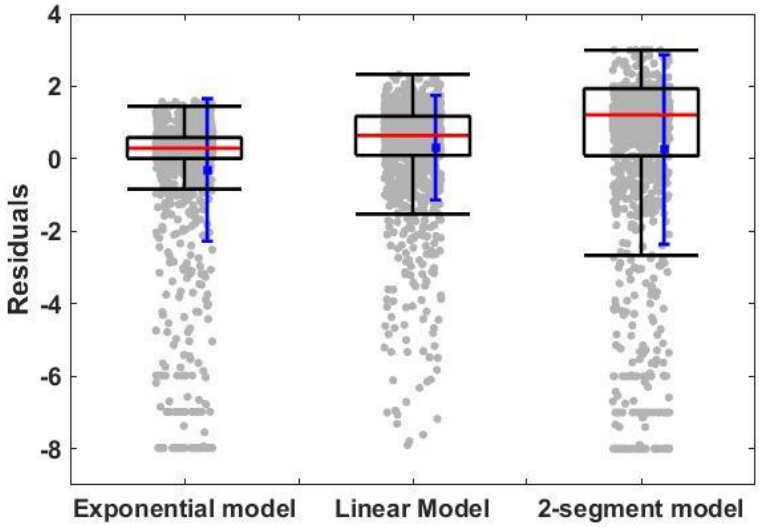

Figure 30: Whisker diagram for the La Habra earthquake residuals (red – median, blue square – mean, error bars – standard deviation).

The quartiles of the present data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. The first quartile is defined as the middle number between the smallest number and the median of the data set. The median of the data is a second quartile. The third quartile is the middle value between the median and the highest value of the data set. The Interquartile Range (IQR) is used here to characterize outliers that skew the data.

Despite the fact that the La Habra data (Figure 30) contains significantly more outliers in the residuals than the Ferndale data (Figure 29). The mean residuals for every model for the La Habra event are much closer to zero (the linear model mean residual is 0.37 MMI units; the exponential model mean residual is -0.25 MMI units; the two-segment model mean residual is 0.28 MMI units) than for Ferndale event (the linear model mean residual is 1.28 MMI units; the exponential model mean residual is 0.98 MMI units; the two-segment model mean residual is 1.57 MMI units).

A visual examination of the residuals has an advantage over numerical model validation methods because it illustrates the complex aspects of the relationship between the model

and the new data. Figures 31-36 show the residuals from the fitted models, providing information on the different features of the model.

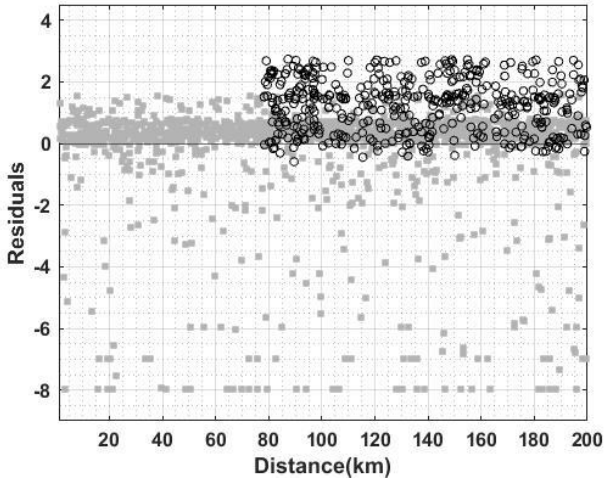

Figure 31: Residuals vs. epicentral distance for the exponential model (grey squares – La Habra earthquake, black circles – Ferndale earthquake).

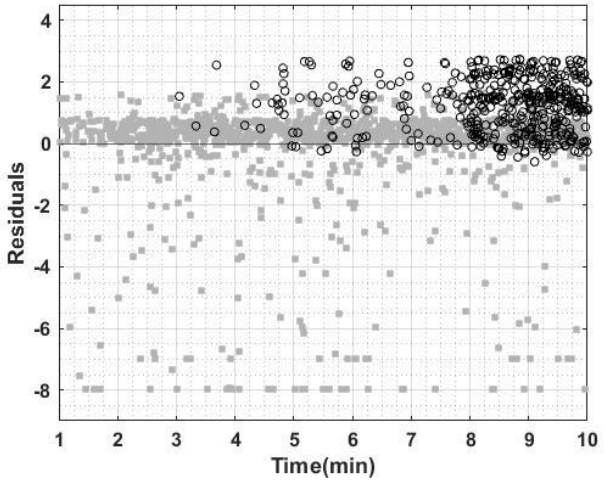

Figure 32: Residuals vs. time for the exponential model (grey squares – La Habra earthquake, black circles – Ferndale earthquake).

The residuals are not randomly distributed around zero, indicating that the linear assumption may be not reasonable (Figure 32). However, even for the exponential model case (Figure 31), the residuals are not distributed normally around zero. In this case, the variances of the error terms at each MMI level are not equal due to the different number

of twitter responses at each level. Moreover, some of the residuals stand out from the basic pattern, confirming there are outliers in the data as indicated by whisker diagrams (Figures 29 and 30). These data points have to be excluded from the calibration procedure detailed below.

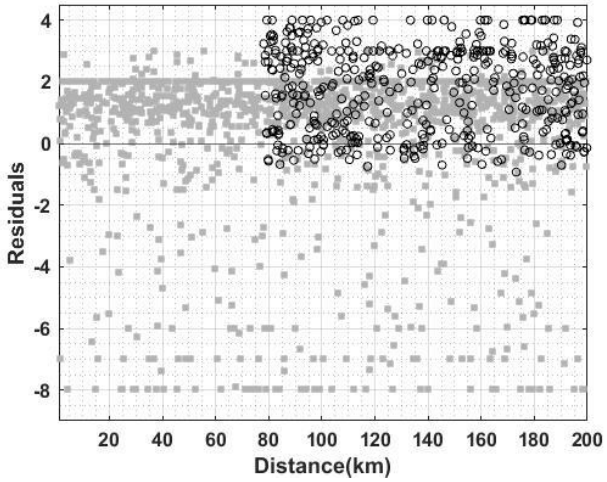

Figure 33: Residuals vs. epicentral distance for the linear model (grey squares – LaHabra earthquake, black circles – Ferndale earthquake).

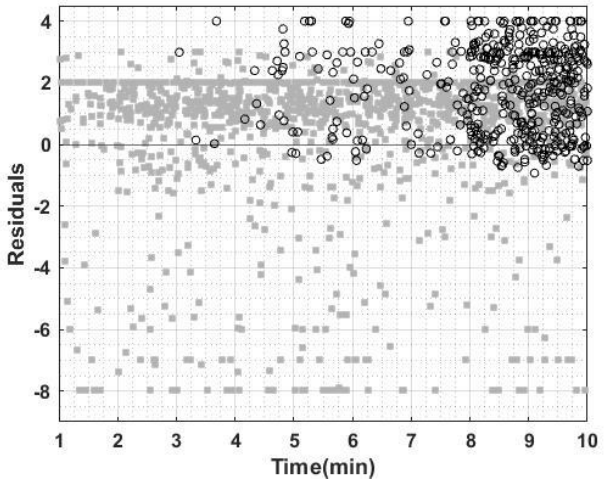

Figure 34: Residuals vs. time for the linear model (grey squares – La Habra earthquake, black circles – Ferndale earthquake).

Intensity attenuates with distance from the epicentre of an earthquake and intensity prediction equations usually rely on some distance metric. For example, some models use

epicentral distance (Bakun and Wentworth, 1997), or closest distance to rupture, or some variant that considers extended fault sources, such as Joyner-Boore distance (Joyner and Boore, 1993). For small magnitude earthquakes, where an earthquake can be approximated by a point source, the difference between point and extended source distance metrics can be minimal.

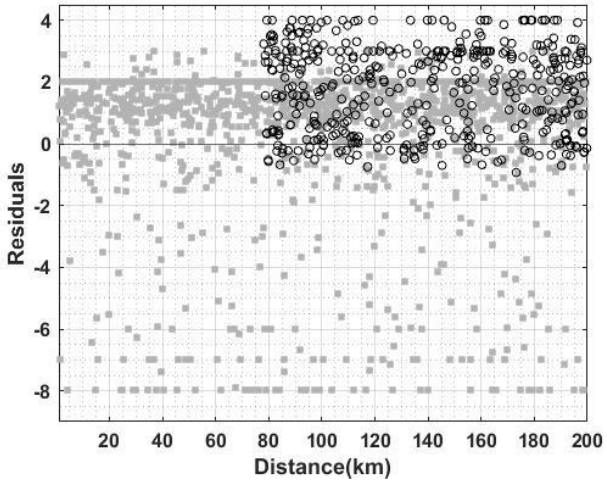

Figure 35: Residuals vs. epicentral distance for the two-segment linear model (grey squares – La Habra earthquake, black circles – Ferndale earthquake).

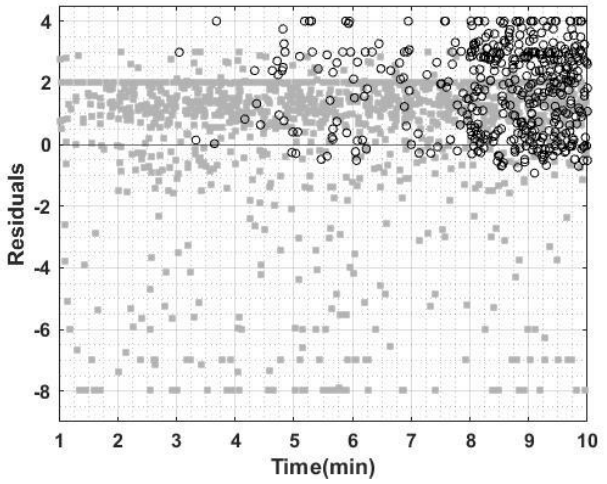

Figure 36: Residuals vs. time for the two-segment linear model (grey squares – La Habra earthquake, black circles – Ferndale earthquake).

However, at larger magnitudes where source finiteness can be significant, prediction equations which use point source distance metrics are not likely to be applicable (Cua et al., 2010). In this evaluation, we consider the point source distance metric epicentral distance as a potential predictor for the equations of Kropivnitskaya et al. (2016). The expected distance metric influence on the predictive equations is based on the assumption that number of positive tweets and tweet rates are expected to attenuate with distance from the epicenter. The reason is in different components of ground motion that may vary with epicentral distance and cause a variation in the type of structural damage. Plots of the residuals versus potential predictor calculated variable epicentral distance as potential predictor (Figures 31, 33, and 35) do not exhibit a systematic structure and confirm that the form of the function cannot be improved using that predictor. To check non-constant temporal variation in the data, the residuals also are plotted versus time (Figures 32, 34, and 36). Intensity level is aggregated over time and we are not able to use time attribute for that dataset. In other words every minute intensity map is updated according to new values that have been received during the last minute. But at the same time all values obtained before are remaining on the map and participating in the overall intensity representation. The residuals plotted versus time do not show any drift in the errors and appear to behave randomly. Both sets of residual plots demonstrate that the delay from zero distance and zero time for Ferndale earthquake, the offshore event occurred 78 kilometers off the coastline. The cumulative error over the ten minute interval is shown on Figure 37. The maximum error for Ferndale earthquake registered at the third minute, the minimum error occurred at the tenth minute. For the La Habra event, the maximum error registered at the first and seventh minutes while the minimum error occurred at the fifth and tenth minute. The model could fit the validation dataset in a better way and calibration process for all models for California region is needed.

Figures 38 and 39 show the data for Chile and Japan earthquakes from Table 6 plotted with the prediction models (Kropivnitskaya et al. 2016). The data distribution follows the same pattern as the models, but it is clear that the models have to be calibrated and shifted to the left to represent the data relationships better. One of the Chile earthquakes has a significantly higher amount of tweets after an earthquake (Figure 38, upward triangles).

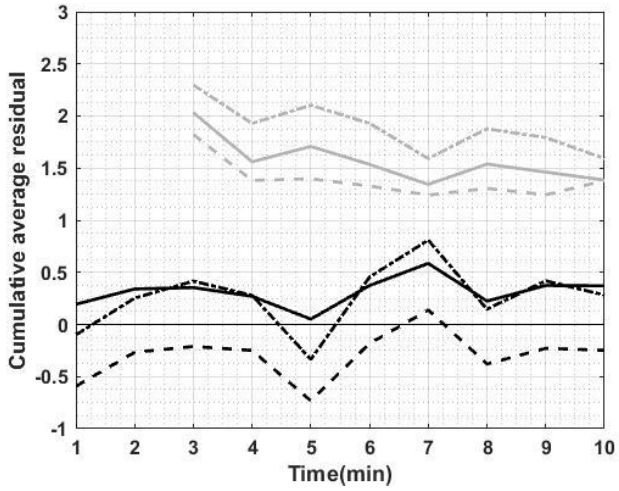

Figure 37: Cumulative average error over ten minutes (grey – Ferndale earthquake, black – La Habra earthquake, solid line – linear model, dashed line – exponential model, dotted line – two-segment linear model).

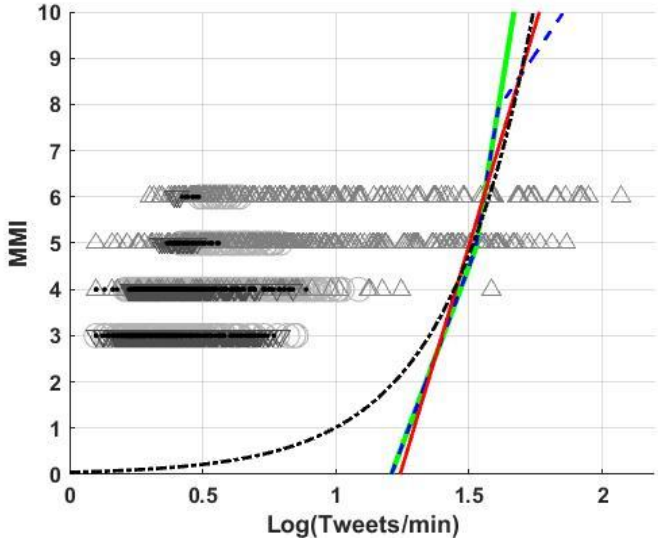

Figure 38: Observed Chile earthquake data (circles - 64km WNW of Iquique, upward triangles - 80km WNW of Iquique, downward triangles - 91km WNW of Iquique, dots - 94km NW of Iquique) with prediction models obtained for Napa earthquake (red line – linear, black line - exponential, green line – two-segment linear, blue line – three segment).

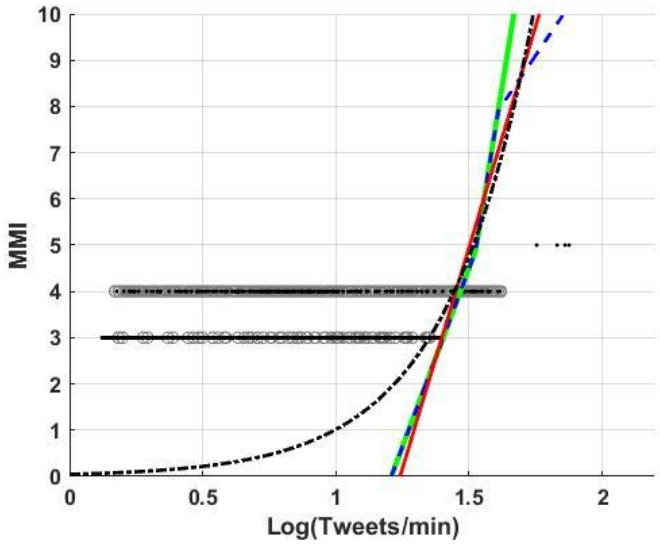

Figure 39: Observed Japan earthquake data (circles – Nago earthquake, dots - Kunisaki-shi earthquake) with prediction models (red line – linear, black line - exponential, green line – two-segment linear, blue line – three segment).

Taking into consideration the fact that all earthquakes have common characteristics, the possible reason could be related to the fact that the earthquake happened the day after the M6.7 event that occurred 64 km WNW of Iquique on March 16, 2014. People were likely alert and their reaction was stronger. The data from Japan does not allow for validation of the empirical relationships at an intensity level higher than V. As a result, the three-segment and two-segment models are excluded from the calibration. Observed data for the Chile region represent the intensity levels up to VI, therefore three-segment model is not taken into consideration.

5.4 Calibration of Existing Relationships

The forward problem that describes the relation between the logarithmic tweets rate and intensity predictions may be represented with an equation in the following form:

$$MMI=f(\theta)+e, \qquad (1)$$

where MMI is the aggregated instrumental intensity level observed, θ is the vector of tweet rates observations, f is the forward equation representing each mathematical model

(linear, exponential, 2-segment, 3-segment), and e is the residuals vector which describes the deviation between the measured and predicted values of intensity (2). Here

$$e=e_o+ e_m, \qquad (2)$$

where e_o accounts for measurement error in the observations, and e_m is the model error.

The calibration here is an inverse procedure. An inverse estimator C that connects the observations MMI to "good" estimates g of the parameters of interest:

$$g=C[f(\theta)+e]=min\sum[MMI-f(\theta)]\hat{}2 \qquad (3)$$

We used known data in the observed relationship between the dependent variable MMI and independent variable logarithmic tweets rate to estimate values for regions other than Napa Valley using new observations from the earthquakes listed in Table 6. Linear regression is used to produce new regression coefficients for each region. We regress instrumental MMI against the logarithmic mean of the number of tweets per minute to obtain new predictive equations (Figure 40) within a legitimate range of values for each model (see Table 7).

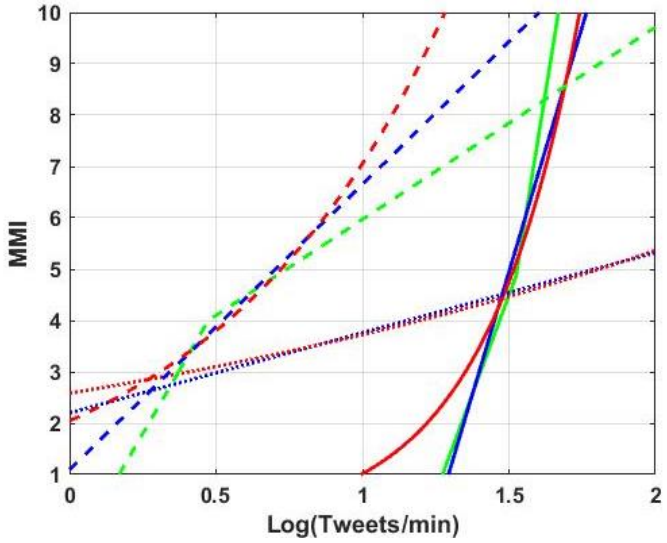

Figure 40: Calibrated models (red line – linear, black line - exponential, green line – two-segment linear, blue line – three segment) for California (solid lines), Japan (dotted lines), Chile (dashed lines).

Table 7: Calibrated Predictive Relationships

Model Name	Equation	RMS error, MMI units	Valid range of values (Ntweets/min)
California			
Linear	$MMI = 19.11\,log(N_{tweets/min}) - 23.72$	0.0033	[17.43; 58.15]
Exponential	$MMI = 0.04636 * e^{3.086*log(N_{tweets/min})}$	0.0035	(0; 55.1]
Two-segment linear	$MMI = 15.41\log(N_{tweets/min}) - 18.63$ $\log(N_{tweets/min}) < 1.52$ $MMI = 35.47\log(N_{tweets/min}) - 49.19$ $1.52 < \log(N_{tweets/min}) < 1.61$	0.0029	[24.4; 46.63]
Japan			
Linear	$MMI = 1.56\,log(N_{tweets/min}) + 2.2$	0.0207	[0.04; 10^5]
Exponential	$MMI = 2.584 * e^{0.3653*log(N_{tweets/min})}$	0.0186	(0; 5064]
Chile			
Linear	$MMI = 1.58\,log(N_{tweets/min}) + 3.241$	0.0143	[0.009; 18 960]
Exponential	$MMI = 3.4 * e^{0.339*log(N_{tweets/min})}$	0.0145	(0; 1517.6]
Two-segment linear	$MMI = 16.06\log(N_{tweets/min}) - 3.5338$ $\log(N_{tweets/min}) < 0.47$ $MMI = 3.59\log(N_{tweets/min}) + 2.32$ $0.47 < \log(N_{tweets/min}) < 0.86$	0.0358	[1.66; 137.8]

We regress the average ground-motion values for specified MMI levels in order to approximately follow the appropriate trend instead of producing a relationship that is overly influenced by the greater statistical volume of data at lower intensities. We applied a least squares solution with 95% confidence bounds for each model.

For California region the lowest root-mean-square (RMS) of 0.0029 MMI units is observed for the two-segment model. For the Napa earthquake the lowest error has been received with three segment model, that is excluded for validation here. The exponential model demonstrates the lowest RMS error of 0.0035 MMI units. For the Japanese data

the difference between the RMS error for the linear and exponential models is also not significant (0.0207 vs. 0.0186). That could be explained by the using the low-mid-valued intensity data in the regression (III - V). For higher lower levels (I-III) the difference will be more significant and for the intensity levels higher than V is less significant (see Figure 39). For the Chile earthquakes, the lowest RMS error of 0.0143 MMI units is observed for the linear model, which is 0.0002 units lower than for exponential model.

Because the RMS error cannot determine whether the model estimates and predictions are biased, we also assessed the residual plots (Figures 41-43).

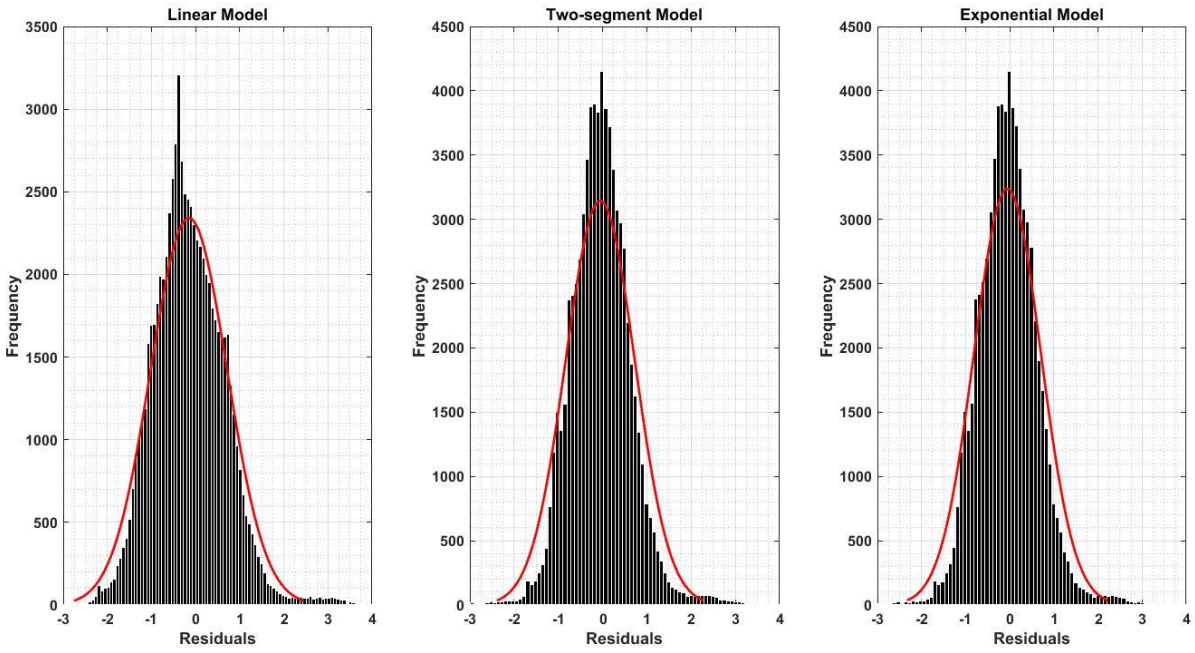

Figure 41: Residuals between calibrated models and observed California data.

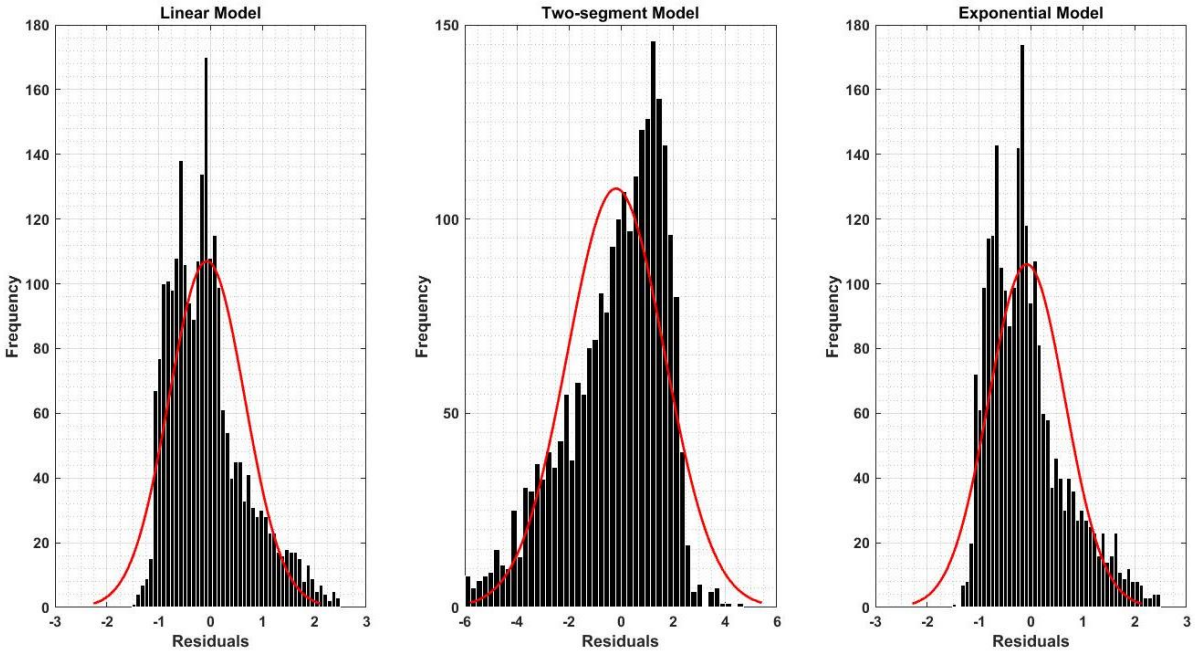

Figure 42: Residuals between calibrated models and observed Chile data.

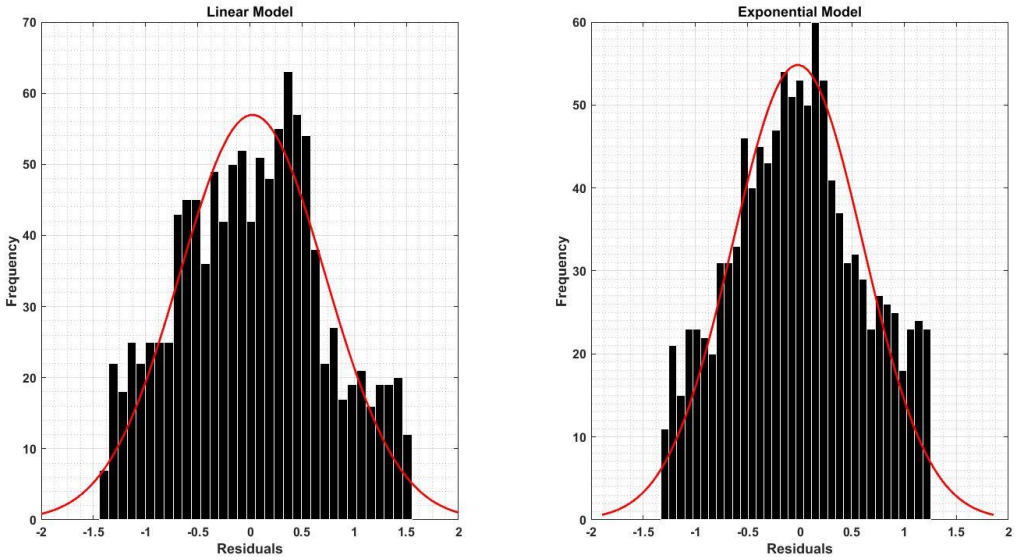

Figure 43: Residuals between calibrated models and observed Japan data.

The residuals between the predicted and observed data shown in Figure 41 for each model in the California region demonstrate normality in every case. The residual distribution for Japan also is nearly normal for both calibrated models. The distribution of

the Chile data residuals is skewed to the left for the two-segment model and skewed to the right for the linear and exponential models. The condition that the error terms are normally distributed is not met in that case.

5.5 Conclusions

Incorporation of social sensor data with traditional data sources using advanced computational processing methods can provide more complete and accurate coverage of damage and loss for rapid hazard response. The prediction equations obtained in this work could be used for real-time seismic hazard mapping and emergency management purposes in California, Japan and Chile using real-time data streaming concept of Twitter data. However, because our previous work had shown because of the the higher level of uncertainty resulting from the use of Twitter data alone comparing to the results, we show here that better results are obtained when it is streamed jointly with instrumental intensity, we propose to use imperical equation together with the data from physical sensors (implementation approach is explained in Kropivnitskaya et al, 2016).

The twitter stream processing here uses data recorded within the ten minutes following an earthquakes (2014). We confirm our earlier hypothesis that the logarithmic number of tweets can be used as a proxy for shaking intensity not just in the California region, but also in other regions with large numbers of Twitter users. In many areas, the importance of this additional data source could be very significant, due to the complete or partial lack of traditional, instrumental data sources as a result of the high cost of their installation and ongoing operation.

5.6 References

Bifet, A., Holmes, G., Kirkby, R., and Pfahringer, B., 2010. MOA: Massive online analysis. *Journal of Machine Learning Research* (JMLR), pp.1601-1604.

Burks, L., Miller, M. & Zadeh, R., 2014. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. s.l., *Tenth U.S. National Conference on Earthquake Engineering Frontiers of Earthquake Engineering*.

Campagne, J., Dux, J., Guyot, P. & Julien, D., 2012. Twitter reaches half a billion accounts - More than 140 million in the U.S.. [Online] Available at: http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J., 2012. Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS* , 17(1), pp. 124-147.

Earle, P., Bowden, D. & Guy, M., 2011. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), pp. 708-715.

Earle, P., Guy M., Buckmaster R., Ostrum C., Horvath S., Vaughan A., 2010. OMG Earthquake! Can Twitter Improve Earthquake Response?. *Seismological Research Letters*, 81(2), pp. 246-251.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer*, 66(4), 568–578. http://doi.org/10.1080/00330124.2014.907699

IBM, n.d. IBM, 2014. IBM Knowledge Center. [Online] Available at: http://www-01.ibm.com/support/knowledgecenter/[Accessed 2015].

Northern California Earthquake Data Center, 2014. UC Berkeley Seismological Laboratory. Dataset. doi:10.7932/NCEDC.

Sakaki, T., Okazaki, M. & Matsuo, Y., 2010. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. *Raleigh, NC, World Wide Web Conference (WWW).*

Schonfeld E., 2009. Mining the thought stream. TechCrunch Weblog Article, http://techcrunch.com/2009/02/15/mining-the-thought-stream/.

Severo, M., Giraud, T., & Pecout, H. (2015). Twitter data for urban policy making: an analysis on four European cities. In C. Levallois (Ed.), *Handbook of Twitter for Research. EMLYON*

Takhteyev Y., Wellman B., Gruzd A., 2012. Geography of Twitter Networks, *Social Networks*, Volume 34, Issue 1, January 2012, pages 73–81

Twitter, 2015. The Twitter Platform Documentation. [Online] Available at: https://dev.twitter.com/overview/documentation [Accessed 2015].

Wald, D.J., Quitariano, V., Heaton, T.H., Kanamori, H., Scrivner C.W., and Worden, c.B. (1999). "Trinet ShakeMaps: rapid generation of peak ground motion and intensity maps for earthquakes in Southern California," *Earthquake Spectra*, 15, 537–555.

Wald, D., Quitoriano, V. & Dewey, J., 2006a. USGS "Did you feel it?" community internet intensity maps: macroseismic data collection via the internet. *Geneva, Switzerland, First European Conference on Earthquake Engineering and Seismology.*

Wald, D., Worden, B., Quitoriano, V. & Pankow, K., 2006b. ShakeMap Manual: Technical Manual, Users Guide, and Software Guide, *Boulder: United States Geological Survey.*

Chapter 6

6 General Discussion and Conclusions

Technological breakthroughs in computing that have taken place over the last few decades have made it possible to achieve emergency management objectives that focus on saving human lives and reduce the associated economic effects. In particular, integrating new information sources from social media, not only with each other, but also with other, existing sources from physical sensors has enabled a vast range of new, real-time applications in the areas of disaster monitoring, hazard mapping and emergency response. Physical sensors such as mobile phones with accelerometers, compasses, and cameras, weather observation stations, location tracking systems with GPS and radio-frequency identification (RFID) and building management systems are continuously producing an enormous amount of information in the form of complex multidimensional data streams. Social services like Twitter and Facebook deliver unstructured streams of real-time data with various observations from human beings (social sensors). Due to the heterogeneous nature of such diverse sources, the analysis of that data is still a difficult task and currently requires special techniques, such as computational streaming techniques. The goal of this work is to demonstrate the efficiency and efficacy of real-time streaming processing of time-dependent data from physical and social sensors for the application in the natural hazards field. To achieve the goal two streaming applications have been developed: one is for near real-time hazard maps production and the other is for joint processing of data from social sensors (Twitter) and physical sensors (strong motion records) for near real-time earthquake intensity calculations. Both tools developed in this work have many practical important applications and could be widely used by policy makers, by insurance companies, by civil engineers and by emergency hazard personnel to assess and plan for areas of maximum loss and death.

6.1 Advantages

1. The parallel streaming implementation of seismic hazard map production proposed in this study addresses the potential for improvement in spatial and temporal

resolution. Computing bottlenecks are removed through parallel decomposition of bottlenecks into sub-tasks. These tasks are separated and solved simultaneously and then combine the results from each set of solutions (Sakr, 2013). The low execution time and high resolution are the biggest advantages of proposed parallel decomposition. Other types of analysis techniques in the geosciences which face similar issues can be improved by using parallel streaming algorithms via platforms such as IBM Streams.

2. Probabilistic seismic analysis involves two main outputs: A seismic source model and empirical ground motion relationships, which are represented by various studies that intrinsically assume that every next generation of models and relationships improves on the previous results. However, there are not enough sensitivity studies showing how the uncertainty in the output of a system can be apportioned to different sources in developed models and how relationships are conducted due to the computational resources needs and lack of appropriate software tools. The eastern Canada hazard maps case study demonstrates how the high-resolution mapping technique developed here can be used for various sensitivity tests of probabilistic seismic hazard outputs for different regions based on the different input parameters of probabilistic seismic hazard analysis.

3. The useof social media data for effective assessment and quantification of the ongoing effects of natural disasters has shown significant potential to date. However, realistic practical applications need additional research, including into the relationship between the properties and parameters of social media data (in our case, tweet rates) and the physical quantities used for hazard and risk measurements (in our case, MMI). From that perspective, the empirical relationships developed between tweet rates and MMI in this study not only can be used in the streaming techniques presented here, but in other hazard applications such as preliminary insurance claim calculations or the planning of resources redistributions.

4. All software tools developed in this study are scalable and perform under an increased or expanding workload through Streams. Moreover, they are extensible to

other types of hazard, such as flood, fire, windor other areas with high level of hazard.

6.2 Disadvantages

1. The rate limitations of Twitter's data stream are restrictive for natural hazard applications. To satisfy such requirements, Twitter Firehose provides access to 100% of the public tweets on Twitter at a price. Firehose data can be purchased through third-party resellers of Twitter data (Kumar et al., 2013). As a result, this study was limited by the public data access restrictions and the developed implementations were used here strictly for demonstration purposes. Although in reality it would require additional effort to implement access to the real Twitter APIs, the resulting potential of near real-time intensity maps for regions with sparse or nonexistent seismic networks is very promising.

2. Previous studies have shown that less than one percent of tweets contain location metadata. The absence of geotagged Twitter data is one of the weakest aspects of this approach. Moreover, the geolocation algorithm used in this study requires manual location references (city names, abbreviations, short names) database extension that is not efficient. These references also may be lost in informal, ungrammatical, and multilingual data, and are therefore non-trivial to identify, affecting the ability to properly exploit user geolocation.

3. Historical Twitter dataset used in this study can have a completeness issue and the model parameters obtained here could only represent the corresponding relationships for specific earthquakes picked for analysis.

4. Another area for improvement is related to this particular streaming. The continuous queries employed here need to be optimized adaptively in order to cope with arbitrary changes in the real stream characteristics and system conditions. The finite list of earthquake and location keywords maybe not sufficient enough to get minimum amount of tweets in reality.

5. Privacy concerns with social networking services. The nature of social media may provide information about a single individual to third parties and that raises privacy concerns, even in the cases of monitoring of social media to better detect or understand catastrope events.

6.3 Future work

The outline of future work can be subdivided into a few categories:

1. All techniques proposed and developed in this study can be applied to other high hazard regions and the associated perils as more physical and social data sources become available. In particular, social sensor data from not only Twitter, but also Facebook, Instagram can help to complement traditional data sources using advanced computational processing methods in the regions with low instrumental coverage.

2. To increase performance, additional decomposition levels could be added to both implementations; e.g. both seismic hazard maps production and intensity calculations.

3. Additional performance tests should be done to define the optimum number of parallel pipelines and minimum/optimum amount of computational resources for the real-time execution.

4. This work could be used as a basis for a universal streams toolkit development (PSHA toolkit and Twitter data processing toolkit).

5. The process for calibrating empirical relationships between tweet rates and MMI could be improved for more precise estimations. Considering each territory has widely diverse features that can be quantified, e.g. the earthquake rates and the size of tweet communities in the area, an automated on-the-fly calibration process based on the live streaming should be developed for implementation into operational methods.

6. As noted above, for the seismic hazard implementation, the current list of earthquake and location keywords maybe not sufficient to get the correct set of tweets. This requires additional study for both various regions to ensure optimal implementation,

in conjunction with efforts to assign more accurate geolocation information. The latter would benefit similar efforts in other hazard areas as well.

7. The text-based geolocation technique and semantic analysis of tweets used in this work could be significantly improved by modern machine and deep learning algorithms.

6.4 References

Kumar S., F. Morstatter, and H. Liu. Twitter Data Analytics. *Springer*, New York, NY, USA, 2013.

Sakr, S. (2013). An introduction to InfoSphere Streams. Retrieved from www.ibm.com: http://www.ibm.com/developerworks/library/bd-streamsintro/bd-streamsintro-pdf.pdf

Appendices

Appendix A: List of algorithms for seismic hazard maps production

Input : File F1 with the description of seismic source zone model from www.seismotoolbox.ca, file F2 with a list of zones (polygon coordinates and seismicity re-occurrence parameters from the Geological Survey of Canada (GSC)), file F3 with other PSHA parameters specified by user, file F4 contains GMPE model
Output: File J encoding map in JSON format, containing ground motion estimation for a specified number of sites

```
1    N1 ← Barrier(F1,F2,F3);
2    I1 ← CatalogInputs(N1);
3    G ← Catalog(I1);
4    N2 ← Barrier(F3, G, F4);
5    I2 ← MapInputs(N2);
6    O ← Map(I2);
7    J ← JSONGen(O);
8    TCPSink(J);
```

Algorithm 1: Production of hazard map

Input : File F1 with the description of seismic source zone model from www.seismotoolbox.ca, file F2 with a list of zones (polygon coordinates and seismicity re-occurrence parameters from the GSC), file F3 with other PSHA parameters specified by user, file F4 contains the GMPE model
Output: File JS encoding map in JSON format containing a ground motion estimation for a specified number of sites

```
1    N1 ← Barrier(F1,F2,F3);
2    I1 ← CatalogInputs(N1);
3    C1, C2, ..., C10 ← Split(I1);
4    G1 ← Catalog(C1);
5    G2 ← Catalog(C2);
     ...
13   G10 ← Catalog(C10);
14   T ← Barrier(G1, G2, ..., G10);
15   U ← TotalCatalog(T);
16   N2 ← Barrier(F3, U, F4);
17   I2 ← MapInputs(N2);
18   M1, M2, ..., M25 ← Split(I2);
19   O1 ← Map(M1);
20   O2 ← Map(M2);
     ...
43   O25 ← Map(M25);
44   J1 ← JSONGen(O1);
45   J2 ← JSONGen(O2);
     ...
68   J25 ← JSONGen(O25);
69   TJ ← Barrier(J1, J2, ..., J25);
70   JS ← TotalJSON(TJ);
71   TCPSink(JS);
```

Algorithm 2: Production of hazard map with parallel execution

Appendix B: List of algorithms for MMI maps production

Input : Files FE, FN in SMC format with the time-series of PGV horizontal (east and north) components at each station
Output: File J encoding map in xyz format, containing MMI estimation in the region

```
1   FLE ← DirectoryScan(FE);// To monitor new data files FE
2   FLN ← DirectoryScan(FN);// To monitor new data files FN
3   FLE_R ← FileSource(FLE);// To transform data files FLE into
    stream FLE_R
4   FLN_R ← FileSource(FLN);// To transform data files FLN into
    stream FLN_R
5   PE ← Custom(FLE_R);      // To parse stream FLE_R into stream
    PE
6   PN ← Custom(FLN_R);      // To parse stream FLN_R into stream
    PN
7   P ← Barrier(PE,PN);      // To synchronize streams PE and PN
8   PGV ← Custom(P);         // To calculate PGV stream
9   MMI ← Custom(PGV);       // To calculate MMI stream
10  MMI_I ← Custom(MMI);     // To interpolate MMI stream
11  FileSink(MMI_I);        // To output results from the MMI_I
    stream into file
```

Algorithm 1: Production of instrumental MMI map

Input : File F contains tweets dataset in JSON format
Output: File J encoding map in xyz format, containing tweets rate estimation in the region

```
1   FL ← DirectoryScan(F); // To monitor new data files F
2   FLC ← FileSource(FL); // To transform data files FL into
    stream FLC
3   JT ← JSONToTuple(FLC); // To parse stream FLC into stream JT
4   PT ← Filter(JT);          // To filter positive tweets into
    stream PT
5   GeoPT ← Custom(PT);     // To parse the location for positive
    tweets
6   GeoPTF ← Filter(GeoPT); // To filter geotagged positive
    tweets
7   TwN ← Custom(GeoPTF);    // To calculate tweets rate
8   TwNI ← Custom(TwN);     // To interpolate tweets rate
9   FileSink(TwIN);          //To output results from TwIN stream
    into file
```

Algorithm 2: Production of tweets rate map

Input : Files FE, FN in SMC format with the time-series of PGV horizontal (east and north) components at each station, file F contains tweets dataset in JSON format
Output: File J encoding map in xyz format, containing combined MMI estimation in the region

```
1   FLE ← DirectoryScan(FE); // To monitor new data files FE
2   FLN ← DirectoryScan(FN); // To monitor new data files FN
3   FL ← DirectoryScan(F);   // To monitor new data files F
4   FLE_R ← FileSource(FLE);// To transform data files FLE into
    stream FLE_R
5   FLN_R ← FileSource(FLN);// To transform data files FLN into
```

```
    stream FLN_R
6   FLC ← FileSource(FL); // To transform data files FL into
    stream FLC
7   JT ← JSONToTuple(FLC); // To parse stream FLC into stream JT
8   PT ← Filter(JT); // To filter positive tweets into stream PT
9   PE ← Custom(FLE_R); // To parse stream FLE_R into stream PE
10  PN ← Custom(FLN_R); // To parse stream FLN_R into stream PN
11  GeoPT ← Custom(PT); // To parse the location for positive
    tweets
12  P ← Barrier(PE,PN); // To synchronize streams PE and PN
13  PGV ← Custom(P); // To calculate PGV stream
14  MMI ← Custom(PGV); // To calculate MMI stream based on PGV
15  GeoPTF ← Filter(GeoPT); // To filter geotagged positive
    tweets
16  TwN ← Custom(GeoPTF); // To calculate tweets rate
17  TwMMI ← Custom(TwN); // To calculate MMI stream based on TwN
    stream
18  MMIC ← Custom(TwMMI,MMI); // To calculate joint MMI stream
19  FileSink(MMIC); //To output results from MMIC stream into
    file
```

Algorithm 3: Production of combined MMI map

Curriculum Vitae

Name: Yelena Kropivnitskaya

Post-secondary Education and Degrees:

Karaganda State Technical University
Karaganda, Kazakhstan
2003 - 2007 Bachelor in Computer Science

Karaganda State Technical University
Karaganda, Kazakhstan
2007-2008 Masters in Computer Science

Western University
London, Ontario, Canada
2012-2016 Ph.D.

Honours and Awards:

Winner of Student Delegate Program for the C4 CatIQ's
2016

Joint Assembly Outstanding Student Paper Award
2015

GSC Pioneers Scholarship
2014

KEGS Foundation Scholarship Award
2013

Mitacs-Accelerate Internship Program
2013

Related Work Experience

Teaching Assistant
Western University
2012-2016

Publications:

Kropivnitskaya, Y., Qin, J., Tiampo, K., & Bauer, M. (2015). Pipelining Implementation for High Resolution Seismic Hazard Maps Production. *International Conference On Computational Science, ICCS 2015 – Computational Science at the Gates of Nature* (pp. 1473-1482). Reykjavik: Procedia Computer Science.

Kropivnitskaya, Y., Tiampo, K., Qin, J., & Bauer, M. (2016). Impact of the Ground Motion Prediction Equation Changes on Eastern Canada Hazard Maps. *2016 CSCE Annual Conference, 51*. London, Canada.

Kropivnitskaya, Y., Tiampo, K., Qin, J., & Bauer, M. (2016). Real-Time Earthquake Intensity Estimation Using Streaming Data Analysis of Social and Physical Sensors. *Pure and Applied Geophysics*, accepted with minor revisions.

Kropivnitskaya, Y., Tiampo, K., Qin, J., & Bauer, M. (2016). The Predictive Relationship between Earthquake Intensity and Tweets Rate for Real-Time Ground Motion Estimation. *Seismological Research Letter*, submitted on 07/19/2016.

Tiampo, K.F., McGinnis, S., Kropivnitskaya, Y., Qin, J., Bauer, M.A. Big data challenges and hazards modeling, invited chapter, *Insurance Catastrophe Risk Modeling*, ed. G. Michel, submitted with minor revisions, 2016