Data-Driven Assessment of Disaster Damage and Recovery Time

Quoc D. Cao

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Youngjun Choe, Chair

Linda N. Boyle

Scott B. Miles

Program Authorized to Offer Degree:
Industrial and Systems Engineering

University of Washington

Abstract

Data-Driven Assessment of Disaster Damage and Recovery Time

Quoc D. Cao

Chair of the Supervisory Committee:
Prof. Youngjun Choe

Although natural hazards may sometimes be predictable, their occurrence is not preventable, especially in low-frequency-yet-high-impact events such as earthquakes and hurricanes. The catastrophic effects of natural hazards can vary vastly from year to year, depending on the seasons, locations, demographics, or resilience of the affected areas. Therefore, improving response system and recovery time is one of the most efficient ways to limit fatality and economic loss from a hazard event. Unfortunately, without being able to gain adequate situation awareness about the damage extent and the potential recovery, we cannot effectively improve these processes.

This dissertation provides a suite of methodological frameworks utilizing statistical tools to aid in the damage assessment and estimation of various infrastructures' recovery to provide emergency managers and stakeholders with timely and extensive situation awareness after a hazard event. The initial step is to assess the actual damage extent immediately after a hazard event so that adequate planning and resources can be allocated. The first methodological framework aims to speed up the post-event damage assessment process. Instead of the more time-consuming and labor-intensive windshield survey method, machine learning algorithms are applied to automatically annotate the damaged and/or flooded buildings on satellite imagery. The annotation results can be used as a proxy for assessing how badly an area is affected. The machine learning algorithms require much less time and resources while still yielding results with reasonable accuracy. Secondly, to improve generalizability and

accuracy of the previous damage assessment framework, a mixed data approach is adopted to combine satellite imagery and other geolocation features such as each building's elevation and proximity to water bodies. Finally, a recovery trajectory estimation framework is introduced to aid in recovery planning for critical infrastructures. The estimation will provide infrastructure management agencies with an idea of the most likely recovery pattern of various critical infrastructures (such as electricity, water, and gas), given different hazard scenarios. This will give them a quantitative assessment of how resilient their infrastructure systems are so that resources can be allocated and necessary investment can be informed effectively. Besides extensive results from numerical studies and empirical data, this dissertation research also contributes two curated datasets to open-access repositories so that others can reproduce and improve the proposed framework.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to my advisor, Professor Youngjun Choe for his countless hours of mentorship and continuous support for my Ph.D. study and related research. His patience, motivation, and immense knowledge greatly guided me to make this dissertation possible. Besides my advisor, I cannot thank enough the rest of my dissertation committee members, Professor Linda N. Boyle, Professor Scott B. Miles, and Professor Faisal Hossain, who generously spent their time to provide me with valuable feedback throughout the journey of the Ph.D. program.

DEDICATION

To my wife T.T., the love of my life.

Chapter 1

INTRODUCTION

Normal operation of infrastructures during and after natural hazard-induced disasters is crucial to the well-being of the public [113]. However, recovery of damaged infrastructures from disasters can cost a lot of time and resources [98]. These problems are especially challenging in low-probability-yet-high-impact events, such as the 2011 Tōhoku earthquake in Japan [93] or the 2017 Hurricane Harvey in the United States [13]. One way to speed up the recovery process is to improve situation awareness of emergency managers and stakeholders so that proper resources and manpower can be allocated to help the victims and restore the infrastructures.

Situation awareness is essential during a community's distressed time. In simple terms, it means for the stakeholders to know "what is going on" relative to their goals, or formally, "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [52]. The term was used to study how fighter pilots are aware of their surroundings in highly stressful and uncertain situations, in which they have to make informed decisions within a short period of time. As shown in its definition, situation awareness can be thought of as having three levels. The first level is perception, in which stakeholders or decision makers try to perceive the current information. The second level is comprehension, where stakeholders process the information to determine what is relevant to their goals. The last level is called projection, which lets the stakeholders achieve the highest level of understanding of the situation to make timely decisions [54].

Situation awareness is therefore important in various phases of the disaster management cycle. Social media platforms, e.g. Twitter, have been common tools studied by disaster

research community to enhance situation awareness to emergency managers, [24, 96, 119]. In recent works, researchers have been trying to provide more context from social media to enrich the perception level of situation awareness. For example, in [120], a GeoView interface was developed to assist emergency responses by providing real-time location-based mapping of social media messages.

This dissertation aims to leverage a different source of information from most of the recent works. Insights will be gathered from publicly available satellite imagery to provide quick and accurate damage assessment frameworks to improve the perception level of situation awareness. In addition, expert judgements will be elicited to estimate an infrastructure recovery trajectory to inform the projection of recovery.

1.1 Research Objectives

The overarching research goal of this dissertation is to provide emergency managers with different modelling frameworks to develop better and timely situation awareness after a hazard event.

Our current ability to estimate infrastructure recovery trajectories is limited, as revealed in the recent resilience planning efforts of U.S. communities, which started in San Francisco, CA [111] and became state-wide initiatives in Washington State [40] and Oregon [39]. These efforts inspired the National Institute of Standards and Technology (NIST)'s Community Resilience Planning Guide [2] as a model for other jurisdictions. Although there is a growing body of literature on computational modeling of recovery, these models are often viewed as resource-intensive black-box approaches and not utilized by communities on the ground.

The NIST Guide defines time to recovery of function as "a measure of how long it takes before a building or infrastructure system is functioning" and "uses time to recovery of function as the primary metric for community resilience." This echoes the widely-recognized importance of characterizing disaster recovery for assessing community resilience [23, 29, 34, 92]. As the quote by Lord Kelvin says "if you cannot measure it, you cannot improve it," the lack of rigorous and sound estimation methods for recovery time impedes the measurable

progress of resilience improvement. These motivations lead to three research directions as follows.

1.1.1 Machine annotation of post-hurricane satellite imagery for identifying damages

When a hurricane makes landfall, situational awareness is one of the most critical needs that emergency managers face before they can respond to the event. To assess the situation and damage, the current practice largely relies on emergency response crews and volunteers to drive around the affected area, which is also known as windshield survey. Another way to assess hurricane damage level is flood detection through synthetic aperture radar (SAR) images (e.g., see the work at the Darthmouth Flood Observatory [3]), or the damage proxy map to identify regional-level damages on the built environment (e.g., the Advanced Rapid Imaging and Analysis (ARIA) Project by Caltech and NASA [1]). SAR imagery is useful in terms of mapping different surface features, texture, or roughness pattern but is harder for laymen to interpret than optical sensor imagery. The resolutions of virtually all SAR images of today are too coarse to permit the building-level (as opposed to regional-level) damage assessment. Also, satellites equipped with SAR sensors are far fewer than those with optical sensors, making timely and frequent data collection challenging. In this work, we focus on using optical sensor imagery as a more intuitive and accessible way to analyze hurricane damage by distinguishing damaged buildings from the ones still intact. From here onwards, we will refer to optical sensor imagery as 'imagery'.

Recently, imagery taken from drones and satellites started to help improve situational awareness from a bird's eye view, but the process still relies on human visual inspection of captured imagery, which is generally time-consuming and unreliable during an evolving disaster. Computer vision techniques, therefore, can be particularly useful. Given the available imagery such as in Figure 1.1, our proposed method can automatically annotate *'Flooded/Damaged Building' vs. 'Undamaged Building'* on satellite imagery of an area affected by a hurricane (Figure 1.2). The annotation results can enable stakeholders (e.g., emergency managers) to better plan for and allocate necessary resources. With decent accu-

racy and quick runtime, this automated annotation process has potential to significantly reduce the time for building situational awareness and responding to hurricane-induced emergencies. We construct this dataset using the imagery before the hazard event to be the *'Undamaged Building'* labels and after the event to be the *'Flooded/Damaged Building'* labels. The dataset and code used in this work are available at my Github repository `https://github.com/qcao10/DamageDetection`. The dataset is also publicly available at the IEEE DataPort (DOI: 10.21227/sdad-1e56).

Figure 1.1: A satellite imagery of the Greater Houston area in 2017.

1.1.2 Post-hurricane damage assessment using satellite imagery and geolocation features

Within the field of damage assessment, deep learning techniques have been showing promising results. A subset of deep learning models called convolutional neural network (CNN) have been applied to detect damage in concrete structures [27, 28, 67], car damage [104, 127], or regional change detection after disaster events [49]. However, these methods rely heavily on the quantity and quality of the labelled dataset, which in some cases might be unavailable or

Figure 1.2: *Flooded/Damaged* samples cropped from the original imagery.

noisy. Sometimes, performance can be capped in some large image dataset such as Imagenet [71] and improvement in performance is marginal, regardless of model architecture.

Before the CNN era, there were established methods to assess flooding hazard risks, such as analyzing precipitation, catchment capacity, or river network analyses [14], generating flood outlines and depth based on topological data [63], or simulating flood spreading [62]. In the seismic risk analysis field, there are also probabilistic risk assessment [51] or defining vulnerability indices for infrastructure systems [105]. These methods are still extremely valuable even in these days as a natural hazard is a natural phenomenon that mostly obeys physical rules.

The above observation inspires us to hypothesize that there could be a potential improvement to the post-hurricane damage assessment process if we can utilize multiple types of data. We propose to utilize the optical sensor satellite imagery and other geolocation features of the individual buildings in our damage annotation framework. This work will open up another possibility to understanding post-disaster damage. For example, for hurricanes, we can combine precipitation level, flooding resilience index, catchment capacity, elevation, or river networks with the imagery data to gain extra performance and also understand which characteristics are more critical to the likelihood of damage. On the other hand, in seismic risk assessment, we can also incorporate the relative distance of the buildings/roads to the epicenter, Ritchter magnitude, ground shaking in the zone through various sensors, or seismic resilience index into the model, in addition to the aerial images. In this work, we present

a mixed data approach to damage annotation by consuming the satellite imagery and other geolocation features such as building elevation and proximity to water bodies (Figure 1.3) to improve the performance and generalizability of our previous work. Our contribution is two fold. First, by considering mixed data, we can leverage more domain knowledge to understand disaster damage assessment better and boost its predictive performance. Second, as an improvement to our previous dataset in [26], we collect the *'Undamaged Building'* labels and *'Flooded/Damaged'* labels from imagery of the same timestamp. Specifically, we manually build the *'Undamaged Building'* labels from the undamaged region of the same imagery captured after the hurricane event. This dataset is more realistic and generalizable since it reflects the actual situation when we want to deploy this damage annotation framework in future events.

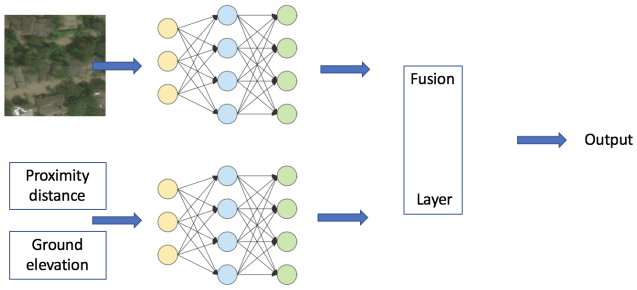


Figure 1.3: Workflow of the mixed data neural network model.

1.1.3 Recovery time estimation in disaster planning

This work proposes a rigorous statistical framework to estimate infrastructure recovery curves (e.g., see Figure 1.4) for a hazard scenario using a combination of expert elicitation and Gaussian process regression (GPR). The two methods complement each other to provide satisfactory solutions to this problem. Estimates gathered from experts will provide initial guidelines on how long it will take for a particular infrastructure to recover to some intermediate functionality levels. GPR will then use these estimates to predict the full recovery

curve while capturing potential uncertainty in its prediction, as well as the uncertainty in the experts' estimates. GPR is also flexible enough to enforce important constraints on its predictions to allow the predicted curve to follow the physical behaviour of the actual recovery curve (e.g., monotonically increasing and bounded between 0 and 100%). This framework can provide disaster stakeholders with a likely projection of recovery pattern so that adequate resource and investment planning can be made. The framework aims to be extensible to various types of infrastructure, while being intuitive and easy to be interpreted by the stakeholders.

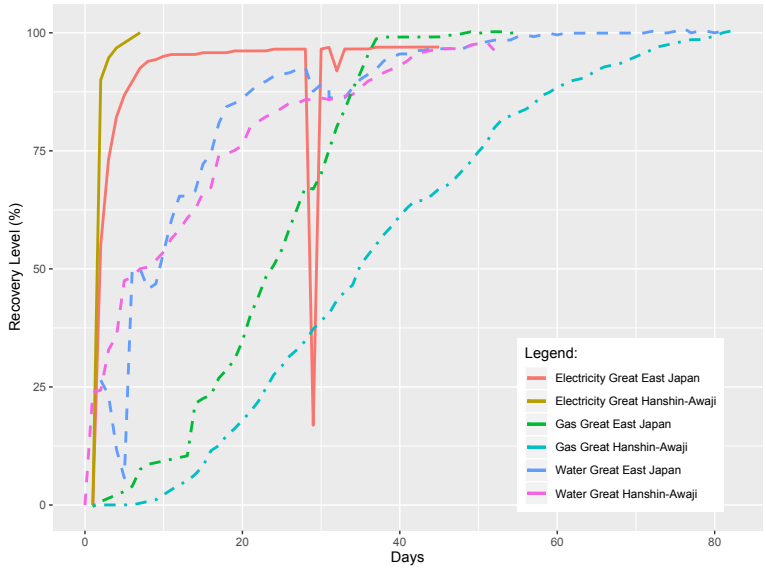


Figure 1.4: Empirical restoration curves of the 1995 Great Hanshin-Awaji Earthquake and the 2011 Great East Japan Earthquake.

1.2 Dissertation Outline

The remainder of this dissertation is organized as follows. Chapter 2 presents an article on using deep learning and satellite imagery to identify damaged buildings after hurricane events [26], which is published in *Natural Hazards*. Chapter 3 about hurricane damage

assessment with satellite imagery and geolocation features, together with Chapter 4 about recovery time estimation using expert elicitation and Gaussian Process Regression are based on two working manuscripts. The dissertation will conclude with some remarks and future research directions in Chapter 5.

Chapter 2

BUILDING DAMAGE ANNOTATION ON POST-HURRICANE SATELLITE IMAGERY BASED ON CONVOLUTIONAL NEURAL NETWORK

2.1 Introduction

After a hurricane, damage assessment is critical to emergency managers for efficient response and resource allocation. One way to gauge the damage extent is to quantify the number of flooded/damaged buildings, which is traditionally done by ground survey. This process can be labor-intensive and time-consuming. In this work, we propose to improve the efficiency of building damage assessment by applying image classification algorithms to post-hurricane satellite imagery. At the known building coordinates (available from public data), we extract square-sized images from the satellite imagery to create training, validation, and test datasets. Each square-sized image contains a building to be classified as either 'Flooded/Damaged' (labeled by volunteers in a crowd-sourcing project) or 'Undamaged'. We design and train a convolutional neural network from scratch and compare it with an existing neural network used widely for common object classification.

We demonstrate the promise of our damage annotation model in the case study of building damage assessment in the Greater Houston area affected by 2017 Hurricane Harvey. The satellite imagery data covers the Greater Houston area before and after Hurricane Harvey in 2017 (Figure 2.1). The flooded/damaged buildings were labeled by volunteers through the crowd-sourcing project, Tomnod [7]. We then process, filter, and clean the dataset to ensure that it has correct labels and can be learned appropriately by a learning algorithm.

By sharing the dataset and code used in this work (see the appendix), we hope that other researchers can build upon this study and help further improve computer vision-based

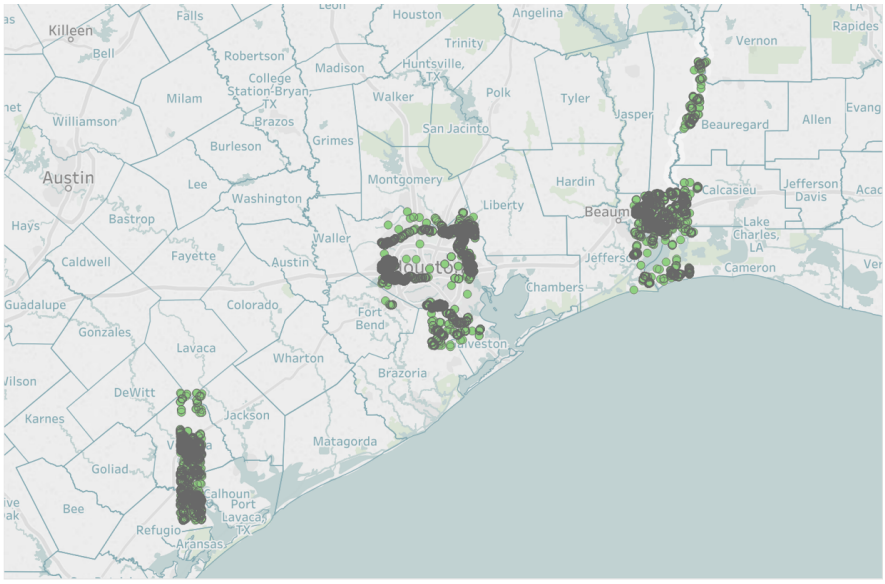


Figure 2.1: The Greater Houston area was affected by Hurricane Harvey in 2017. The green circles represent the coordinates of flooded/damaged structures tagged by Tomnod volunteers.

damage assessment process. The shared code includes a pre-trained deep-learning architecture that achieves the best classification accuracy (detailed in Section 2.4). It can facilitate transfer learning either in feature extraction, fine-tuning, or as a baseline model to speed up the learning process for future hurricane events.

The remaining of this chapter is organized as follows. In Section 2.2, we present a brief review of convolutional neural network, machine learning-based damage annotation work on post-hurricane satellite imagery, and challenges in the damage annotation on satellite imagery. Section 2.3 describes our proposed methodological framework for the damage annotation. Details of the implementation and discussion of the results are presented in Section 2.4. Finally, Section 2.5 concludes this work and draws some future research directions.

2.2 Background

2.2.1 Convolutional neural network

The convolutional neural network (CNN) [75] often yields outstanding results over other algorithms for computer vision tasks such as object categorization [66], image classification [35, 71], and object recognition [36]. Variations of CNN have been successfully applied to remote sensing image processing tasks [128] such as aerial scene classification [79, 87, 123], SAR imagery classification [129], or object detection in unmanned aerial vehicle imagery [19].

Structurally, CNN is a feed-forward network that is particularly powerful in extracting hierarchical features from images. The common structure of CNN has three components: the convolutional layer, the sub-sampling layer, and the fully connected layer as illustrated in Figure 2.2.

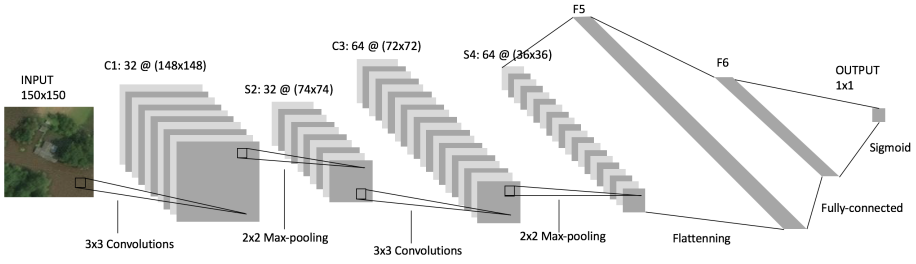


Figure 2.2: A convolutional neural network inspired by LeNet-5 architecture in [75]; C: Convolutional layer, S: Sub-sampling layer, F: Fully connected layer; 32@(148x148) means there are 32 filters to extract features from the input image, and the original input size of 150x150 is reduced to 148x148 since no padding is added around the edges during 3×3 convolution operations so 2 edge rows and 2 edge columns are lost; 2x2 Max-pooling means the data will be reduced by a factor of 4 after each operation; Output layer has 1 neuron since the network outputs the probability of one class ('Flooded/Damaged Building') for binary classification.

In the convolutional layer (C in Figure 2.2), each element (or neuron) of the network in a layer receives information from a small region of the previous layer. A 3x3 convolutional filter will take a dot product of 9 weight parameters with 9 pixels (3x3 patch) of the input, and the resulting value is transformed by an activation function to become a neuron value in the next layer. The same region can yield many information maps to the next layer through many convolutional filters. In Figure 2.2, at convolutional layer C1, we have 32 filters that represent 32 ways to extract features from the previous layers and form a stack of 32 feature matrices. Another advantage of CNN is its robustness to shift of features in the input images [57]. This is crucial since in many datasets, objects of interest are not necessarily positioned right at the center of the images and we want to learn the features, not their positions.

In the sub-sampling layer (S in Figure 2.2), the network performs either local averaging or max pooling over a patch of the input. If the sub-sampling layer size is 2x2 such as S2, local averaging will yield the mean of the 4 nearby convoluted pixel values, whereas max pooling will yield the maximum value among them. Essentially, this sub-sampling operation reduces the input feature matrix to half its number of columns and rows, which helps to reduce the resolution by a factor of 4 and the network's sensitivity to distortion.

After the features are extracted and the resolution reduced, the network will flatten the final stack of feature matrices into a feature vector and pass it through a sequence of fully connected layers (F in Figure 2.2). Each subsequent layer's output neuron is a dot product between the feature vector and a weight vector, transformed by a non-linear activation function. In this work, the last layer has only 1 neuron, which is the probability of a reference class ('Flooded/Damaged building').

As mentioned, the dot products are transformed by an activation function. This gives a neural network, with adequate size, the ability to model any function. Some common activation functions include sigmoid $f(x) = \frac{1}{1+e^{-x}}$, rectified linear unit (ReLU) $f(x) = max(0, x)$, and leaky ReLu $f(x) = max(\alpha x, x)$, with $0 < \alpha \ll 1$. There is no clear reason to choose any specific function over the others to improve performance of a network. However,

using ReLU may speed up the training of the network without affecting the performance [59].

2.2.2 Machine learning-based damage annotation on post-hurricane satellite imagery

Machine learning on remote sensing imagery is actively researched to assess damage from or susceptibility to various hazards such as earthquake [106], landslide [11,64], tsunami [88], and wildfire [83]. Such methods showed remarkable promise. However, often leveraging unique characteristics of each hazard type, they are not directly applicable to damage annotation on post-hurricane imagery and sometimes require extensive pre-processing. For example, the work in [106] for post-earthquake damage assessment closely resembles our work. Both pre-event and post-event are used to extract building's roof and texture features. This process requires input from expert operators and rely on texture feature of the buildings in the imagery, which sometimes may not be available depending on means of collection. In another work, [88] classifies regions of pixels into water, vegetation, urban, and bare land, which does not provide the building level granularity as we pursue. Furthermore, since we rely purely on the widely available optical sensor imagery, convolutional neural network is one of the most suitable model classes due to its flexible feature extraction capability and architecture.

Some recent studies used machine learning to assess post-hurricane damages on satellite imagery. A small project studied detecting *flooded roads* by comparing pre-event and post-event satellite imagery [68] but the method is not applicable to other types of damages. Two commercial vendors of satellite imagery also separately developed unsupervised algorithms to detect flooded area using spectral signature of impure water (which is not available from the pansharpened satellite images in our data) [9, 10]. Before deep learning era, a method using a pattern recognition template set was applied to detect hurricane damages in *multispectral* images [17] but the method is not applicable to our pansharpened images.

2.2.3 Challenges in damage annotation on satellite imagery

There are multiple challenges in damage annotation on satellite imagery. First, satellite imagery resolution is not as high as various benchmark datasets commonly used to train neural network (NN) (e.g., ImageNet [71] and traffic signs [36]) with respect to the objects of interest. Dodge & Karam [48] studied the performance of NNs under quality distortions and highlighted that NNs could be prone to errors in blurry and noisy images. Although our dataset is of relatively high resolution (e.g., one of the satellites capturing the imagery is GeoEye-1, which has 46cm panchromatic resolution [4]), it is still far from the resolution of common-object detection datasets (e.g., animals, vehicles). In fact, the labeling task on satellite imagery is hard even with human visual inspection, which leads to another challenge. The volunteers' annotation could be erroneous. To limit this, the crowd-sourcing platform has a proprietary system that computes the agreement score of each label. In this work, we ignore this information to gather as many labels as possible and take the given labels as ground truth since limited size of training data could be a critical bottle-neck for models with many parameters to learn such as NNs. Third, there are some inconsistencies in image quality. Since the same region can be captured multiple times on different days, the same coordinate may have multiple images of different qualities (e.g., due to pre-processing), as shown in Figure 2.3. In summary, effective learning algorithms should overcome the challenges from low-resolution images, noisy labels, and inconsistent image qualities.

2.3 Methodology

In this section, we describe our end-to-end methodological framework from collecting, processing, featurizing data to building the convolutional neural network to classify whether a building in a satellite image is flooded/damaged or not.

(a) Lower-quality orthorectification (b) Higher-quality orthorectification

(c) More blurry (d) Less blurry

Figure 2.3: Different orthorectification and pre-processing quality of the same location on different days.

2.3.1 Data description

The satellite imagery of the Greater Houston area was captured by optical sensors with sub-meter resolution, preprocessed (e.g., orthorectification and atmospheric compensation), and pansharpened by the image provider. The raw imagery consists of around four thousand

image strips taken on multiple days (each strip is roughly 1GB and has around 400 million pixels with RGB bands). Some strips overlap and have black pixels in the overlapped region. Some images are also covered fully or partially by clouds. Figure 2.4 shows a typical strip in the dataset and Figure 2.5 shows some examples of *low quality* images (from the perspective of model training) that we chose to discard.

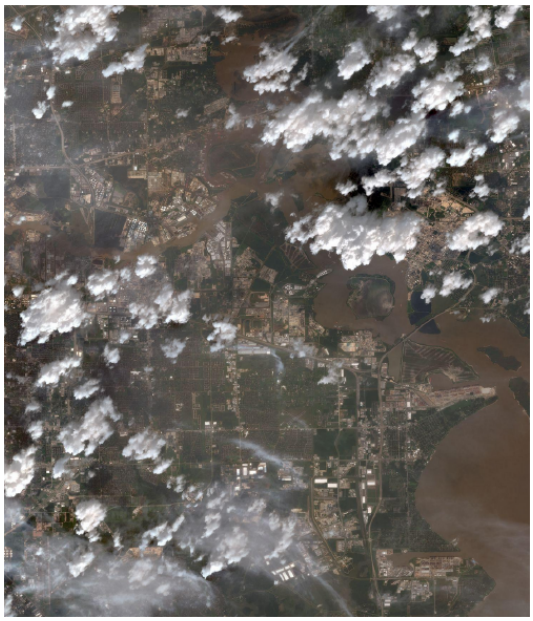


Figure 2.4: A typical strip of image in the dataset.

2.3.2 Damage annotation

We present here our methodological framework (Figure 2.6) that starts from raw data input to create damage annotation output. The first step is to process the raw data to create training-ready data by using a cropping window approach. Essentially, the building coordinates, which can be easily obtained from public data (e.g., OpenStreetMap [6]), can be used as the centers of cropping. We use the building coordinates already associated with the damage labels from Tomnod. A window is then cropped from the raw satellite imagery to create a data sample. Tomnod volunteers' annotation of flooded/damaged buildings is taken

(a) Blacked out partially

(b) Covered by cloud partially

(c) Covered by cloud mostly

(d) Covered by cloud totally

Figure 2.5: Examples of discarded images during the data cleaning process due to their potential to hinder model training.

as the ground truth for the positive label, 'Flooded/Damaged building'. At the same coordinates, we crop windows from the imagery captured before the hurricane to create negative data samples, labeled 'Undamaged building'.

The optimal window size depends on various factors including the image resolution and

building footprint sizes. Too small windows may limit the background information contained in each sample, whereas too large ones may introduce unnecessary noise. We keep the window size as a tuning hyper-parameter in the model. A few sizes are considered such as 400x400, 128x128, 64x64, and 32x32.

The cropped images are then manually filtered to ensure the high quality of the dataset. To let the model generalize well, we only discard the images that can obviously hamper the algorithm's learning process, such as the example images in Figure 2.5. The cleaned images are then split into training, validation, and test sets and fed to a convolutional neural network for damage annotation as illustrated in Figure 2.6. Validation accuracy is monitored to tune the necessary hyper-parameters (including the window size).

2.3.3 Data processing

As described above, the data generation starts from a building coordinate. Since there are multiple raw images containing the same coordinates, there are duplicate images with different quality. This can potentially inflate the prediction accuracy as the same coordinate may appear in both the training and test sets. We maintain a set of the available coordinates and make sure each coordinate is associated with a unique, "good-quality" image in the final dataset through a semi-automated process. We first automatically discard the totally blacked out images for each coordinate, and keep the first image we encounter that is not totally black. The resulting set of images are manually filtered to eliminate the images that are partially black or covered by clouds.

2.3.4 Data featurization

Since we control the window size based on physical distance, there could be round-off errors when converting the distance to the number of pixels. Therefore, we project them into the same feature dimension. For instance, both a 128x128 image and a 127x129 image are projected into 150x150 dimension. The images are then fed through a CNN to further extract useful features, such as edges, as illustrated in Figure 2.7.

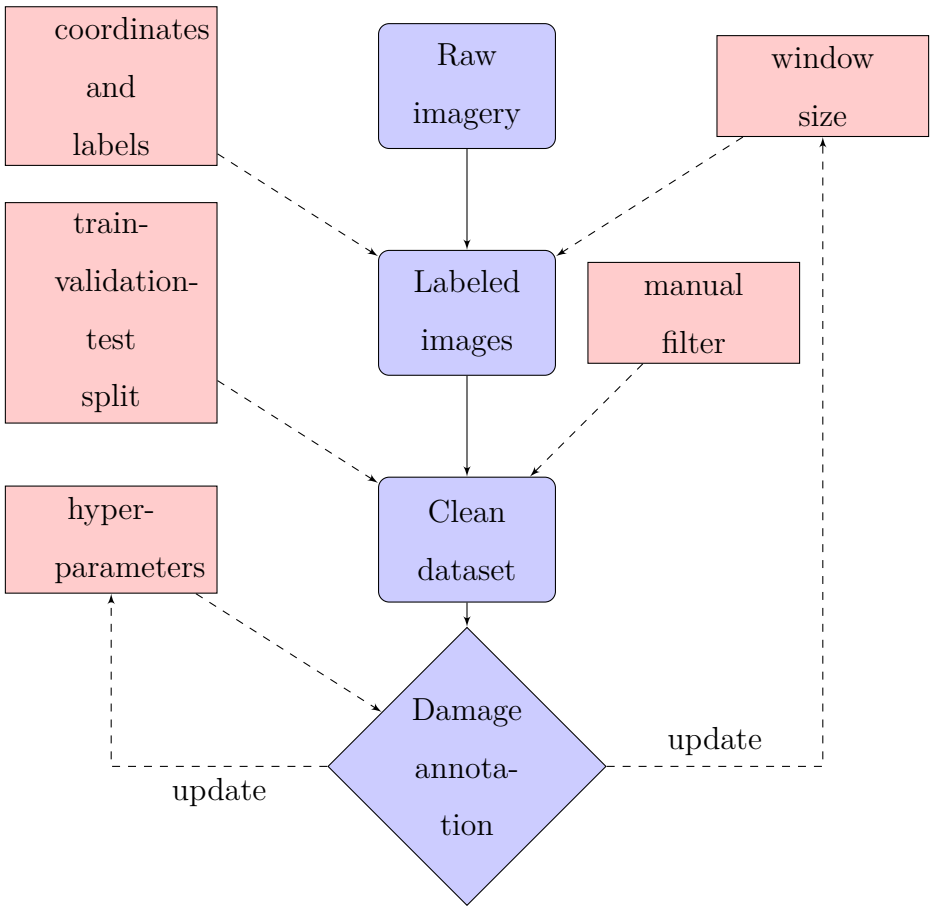
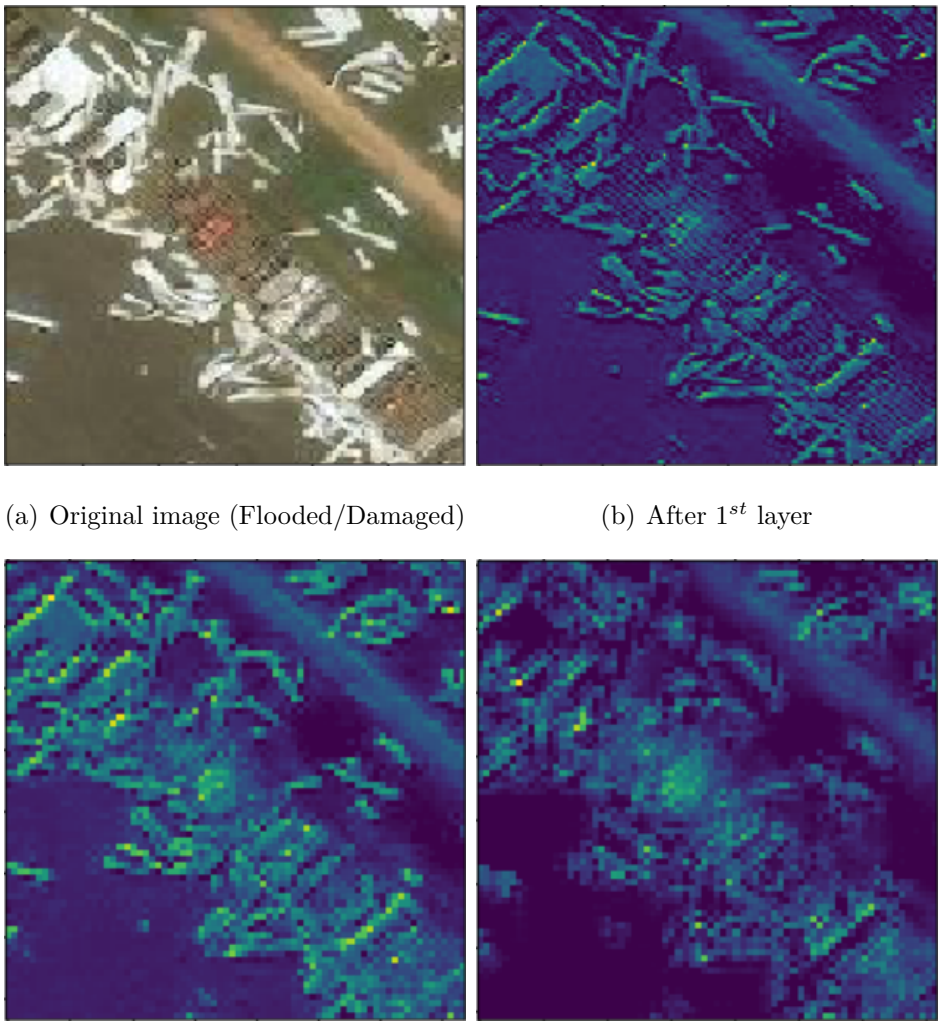


Figure 2.6: The damage annotation framework.



(a) Original image (Flooded/Damaged)

(b) After 1^{st} layer

(c) After 2^{nd} layer

(d) After 3^{rd} layer

Figure 2.7: Information flow within one filter after each convolutional layer. The initial layers act as a collection of edge extraction. At a deeper layer, the information is more abstract and less visually intepretable.

How to construct the most suitable CNN architecture is an ongoing research problem. The common practice, known as *transfer learning*, is starting with a known architecture and fine-tuning it. We experiment with a well-known architecture, VGG-16 [114], and modify

the first layer to suit our input dimension. VGG-16 is known to perform very well on the ImageNet dataset for common object classification.

However, because of the substantial differences between the common object classification and our flooded/damaged building classification, we also build our own network from scratch. We carefully consider proper hyper-parameters, as similarly done in [76]. Our basis for determining the size and depth of a customized network is to monitor the information flow through the network and stop enlarging the network when there are too many *dead* filters (*i.e.*, blank filters that do not carry any further information to the subsequent layers in the network). Due to the nature of the rectified linear unit (ReLU), which is defined as $\max(0, x)$, there will be many zero weights in the hidden layers. Although sparsity in the layers can promote the model to generalize better, it may cause the problem on gradient computation at 0, which in turns does not update any parameters, and hurt the overall model performance [59,124]. We see that in Figure 2.8 after four convolutional layers, about 30% of the filters are dead and will not be activated further. This is a significant stopping criterion since we can avoid a deep network such as VGG-16 to save the computational time and safeguard satisfactory information flow in the network at the same time.

We present our customized network architecture that achieves the best result in Table 2.1. The network begins with four convolutional and max pooling layers and ends with two fully connected layers.

In our CNN structure, with four convolutional layers and two fully connected layers, there are already about 3.5 million parameters to train, given $67,500$ pixels as an input vector for each image. The VGG-16 structure [114], with thirteen convolutional layers, has almost 15 million trainable parameters, which can over-fit, require more resources, and reduce generalization performance on the testing data. In addition, as discussed in [76], the network depth should depend on the complexity of the features to be extracted from the image. Since we have only two classes of interest, a shallower network can be favourable in terms of training time and generalization.

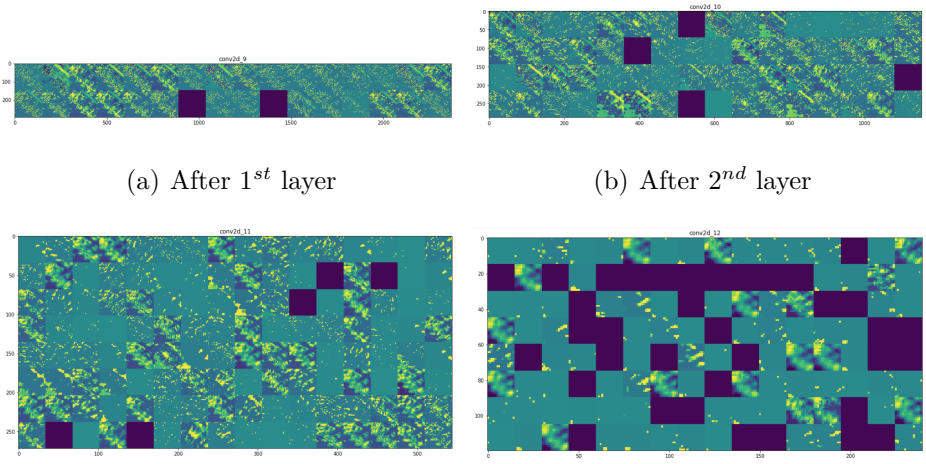
Table 2.1: Convolutional neural network architecture that achieves the best result.

Layer type	**Output shape**	**Number of trainable parameters**
Input	3@(150x150)	0
2-D Convolutional 32@(3x3)	32@(148x148)	896
2-D Max pooling (2x2)	32@(74x74)	0
2-D Convolutional 64@(3x3)	64@(72x72)	18,496
2-D Max pooling (2x2)	64@(36x36)	0
2-D Convolutional 128@(3x3)	128@(34x34)	73,856
2-D Max pooling (2x2)	128@(17x17)	0
2-D Convolutional 128@(3x3)	128@(15x15)	147,584
2-D Max pooling (2x2)	128@(7x7)	0
Flattening	1x6272	0
Dropout	1x6272	0
Fully connected layer	1x512	3,211,776
Fully connected layer	1x1	513

Note: The total number of trainable parameters is 3,453,121. C@($A \times B$) is interpreted as that there are a total of C matrices of shape ($A \times B$) stacked on top of one another to form a three-dimensional tensor. 2-D Max pooling layer with (2×2) pooling size means that the input tensor's size will be reduced by a factor of 4.

2.3.5 Image classification

Due to the limited availability of pre-event images and the exclusion of some images (e.g., due to cloud coverage) in the *Flooded/Damaged* and *Undamaged* categories, our dataset is



(a) After 1^{st} layer

(b) After 2^{nd} layer

(c) After 3^{rd} layer

(d) After 4^{th} layer

Figure 2.8: Information flow in all filters after each convolutional layer. The sparsity increases with the depth of the layer, as indicated by the increasing number of dead filters.

unbalanced with the majority class being *Flooded/Damaged*. Thus, we split the dataset into training, validation, and test datasets as follows. We keep the training and validation sets balanced and leave the remaining data to construct two test sets, a balanced set and an unbalanced (with a ratio of 1:8) set.

The first performance metric is the classification accuracy. In contrast to the balanced test set, we note that the baseline accuracy for the *unbalanced* test set is $8/9 = 88.89\%$ (greater than the random guess accuracy, 50%), which can be achieved by annotating all buildings as the majority class *Flooded/Damaged*. In addition, as the classification accuracy is sometimes not the most pertinent performance measure, we also monitor the area under the receiver operating characteristic curve (AUC), which is a widely-used criterion to measure the classification ability of a binary classifier under a varying decision threshold [44].

2.4 Implementation and Result

We train the neural network using the *Keras* library with TensorFlow backend with a single NVIDIA K80 Tesla GPU. The network weights are initialized using Xavier initializer [58]. The mini batch size for the stochastic gradient descent optimizer is 32.

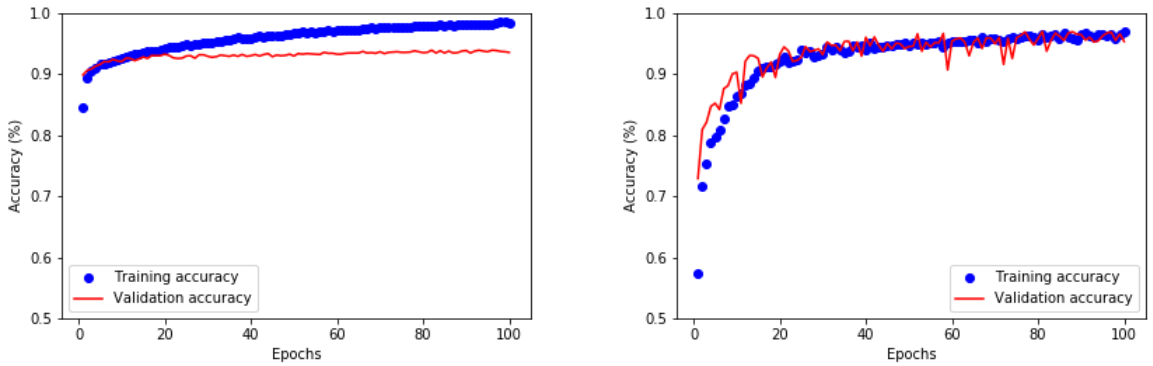
After the data cleaning process, our dataset contains 14,284 positive samples (*Flooded/Damaged*) and 7,209 negative samples (*Undamaged*) at unique geographical coordinates. 5,000 samples of each class are in the training set. 1,000 samples of each class are in the validation set. The rest of the data are reserved to form the test sets, i.e. in the balanced test set, there will be 1,000 samples of each class, and in the unbalanced test set, there will be 8,000 samples of *Flooded/Damaged* class and 1,000 samples of *Undamaged* class.

Due the expensive computational cost of training the CNN, we investigate selected combinations of the hyper-parameters in a greedy manner, instead of tuning all the hyper-parameters through a full grid search or full cross-validation. For example, we investigate the performance of a model with multiple window sizes (400x400, 128x128, 64x64, and 32x32) and select the 128x128 window size.

We also implement a logistic regression (LR) on the featurized data to see how it compares to fully connected layers. Although LR under-performs in most cases, it still achieves good accuracy (little over 90% in Table 2.2). This illustrates that the image featurization through the network works well enough that a simple algorithm like LR can perform well on this data.

For activation functions in the CNN, a rectified linear unit (ReLU) is a common choice, thanks to its simplicity in gradient computation and prevention of vanishing gradient, which is common with other activation functions such as sigmoid or hyperbolic tangent. But, as seen in Figure 2.8, clamping the activation at 0 could potentially cause a lot of filters to be dead. Therefore, we also consider using a leaky ReLU activation with $\alpha = 0.1$ based on the survey in [124]. However, leaky ReLU turns out to not significantly improve the accuracy in our implementation (Table 2.2).

To counter over-fitting, which is a recurrent problem of deep learning, we also adopt data augmentation in the training set through random rotation, horizontal flip, vertical and horizontal shift, shear, and zoom. This can effectively increase the number of training samples to ensure better generalization and achieve better validation and test accuracy (Note that we do not perform data augmentation in the validation and test sets). Furthermore, we also employ 50% dropout and L2 regularization with $\lambda = 10^{-6}$ in the fully connected layer. Dropout [116] is an effective method to prevent over-fitting, especially in neural network with many neurons. The method prevents neurons from remembering too much training data by dropping out randomly chosen neurons and their connections during the training time. L2 regularization is one of the regularization techniques that has been shown to perform better on ill-poised problems or noisy data. Early application of the regularization in computer vision can be traced back to edge detection in images where the changes in intensity in an image are considered noisy [21]. These measures are shown to fight over-fitting effectively and significantly improve the validation accuracy in Figure 2.9.

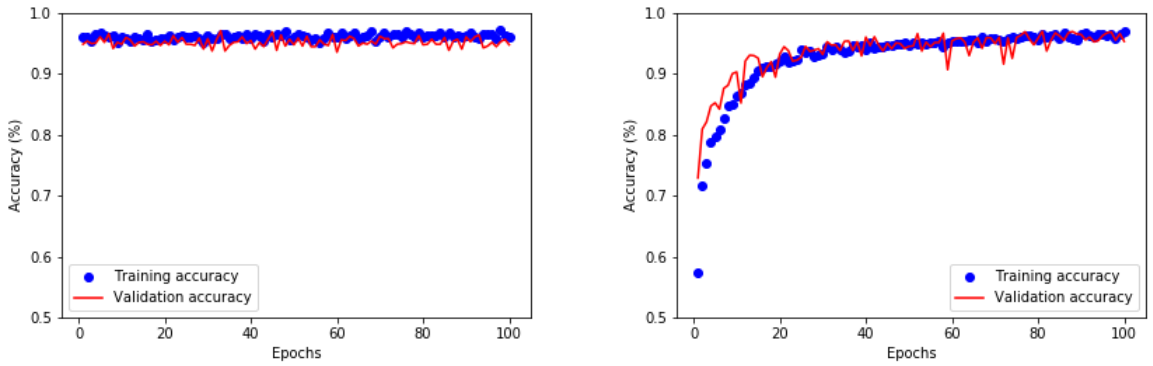


(a) Without drop-out and image augmentation, the validation accuracy separates from the training accuracy.

(b) No apparent sign of over-fitting can be seen as over-fitting seems to happen after about 10 epochs as the validation accuracy follows the training accuracy.

Figure 2.9: Over-fitting is prevented using data augmentation, drop-out, and regularization.

As mentioned in Section 2.3.4, we consider using a pre-built architecture VGG-16 (transfer learning) and building a network from scratch. In Figure 2.10, we see that the deeper and larger network can achieve a high-level validation accuracy earlier, but the accuracy pretty much plateaus (*i.e.*, over-fitting happens) after a few epochs. Our simpler network can facilitate learning gradually, where the validation accuracy keeps increasing to achieve a higher value than the deeper network, and takes about 75% less training time.



(a) Transfer learning using pre-built network

(b) Custom network

Figure 2.10: Comparison between using a pre-built network and our network. The two networks almost have the same level of performance except our network achieves a slightly better accuracy with a much smaller network size. It is also noticeable that due to large number of pre-trained parameters, the bigger network achieves high accuracy right at the beginning but fails to improve subsequently.

We use two adaptive, momentum-based optimizers, RMSprop and Adam [70], with the initial learning rate of 10^{-4}. Adam generally leads to about 1% higher validation accuracy and less noisy learning in our implementation.

Table 2.2 summarizes the performances of various models. The best performing model is our customized network with data augmentation and dropout using Adam optimizer, which can achieve 97.08% accuracy on the unbalanced test set. The AUC metric is also computed

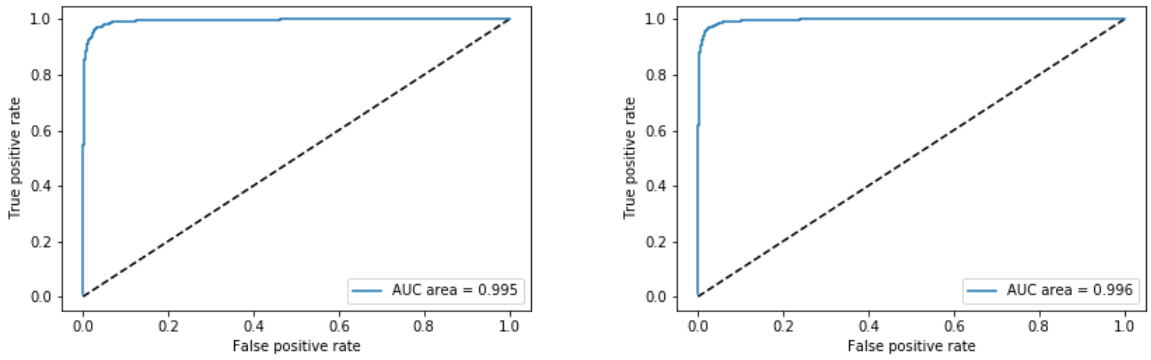
and shows a satisfying result of 99.8% on the unbalanced test set.

Table 2.2: Model performance.

Model	**Validation Accuracy**	**Test Accuracy (Balanced)**	**Test Accuracy (Unbalanced)**	**F1 Score**
CNN	95.8%	94.69%	95.47%	0.9575
Leaky CNN	96.1%	94.79%	95.27%	0.9558
CNN + DA + DO	97.44%	96.44%	96.56%	0.9674
CNN + DA + DO (Adam)	**98.06%**	**97.29%**	**97.08%**	**0.9723**
Transfer + DO	93.45%	92.8%	92.8%	0.9304
Transfer + DA + DO	91.1%	88.49%	85.99%	0.8800
LR + L2	93.55%	92.2%	91.45%	0.7713
SVM + L2	92.02%	91.85%	90.95%	0.7002
Transfer + DA + FDO	96.5%	95.34%	95.73%	0.9594
Leaky + Transfer + DA + FDO +L2	96.13%	95.59%	95.68%	0.9598
Leaky + Transfer + DA + FDO + L2(Adam)	97.5%	96.19%	96.21%	0.9643

Legend: CNN: Convolutional Neural Network; Leaky: Leaky ReLU activation function, else, the default is ReLU; DA: Data Augmentation; LR: Logistic Regression with features built by convolutional operations; L2: L2 regularization; SVM: Support vector machine classifier; (Adam): Adam optimizer, else, the default is RMSprop optimizer; DO: 50% dropout only in the fully connected layer; FDO: Full dropout, *i.e.,* 25% dropout after every max pooling layer and 50% in the fully connected layer; Transfer: Transfer learning using VGG-16 architecture.

Although the overall result is satisfactory, we also investigate a few typical cases where



(a) AUC of balanced test set

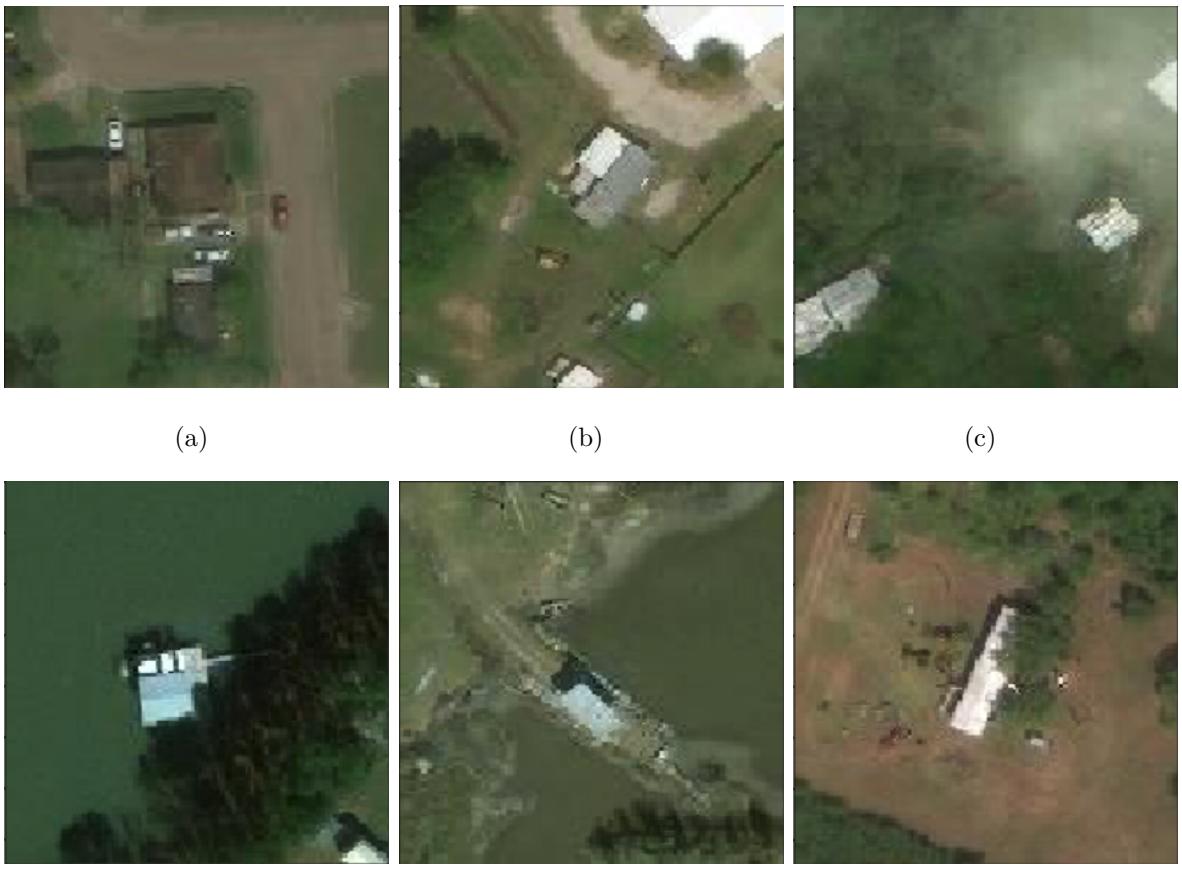
(b) AUC of unbalanced test set

Figure 2.11: AUC for the balanced and unbalanced test sets using our best performing model—CNN + DA + DO (Adam)—in Table 2.2.

the algorithm makes wrong classification to see if any intuition can be derived. Figure 2.12 shows some of the false positive cases. We hypothesize that the algorithm could predict the damage through flood water and/or debris edges. Under such hypothesis, the cars in the center of Figure 2.12(a), the lake water in Figure 2.12(b), the cloud covering the house in Figure 2.12(c), and the trees covering the roof in Figure 2.12(f) can potentially mislead the model. For the false negative cases in Figure 2.13, it is harder to make sense out of the prediction. Even through careful visual inspection, we cannot see Figures 2.13(a)(b) as being flooded/damaged. These could potentially be labeling mistakes by the volunteers. On the other hand, Figures 2.13(e)(f) are clearly flooded/damaged, but the algorithm misses them.

2.5 Summary

We demonstrated that convolutional neural network can automatically annotate flooded/damaged buildings on post-hurricane satellite imagery with high accuracy. While our data is specific to the geographical condition and building properties in the Greater Houston area during Hurricane Harvey, the model can be further improved and generalized to other future hurri-



(a) (b) (c)

(d) (e) (f)

Figure 2.12: False positive examples (the label is *Undamaged*, whereas the prediction is *Flooded/Damaged*).

cane events in other regions by collecting more positives samples from other past events and negative samples from other areas. From more data, we can obtain both more robust models as well as sets of hyper-parameters. As mentioned in Section 3, the cropping window size is dependent on the resolution of the available imagery (the mode of collection), as well as the typical building footprint size in the region. From our grid search on the window size, as long as the window captures the building adequately, the performance does not vary significantly so the data collection operator can still have some flexibility about the collection methods

(a) (b) (c)

(d) (e) (f)

Figure 2.13: False negative examples (the label is *Flooded/Damaged*, whereas the prediction is *Undamaged*).

(e.g., using drones).

For faster disaster response, a model should be able to process and annotate on low-quality images. For example, images taken right after a hurricane landfall can be covered largely by cloud. Also, image providers might not have enough time to pre-process images well due to the urgency of situation. We will investigate how a model can be made robust against such noise and distortion to reliably annotate damages.

Although the current work extracts the positive and negative samples based on different

temporal information of the dataset, it would be more realistic to gather the samples from different spatial information using the same timestamp after the event happens. This would make the data closer to the actual scenario when the method is targeted to be deployed after an event happens. This direction warrants further investigation in future work. We also wish to extend the model to the annotation of road damages and debris, which could help plan effective transportation routes of medical aids, foods, or fuels to hurricane survivors.

Appendix: dataset and code

The dataset and code used in this work are available at the first author's Github repository `https://github.com/qcao10/DamageDetection`. The dataset is also available at the IEEE DataPort (DOI: 10.21227/sdad-1e56).

Chapter 3

POST-HURRICANE DAMAGE ASSESSMENT USING SATELLITE IMAGERY AND GEOLOCATION FEATURES

3.1 Introduction

Damage assessment after hurricane makes landfall is increasingly gaining attention within the disaster research community. The current practice of windshield survey, which relies on emergency response crews and volunteers to drive around the affected area is known to be costly and time consuming. To speed up the process, several researches have been conducted to reduce data collection time or assist the visual inspection. One notable direction is using deep learning to detect whether a building is damaged or not after a hurricane event. In our previous work [26], we have shown that using satellite imagery from crowdsourcing campaign can achieve state of the art performance on classifying damaged building based on several metrics such as accuracy, precision-recall, and F1 score. Besides satellite imagery in Red-Green-Blue (RGB) band, other studies have explored the use of deep learning in flood risk assessment using IKONOS-2 imagery [122], or time series of satellite imagery [115]. In this work, we continue to use the high-resolution optical sensor RGB imagery since they are easier to be interpreted.

The remaining of this chapter is organized as follows. In Section 3.2, we present a brief review of current literature on convolutional neural network and its applications using multimodal data . Section 3.3 describes our proposed methodology in dataset construction and model architecture. Details of the implementation and discussion of the results are presented in Section 3.4. Finally, Section 3.5 concludes this work and draws some future research directions.

3.2 Background

3.2.1 Relevant flood damage assessment practice

In this section, we review some of the state-of-the-art methods in flood damage assessment. There has always been a close relationship between flooding properties, ground topography and damage quantification. Within the field of hydrology, the role of bare-earth topography elevation is so important to hydraulic modeling of water flows that the elevation estimation methods have been studied in various works such as [103, 126]. In [73], using moderate resolution imaging spectrometer (MODIS) time-series imagery, the crop damage extend at each map grid pixel (approximately 500m) is studied as a function of flood depth and flood duration. Similarly, a study by [88] attempts to classifies regions of pixels into water, vegetation, urban, and bare land at the coastal areas of Japan after the earthquake triggered tsunami. In a separate study, relative frequency of flood inundation is shown to exhibit the same probabilistic distribution as relative water depth, which is characterized by bed elevation [115].

The above methods mostly utilize variants of synthetic aperture radar (SAR) imagery and/or earth elevation from digital elevation model (DEM) in their work. SAR imagery has its own advantages in mapping surface features, or roughness pattern. However, it could be harder for laymen (e.g emergency managers and first responders) to interpret than optical sensor imagery. In addition, there are fewer satellite equipped with SAR sensors than optical sensors. Our approach in this work is to leverage the availability and interpretability of high-resolution RGB satellite imagery. Our goal is to create a framework that can be readily deployed through crowd-sourcing campaigns such as the one used in this work or aerial imagery collected from drones.

3.2.2 Deep learning with mixed data

Recent studies have demonstrated that the information is much richer when combining images and other modality of data. In [118], image classification can be improved by including

location context, derived from Global Position System (GPS) tags, of the images. Similarly, [72] also successfully learn that images on Instagram and text caption can interact with each other to inform a more complex meaning that can explain their intent, contextual, and semiotic relationships. Perhaps most related to our work are the studies showing promising results in predicting housing price using traditional housing attributes, such as area, number of rooms, zipcode, etc. combining with the house interior/exterior photo [12], or the neighbourhood street and aerial views [74]. There is still a relatively smaller number of researches about using multimodal data compared to pure image feature. As highlighted by many authors in the field, there are various levels of challenges in collecting data and how to incorporate the non-image features effectively into the (CNN) model. In our work, we also encounter similar issues and data collection and preprocessing easily take up a major amount of work. Nonetheless, the result is really rewarding for us to achieve state-of-the-art performance in post-hurricane damage assessment. The computational cost, given the data is available, is still much more efficient than physical data collection and site survey.

3.3 Methodology

In this section, we describe our dataset and model architecture.

3.3.1 Data description

The data we used are the publicly available imagery captured after Hurricane Harvey event (post-event data), plus the coordinates annotated by crowd-sourcing volunteers as they think a building is damaged or flooded, made available by DigitalGlobe [7]. The raw imagery data covering the Greater Houston area was captured in about four thousand strips ($\sim$ 400 million pixels ($\sim$ 1GB) with RGB bands per strip) in different days. In our previous work [26], we used the post-event imagery to crop the images at those coordinates to build the positive labels (*Damaged*), and pre-event data at the same coordinates to build the negative labels (*Undamaged*). This approach using temporal difference to separate the data presents some limitations in terms of modelling and usability in the future. As can be seen in Figure 3.1a

and Figure 3.1b, the positive and negative labels from different timestamps may have different color scale, hue, or saturation. In addition, there is flood water almost in every positive labels, which might lead the model into water detection rather than actual damage detection.

In current work, we extract the data from the same post-event imagery (Figure 3.1c and Figure 3.1d, where it is inherently more difficult for the model to distinguish between damaged and undamaged building where flood water already invades most of the area. The color scale, hue, or saturation are also more consistent across the whole dataset to eliminate undesirable learning by the color scale.

(a) *'Undamaged Building'* using pre-event imagery

(b) *'Flooded/Damaged'* using post-event imagery

(c) *'Undamaged Building'* using post-event imagery

(d) *'Flooded/Damaged'* using post-event imagery

Figure 3.1: Two different ways to construct the dataset. The first way (a and b) uses different temporal information of the same location to label the data. The second way (c and d) uses different spacial information of the same timestamp to label the data

Since the coordinates provided by DigitalGlobe only consists of positive labels, we need to manually collect the negative labels ourselves as shown in Figure 3.2. From the OpenStreetMap API [6], we divide the area into customized, much smaller strips to extract buildings coordinates that does not share the same footprint with the coordinates provided by DigitalGlobe's volunteers. To this end, we are assuming that the search by the volunteers are exhaustive, and every building coordinate not found by the volunteers is considered as

undamaged.

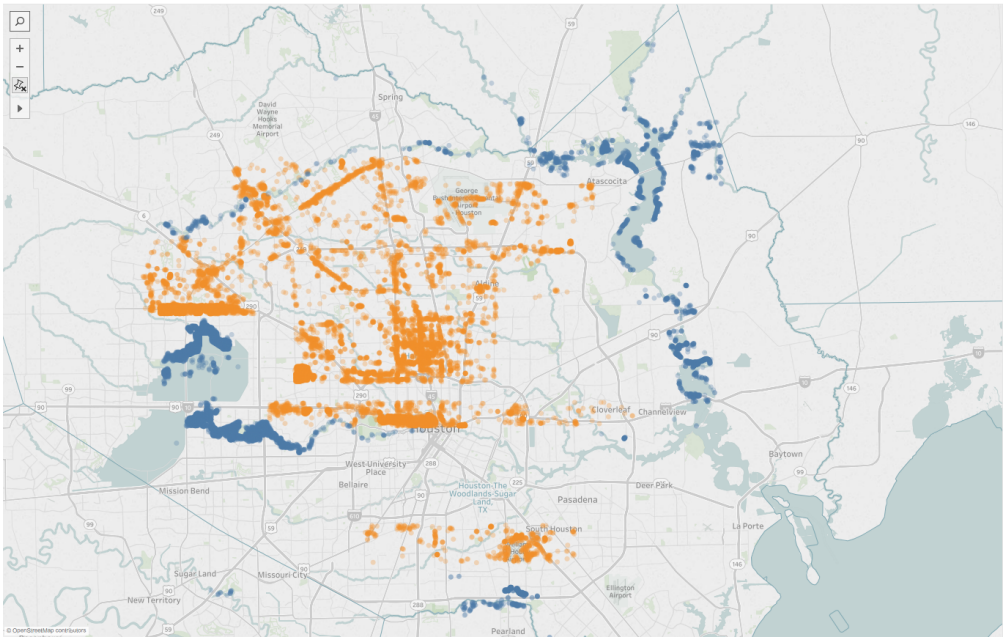
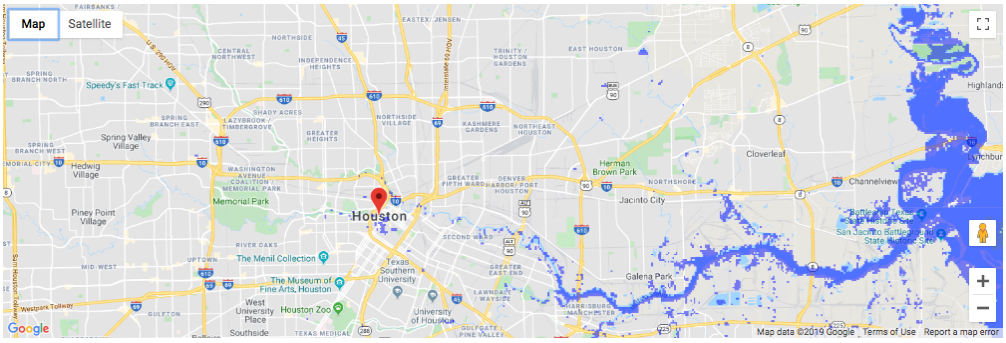


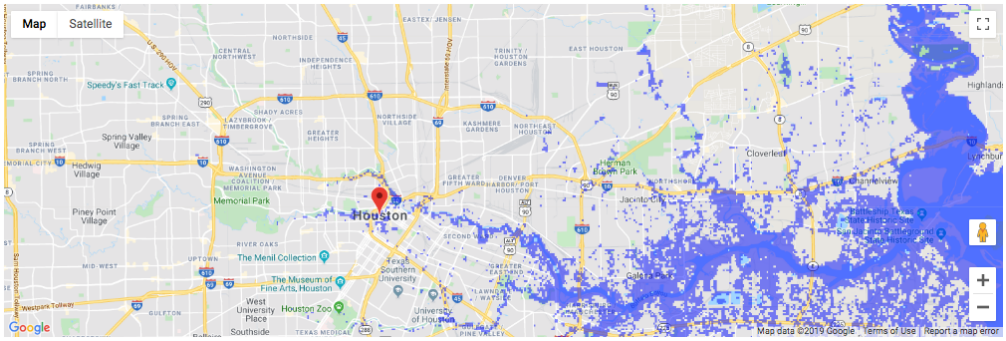
Figure 3.2: *'Flooded/Damaged'* and *'Undamaged'* building locations after the events.

After collecting the set of coordinates for both labels, we use Google Map Developers API [5] to get the elevation at these coordinate to build the elevation feature for the dataset. From the same set of coordinates, we use QGIS GRASS API to find the distance from each coordinate to their nearest water body. The raster data for Texas area water bodies are provided by the United States Geological Survey (USGS) Geographic Information System Data [8].

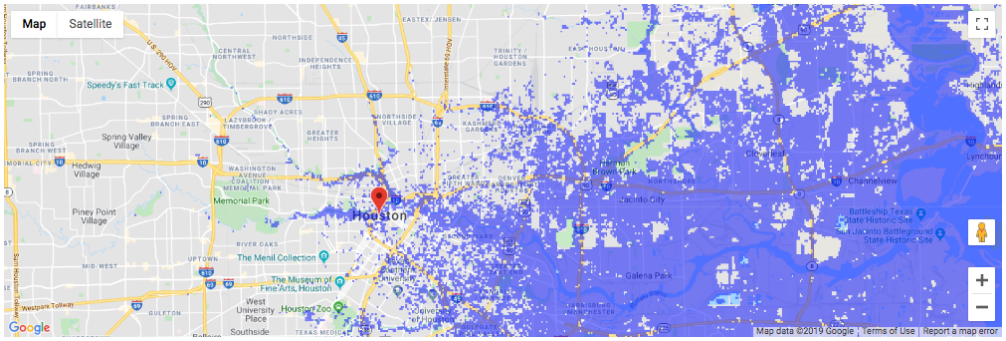
The rationale behind choosing distance from water bodies and elevation as extra geolocation features comes from some visualization and analyses. As can be seen in Figure 3.2, most of the *'Flooded/Damaged'* buildings are very close to the major water bodies in the region. This is further confirmed through a flood simulation in the Houston area as shown in Figure 3.3. As we increase the flood depth, area around major water bodies have higher likelihood of being flooded.



(a) Simulation of 5-meter flood in Houston area



(b) Simulation of 10-meter flood in Houston area



(c) Simulation of 15-meter flood in Houston area

Figure 3.3: Simulation of different levels of flooding in the Houston area .

For the second geolocation feature, we also make some strategic comparison between choosing elevation and individual building coordinates. Initially, coordinates seems to be a

logical choice to encode the neighbouring relationship of building cluster, which share similar disaster resilience characteristics and tend to be affected together. However, elevation can be even more informative. First, it can be used as an encoder for neighbouring houses as well, as nearby houses tend to not differ much in elevation. Second, we can capture an obvious physical behaviour, in which lower elevation may result in higher likelihood of getting flooded. Last but not least, elevation may be used to generalize to other regions, whereas coordinates practically cannot. Another region may have a different elevation, and different flood catchment capacity but as long as their relative difference in elevation still prevail, we can still deploy the model trained using Houston information to quickly perform damage assessment over there.

3.3.2 Model description

The models presented in Section 3.4 are based on deep convolutional neural network. The model consists of an image encoder for the imagery, some fully connected layers to encode the geolocation features, some fusion layers to combine the two encoded information, and a class prediction layer.

For image encoder, the same convolutional setup in [26] is adopted with a sequence of convolution layers, max pooling layers, followed by a fully connected layer. At the end, the image encoder yields a 4 dimensional embedding for the imagery. For the geolocation encoder, two layers of fully connection results in a 4 dimensional embedding. These two embedding are concatenated to form a common embedding dimension of 8 in the fusion layers, which yields the final single node for class prediction after a few more fully connected layer. (Figure 3.4).

There are some hyper-parameter tuning works in the embedding size. We would like to investigate the effect of giving the same embedding sizes to the imagery and the geolocation features. From our previous work, we know that the image encoding works quite well with the image feature alone so there is no issue with using a small embedding size (e.g 4 dimensions). The question remains whether to give the geolocation the same or smaller embedding size

since we start with only two features. This decision is informed through analyzing the performance of the model using purely imagery and geolocation features. As can be seen in Section 3.4, geolocation features already provide a good signal to the likelihood of building damage, almost comparable to the imagery, which leads us to give the the equal embedding size in the combined model.

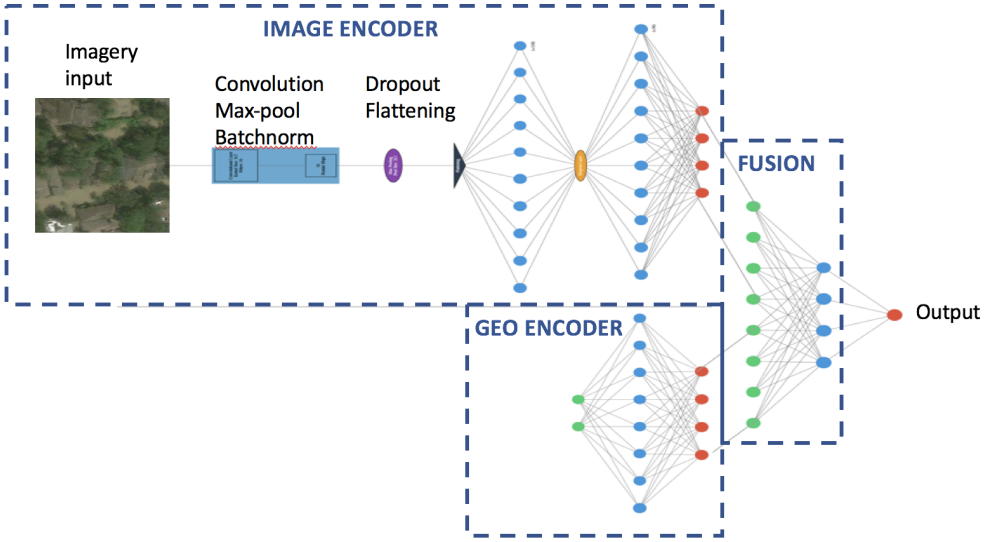


Figure 3.4: The mixed data neural network model that utilizes both satellite imagery and geolocation features to detect damaged building.

3.4 Implementation and Result

The network is trained to optimize with the Adam optimizer using the cross-entropy loss. We run all experiments on CentOS 7.7 (64-bit Linux) by training and validating our model using 5-fold cross-validation for 70 training epochs. In total, we spent 25 GPU hours to run all experiments. The model is built through the *Keras* library with TensorFlow backend with a single NVIDIA K80 Tesla GPU with 64GB memory on a quad-core CPU machine.

The individual images are cropped from the original imagery at the window size of

256x256 due to its better performance in the mixed data model. We experimented with two different cropping window sizes, 128x128 and 256x256 pixels, as we observed they yielded the highest performance metrics in our previous work.

This method relies heavily on the availability and quality of data and therefore poses some potential limitations. First, the imagery data was taken as a time series, with a lot of pan-sharpening and cloud coverage issues. It takes multiple iterations to get a reasonbale amount of usable data. Second, the geolocation data comes from different sources. Some data sources do not have complete data and/or use different formats. Preprocessing is intensive to join all the data together based on their geo-coordinates. Nevertheless, the entire process can be done computationally and added extra time and manpower saving to traditional damage assessment practice such as post-event windshield survey.

After cleaning and manual filtering, we are left with 13,993 positive samples (*Damaged*) and 10,384 negative samples (*Undamaged*) of unique coordinates. The dataset is split to have 67% of the data as training data as 33% as test data. The class distribution is preserved similarly to the original distribution in training and test data. The split is repeated 5 times to form 5 cross-validation sets, in each of which we train and test using the same model architecture to get both the mean and standard deviation of performance.

We present our model performance results on the test data in Table 3.1 based on the probability threshold of 50% to determine a class prediction. Due to class imbalance in our dataset, the metrics used here are accuracy, F1 score, precision, and recall. Image feature yields better precision. On the other hand, geolocation features seem to provide better recall than precision so it is more effective in detecting *Damaged* samples. This is not surprising since the geolocation features are carefully designed and we expect building damage to follow physical laws. Generally, combining different types of data yields quite balanced performance and improves all metrics. Depending on priority of the model users, probability threshold can be adjusted to trade for more recall in order to identify more *Damaged* samples.

Table 3.1: Performance metrics across models

Method	Metrics			
	ACC	Precision	Recall	F1 score
Img only	$79.5 \pm 8.3\%$	0.88 ± 0.03	0.64 ± 0.30	0.68 ± 0.22
Geo only	$88.6 \pm 1.4\%$	0.86 ± 0.03	0.97 ± 0.003	0.91 ± 0.02
Img + Geo	$97.47 \pm 2.5\%$	0.91 ± 0.14	0.99 ± 0.003	0.94 ± 0.08

Remarks: Img only: model trained on image feature only; Geo only: model trained on geolocation features only; Img + Geo: model trained on both types of data. Each performance metric reported here shows the mean ± standard deviation across 5 cross-validation sets.

3.5 Summary

We have demonstrated that damaged buildings can be detected with decent accuracy. Geolocation information can substantially improve the performance of CNN, and reduce the hyper-parameter tuning work. The model can be generalized to other regions and events as more data from more hazard events is aggregated. Since the geolocation features used are carefully chosen as relative elevation and relative proximity to water body, the model can be adapted to deploy to other regions and events without retraining. It could be that the relationship between elevation and flood likelihood is specific to regions, but we are trying to capture the neighboring representation of the buildings so to apply to another region, we might only need to adjust the output likelihood linearly, to gain more recall, if necessary. However, there is be a trade off between using too specific features to a disaster type (e.g. hurricane) such as the proximity to water bodies and generalization to other types (e.g earthquakes). It could be helpful for the disaster management community to train a few models with performance validated on past events, that can be ready when another event makes landfall.

Chapter 4

INFRASTRUCTURE RECOVERY CURVE ESTIMATION USING GAUSSIAN PROCESS REGRESSION ON EXPERT ELICITED DATA

4.1 Introduction

Infrastructure recovery time estimation is critical to disaster management and planning. Inspired by recent resilience planning initiatives, we consider a situation where experts are asked to estimate the time for different infrastructure systems to recover to certain functionality levels after a scenario hazard event. We propose a methodological framework to use expert-elicited data to estimate the expected recovery time curve of a particular infrastructure system. This framework uses the Gaussian process regression (GPR) to capture the experts' estimation-uncertainty and satisfy known physical constraints of recovery processes.

While more data would generally yield a more accurate estimate, there is a practical limitation on collecting expert-elicited data. We study how to balance between the cost of collecting data from expert elicitation and the estimation accuracy of GPR. We consider multiple expert elicitation schemes to identify the best way to estimate the recovery curve with a reasonable cognitive burden on experts while maintaining good estimation accuracy.

We simulate expert-elicited data by randomly generating expert estimates, which are assumed to be generally close to the empirical recovery curve observed in a case study event. We evaluate the proposed estimation method based on different empirical recovery curves from different prefectures and infrastructures after the 1995 Great Hanshin-Awaji Earthquake and the 2011 Great East Japan Earthquake [100].

The rest of this chapter is organized as follows. Section 4.2 briefly reviews relevant literature on expert elicitation and GPR. Section 4.3 presents the proposed estimation methodol-

ogy. Section 4.4 shows the performance of this method through extensive numerical studies and sensitivity analyses. Section 4.5 draws insights for potential users of this method and concludes the research.

4.2 Background

4.2.1 Expert Elicitation for Disaster Recovery Estimation

Participatory methods, especially expert elicitation, have been used extensively in disaster research especially in the areas where empirical data are scarce [90, 92]. The study in [30] elicits from experts infrastructure recovery estimates (at 0 hours, 72 hours, and 2 weeks from a hypothetical event) and qualitative inter-dependencies between those infrastructures. However, the study limits itself to short-term restoration and does not factor uncertainties into the recovery time estimation.

Expert elicitation itself is a well-established research domain [41, 61]. One of the most well-known elicitation approaches is the Delphi method [22, 47] characterized by its iterative, anonymous approach for developing consensus among experts. This method has been used widely in governments and industries [45, 46]. Another approach is the Cooke Classical Model [37, 41], also known as Cooke's method, which is one of the most established methods in expert elicitation literature. This method uses calibration questions, for which true values are known to the facilitator, to measure both accuracy and informative-ness of an individual expert's judgement. These performance measurements, called calibration score and information score, respectively, are used as weights for aggregating multiple experts' judgements. Although developing calibration questions requires extra efforts, this performance-based weighting scheme has empirically proven effective [43] and represents the state-of-the-art among various weighting schemes [15, 38, 42]. In this work, we propose to elicit data from the expert panel using both Delphi and Cooke's methods. The Delphi method is used to estimate a crucial quantity that needs a consensus across experts. The Cooke's method is used to aggregate recovery estimates across experts according to

performance-based weights.

Although many studies elicit point estimates or probability distributions from experts, there are only a few studies on eliciting functions (e.g., recovery curve) from experts [20, 50, 132]. Arguably, the most systematic expert elicitation approach to functional estimation is developed in [69]. This study estimates seismic collapse fragility functions by eliciting quantiles of probability distributions, which encompass uncertainties of both seismic shaking intensity and resulting building collapse, from earthquake-engineering professionals. The reported estimates therein are created by first fitting lognormal distributions to the elicited probability estimates and then aggregating the distributions using Cooke's method. While this approach using the lognormal distribution (often used to model collapse fragilities) is defensible for this study, generalizing the approach to other functional estimation (especially recovery time estimation) has a major drawback. Using a parametric distribution like lognormal is too restrictive to reflect the uncertainties underlying the complex recovery processes being modeled. Thus, this study uses GPR, which allows us to nonparametrically model recovery curves and the associated uncertainties.

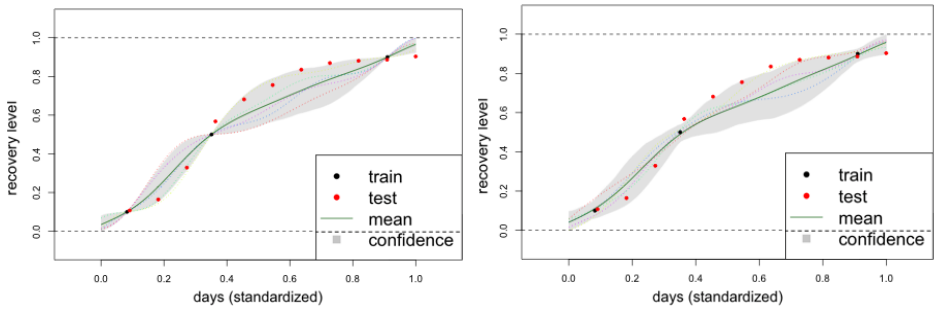
Integration of expert judgements and empirical data is briefly mentioned in the NIST Guide [2], but no specific guideline is provided on the integration. The Oregon Resilience Plan [39] was the only resilience planning initiative that explicitly used both expert judgements and past event data, but the estimation process was still ad-hoc. Currently, to our best knowledge, there is no systematic statistical inference method being used for expert-based recovery time estimation in practice. This gap inspired us to develop the proposed method.

4.2.2 Gaussian Process Regression

Gaussian Process (GP) is a nonparametric model that offers the flexibility to model a stochastic process. It has been used successfully in many applications, such as engineering, physics, biology, economics, or other fields, in both regression and classification problems [60, 97, 99, 110]. It specifies a prior distribution over function spaces, where the relationships over data are encoded in the covariance functions of multivariate Gaussian distri-

butions. Once the input data is available, GP can model the posterior over function spaces. The covariance will determine properties or constraints of the process, such as characteristic length scale, smoothness, or variance [107]. In this study, we are estimating recovery curves, so we will focus on Gaussian process regression (GPR).

Besides its low bias towards any functional form, GPR is also more suitable to our task than other parametric methods. It can capture both the uncertainty in the region where training data is not available and the variability in the training data itself. As a well-known issue in judgement-based forecasting, no matter how rigorous the elicitation process is, the results still depend on the experts' ability to estimate the quantity of interest. Because the expert estimates are noisy, GPR will capture the variability as an extra source of uncertainty during the inference step. Figure 4.1 shows two different ways to fit GPR to estimate a recovery curve, with or without noise in the training data.



(a) GPR with Noise-free observations (b) GPR with Noisy observations

Figure 4.1: GPR fitting with noise-free and noisy observation for Fukushima electricity recovery.

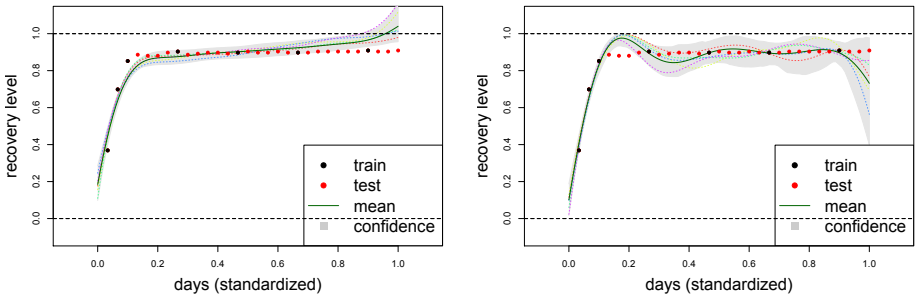
Due to the physical nature of the recovery curve, we also need to impose some constraints on the GPR model. First, the functionality level should be between 0% and 100%. Therefore, we will bound the prediction of the GPR model to be strictly between 0 and 1. Second, although it is possible that functionality level may temporarily decrease in reality (e.g., due to an aftershock), it should generally increase over time. Hence, to capture this behaviour and

reduce the prediction error, we also enforce that the curve is monotonically increasing with respect to time. Montonicity and boundedness are the linear inequality constraints actively researched in the GP framework [81, 82, 84, 85, 110]. In Figure 4.2, we show the effects of imposing only monotonicity, only [0,1] boundedness, and both constraints in the model for the Fukushima prefecture electricity recovery using the R package `lineqGPR` [81, 82]. It is helpful to have both constraints in the model. Otherwise, the model may behave in contrast to the expected physical behaviour of infrastructure recovery. In addition, the constraints will help to reduce the variance of the prediction. However, imposing these constraints may be potentially too rigid to capture the flat region near 0% and 90% of recovery. We can alleviate this issue by eliciting the boundary points so that the GPR is only interpolating between the elicited data points. Furthermore, the recovery curve can be constructed up to a functional level below 100% (e.g., 90%) as suggested by the NIST Guide [2]. We will apply these measures in Section 4.4 for the numerical studies.

4.3 Method

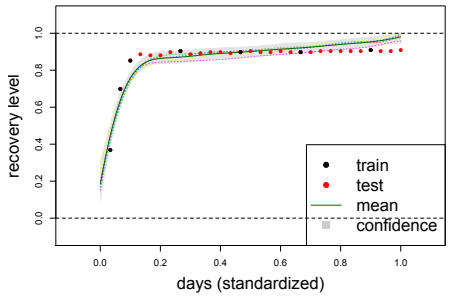
4.3.1 Consideration in recovery curve estimation

Our goal in this study is to estimate the infrastructure recovery curve from point estimates given by experts using GPR. The two steps (i.e., expert elicitation and GPR) are not designed totally independently. We carefully design the whole framework considering the logistical, computational, and theoretical constraints of both steps. The curve is characterized by two dimensions, namely, the recovery level measured in percentage (100% means the system is fully functional) and the recovery time measured in either days or hours from the disruption. GPR, similarly to other regression methods, is more suitable for interpolation between training data (as opposed to extrapolation). To achieve better performance, it is desired for the expert elicited data to possess two properties. First, it should be as evenly distributed as possible so that the interpolated prediction does not exhibit too much uncertainty. Second, it should cover the boundary values to avoid predicting values beyond the range of the given



(a) GPR with monotonicity constraint only.

(b) GPR with boundedness constraint only.



(c) GPR with both monotonicity and boundedness constraints.

Figure 4.2: Different GPR constraints for the Fukushima earthquake event.

data. It would be of our best interest to elicit the data in both the range of recovery level (e.g 10%, 30%, 50%, 70%, 90%) and recovery time (e.g $2.5D/10$, $5D/10$, $7.5D/10$, $9D/10$), where D is the (estimated) earliest time for the infrastructure to recover to 100% or another high functional level (e.g. 90%) depending on what kind of recovery curve we want to construct, to follow the NIST Planning Guide [2]) given by the expert. However, it may not be very intuitive to elicit recovery levels at some certain time, e.g. "What is the estimated recovery level at day 11 after the event?". Therefore, in our proposed method below and the numerical studies in Section 4.4, we only elicit the recovery time at some certain recovery

levels.

In addition to eliciting recovery times at different recovery levels, we also want to elicit D, as introduced above, so that we can normalize the recovery time to be in the range of $[0, 1]$ for the following reasons:

1. The constructed curve could be more generalizable to future disaster events. If we face another similar events in the future, where similarity is defined by dominant characteristics of the events (e.g., Richter magnitude and earthquake resilience of the area, or Saffir–Simpson scale and hurricane resilience of the area), we can significantly reduce the elicitation effort by either using the existing recovery curves or simply eliciting the earliest full recovery time D and scaling the recovery time based on the particular D values of new events.

2. It is easier to compare different recovery curves of different natures on the same scale in the range of $[0, 1]$.

3. We can offer some insights from the shape or pattern (e.g., for hurricane category 1 vs. 5; magnitude 6 vs. 8; power vs. water; urban vs. rural) of recovery, which has formed consensus across many communities, so that other communities lacking the opportunity/resource to conduct extensive elicitation procedure can still use these curves as references of possible recovery trends.

4. In terms of GPR modelling, we want to have both axes in the process to be between $[0, 1]$ in the fitting and inference following the implementation in [81, 82]. The actual unit of recovery time can be easily scaled back to days or hours after the inference procedure.

4.3.2 Challenges in Expert Data Elicitation and Modelling

From the design considerations above, we anticipate some challenges in the elicitation process as follows:

1. Obtaining the earliest time to full recovery D: D should be universal across all experts. One way to obtain this is to have an open discussion among experts until they reach a general consensus on how long D should be. Another way is to employ point-based expert elicitation methods, for example in [101], to estimate the probability distribution of D. Another possibility is to use the individual expert's D value to normalize their own recovery time estimate.

2. Obtaining the input noise level σ: This is required for the statistical modelling process. This can be interpreted as how uncertain the experts' estimates are. We may gather the data and estimate the uncertainty based on their data after the elicitation process. This σ will account for both within-expert and across-expert uncertainty. The GPR framework assumes that one type of noise is present in the data, which accounts for all the uncertainty, and that the noise level is constant across all levels of input. In case we want to decompose the uncertainty further, it is more straightforward to estimate the across-expert uncertainty since we have different expert data at each recovery level. However, within-expert uncertainty estimation is tricky. One way to estimate it is through Cooke's method where calibration questions are used to measure the inherent estimation uncertainty. However, one can challenge the underlying assumption that the estimated uncertainty based on calibrating questions remains the same as the uncertainty for main questions. Regardless, it is reasonable to assume that within-expert variability is negligible compared with across-expert variability.

3. Obtaining more elicited data: While we may drive down the estimation uncertainty by collecting more data, this would impose more logistical burden to the experts. In addition, the experts may have some cognitive difficulty to distinguish between smaller difference in recovery levels, e.g 10% and 20%.

On the other hand, there are also some challenges in the modelling and inference process:

1. If we ask each expert to give an estimate of D, it is challenging to determine which D

to use and how to normalize the time.

2. If we have noise/uncertainty in both dimensions (input and output), it does not follow the conventional GPR framework, in which $y = f(x) + \epsilon$, where ϵ follows $N(0, \sigma^2)$. To further elaborate this point, in the GPR framework, we assume that the input is fixed, i.e. if we want to predict the recovery time at each functionality level, we fix the functionality level and the prediction of recovery time will exhibit some level of uncertainty. This is consistent with our experiment implementation in Section 4.4.

4.3.3 Recovery curve estimation framework

In terms of workshop design, we can adopt the Cooke Classical Model [41] to perform elicitation of expert judgments. There are several ways to aggregate experts' estimates, such as linear pooling or performance based weighting. Linear pooling, although with its least logistical cost of designing calibration questionnaire, is shown to under-perform other performance based methods [43]. Although questionnaire design is beyond the scope of this work, we outline one way to perform calibration on the expert judgments following the performance based weighting methods. There should be a set of calibration questions, which is closely related to the quantity of interest we are trying to estimate. An example question could be, given a functionality level, what is the estimated time that the expert thinks an infrastructure can take to achieve. An expert will be asked to give different quantile estimates on the quantity, and they form their subjective probability mass about such quantity. Under the Cooke Classical model, there will be two types of scores being generated from this calibration exercise. The first is an information score, or how confident an expert is about her estimates. The second score is a calibration score, which is the likelihood that her judgement corresponds to the actual results. A product of the two quantities can be used as a general score to determine the performance weight, which is then used to take the weighted average of experts' estimates. To further optimize for performance, we can vary the selection threshold, below which will render an expert's weight to 0, to get the best

performance metric on the calibration questions. Then, that set of optimized weights can be used to elicit the quantity of interest.

We also consider a few elicitation schemes. Each scheme has its own advantage and disadvantage, which are also presented individually.

Scheme 1: Maximum amount of elicited data and elicit in parallel.

1. Ask each expert for the maximum amount of time to recovery D, recovery times at fixed functionality levels (10%, 30%, 50%, 70%, 90%), and functional levels at fixed recovery times (e.g $2.5D/10$, $5D/10$, $7.5D/10$, $9D/10$).

2. Use the sample mean/median (across experts) of all elicited data as the training data, with Cooke's method weighting if necessary.

3. Use all estimates (across experts) at each level to estimate the noise level.

4. Fit GPR and make inference

5. Repeat the process in other events

Advantage: Full range of data over both dimensions. Impose less burden in elicitation logistics than scheme 2. Elicitation can finish in one stage.

Disadvantage: Uncertainty in D estimation can lead to erroneous and high uncertainty in prediction. Furthermore, as mentioned above, the GPR framework assumes one dimension as fixed input. Eliciting in both dimension could violates this assumption.

Scheme 2: Two-stage elicitation. The maximum recovery time D will be iteratively discussed among the experts until reaching consensus.

1. Stage 1: Ask each expert the maximum amount of time to recovery D.

2. Show all the expert the (range of) elicited D values.

3. Ask expert to revise their D estimate until reaching agreement.

4. Stage 2: Ask each expert a full range of fixed recovery level (10%, 30%, 50%, 70%, 90%) and recovery time (e.g $2.5D/10$, $5D/10$, $7.5D/10$, $9D/10$), with Cooke's method weighting if necessary.

5. Use the mean/median of each estimate as the training data.

6. Use all estimate at each data point to estimate the noise level.

7. Fit GPR and make inference

8. Repeat the process other events

Advantage: Full range of data over both dimension. Reduce uncertainty in the maximum recovery time.

Disadvantage: Two-stage elicitation will require more effort from the expert.

Scheme 3: Using either scheme 1 or scheme 2 but with less number of elicited data (e.g., 3 points for each dimension)

Advantage: Less burden on the expert.

Disadvantage: May result in a sub-optimal fit and prediction.

Scheme 4: Elicit on only one dimension, fixing recovery levels and ask for recovery time.

1. Obtain estimate of D, following either scheme 1 or scheme 2. The numerical studies will use scheme 4 with each expert's estimate of recovery time being normalized by her own estimate of D.

2. (OPTIONAL) Repeat the process for the 3 scenarios (worst, best, most likely)

Advantage: Straightforward in modelling. Simple to interpret and implement.

Disadvantage: Data may be sparse. In some events (e.g the Fukushima electricity recovery in Section 4.4), the recovery is expected to be extremely fast in the first few hours. The expert may say the recovery is up to 70% in the first day and 90% the next day. In this

case, the GRP model may not provide much additional values to stakeholders in recovery planning.

Scheme 1 will speed up the elicitation process since we can elicit on both dimensions. However, the question is whether we need to elicit in both ways (fix the level then elicit the time, and fix the time then elicit the level). Scheme 2 is almost identical to scheme 1, except with the elicitation of the earliest full recovery time D to reach either 100% or 90% to normalize the time axis. Scheme 3 is a less resource-demanding version of Scheme 1 and 2. In Section 4.4, we will study the optimal number of elicitation levels through sensitivity analysis. Although we can try to elicit in both ways, to be consistent with the GPR framework, we can only use one dimension (either recovery time or recovery level) as input and predict the remaining dimension. Instead of spending experts' resources on eliciting in both ways, we can use their effort to elicit more recovery time at higher granularity of functionality level or elicit more scenarios (best, worst, most likely). In view of the above considerations, we will demonstrate the framework of Scheme 4 in the numerical studies in Section 4.4.

4.4 Numerical Studies

To demonstrate the performance of the framework, we try to evaluate it on different empirical recovery curves from different prefectures and infrastructures in the 2011 Great East Japan Earthquake Disaster and the 1995 Great Hanshin-Awaji Earthquake Disaster. The framework is designed to be applied where an expert elicitation workshop is run in conjunction with statistical modelling. For demonstration purpose in this work, we will simulate the expert opinion. Assuming that the experts are capable of estimating the true recovery curve with a reasonable accuracy, we use the entire available empirical data (such as those in Figure 1.4) to fit the polynomial regression model as a surrogate to the expert opinion. The simulated expert can be queried for recovery time given a functionality level and vice versa. In Scheme 4, we provide a functionality level as an input to the simulated expert and obtain

the recovery time estimate as the output. Each expert can be modelled using Eq. (4.1):

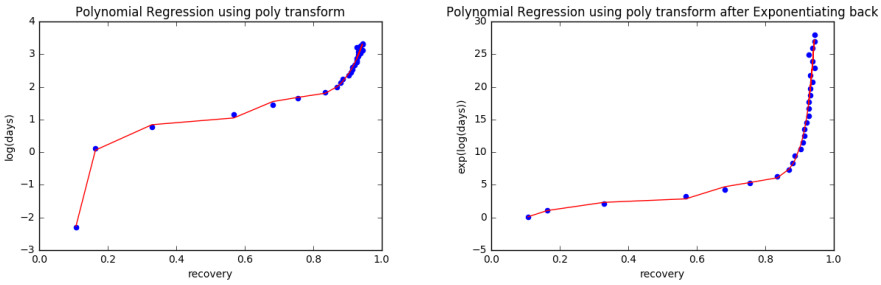
$$days = f_{poly}(recovery) + \epsilon_{days}, \tag{4.1}$$

where ϵ follows $N(0, \sigma^2)$ as per standard polynomial regression notation. σ^2 will capture the variability within each expert.

However, it is better for the model to utilize the fact that the output (i.e., recovery time) is always positive by taking log transformation on the output variable. Thus, the expert model in Eq. (4.1) becomes

$$\log(days) = g_{poly}(recovery) + \epsilon_{log(days)} \tag{4.2}$$

The fitted polynomial regression function g_{poly} is demonstrated in Figure 4.3.



(a) Polynomial fit in the log scale of the days/recover time

(b) Predicting recovery time in the original scale

Figure 4.3: Expert simulated model is built based on all the available data of a past event using polynomial regression, which will allow the sampling from the curve will be very close to the actual values. The model represents an average prediction across multiple replications and multiple expert. In other words, if we have infinite amount of experts, we assume that their average prediction will converge to the fitting curve or the actual data.

Furthermore, it is desirable to model two distinct sources of the estimation variability $\epsilon_{log(days)}$. Thus, a layer of Gaussian noise ϵ_1 is added to the output to model within-expert

variability. A second layer of Gaussian noise ϵ_2 is then added to model across-expert variability. The two noise terms are additive in the log-transformed model because we model the errors to be multiplicative in the original scale. The log transformation will then make the multiplicative errors become additive, to be consistent with the polynomial regression framework. The multiplicative errors are intuitive. For example, consider a scenario event that makes the recovery estimation challenging for all experts (i.e., high across-expert variability, $Var(\epsilon_2)$). Then, the individual expert's large uncertainty perhaps due to lack of experience (i.e., high within-expert variability, $Var(\epsilon_2)$) will *amplify* the effect of the challenging estimation problem, thus resulting in highly variable recovery time estimates. In summary, the elicited recovery time estimates are simulated using

$$days_{simulated} = \exp(g_{poly}(recovery) + \epsilon_1 + \epsilon_2) \tag{4.3}$$

As an implementation note, due to the randomness from ϵ_1 and ϵ_2, sometimes the sequence of simulated expert's estimates could be non-monotonic. How likely it happens depends on the variance of the errors. Since we assume the experts are only providing estimates for a monotonic recovery curve (i.e., no deterioration of infrastructure functionality in the midst of recovery, for example, due to aftershocks), they will only provide monotonically increasing recovery time estimates with respect to the functionality level. In our simulation, to ensure that the simulated recovery estimates satisfy this assumption, we reject the non-monotonic estimate paths until a monotonic sequence is generated.

Using the simulated data, the GPR model with monotonicity and boundedness is fit as follows:

$$recovery = GPR(days_{simulated}) \tag{4.4}$$

Figure 4.4 shows the performance of this framework on different prefectures (Miyagi electricity, Fukushima electricity, and Iwate electricity) and different infrastructures within the same prefectures (Great Hanshin water and gas). We simulate the process of eliciting from 5 experts, asking for recovery time at 10%, 30%, 50%, 70%, 90% functionality levels,

averaging their estimates, and construct the GPR curve.

It is observed that the method is very flexible. In Fukushima electricity recovery, although the actual recovery started at about 40% in day 1, we can still capture the rest of the recovery curve simply by eliciting from 30% onward. This translates to some freedom to the experts in actual workshops. They can skip some levels if they think it does not make sense to estimate when they think the recovery actually will happen quickly initially.

In Figure 4.5, it may seems that the recovery does not capture the initial recovery stage very well. This often happens in lagging infrastructure such as gas, which is usually recovered after electricity and water. We do not view this as an issue since we believe the portion of the recovery curve between 10% and 90% should deserve the most attention, which the model can capture quite well with high confidence.

We also investigate how sensitive the estimation framework is to the number of experts by monitoring the root mean square error (RMSE) of prediction on the available test data (different from the recovery levels elicited from the experts). We first perform simulation to measure the performance in terms of RMSE of the framework with 1, 3, 5, 7, 9, 11 experts based on Miyagi electricity recovery to see if there is an "elbow" of performance change point to balance the logistics of elicitation and accuracy, as shown in Figure 4.6. In Table 4.1, we vary the number of simulated experts to be 3, 5, or 10. Given a fixed noise level within and across experts, it seems that the result is quite stable with 5 experts. We acknowledge that in this simulation, all the experts are modelled to exhibit the same level of uncertainty, which is not realistic in practice. In fact, in usability testing experiments in [55], where the participatory performance, involving both expert and novice users, is measured in a group of 5 and beyond, the study shows that some randomly selected group of 5 participants can perform relatively well although the risk is that the performance variance is high. However, in actual workshops, there could be more than 5 experts (among whom, the expertise level is theoretically more consistent than the study in [55]), and as long as their opinions converge to some underlying quantity, the estimation still can provide a reasonable recovery curve.

We conduct another analysis to measure how the framework performs with different

levels of elicitation. Our initial hypothesis is that performance will improve as we elicit more data, which may increase more logistical burden to the expert. The hypothesis is generally confirmed from Figure 4.7. It also does not penalize performance very much to have custom spacing of levels, so we can focus more on asking the experts at more intuitive recovery levels.

Table 4.1: Sensitivity analysis on the framework performance to the number of experts. In this table, the experts are simulated to have equal contribution to their estimate, and the simulated noise variance in equation 4.3 is $Var(\epsilon_1) = Var(\epsilon_2) = 0.1$. Note that the unit for RMSE is the fraction of recovery level. The RMSE presented is the average across 100 simulation replications.

Prefacture/ Infrastructure	Number of Experts	RMSE
Fukushima electricity	3	0.0637
	5	0.0567
	10	0.0524
Miyagi electricity	3	0.0405
	5	0.0340
	10	0.0297
Iwate electricity	3	0.0457
	5	0.0398
	10	0.0373
Great Hanshin water	3	0.0447
	5	0.0352
	10	0.0334
Great Hanshin gas	3	0.0542
	5	0.0516
	10	0.0497

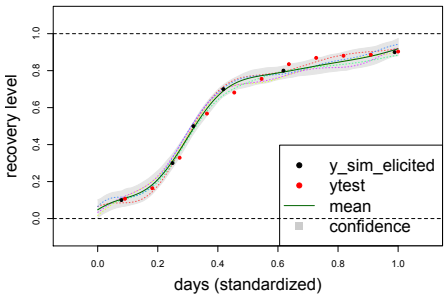
Table 4.2: Sensitivity analysis on the framework performance to the uncertainty in expert estimation ($Var(\epsilon_1), Var(\epsilon_2)$ in Equation 4.3. In this table, data is simulated from 5 experts for 100 simulation replications.

Prefacture/ Infrastructure	$Var(\epsilon_1), Var(\epsilon_2)$	RMSE
Fukushima electricity	0.1	0.0567
	0.3	0.0600
	0.5	0.0985
Miyagi electricity	0.1	0.0340
	0.3	0.0583
	0.5	0.0811
Iwate electricity	0.1	0.0398
	0.3	0.0549
	0.5	0.0686
Great Hanshin water	0.1	0.0352
	0.3	0.0427
	0.5	0.0546
Great Hanshin gas	0.1	0.0484
	0.3	0.0636
	0.5	0.0916

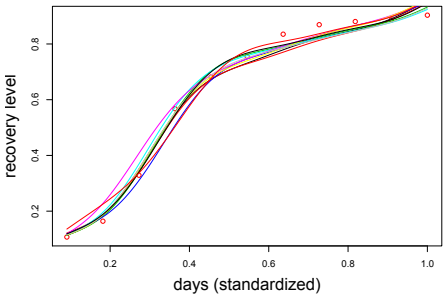
4.5 Summary

We demonstrate in this research a framework to assist the community resilience planning through constructing potential infrastructure recovery curve. The method combines experts' opinions and Gaussian process regression to unify domain knowledge and uncertainty quantification in the curve. To understand the method better, we perform extensive sensitivity analysis to draw insights in various elicitation schemes, the effect of number of experts,

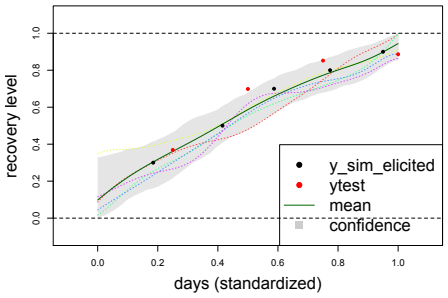
number of elicited points, at which levels, to predictive performance. Although the application of the framework is infrastructure recovery after natural hazards domain, it can be generalized to other fields such as reconstruction of building stock or housing recovery after a disaster. We do not consider inter-dependency in this study and treat infrastructures as having independent recovery process. One future research direction could be conducted to model their structural dependency to improve the performance and/or reduce the reliance on expert estimate.



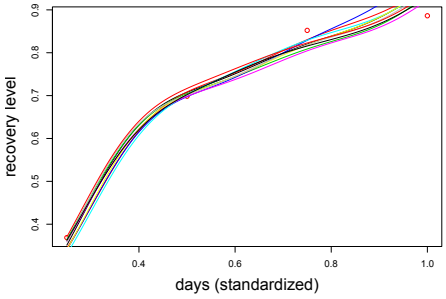
(a) Miyagi electricity recovery curve with 95% Confidence interval



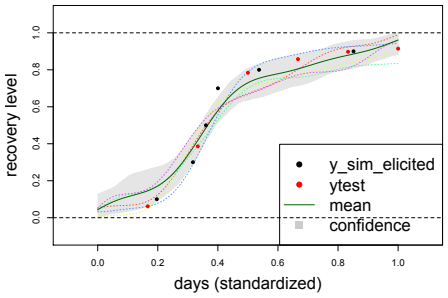
(b) Miyagi electricity recovery curve with 10 mean predictions



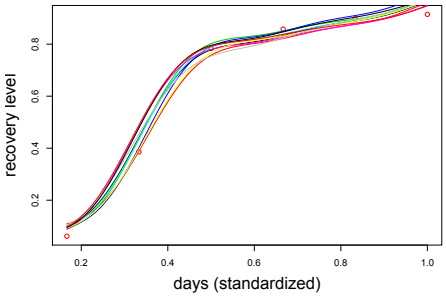
(c) Fukushima electricity recovery curve with 95% Confidence interval



(d) Fukushima electricity recovery curve with 10 mean predictions

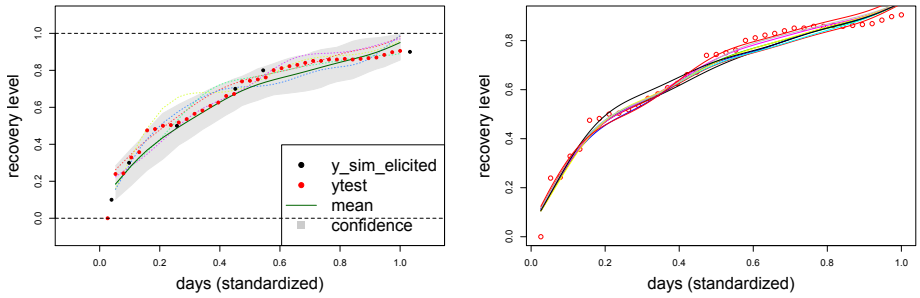


(e) Iwate electricity recovery curve with 95% Confidence interval



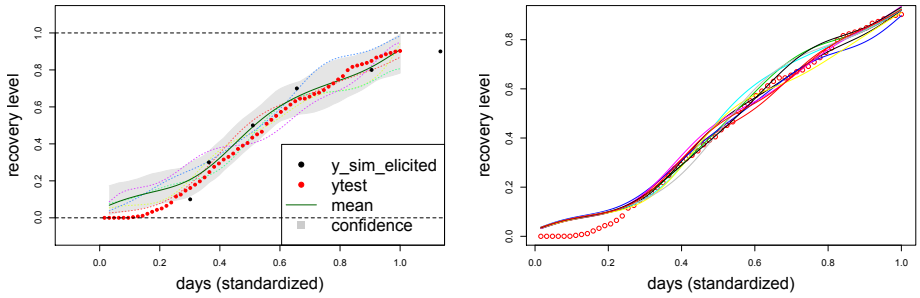
(f) Iwate electricity recovery curve with 10 mean predictions

Figure 4.4: Numerical results on different prefectures (Miyagi, Fukushima, and Iwate). The figures on the left column show the result of GPR model built on one simulated draw of expert opinion. The grey bands show the 95% confidence interval to capture the uncertainty around the predicted curves. The figures on the right column show different mean predictions based on different simulated draws of expert opinion. In all cases, we simulate the process of elicitation from 5 experts.



(a) Great Hanshin water recovery curve with 95% Confidence interval

(b) Great Hanshin water recovery curve with 10 mean predictions



(c) Great Hanshin gas recovery curve with 95% Confidence interval

(d) Great Hanshin gas recovery curve with 10 mean predictions

Figure 4.5: Numerical results on different Water supply and Natural gas infrastructures. The data simulates the process of elicitation from 5 experts.

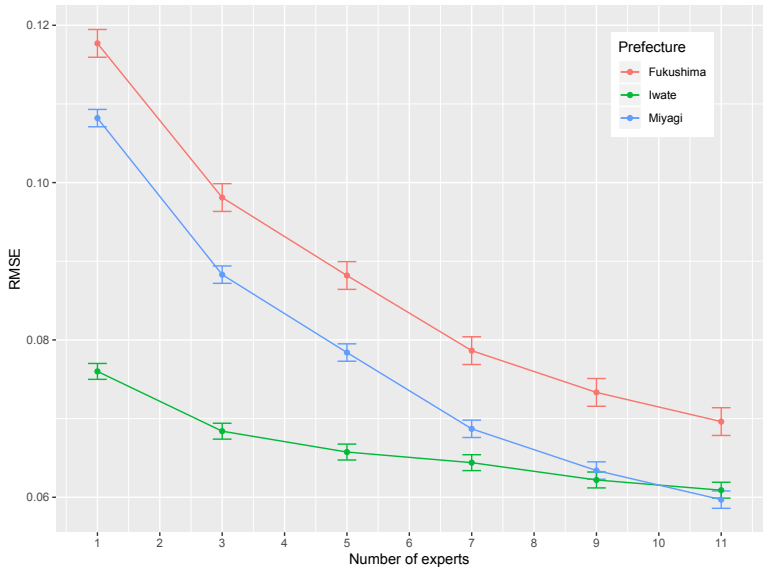


Figure 4.6: The plot shows the performance of the framework for electricity recovery at Fukushima, Miyagi, and Iwate prefactures as a function of the number of experts. The error bar at each level of experts shows the 95% confidence interval of test RMSE in 100 simulation replications. Although the more number of experts involving in the elicitation process results in better performance, it is observed that there is a diminishing marginal return as the number of experts increases in 2 out of 3 cases. The rate of performance gain is fastest when engage from 1 to 3 experts. The rate is slower from 3 to 7 experts. It drops to the slowest rate if we increase from 7 to 11 experts.

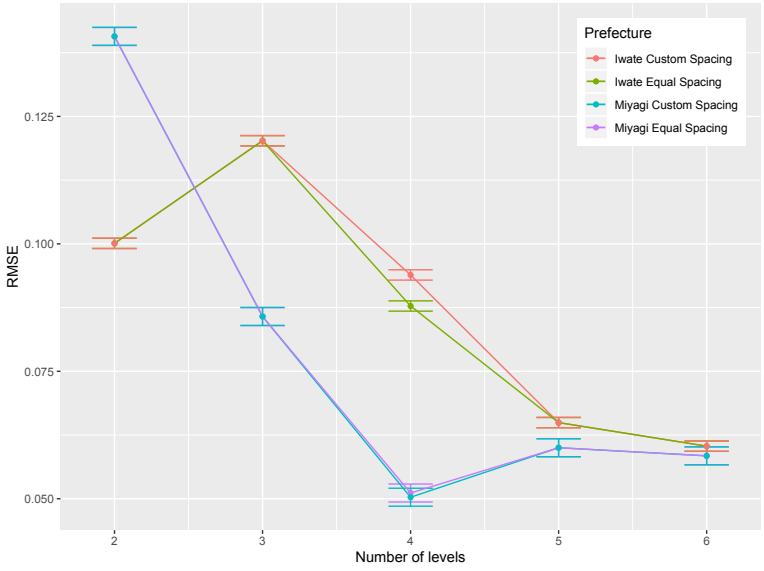


Figure 4.7: The plot shows the performance of the framework in terms of test RMSE in 100 simulation replications for electricity recovery at Miyagi and Iwate prefectures as a function of the number of elicitation levels. We evaluate the performance when eliciting 2, 3, 4, 5, 6 levels from the experts. Custom spacing means we fix the elicited recovery levels at intuitive levels to the experts such as 10%, 30%, 50%, etc, regardless of the number of levels. Equal spacing means we get the levels by equally divide the range from 10% to 90% by the number of levels, which results in some odd levels, such as 10%, 36.67%, 63.33%, 90% at 4 elicitation levels. The plot shows some general trends that at 4 to 5 elicitation levels, the performance can be satisfactory.

Chapter 5

CONCLUSION

This dissertation develops three fast and efficient frameworks to improve situation awareness of stakeholders, especially emergency managers and first responders, after a hazard event. Situation awareness is valuable to make effective resource planning and timely rescue actions. However, the cost of obtaining this information is usually not trivial. This is due to the fact that: (a) it could be time-consuming to collect data, especially right after the event, (b) disaster management is usually a collective effort from many organizations, so the data collected has to be processed to synchronize and understand the findings, and (c) natural hazards do not happen frequently so it is not obvious how to leverage emergency managers' experience to react better in future events. The following chapters address these challenges in a data-driven manner while carefully considering the logistics cost and interpretability so that they can be adopted widely among the disaster management community.

Chapter 2 proposes a semi-automated framework to quantify damage of the affected region after a hurricane event. We show that through publicly available satellite imagery, the convolutional neural network (CNN) model can perform building damage annotation with reasonable accuracy. Although some level of manual data processing is required, this method provides a great saving in terms of time and labor to conventional methods such as windshield surveys to quantify damage. Recognizing the limitations in Chapter 2, Chapter 3 proposes several improvements in both dataset construction and methodology. Instead of extracting the damaged and undamaged building images based on different temporal information of the dataset, we gather data from different spatial information using the same timestamp after the event happens. This makes the dataset closer to the actual scenario when the method is targeted to be deployed after an event. In addition to using satellite imagery information,

leveraging geolocation information can significantly improve the performance of CNN, and reduce the hyper-parameter tuning effort. This method shows a promising direction to combine other disaster damage quantification features that have been used before the CNN era, such as area flooding resilience index, or distance from epicenters, to remotely sensed images. Since the geolocation features used are carefully chosen as relative elevation and relative proximity to water body, the model can be adapted to deploy to other regions and events without retraining. These two chapters can help improve the perception level of disaster situation awareness so that stakeholders can make informed decisions to assist the victims. The results obtained, a set of bird's-eye view of buildings in the affected region and their damage status, are easy to interpret and verify.

Chapter 4 aims to improve the projection level of situation awareness by providing a reliable estimate of potential infrastructure recovery time. Our method combines experts' opinions and Gaussian process regression to unify domain knowledge and uncertainty quantification in the estimated recovery curves. Our extensive numerical studies and sensitivity analyses demonstrate that the method is robust against different elicitation schemes such as number of experts, number of elicited points, and elicitation levels. Although this work is built around infrastructure recovery, it can be easily extensible to other other recovery, such as for different capitals and services that are important for community resilience as mentioned in [91], as long as we can organize the necessary workshops to conduct elicitation. A caveat to the framework is that, it could be difficult to build expertise around an overarching disaster recovery pattern due to low frequency and small similarity of extreme events. However, the recovery process of individual infrastructure, such as the 16 critical infrastructure sectors considered by the Department of Homeland Security, can be familiar to the infrastructure management experts. They have built experience and collected data over years of observing their respective infrastructure affected and recovered from various events. The uncertainty of their estimate about an unseen event can be quantified and reduced through Gaussian process regression and the elicitation workshop design. Furthermore, it is unclear how to define an overarching disaster recovery curve as it depends on specific use cases of

decision makers. For example, one could define it as the recovery of all infrastructures, which may be practically too long for the citizen and the government to base their decisions on. Another possible interpretation could be the fastest time until the basic infrastructure (e.g electricity) can recover. That also has its own drawback. The fact that electricity comes back in two weeks does not mean the displaced populations can come back after the same period as they need some certain infrastructure networks to recover, such as water and gas supply, so that their homes can be habitable. Therefore, estimated recovery curves of individual infrastructure would be of better use to decision makers as they provide more flexibility and interpretability.

From a disaster management perspective, there are numerous ways to assist emergency managers and stakeholders in their planning, decision making, and performance monitoring efforts and the situation awareness framework presented in [52] helps fit all the related research contributions in their logical and intuitive order. From the above three research objectives, we can now present a full roadmap of how this dissertation can be applied in practice.

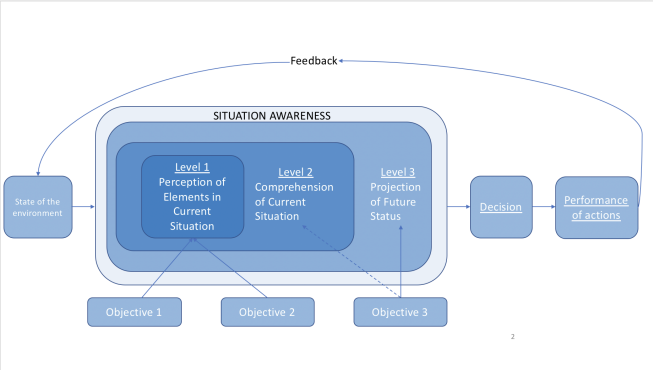


Figure 5.1: An adaptation of the situation awareness framework in dynamic decision making.

As we can see from an adaptation of the situation awareness framework in dynamic decision making presented in [52] (Figure 5.1), there is a loop of feedback to constantly improve the current state information and situation awareness of stakeholders to make better

decisions overtime. A potential use case of the research objectives is as follows. Shortly after a hurricane makes landfall, we can start deploying the models, as presented in Objectives 1 and 2, to quickly inform emergency managers of the damage extent of the region to provide Perception level of situation awareness. Depending on the related works or extension of this work, there can be other damage assessment frameworks available, and they can be combined altogether to provide a bigger picture about the whole post-event situation to provide Comprehension level of situation awareness. In Objective 3, the expert elicitation workshop can be another opportunity to improve Comprehension level as well, as the experts have a chance to discuss their prior knowledge and gather all the available information, which potentially could be the damage assessment results from Objective 1 and 2. For instance, after the damage assessment, the percentage of damaged or flooded buildings within the region can be made known to the experts to inform them about the actual severity of the event. As a natural next step in Objective 3, the critical infrastructure recovery trajectory can assist emergency managers with the projection of how fast or likely the infrastructures and community can recover from this event. After decisions are made and model performances are measured for this event, we can gather further information, either in the form of more samples for damage assessment models, more recovery data points, or better knowledge to the experts. Then, we can improve all the models and make more informed decisions in future events, as we follow the feedback loop in Figure 5.1.

Alternatively, we can perform the above iteration continuously within the same event. That actually provides several benefits. For high-impact events, there usually is a dedicated team of emergency managers, including infrastructure management experts, working together, which could facilitate more frequent expert elicitation workshops. For example, the entire situation awareness building process can be repeated every month as more information about the situation unwraps. As the experts' knowledge could potentially evolve as more data is available, the variability in their estimates could be reduced. Furthermore, as time goes, we would have more anchor points, i.e data with the actual recovery data and no more uncertainty (e.g., electricity may recover quickly to above 80% after 1 month), and we can

update the recovery curve estimate for the remaining functionality of the infrastructure.

In the future, there are two potential directions that can be pursued following our findings from Chapter 2 and 3. The first direction is rapid, real-time damage mapping of damaged buildings. From recent efficient and instant object detection algorithms such as [108, 109], it would be possible to gather damage status of buildings through more accessible devices such as drones, freeing the reliance on satellite imagery. Since satellite imagery are considered to be complex to process due to their size, containing several objects at different scales [117], incorporating geolocation features to existing object detection algorithms can improve its metrics such as precision and recall. A potential challenge of this direction is the amount of labelled data required is usually quite large, which grows together with how complex the model is. This poses another issue of noisy or wrong labels, which leads to a second potential direction, label refinement. Recent studies have highlighted the needs of label refinement in the presence of noisy or wrong labels in large scale datasets [16], or in remote sensing data [112]. From our studies, geolocation features alone already inform a substantial prior knowledge about the damage likelihood. We can use that information to refine the imagery data and correct the wrong labels as necessary.

From the work in Chapter 4, we observe that there is an inherent relationship among different infrastructure recoveries. For example, electricity tends to recover first, followed by water supply, and then gas supply. It would be interesting to investigate the dependencies between infrastructures. For instance, the study in [30] models infrastructure dependencies qualitatively and limits to short-term projection without uncertainty quantification. We may directly model their dependencies to either improve the prediction, narrow down the uncertainty, or reduce the reliance on expert estimates. Although these dependencies may be already captured in the experts' estimate, this direction can help to generate recovery curves of other infrastructures when we do not have access to the experts' opinion.

BIBLIOGRAPHY

[1] Advanced Rapid Imaging and Analysis (ARIA). https://aria.jpl.nasa.gov/about, visited 2019-04-27.

[2] Community resilience planning guide for buildings and infrastructure systems. https://dx.doi.org/10.6028/NIST.SP.1190v1, visited 2019-04-27.

[3] Dartmouth Flood Observatory (DFO). https://floodobservatory.colorado.edu/, visited 2019-04-27.

[4] GeoEye-1 satellite sensor. https://www.satimagingcorp.com/satellite-sensors/geoeye-1/, visited 2019-04-27.

[5] Google Maps Platform. https://cloud.google.com/maps-platform/maps/, visited 2020-09-14.

[6] OpenStreetMap. https://www.openstreetmap.org/, visited 2019-04-27.

[7] Tomnod. http://blog.maxar.com/news-events/2019/in-the-blink-of-an-eye-looking-back-on-nine-years-with-tomnod, visited 2019-04-27.

[8] United States Geological Survey. https://www.usgs.gov/products/data-and-tools/gis-data/, visited 2020-09-14.

[9] Anatomy of a catastrophe. https://www.planet.com/insights/anatomy-of-a-catastrophe/, visited 2018-01-12, 2017.

[10] Unsupervised flood mapping. http://gbdxstories.digitalglobe.com/flood-water/, visited 2018-03-10, 2017.

[11] Metehan Ada and B. Taner San. Comparison of machine-learning techniques for landslide susceptibility mapping using two-level random sampling (2LRS) in Alakir catchment area, Antalya, Turkey. *Natural Hazards*, 90(1):237–263, Jan 2018.

[12] Eman Ahmed and Mohamed Moustafa. House price estimation from visual and textual features. *arXiv preprint arXiv:1609.08399*, 2016.

[13] Kimberly Amadeo. Hurricane harvey facts, damage and costs. *The Balance. Retrieved from https://www. thebalance. com/hurricane-harvey-facts-damage-costs-4150087*, 2018.

[14] Heiko Apel, Annegret H Thieken, Bruno Merz, and Günter Blöschl. Flood risk assessment and associated uncertainty. 2004.

[15] WP Aspinall and RM Cooke. Quantifying scientific uncertainty from expert judgement elicitation, in. *Risk and uncertainty assessment for natural hazards*, page 64, 2013.

[16] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.

[17] C. F. Barnes, H. Fritz, and J. Yoo. Hurricane disaster assessments with image-driven data mining in high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6):1631–1640, June 2007.

[18] Luke Barrington, Shubharoop Ghosh, Marjorie Greene, Shay Har-Noy, Jay Berger, Stuart Gill, Albert Yu-Min Lin, and Charles Huyck. Crowdsourcing earthquake damage assessment using remote sensing imagery. *Annals of Geophysics*, 54(6), 2011.

[19] Y. Bazi and F. Melgani. Convolutional SVM networks for object detection in UAV imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3107–3118, June 2018.

[20] Francesca Beccacece, Emanuele Borgonovo, Greg Buzzard, Alessandra Cillo, and Stanley Zionts. Elicitation of multiattribute value functions through high dimensional model representations: Monotonicity and interactions. *European Journal of Operational Research*, 246(2):517–527, 2015.

[21] M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, Aug 1988.

[22] Bernice B Brown. Delphi process: a methodology used for the elicitation of opinions of experts. Technical report, Rand Corp Santa Monica CA, 1968.

[23] Michel Bruneau, Stephanie E. Chang, Ronald T. Eguchi, George C. Lee, Thomas D. O'Rourke, Andrei M. Reinhorn, Masanobu Shinozuka, Kathleen Tierney, William A. Wallace, and Detlof Von Winterfeldt. A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities. *Earthquake Spectra*, 19(4):733–752, 2003.

[24] Mark A Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web*, pages 695–698, 2012.

[25] Quoc Dung Cao and Youngjun Choe. Deep learning based damage detection on post-hurricane satellite imagery. *arXiv preprint arXiv:1807.01688*, 2018.

[26] Quoc Dung Cao and Youngjun Choe. Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks. *Natural Hazards*, pages 1–20, 2020.

[27] Young-Jin Cha, Wooram Choi, and Oral Büyüköztürk. Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5):361–378, 2017.

[28] Young-Jin Cha, Wooram Choi, Gahyun Suh, Sadegh Mahmoudkhani, and Oral Büyüköztürk. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, 33(9):731–747, 2018.

[29] Stephanie E. Chang. Urban disaster recovery: A measurement framework and its application to the 1995 Kobe earthquake. *Disasters*, 34(2):303–327, 2010.

[30] Stephanie E. Chang, Timothy Mcdaniels, Jana Fox, Rajan Dhariwal, and Holly Longstaff. Toward disaster-resilient cities: Characterizing resilience of infrastructure systems with expert judgments. *Risk Analysis*, 34(3):416–434, 2014.

[31] S. Chen, H. Wang, F. Xu, and Y. Jin. Target classification using the deep convolutional networks for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4806–4817, Aug 2016.

[32] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2811–2821, May 2018.

[33] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11 – 28, 2016.

[34] Gian Paolo Cimellaro, Andrei M. Reinhorn, and Michel Bruneau. Framework for analytical quantification of disaster resilience. *Engineering Structures*, 32(11):3639–3649, 2010.

[35] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1237–1242. AAAI Press, 2011.

[36] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber. A committee of neural networks for traffic sign classification. In *The 2011 International Joint Conference on Neural Networks*, pages 1918–1921, July 2011.

[37] Robert T. Clemen. Comment on cooke's classical method. *Reliability Engineering & System Safety*, 93(5):760 – 765, 2008. Expert Judgement.

[38] Robert T Clemen and Robert L Winkler. Calibrating and combining precipitation probability forecasts. In *Probability and Bayesian statistics*, pages 97–110. Springer, 1987.

[39] OSSPAC (Oregon Seismic Safety Policy Advisory Commission). The oregon resilience plan: Reducing risk and improving recovery for the next cascadia earthquake and tsunami. 2013.

[40] Washington (State). Seismic Safety Committee. *Resilient Washington State: A Framework for Minimizing Loss and Improving Statewide Recovery After an Earthquake: Final Report and Recommendations, November 2012*. Washington State Seismic Safety Committee, 2012.

[41] Roger Cooke and Others. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand, 1991.

[42] Roger M Cooke, Susie ElSaadany, and Xinzheng Huang. On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering & System Safety*, 93(5):745–756, 2008.

[43] Roger M Cooke and Louis LHJ Goossens. Tu delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5):657–674, 2008.

[44] Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 313–320, Cambridge, MA, USA, 2003. MIT Press.

[45] Siddhartha Dalal, Dmitry Khodyakov, Ramesh Srinivasan, Susan Straus, and John Adams. ExpertLens: A system for eliciting opinions from a large pool of non-collocated experts with diverse knowledge. *Technological Forecasting and Social Change*, 78(8):1426–1444, 2011.

[46] Norman Dalkey and Bernice Brown. Comparison of group judgment techniques with short-range predictions and almanac questions. Technical report, RAND CORP SANTA MONICA CA, 1971.

[47] Norman C Dalkey. The Delphi method: An experimental study of group opinion. Technical report, RAND CORP SANTA MONICA CALIF, 1969.

[48] S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2016.

[49] Jigar Doshi, Saikat Basu, and Guan Pang. From satellite imagery to disaster insights. *arXiv preprint arXiv:1812.07033*, 2018.

[50] Ian Durbach, Bruno Merven, and Bryce McCall. Expert elicitation of autocorrelated time series with application to e3 (energy-environment-economic) forecasting models. *Environmental Modelling & Software*, 88:93–105, 2017.

[51] Bruce R Ellingwood. Earthquake risk assessment of building structures. *Reliability Engineering & System Safety*, 74(3):251–262, 2001.

[52] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human factors*, 37(1):32–64, 1995.

[53] Mica R Endsley and Daniel J Garland. *Situation awareness analysis and measurement*. CRC Press, 2000.

[54] Mica R Endsley, Daniel J Garland, et al. Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 1(1):3–21, 2000.

[55] Laura Faulkner. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3):379–383, 2003.

[56] Federal Emergency Management Agency. Damage assessment operations manual. Technical report, The U.S. Department of Homeland Security, April 2016.

[57] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, Apr 1980.

[58] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[59] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[60] S. Golchi, D. R. Bingham, H. Chipman, and D. A. Campbell. Monotone emulation of computer experiments. *SIAM-ASA Journal on Uncertainty Quantification*, 3(1):370–392, 2015.

[61] Woo Gordon. *Calculating catastrophe*. World Scientific, 2011.

[62] B Gouldby, P Sayers, J Mulet-Marti, MAAM Hassan, and D Benwell. A methodology for regional-scale flood risk assessment. In *Proceedings of the Institution of Civil Engineers-Water Management*, volume 161, pages 169–182. Thomas Telford Ltd, 2008.

[63] Jim W Hall, RJ Dawson, PB Sayers, C Rosu, JB Chatterton, and R Deakin. A methodology for national-scale flood risk assessment. In *Proceedings of the Institution of Civil Engineers-Water and Maritime Engineering*, volume 156, pages 235–247. Thomas Telford Ltd, 2003.

[64] Haoyuan Hong, Himan Shahabi, Ataollah Shirzadi, Wei Chen, Kamran Chapi, Baharin Bin Ahmad, Majid Shadman Roodposhti, Arastoo Yari Hesar, Yingying Tian, and Dieu Tien Bui. Landslide susceptibility assessment at the Wuning area, China: a comparison between multi-criteria decision making, bivariate statistical and machine learning methods. *Natural Hazards*, Nov 2018.

[65] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2042–2050. Curran Associates, Inc., 2014.

[66] Fu Jie Huang and Yann LeCun. Large-scale learning with SVM and convolutional nets for generic object categorization. In *Proceedings - 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, volume 1, pages 284–291, 2006.

[67] Hong-wei Huang, Qing-tong Li, and Dong-ming Zhang. Deep learning based image recognition for crack and leakage defects of metro shield tunnel. *Tunnelling and Underground Space Technology*, 77:166–176, 2018.

[68] K. Jack. Road inspector using neural network. `https://github.com/jackkwok/neural-road-inspector`, visited 2019-04-27, 2017.

[69] Kishor Jaiswal, David J Wald, David M Perkins, Willy P Aspinall, and Anne S Kiremidjian. Estimating structural collapse fragility of generic building typologies using expert judgment. 2014.

[70] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[71] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[72] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*, 2019.

[73] Youngjoo Kwak, Badri Bhakta Shrestha, Atsuhiro Yorozuya, and Hisaya Sawano. Rapid damage assessment of rice crop after large-scale flood in the cambodian floodplain using temporal spatial data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3700–3709, 2015.

[74] Stephen Law, Brooks Paige, and Chris Russell. Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–19, 2019.

[75] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, pages 319–, London, UK, UK, 1999. Springer-Verlag.

[76] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen. Medical image classification with convolutional neural network. In *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*, pages 844–848, Dec 2014.

[77] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma. Large-scale remote sensing image retrieval by deep hashing neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):950–965, Feb 2018.

[78] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing.

[79] Y. Liu, Y. Zhong, and Q. Qin. Scene classification based on multiscale convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–13, 2018.

[80] Y. Long, Y. Gong, Z. Xiao, and Q. Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, May 2017.

[81] Andres F. Lopez-Lopera, Franccois Bachoc, Nicolas Durrande, and Olivier Roustant. Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM-ASA Journal on Uncertainty Quantification*, 6(3):1224–1255, 2018.

[82] Andrés F. López-Lopera, François Bachoc, Nicolas Durrande, Jérémy Rohmer, Déborah Idier, and Olivier Roustant. Approximating Gaussian Process Emulators with Linear Inequality Constraints and Noisy Observations via MC and MCMC. (Mc), 2019.

[83] Jiazheng Lu, Yu Liu, Guoyong Zhang, Bo Li, Lifu He, and Jing Luo. Partition dynamic threshold monitoring technology of wildfires near overhead transmission lines by satellite. *Natural Hazards*, 94(3):1327–1340, Dec 2018.

[84] Hassan Maatouk. Finite-dimensional approximation of Gaussian processes To cite this version : HAL Id : hal-01533356 Finite-dimensional approximation of Gaussian processes with inequality constraints. 2017.

[85] Hassan Maatouk and Xavier Bay. Gaussian Process Emulators for Computer Experiments with Inequality Constraints. *Mathematical Geosciences*, 49(5):557–582, 2017.

[86] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, Feb 2017.

[87] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. High-resolution aerial image labeling with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7092–7103, Dec 2017.

[88] Akansha Mehrotra, Krishna Kant Singh, M. J. Nigam, and Kirat Pal. Detection of tsunami-induced changes using generalized improved fuzzy radial basis function neural network. *Natural Hazards*, 77(1):367–381, May 2015.

[89] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8):1778–1790, Aug 2004.

[90] Scott B. Miles. Participatory model assessment of earthquake-induced landslide hazard models. *Natural Hazards*, 56(3):749–766, 2011.

[91] Scott B Miles. Foundations of community disaster resilience: Well-being, identity, services, and capitals. *Environmental Hazards*, 14(2):103–121, 2015.

[92] Scott B. Miles and Stephanie E. Chang. Modeling community recovery from earthquakes. *Earthquake Spectra*, 22(2):439–458, 2006.

[93] Nobuo Mimura, Kazuya Yasuhara, Seiki Kawagoe, Hiromune Yokoki, and So Kazama. Damage from the great east japan earthquake and tsunami-a quick report. *Mitigation and adaptation strategies for global change*, 16(7):803–818, 2011.

[94] Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International conference on machine learning (ICML-12)*, pages 567–574, 2012.

[95] L. Mou, P. Ghamisi, and X. X. Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655, July 2017.

[96] Alivelu Mukkamala and Roman Beck. Enhancing disaster management through social media analytics to develop situation awareness what can be learned from twitter messages about hurricane sandy? In *PACIS*, page 165, 2016.

[97] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[98] Javad Najafi, Ali Peiravi, Amjad Anvari-Moghaddam, and Josep M Guerrero. Power-heat generation sources planning in microgrids to enhance resilience against islanding due to natural disasters. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, pages 2446–2451. IEEE, 2019.

[99] Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

[100] N Nojima. Restoration processes of utility lifelines in the great east japan earthquake disaster, 2011. In *15th World Conference on Earthquake Engineering (15WCEE)*, pages 24–28, 2012.

[101] JE Oakley and A O'Hagan. Shelf: the sheffield elicitation framework (version 2.0). *Sheffield, UK: School of Mathematics and Statistics, University of Sheffield*, 2010.

[102] Ferda Ofli, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, Pauline Millet, Friedrich Reinhard, and Matthew Parkan. Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data*, 4(1):47–59, 2016.

[103] FE O'Loughlin, RCD Paiva, M Durand, DE Alsdorf, and PD Bates. A multi-sensor approach towards a global vegetation corrected srtm dem product. *Remote Sensing of Environment*, 182:49–59, 2016.

[104] Kalpesh Patil, Mandar Kulkarni, Anand Sriraman, and Shirish Karande. Deep learning based car damage classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 50–54. IEEE, 2017.

[105] Kyriazis Pitilakis, Maria Alexoudi, Sotiris Argyroudis, Olivier Monge, and Christophe Martin. Earthquake risk assessment of lifelines. *Bulletin of Earthquake Engineering*, 4(4):365–390, 2006.

[106] Hamid Reza Ranjbar, Alireza A. Ardalan, Hamid Dehghani, and Mohammad Reza Saradjian. Using high-resolution satellite imagery to provide a relief priority map after earthquake. *Natural Hazards*, 90(3):1087–1113, Feb 2018.

[107] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[108] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[109] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[110] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. *Journal of Machine Learning Research*, 9:645–652, 2010.

[111] San Francisco Bay Area Planning and Urban Research Association (SPUR). The resilient city: Defining what san francisco needs from its seismic mitigation policies, 2009.

[112] Ronghua Shang, Junkai Lin, Licheng Jiao, and Yangyang Li. Sar image segmentation using region smoothing and label correction. *Remote Sensing*, 12(5):803, 2020.

[113] Ge Shaoyun, Li Jifeng, Liu Hong, Cao Yuchen, Yang Zan, and Yan Jun. Assessing and boosting the resilience of a distribution system under extreme weather. In *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2019.

[114] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository*, abs/1409.1556, 2014.

[115] Sergii Skakun, Nataliia Kussul, Andrii Shelestov, and Olga Kussul. Flood hazard and flood risk assessment using a time series of satellite images: A case study in namibia. *Risk Analysis*, 34(8):1521–1537, 2014.

[116] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[117] Jérémie Sublime, Andrés Troya-Galvis, and Anne Puissant. Multi-scale analysis of very high resolution satellite images using unsupervised techniques. *Remote Sensing*, 9(5):495, 2017.

[118] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, pages 1008–1016, 2015.

[119] Brian Tomaszewski. Situation awareness and virtual globes: Applications for disaster management. *Computers & Geosciences*, 37(1):86–92, 2011.

[120] Ming-Hsiang Tsou, Chin-Te Jung, Christopher Allen, Jiue-An Yang, Su Yeon Han, Brian H Spitzberg, and Jessica Dozier. Building a real-time geo-targeted event observation (geo) viewer for disaster management and situation awareness. In *International Cartographic Conference*, pages 85–98. Springer, 2017.

[121] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[122] CJ Van der Sande, SM De Jong, and APJ De Roo. A segmentation and classification approach of ikonos-2 imagery for land cover mapping to assist flood risk and flood damage assessment. *International Journal of applied earth observation and geoinformation*, 4(3):217–229, 2003.

[123] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, July 2017.

[124] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *Computing Research Repository*, abs/1505.00853, 2015.

[125] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. *IEEE intelligent systems*, (6):52–59, 2012.

[126] Ting Yuan, Hyongki Lee, Hanwen Yu, Hahn Chul Jung, Austin Madson, Yongwei Sheng, and Edward Beighley. Mapping forested floodplain topography using insar and radar altimetry. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12):5189–5198, 2019.

[127] Daniel Zhang, Yang Zhang, Qi Li, Thomas Plummer, and Dong Wang. Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1221–1232. IEEE, 2019.

[128] L. Zhang, L. Zhang, and B. Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, June 2016.

[129] Z. Zhang, H. Wang, F. Xu, and Y. Jin. Complex-valued convolutional neural network and its application in polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7177–7188, Dec 2017.

[130] Z. Zhong, J. Li, Z. Luo, and M. Chapman. Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):847–858, Feb 2018.

[131] Wubai Zhou, Chao Shen, Tao Li, Shu-Ching Chen, and Ning Xie. Generating textual storyline to improve situation awareness in disaster management. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, pages 585–592. IEEE, 2014.

[132] Kirsten Zickfeld, Anders Levermann, M Granger Morgan, Till Kuhlbrodt, Stefan Rahmstorf, and David W Keith. Expert judgements on the response of the Atlantic meridional overturning circulation to climate change. *Climatic Change*, 82(3-4):235–265, 2007.