

**STATISTICAL MODELING OF WATER PIPELINE DAMAGE IN
EARTHQUAKES**

by

Adam Yusuf Bagriacik

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Civil Engineering

Spring 2017

© 2017 Adam Yusuf Bagriacik
All Rights Reserved

**STATISTICAL MODELING OF WATER PIPELINE DAMAGE IN
EARTHQUAKES**

by

Adam Yusuf Bagriacik

Approved: _____
Rachel Davidson, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Harry W. Shenton III, Ph. D.
Chair of the Department of Civil and Environmental Engineering

Approved: _____
Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____
Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

First, I would like to thank Professor Davidson for all your support during the past two years. I entered Graduate School as a novice in this subject, but your faith and patience allowed me to sufficiently grasp the concepts and overcome all obstacles.

To the Professor Davidson office group – Kun Yang, Zeinab Yahyazadeh, Di Ha, and, most importantly, Grandma (Dong Wang), thank you for your support and for putting up with me for 6 or so hours a day. You provided a social environment to the graduate office so research never became overwhelmingly stressful.

Finally, I want to acknowledge my friends and family for all their love. To my roomie, Kokeb Abera, thanks for paying half the rent the past two years. Thank you Nina King, Mom (Amy Dugan), Grandmom (Amy Serafine), and Pop-pop (Glen Serafine) for literally everything. I can attribute all of my success and accomplishments to you four. I love you all so much!

TABLE OF CONTENTS

LIST OF TABLES	v
LIST OF FIGURES	vi
ABSTRACT	vii
Chapter	
1 INTRODUCTION	1
2 AVAILABLE MODELS OF EARTHQUAKE DAMAGE TO WATER PIPELINES.....	3
3 COVARIATES.....	8
3.1 Hazard-Related Covariates	9
3.2 Pipe Attributes	10
4 DATA	13
5 MODEL TYPES.....	21
5.1 Repair Rate Method.....	21
5.2 Logistic Regression (Logit).....	22
5.3 Boosted Regression Trees (BRT).....	23
5.4 Random Forests (RF)	25
6 ANALYSIS	27
6.1 Balancing Data	27
6.2 Evaluation Methods.....	28
7 RESULTS.....	33
7.1 Balancing Methods	33
7.2 Model Results	34
7.2.1 Total city level prediction.....	35
7.2.2 Individual pipe classification.....	36
7.2.3 Suburb-level prediction	37
7.3 Importance of Covariates	41
8 CONCLUSIONS	47
REFERENCES	49

LIST OF TABLES

Table 2.1:	Statistical models developed in previous literature studies using the RR method.....	5
Table 3.1:	Summary of hypothesized effects of different factors on earthquake damage to pipes.....	8
Table 4.1:	Descriptive statistics for numerical covariates.....	16
Table 4.2:	Descriptive statistics for Pipe Material covariate, with reference level indicated.....	17
Table 4.3:	Descriptive statistics for Pipe Type covariate with reference level indicated.....	17
Table 4.4:	Descriptive statistics for Trench Type covariate, with reference level indicated.....	18
Table 4.5:	Three Datasets Considered for Analysis.....	19
Table 4.6:	Preliminary Results of Three Datasets.....	20
Table 6.1:	Evaluation metrics.....	31
Table 6.2:	Example confusion matrix.....	32
Table 7.1:	Comparison of predictive performance for different sampling methods using cross validation.....	34
Table 7.2:	Comparison of predictive performance for four model types using cross validation with the February earthquake data and June earthquake holdout validation.....	35
Table 7.3:	Logit results for full and reduced models.....	44
Table 7.4:	Predictive performance of reduced models compared to full models.....	46

LIST OF FIGURES

Figure 4.1: Christchurch Location in New Zealand.....	14
Figure 4.2: Conceptual framework of water pipeline earthquake damage	15
Figure 4.3: Pipe locations for (a) Casewise Deletion, (b) LRI Imputation, and (c) PGV Limited datasets.....	19
Figure 7.1: Number of damaged pipes in each suburb for February earthquake, (a) Observed and in each model: (b) Logit, (c) BRT, (d) RF smote, and (e) RR	38
Figure 7.2: Predicted vs. observed number of damaged pipes in each suburb for February earthquake, for each model: (a) Logit, (b) BRT, (c) RF smote (note y-axis range is different), and (d) RR	39
Figure 7.3: Number of damaged pipes in each suburb for June earthquake, (a) Observed and in each model: (b) Logit, (c) BRT, (d) RF smote, and (e) RR	40
Figure 7.4: Predicted vs. observed number of damaged pipes in each suburb for June earthquake, for each model: (a) Logit, (b) BRT, (c) RF smote (note y-axis range is different), and (d) RR.....	41

ABSTRACT

A large dataset of water pipeline damage from the February and June 2011 earthquakes in Christchurch, New Zealand is used to fit four mathematical model types—logit, boosted regression trees (BRT), and random forest (RF), and the repair rate (RR) method common in the literature. Cross validation and holdout validation are used with multiple metrics to fully evaluate the models' ability to accurately predict the total number and approximate spatial distribution of damaged pipes; to correctly classify each individual pipe as damaged or not, and to describe the relative importance of pipe and earthquake attributes in predicting damage. Results suggest that while BRT offers the best overall performance, logit offers the advantages of a closed form solution and an ability to compare pipe materials explicitly, and the far simpler RR method is very good at predicting the total number of damaged pipes, though less capable of prediction at the individual pipe or suburb level. The analysis provides evidence that “modified” PVC (MPVC), UPVC, Polyethylene 80B (PE80B), High Density Polyethylene (HDPE), and Cast Iron (CI) were associated with the least damage, and Galvanized Iron (GI) with the most; and that the more recent the type of trench it is in, the less likely a pipe is to be damaged, even when controlling for the pipe age. The analysis highlights the need to compute and report the predictive errors of different types and acknowledge them in using the models for subsequent analysis.

Chapter 1

INTRODUCTION

Earthquakes can cause extensive damage to buried water pipelines, severely disrupting a community's supply of water for firefighting, drinking, cleaning, industrial processes, and other uses. Being able to manage that risk requires an understanding of the likely amount and spatial distribution of damage future earthquakes are likely to cause, and what attributes of the pipes and/or ground motion are most associated with damage. Previous research has produced multiple mathematical models of pipe damage in earthquakes, and has identified the primary factors associated with increased damage (O'Rourke and Liu 2012). In this paper, we add to that body of knowledge by using damage data from the February 22, 2011 (Mw = 6.2) earthquake in Christchurch, New Zealand to fit new mathematical models of earthquake-caused damage to water pipelines, while data from the June 13, 2011 Christchurch earthquake is used to assess the predictive power of the models. Since all models will not serve all purposes equally well, it is important to specify the intended uses of the models *a priori*. Uses of these models include describing the magnitude of risk to the water supply system, supporting emergency response planning (e.g., repair resources needed), supporting mitigation planning (e.g., pipe materials that should/should not be used), and providing damage maps to use as input for models of service disruptions and societal impact. To support these applications, we had four specific goals. For a specified earthquake, the models should accurately predict the (1) total number and (2) approximate spatial distribution of damaged pipes. They should

also aim to (3) correctly classify each individual pipe as damaged or not, and (4) describe the relative importance of various pipe and earthquake attributes in predicting damage, especially material type, trench type, and other characteristics that could be modified as part of a mitigation program.

The study presented offers contributions related to the data, analyses, and evaluation methods used. First, we employ a uniquely large and complete dataset from an earthquake that caused extensive damage to a modern water supply system. The dataset includes observations on 83,746 pipes (2,311 of which experienced damage) with multiple relevant characteristics of each. Second, we employ and compare statistical and machine learning models—logistic regression (logit), boosted regression trees (BRT), and random forests (RF)—that promise multiple benefits and are well-developed though new to this application. These model types allow investigation of multiple covariates and interactions among them. They can use each length of pipe as a unit of analysis rather than repair rate for a region, ensuring that the covariates refer more directly to a specified pipe rather than being smoothed over a region. We compare the model types to the simpler approach used in the literature. Third, in this study, for the first time, we use multiple metrics to fully evaluate and compare the models' ability to predict damage in future events and achieve the four stated goals—total count and spatial distribution of damaged pipes, classification of individual pipes, and relative importance of covariates. After summarizing the empirical literature on models of earthquake damage to water pipelines in Chapter 2, we summarize previous findings on influential covariates in Chapter 3. The data, model types, analyses, and results are described in Chapters 4, 5, 6, and 7, respectively.

Chapter 2

AVAILABLE MODELS OF EARTHQUAKE DAMAGE TO WATER PIPELINES

Many analyses have been conducted to examine performance of buried pipelines in earthquakes—Physical experiments (e.g., Abdoun et al. 2009, Tsai et al. 2000), analytical (e.g., Davis et al. 2007), numerical (e.g., Vazouras et al. 2010), and empirical or statistical curve fitting (e.g., Jeon and O’Rourke 2005). The focus here is on empirical models, i.e., mathematical relationships fitted to damage data recorded in previous earthquakes (e.g., O’Rourke and Ayala 1993, O’Rourke and Jeon 1999, Pineda-Porras and Ordaz 2010, Kimishima et al. 2011, O’Rourke et al. 2012). O’Rourke and Liu (2012), Pineda-Porras and Najafi (2010), and Lanzano et al. (2014), respectively, provide overviews of pipeline performance in earthquakes in general and of statistical models specifically.

The empirical modeling efforts have typically used the same general approach, which for convenience we call the repair rate method (RR). They have aimed to develop a curve that relates *repair rate*, RR (number of repairs (i.e., damage locations) per km. of pipe) to a measure of ground motion, ground deformation, or strain. In some cases, they present different curves for different groups of pipe based on their material, diameter, or other characteristic. Isoyama et al. (2000), for example, presents four different equations, one for each combination of pipe material (ductile cast iron and cast iron) and a ground motion measure (PGV and PGA). The approach typically involves first dividing the affected geographic area into regions of approximately equal ground motion intensity. For each ground motion contour, a single value of repair rate, RR , is computed (total number of repairs/length of pipe), producing a data

pair of RR and ground motion level. The observations of paired data are plotted, a least squares line is fitted to them (Table 2.1).

In this approach, each earthquake produces a relatively small number of observations (typically 5 to 25). Fitting separate curves for different pipe materials or other subsets of pipes can further reduce the number of observations. Many papers report an R^2 value as an indication of goodness-of-fit, and they are typically relatively high (e.g., 0.84 and 0.98 in Jeon and O'Rourke 2005 and Milashuk and Crane 2012, respectively). It is important to note, however, that these R^2 values are measuring *ecological correlation* rather than *individual correlation*, because the observations are based on groups of pipes rather than individual pipes. As the seminal paper Robinson (1950, p339) explained, "there need be no correspondence between the individual and the ecological correlation." Thus, while the high R^2 values reported seem to suggest high quality models showing strong correlations between repairs and ground motion, they may be misleading. Most studies have used a comparison to previous models and the R^2 values as forms of assessment. They have not assessed out-of-sample predictive power, i.e., the models' ability to correctly predict damage for observations not in the sample used to fit the model. Finally, previous studies have not typically examined multiple pipe attributes simultaneously. Since the pipe attributes are not independent (e.g., most trunks are one of a few material types, and have relatively large diameters), it is unclear whether they are capturing the attribute specified or something related to it. In the one exception we found, Maruyama et al. (2015) fitted a logit model to water pipeline damage from the 2011 Tōhoku earthquake. The model considered four covariates, but ultimately underestimated actual damage because liquefaction was not considered.

Table 2.1: Statistical models developed in previous literature studies using the RR method

Reference	Form of model ^a	Different models ^a	Earthquake(s) source for data	Num. obs.
O'Rourke and Ayala 1993	$RR = aPGV^b$	One model for each <i>CI</i> , <i>AC</i> , <i>CONC</i> , <i>PC</i>	4 U.S. and 2 Mexican, 1965-1989	11
Eidinger and Maison 1995/ALA 2001	$R_{1000} = KaPGV$ $R_{1000} = KbPGD^c$	The prediction is the maximum result of the two eqns. K factor depends on pipe material, joint, soil	5 U.S. and 2 Mexican, 1965-1989	20
O'Rourke and Jeon 1999	$\log(RR) = a\log(IM) +$ $RR = c(PGV/D^d)^e$	For first eqn., 6 models, for all combinations of (<i>CI</i> , <i>DI</i> , and <i>AC</i>), (<i>PGV</i> and <i>D</i>). For second eqn., one model for <i>CI</i> , <i>DI</i>	1994 Northridge	3 – 17
Isoyama et al. 2000	$RR = C_p C_d C_g (IM - a)^b$	For first eqn., 4 models, for all combinations of (<i>CI</i> and <i>DI</i>), $IM=(PGA$ and $PGV)$	1995 Hyogoken-nanbu (Kobe)	12 – 17
Chen et al. 2002	$RR = aIM^b$	S6 models for all combinations of (small/large <i>D</i>), $IM=(SI$, PGA and $PGV)$	1999 Ji-Ji (Chi-Chi)	-
Pineda-Porras and Ordaz 2003	$RR = \Phi(PGV - \mu/\sigma)$ $RR = aPGV +$	TFirst eqn. for lower <i>PGV</i> . Second eqn. for higher <i>PGV</i>	1985 Michoacan	5 – 20
O'Rourke and Deyoe 2004	$RR = aIM^b$	T3 models. One for each $Var=PGV_{SW}$, PGV_{BW} and ϵ .	4 U.S. and 2 Mexican, 1965-1989	9 – 23
Jeon and O'Rourke 2005	$\ln(RR) = a\ln(IM) +$	T3 models. One for each $Var=PGV$, $VM PGV$ and $GMPGV$. Only <i>CI</i> pipes.	1994 Northridge	6 – ~80
Pineda-Porras and Ordaz 2007	$RR = a$ $RR = b(PGV^2/PGA) + c$	The first eqn. is used for lower values of PGV^2/PGA . The second eqn. is used for higher values of PGV^2/PGA	1985 Michoacan	5 – 20
Maruyama et al. 2008	$RR = C_p C_d C_g^*$ $\Phi((\ln(PGV) - \lambda)/\zeta)$	2 models for different extent of damage	2003 Northern-Miyagi, 2003 Tokachi-oki, and 2004 Mid-Nigata	12 – 14
Maruyama and Yamakazi 2010; Maruyama et al. 2010	$RR = C_p C_d C_g^*$ $\Phi((\ln(PGV) - \lambda)/\zeta)$	2 models. One used for (<i>CI</i> and <i>V</i>) and another for <i>DI</i>	4-5 Japanese	12 – 17

Table 2.1 continued

Reference	Form of model ^a	Different models ^a	Earthquake(s) source for data	Num. obs.
Pineda-Porras and Ordaz 2010	$RR = a$ $RR = b \frac{PGV^2}{PGA} + c$	6 models. Three sets of the two eqns. One set for each of the following: low <i>DGS</i> , high <i>DGS</i> , unknown <i>DGS</i> . The first eqn. is used for lower values of <i>PGV</i> ² / <i>PGA</i> . The second eqn. is used for higher values of <i>PGV</i> ² / <i>PGA</i>	1985 Michoacan	4
Kimishima et al. 2011	$RR = C_p C_d C_g^* \Phi((\ln(PGV) - \lambda)/\zeta)$	2 models. One used for (<i>CI</i> and <i>V</i>) and another for <i>DI</i>	2007 Niigata-Ken Chuetsu-Oki Japan	11 – 12
Milashuk and Crane 2012	$\ln(RR) = a \ln(PGV) +$	3 models. One each for <i>CI</i> , <i>AC</i> , <i>PVC-MPVC</i>	2010 Darfield and 2011 Christchurch, plus 4 older earthquakes	5 – 14
O'Rourke et al. 2012	$\log(RR) = a \log(GMPGV) + b$	2 models. One each for <i>AC</i> , <i>CI</i>	2010 Darfield and 2011 Christchurch, plus 4 U.S. earthquakes	13
O'Rourke et al. 2014	$\log(RR) = a \log(GMPGV) +$	2 models. One each for <i>AC</i> , <i>CI</i>	2011 Christchurch, plus 4 U.S. earthquakes	9 – 14
Cubrinovski et al. 2014	$\ln(RR) = a \ln(PGA) +$	5 models. One for each LRI level (0-4). Only <i>AC</i> materials used.	2011 Christchurch	10 – ~40
Bouziou and O'Rourke 2015	$\log(RR) = a \log(GMPGV) +$	2 models. One for each <i>AC</i> , <i>CI</i>	2011 Christchurch, plus 4 U.S. earthquakes	6 – 13
Maruyama et al. 2015	$P(D) = \frac{\exp(\vec{x}_i^T \vec{\beta})}{1 + \exp(\vec{x}_i^T \vec{\beta})}$	One model with <i>PGV</i> , pipe length, pipe material vulnerability, and ground condition vulnerability	2011 Tōhoku	-

^a *a*, *b*, *c*, *d*, *e* are constants; *RR*=repair rate; *D*=Diameter; *K*=factor for pipe material, soil, and joint; *IM*=ground motion intensity measure; *PGV*=peak ground velocity; *PGA*=peak ground acceleration; *CI*=cast iron; *DI*=ductile iron; *AC*=asbestos cement; *CONC*=concrete; *PC*=prestressed concrete; *PGD*=peak ground displacement; *C_p*=correction factor for pipe material; *C_d*=correction factor for diameter; *C_g*=correction factor for topography; *SI*=spectrum intensity; *R₁₀₀₀*=repairs per 1000 ft; *ε*=ground strain; *V*=vinyl; *VM PGV*=maximum vector magnitude peak ground velocity; *GMPGV*=geometric mean peak ground velocity; *DGS*=differential ground subsidence; *P(D)*=probability of pipe damage

The study presented herein adds to the literature by using individual pipes as the unit of analysis, thus focusing on the individual correlation that is truly of interest; by investigating multiple covariates simultaneously to more precisely identify the characteristics most directly associated with damage; and by explicitly evaluating the out-of-sample predictive power of the new and repair rate models through cross validation and holdout validation.

Chapter 3

COVARIATES

Several factors have been investigated to determine their influence on pipe damage in earthquakes, including ground shaking, permanent ground deformation, pipe material, pipe diameter, year laid, pipe type (e.g., trunk, main), and trench backfill type. Previous findings on the hazard- and pipe-related covariates are presented in Sections 3.1 and 3.2, respectively, and Table 3.1 summarizes the resulting hypotheses for our analysis.

Table 3.1: Summary of hypothesized effects of different factors on earthquake damage to pipes

Factor	Hypothesized effect	Strength of evidence
Ground shaking	Stronger shaking → more damage	Strong consensus
Permanent ground deformation	Larger displacement → more damage	Strong consensus
Pipe material	Less ductile (AC, CI) → most damage PVC → middle level of performance More ductile (DI, S) → least damage	Moderate
Pipe diameter	Smaller diameter → more damage	Strong consensus
Year laid	Older → More damage	Little
Pipe type (e.g., trunk, main)	Unclear	Little
Trench backfill type	Depends on pipe material and PGD	Little

3.1 Hazard-Related Covariates

All previous empirical models have used some measure of hazard—ground shaking or permanent ground displacement—as the primary covariate and found evidence that more intense ground motion is associated with more damage. Pineda-Porras and Najafi (2010) identifies at least nine ground motion metrics employed (Modified Mercalli Intensity (MMI), peak ground acceleration (PGA), peak ground velocity (PGV), peak ground displacement (PGD), arias intensity (AI), spectral intensity (SI), maximum ground strain (ϵ_g), and the composite PGV^2/PGA). MMI was used in the 1980s and 1990s (e.g., Eguchi 1991) because of its availability for earlier events, but it was then phased out in favor of instrument-based measures. PGA was widely used until 2000 since it was easier to compute than PGV and ground motion prediction equations for PGA were more available than those for PGV. Since then, however, PGV has become the most common measure because it has been shown to lead to better fitting models than PGA (e.g., Isoyama et al. 2000), and it is directly related to but easier to compute than ground strain, the main cause of damage (Pineda-Porras and Najafi 2010). More recently, geometric mean peak ground velocity (GMPGV) has been considered as well (Bouziou and O'Rourke 2015).

In addition to transient ground motions, pipe damage can be caused by permanent ground deformation due to liquefaction, landslide, fault displacement, or settlement. Approximately a dozen studies have investigated and all have found evidence that an increase in permanent ground displacement or occurrence of liquefaction is associated with increased damage. Many studies have measured liquefaction in terms of a binary or three-level categorical covariate (none, partial, or total liquefaction) (e.g., Isoyama 2000, Shirozu and Isoyama 1998, O'Rourke and Toprak 1997). Others have used permanent ground displacement (PGD) as a

continuous measure (e.g., Terzi et al. 2007, Hwang and Lin 1997, Heubach 1995). O'Rourke and Deyoe (2004) use ground strain, and more recently, Cubrinovski et al. (2011) developed the Liquefaction Resistance Index (*LRI*), a five-level interval covariate with 0 to 4 corresponding to most to least susceptible to liquefaction (see definition in Chapter 4).

3.2 Pipe Attributes

Material is the pipe attribute most commonly investigated. No clear consensus has emerged, however, on an ordering of materials by vulnerability to earthquake damage. Synthesizing the previous work is challenging because studies do not all consider the same set of materials or define them the same way, they use data spanning multiple countries and decades, and they often do not control for other factors that could be confounding conclusions about the influence of material, such as, pipe diameter. Despite these challenges, there does seem to be some evidence from empirical models that brittle pipes, including asbestos cement (AC) and cast iron (CI), are the most likely to be damaged; ductile pipes, such as, ductile iron (DI) and steel (S) are the least likely; and the performance of polyvinyl chloride (PVC) pipes is in the middle (e.g., Eguchi 1991, Eiding et al. 1995, Isoyama et al. 2000, Maruyama and Yamakazi 2010, Tsai et al. 2011, Cubrinovski et al. 2014, O'Rourke et al. 2014). There is less evidence on galvanized iron (GI), concrete lined steel (CLS), polyethylene (PE), or more specific types of PVC and PE. Cubrinovski et al. (2014, p.23), based on the same data as the study herein, addresses these material types, finding that “modified PVC (MPVC), PVC, medium-density polyethylene (MDPE80) and DI had <1 repairs/km, HDPE had 1.9 repairs/km, AC and S had 2.3 and 2.7

repairs/km, respectively, CI and CLS both had 3.2 repairs/km, and GI had the highest rate of 8.9 repairs/km.”

With the exception of Chen et al. (2002), most studies addressing the effect of pipe diameter have found evidence that a smaller pipe diameter is associated with increased damage (e.g., Katayama et al. 1975, Honegger 1995, Eidinger et al. 1995, Isoyama et al. 2000, Tsai et al. 2011). Considering in particular results from the Loma Prieta (1989), Northridge (1994), and Kobe (1995) earthquakes, ALA (2001, p.44) concludes that there is “not enough empirical evidence to prove a diameter effect exists for all pipe materials in any given water system. However, the empirical evidence strongly indicates that some relationship does exist, and that the largest pipes, those over 12 in. diameter, have lower damage rates than do common diameter distribution pipes of 4 in. to 12 in. diameter.” O’Rourke and Liu (2012) suggests a reason for a diameter effect is that as diameter increases, so does the joint embedment depth and therefore the ability to accommodate axial compression or extension without damage. Eidinger and Avila (1999) offer as additional possible reasons that larger diameter pipes are associated with lack of attachments, placement in better soils, more careful design and construction, and thicker walls.

The year a pipe was laid may be another important indicator of pipe condition, especially for materials subject to corrosion, and may represent differences in installation practices, materials, or other pipe attributes not captured in the other covariates (Eidinger and Avila 1999, Tromans 2004). Pipe type (e.g., trunk, main) also may also be relevant since the water pressure, orientation, and burial depth can vary by type (Cubrinovski et al. 2014). Finally, a couple sources discuss the soil/trench backfill type. O’Rourke and Liu (2012) note that for iron-based materials, like CI and

S, corrosive soils are associated with more damage than non-corrosive soils.

Examining repair rates for four backfill types (native soil, imported gravels, AP20, AP40) in the Christchurch earthquake, Cubrinovski et al. (2014) found different types were associated with more damage depending on the combination of pipe material and liquefaction susceptibility (LRI).

Chapter 4

DATA

The water supply system in Christchurch, New Zealand experienced a sequence of earthquakes in 2010 and 2011. The largest ones occurred on September 4, 2010 ($M_w=7.1$), February 22, 2011 ($M_w=6.2$), June 13, 2011 ($M_w=6.0$), and December 23, 2011 ($M_w=5.9$). The dataset used in this analysis includes damage observations from both the most damaging February 2011 earthquake and the subsequent June earthquake. We assumed damage recorded between the two was caused by the February event, and damage recorded from June to December was caused by the June event. The study area, which includes Christchurch city and the Banks Peninsula District within Lyttelton harbour, covers approximately 450 km² with a population of 350,000 (Cubrinovski et al. 2011 and 2014) (Fig. 4.1). The unit of analysis is an individual pipe segment with endpoints defined at nodes where there are significant changes in direction (e.g., street intersections) or connections with other pipes. The response variable, y , is binary (pipe is damaged or not), with 2.76% and 1.23% of pipes damaged in the February and June datasets, respectively.



Figure 4.1: Christchurch Location in New Zealand

The choice of covariates was guided by a simple conceptual framework (Fig. 4.2), findings from the literature (Chapter 3), and data availability. Pipe damage depends on the interaction of hazard, exposure, and vulnerability. Hazard includes transient and permanent ground deformations. The latter, in turn, depends on transient ground motions, liquefaction susceptibility of the soil, and groundwater depth. Pipe vulnerability is thought to depend on pipe material, diameter, year laid, pipe type (e.g., trunk, main), and trench backfill type.

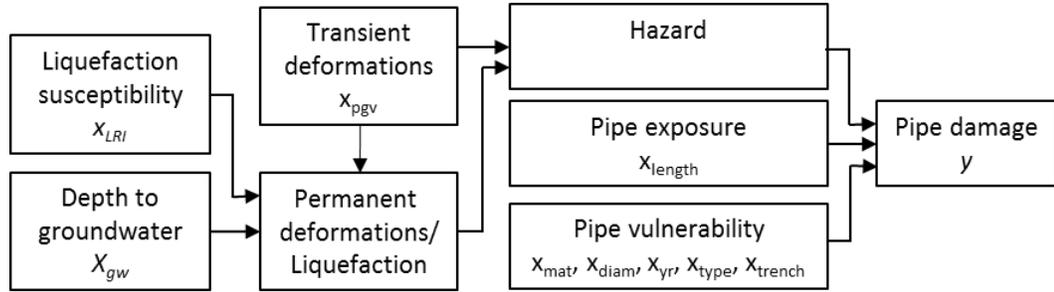


Figure 4.2: Conceptual framework of water pipeline earthquake damage

Peak ground velocity data was developed based on actual ground motion station recordings and the general spatial pattern of ground motion throughout the region (Bradley and Hughes 2012). At each pipe location, the method provides a probability distribution of PGV, with the uncertainty increasing with distance to the recording stations. In this analysis, we use the median PGV at each pipe centroid. The Liquefaction Resistance Index (LRI) provides a measure of a soil’s resistance to liquefaction (Cubrinovski et al. 2011). For each site, the demand was estimated based on recorded PGA, water table depth, and earthquake magnitude; the factor of safety was estimated from the observed severity of liquefaction; and the resistance was then back-calculated based on the liquefaction evaluation procedure in Youd et al. (2011). The resistances were discretized into five levels from 0 to 4 (smaller to greater liquefaction resistance). We treated LRI as a continuous covariate (Table 4.1).

Table 4.1: Descriptive statistics for numerical covariates

Covariate	Definition	Mean	C.O.V.	Median	Min.	Max.
x_{pgv}	Peak ground velocity (Feb.), cm/s	49.7	0.284	50.0	21.6	80.3
	Peak ground velocity (June), cm/s	30.7	0.372	30.3	10.6	62.5
x_{LRI}	Liquefaction Resistance Index, LRI ^a	2.37	0.452	3	0	4
x_{gw}	Groundwater table depth, m ^a	2.54	0.794	2.0	-0.42	13.54
x_{length}	Pipe length, m ^a	24.5	1.535	8.8	0.10	1,046
x_{diam}	Diameter, mm ^a	84.2	0.792	50	13	600
x_{yr}	Year laid, year ^a	1982	0.010	1987	1900	2011

^aData for all covariates except PGV are for the February dataset. Values are slightly different for the June dataset (means are within 1%, except for x_{LRI} and x_{gw} , which are 2.8% higher and 8.2% lower in June, respectively).

Data on the exposure and vulnerability covariates that describe the water supply system were obtained from Christchurch City Council (CCC) and the Stronger Christchurch Infrastructure Rebuild Team (SCIRT). There are four pipe types and four trench types (Tables 4.3 and 4.4). *Trunks* deliver water from aquifer-pumped bores to the *mains*, which follow roads. *Crossovers* deliver water from mains to *submains*, which are typically under footpaths and connect to individual properties. Crossovers typically run perpendicular to the mains and submains and often connect to mains at fire hydrants (Cubrinovski et al. 2014). Before 1984, trench backfill material was either locally excavated soil material or if the local material was inappropriate, imported from quarries. From 1984 on, trench construction was standardized. All pipes were put in a sand layer and covered with gravel mix (AP40 until 2005, AP20 after) (Cubrinovski et al. 2014). Overlaying the data in a geographic information system (GIS), a dataset was compiled with an observation for each length of pipe, and

a value for the response variable, y , and each covariate in Figure 4.2 and Tables 4.2, 4.3 and 4.4.

Table 4.2: Descriptive statistics for Pipe Material covariate, with reference level indicated

	Pipe material ^a , x_{mat}	Number ^b
AC	Asbestos cement	15,511
CI	Cast iron	5,201
DI	Ductile iron	1,901
GI	Galvanized iron [Reference]	10,298
S	Steel	703
CLS	Concrete-lined steel	1,232
UPVC	Unplasticised polyvinyl chloride	6,904
MPVC	Modern polyvinyl chloride	2,251
PE80B	Polyethylene 80	13,033
PE100	Polyethylene 100	196
HDPE	High-density polyethylene	26,348
LDPE	Low-density polyethylene	168

^aDI includes ductile iron, concrete-lined ductile iron, and M-lined ductile iron; UPVC includes UPVC and PVC; PE80B includes MDPE80, PE if submain after 2000; PE100 includes PE100, MDPE100, PE if main; HDPE includes HDPE and PE if submain before 2000

^bData for all covariates are for the February dataset. Values are slightly different for the June dataset (within 4%, except for CLS, which is 7% different).

Table 4.3: Descriptive statistics for Pipe Type covariate with reference level indicated

Pipe type, x_{type}	Number ^b
Trunk	168
Main [Reference]	32,283
Submain	28,040
Crossover	23,255

^bData for all covariates are for the February dataset. Values are slightly different for the June dataset (within 4%, except for CLS, which is 7% different).

Table 4.4: Descriptive statistics for Trench Type covariate, with reference level indicated

Trench type, x_{trench}	Number ^b
Pre-1984, local [Reference]	10,085
Pre-1984, import	25,130
1984-2005, AP40	42,428
Post-2005, AP20	6,103

^bData for all covariates are for the February dataset. Values are slightly different for the June dataset (within 4%, except for CLS, which is 7% different).

In the original dataset of 111,389 observations, 27,643 (25%) did not have data for the LRI covariate. Those observations with missing data were not randomly distributed; rather, most were in the less damaged Western part of the city. In fact, only 0.3% of the observations without LRI data were damaged versus 2.8% of those with the LRI data. We considered three methods of handling the missing data—casewise deletion, imputation, and a PGV-limited analysis. In the first method, all observations with missing data were deleted, leaving 83,746 observations. This is the simplest method, but removes 25% of the data and could introduce bias to the analysis since the missing data are not random. In the second method, we imputed the missing LRI values using spatial interpolation, which results in a complete dataset, but introduces error due to the imputation. Finally, we restricted the analysis to the region that experienced $PGV \geq 45$ cm/s, for which all observations were complete, which avoids potential bias due to missing data and imputation error, but results in many fewer observations (50,582) and a model that is only applicable for those more intense ground motions (Table 4.5). Figure 4.3 is the spatial location of the pipes for each dataset related to Christchurch. Preliminary logistic regression model (Section 5.2) results are similar across the three datasets in terms of goodness of fit, estimated

coefficients, significant levels, and prediction errors (Table 4.6). The metrics used for comparison are defined in Table 6.1. We proceeded with casewise deletion method, which had small errors and was most reliable.

Table 4.5: Three Datasets Considered for Analysis

Dataset	Description	Total Number of Pipes	Total Number of Repairs	% Pipes Need Repairs
Casewise Deletion	No Changes	83,756	2,312	2.76%
LRI Imputation	Estimated LRI for 27,663 pipes	108,668	2,592	2.39%
PGV-Limited	$PGV \geq 45$ cm/s	50,852	1,720	3.38%

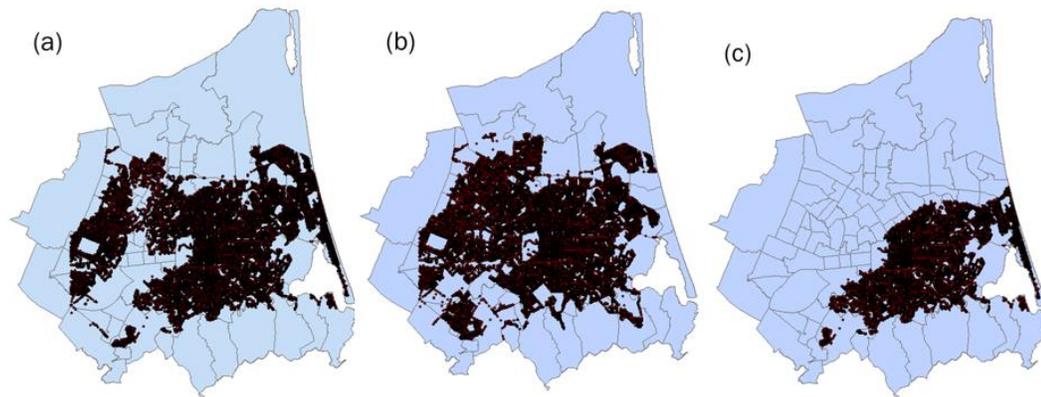


Figure 4.3: Pipe locations for (a) Casewise Deletion, (b) LRI Imputation, and (c) PGV Limited datasets

Table 4.6: Preliminary Results of Three Datasets

Dataset	TE ^a	TPE	RMSE	MAE	MASE	MPSE
Casewise Deletion	0	0.01%	0.157	0.049	3.2	35%
LRI Imputation	0	0.01%	0.148	0.043	3.9	47%
PGV-Limited	0	0.01%	0.174	0.060	4.0	25%

^aTable 6.1 provides the definition for each metric. Lower values are preferred for TE, TPE, RMSE, MAE, MASE, and MPSE.

Chapter 5

MODEL TYPES

The purpose of this analysis is to develop models to predict if an individual pipe will be damaged in a specified earthquake as a function of characteristics of the pipe and ground motion at the site. Three model types were considered—logistic regression (logit), boosted regression trees (BRT), and random forests (RF). The first is a nonlinear parametric statistical model; the latter two are nonparametric machine learning techniques. They are all appropriate for situations with a binary response variable, like this one, and have all been widely used in the literature, although never for this application. The models were fitted using *R* software v3.3.1 with default settings except where noted. Logit, BRT, and RF models were fitted using the *glm* {stats}, *gbm* {gbm v2.1.1}, and *randomForest* {randomForest v4.6-12} functions (in the noted package), respectively (R Core Team 2016, Ridgeway et al. 2015, Liaw and Wiener 2002).

5.1 Repair Rate Method

As described in Chapter 3, the repair rate (RR) models required first dividing the study area into regions of approximately equal ground motion intensity. For each ground motion contour, a single value of repair rate, *RR*, was computed (total number of repairs/length of pipe), producing a data pair of *RR* and ground motion level. In this analysis, we used 5 cm/s contours of PGV from 25 cm/s to 85 cm/s. Applying $\ln()$ transforms of the resulting 12 observations of paired data, linear regression was used to the model: $\ln(RR) = 2.188 + 15.627\ln(PGV)$. We approximated repair rate as the number of damaged pipes divided by length of pipe, assuming that each pipe had no more than one repair. The resulting underestimation of repair rate is negligible. Based

on the Poisson assumption and repair rates by pipe material, only 10% of pipe have more than a 1% chance of more than one repair.

5.2 Logistic Regression (Logit)

Logistic regression is the most commonly used model type when the response variable is binary. A specific case of a generalized linear model (GLM) in which the response variable, Y , is assumed to follow a binomial distribution and a logit link function is used (Agresti 2007), the model is written:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \vec{x}_i^T \vec{\beta} = \beta_0 + \sum_{j=1}^m \beta_j x_{ji} \quad (5.1)$$

where p_i is the probability $Y_i = 1$ (in this case, the probability pipe i is damaged), \vec{x}_i is a vector of covariates for pipe i , and $\vec{\beta}$ is a vector of regression coefficients to be estimated, typically using the maximum likelihood method. The left side of Eq. 5.1, called the logit function of p_i , is the log of the odds ratio and can be any real number. (Logistic regression models are often called logit models.) Note that unlike in linear regression, β_j represents the change in the logit of the probability associated with a unit change in the j^{th} covariate, holding all other covariates constant. Solving for p_i , we obtain:

$$p_i = \frac{\exp(\vec{x}_i^T \vec{\beta})}{1 + \exp(\vec{x}_i^T \vec{\beta})} \quad (5.2)$$

Compared to the other model types considered, logistic regression has the benefit of offering a closed form equation, making it easy to share, and allows examination of the relative effects of different levels of categorical covariates (e.g., relative effects of different pipe materials). However, as a parametric model, including interactions among the covariates or additional nonlinear relationships between the

logit and covariates requires specifying them *a priori*, and comparing the relative importance of numerical and categorical covariates to each other is difficult.

5.3 Boosted Regression Trees (BRT)

Instead of fitting a single complex model, BRT creates many simple models and linearly combines them. The method integrates two algorithms: classification and regression trees (CART), used to develop individual tree models, and boosting, which combines them to maximize predictive performance. The CART method works by partitioning the full set of observations based on their covariate values into groups that have similar response variable values. Choosing a splitting variable and a splitting value of that variable, observations that are below the splitting value are placed in one group; observations above are in another group. For each of the two groups, a new splitting variable and value are selected, and the observations are partitioned again. The process continues until a stopping criterion is reached. In each case, the splitting variable and value are chosen so as to minimize prediction errors. In each terminal node, for regression trees, each observation is assigned the mean response of all observations in the node (Elith et al. 2008).

Boosting is a forward, stagewise approach that fits many smaller models (in this case, regression trees) and linearly combines them so as to minimize a specified loss function (Elith et al. 2008). The first tree is fitted to the original data, and each subsequent tree is fitted to the residuals of the model based on the previous set of trees, attempting to fine-tune the model by focusing on the observations that are most difficult to predict. Since the response variable in this case is binary, we use the default Bernoulli deviance as the loss function (Ridgeway 2012). Boosting can be thought of as a form of functional gradient descent (Elith et al. 2008). Each tree is

fitted using only a randomly sampled specified percentage of the available data (default is 50%). This speeds the procedure and adds a random component that improves predictive performance.

Three parameters must be set in the BRT method. The learning rate/shrinkage, lr , is a value less than one that determines the contribution of each added tree. The smaller the lr , the less each successive tree contributes to the model and the more trees are required. Tree complexity, tc , also known as interaction depth, indicates the number of splitting nodes in each tree (typically, $tc \leq 5$). The maximum number of trees to be fitted, nt , is usually in the thousands. Since BRT can overfit data when too many trees are included, however, within each run, the actual number of trees included in the result is less than nt . Using the 10-fold cross validation option of *gbm*, the number of trees that minimizes the cross validation prediction error is identified and the associated results are used as the solution (Ridgeway 2012, Elith et al. 2008). To set the values of the three parameters, we fitted 45 models using all combinations of $lr = (0.01, 0.005, 0.001)$, $tc = (1, 3, 5)$, and $nt = (1000, 2500, 5000, 7500, 10000)$, and compared the resulting prediction errors (Elith et al. 2008). Based on this analysis, we used $lr = 0.01$, $tc = 5$, $nt = 7,500$.

BRT has the advantages that it can have better predictive performance at least in some cases, nonlinearities and interactions are modeled automatically, it is insensitive to outliers and inclusion of extra covariates, the relative importance of covariates are based on reduction in out-of-sample error and can be compared across numerical and categorical covariates (Elith et al. 2008).

5.4 Random Forests (RF)

Like BRT, random forests is a method that involves the combination of many trees (Breiman 2001). While in boosting, the trees evolve over time and are weighted, in random forests, which uses a modified version of bagging, the individual trees are identically distributed and equally weighted (Hastie et al. 2013). The idea of bagging is to reduce the variance in prediction by averaging many noisy but approximately unbiased trees. Random forests improves on the variance reduction by adding randomness in the trees so as to reduce the correlation between them and thus the overall variance, without increasing the variance in each one too much. Each tree is grown using a bootstrap sample of the data drawn with replacement. To add randomness and thus reduce correlation among the trees, at each splitting node, only a randomly selected subset of covariates are considered as candidate splitting variables. The process is repeated until a stopping criterion is reached, either the maximum number of nodes in the tree or the minimum number of observations in a terminal node. Since we have a binary response, we use classification trees, which means that in each terminal node of a tree, all observations are assigned the majority class if more than a specified percentage of the observations in the node are in the majority class (default is 50%; in this analysis we use 2.76%, probability of majority class (damaged pipe) in the dataset); and they are assigned to the minority class otherwise (Hastie et al. 2013). Similarly, if more than the specified percentage of trees classify a new observation as majority class, it is classified as majority class; otherwise it is minority class.

Random forests requires setting three parameters. The most important is the number of covariates to randomly evaluate at each splitting node, *mtry*. A balance must be found between a smaller *mtry*, which decreases the correlation between trees and thereby improves prediction, and a larger *mtry*, which improves the strength of the individual trees (Breiman 2001). The number of trees, *ntree*, should be set to be large enough that the prediction error has stabilized. Since bootstrap samples are used to fit each tree, about a third of the data is omitted from each tree and can be used compute out of bag (OOB) errors along the way, as the trees are being added (Hastie et al. 2013). Adding more trees does not typically lead to overfitting (Hastie et al. 2013). To ensure that trees do not get too large, a stopping criterion must be set, either the maximum number of nodes in each tree, *maxnodes*, or the minimum number of observations permitted in a terminal node, *nodesize*. We first set *ntrees*=1000 because preliminary results showed that value was large enough for the OOB error to stabilize. To set the other two parameters, we fitted 16 models using all combinations of *mtry* = (2, 4, 6, 8) and *maxnodes* = (5, 10, 15, 20). Based on a comparison of the results, we set *mtry* = 4 and *maxnodes* = 10. Random forests is popular because the minimal error can typically be reached using fewer trees than a BRT, they tend not to overfit data, they require little tuning, and like BRT, they automatically include interactions and nonlinearities (Hastie et al. 2013).

Chapter 6

ANALYSIS

6.1 Balancing Data

The dataset is imbalanced, meaning that one class of y (damaged) is far less prevalent than the other (not damaged). Specifically, only 2,311 (2.76%) of 83,746 pipes in the dataset are damaged, creating a situation of intrinsic, relative (though not absolute) rarity (Weiss 2004). Although logistic regression works as long as there is no absolute rarity, and thus the small sample bias is avoided (Allison 2012), imbalanced data requires special care in machine learning analyses and in the choice of evaluation metrics to compare models. Imbalance can compromise performance of learning algorithms (He and Garcia 2009, Lopez et al. 2013, and Chawla 2010). In trees, for example, successive partitioning of the dataset reduces the number of minority instances in each branch, which can result in concepts going unlearned.

To address the imbalance for the BRT and RF models, we apply three commonly used sampling methods (He and Garcia 2009)—random undersampling, random oversampling, and the Synthetic Minority Oversampling Technique (SMOTE). In undersampling, we randomly remove observations from the majority class (no damage). In oversampling, we randomly sample from observations in the minority class (damage), then duplicate and add them to the dataset. In SMOTE, artificial minority class observations are created based on the k -nearest neighbors and added to the dataset. In all cases, we modified the dataset so that the number of observations with damage equals the number with no damage. Each method is widely used but has some potential drawbacks. Undersampling removes information about the majority class, potentially causing the algorithm to miss important concepts related to

the majority class. Oversampling can lead to overfitting, and SMOTE can create overgeneralization because it generalizes the minority class without consideration of the majority class leading to overlapping of the two (He and Garcia 2009). To correct for the bias introduced by these sampling methods (Weiss and Provost 2003), the damage probabilities they produce, p' , were adjusted using the following equation from Dal Pozzolo et al. (2015), $p = \alpha p' / (\alpha p' - p' + 1)$, where p is the final bias-corrected $P(y = 1)$ and α is the ratio of the numbers of observations in the majority class to the number in the minority class in the original dataset.

6.2 Evaluation Methods

In evaluating and comparing the models, we are most concerned with their ability to predict damage for future earthquakes, so we used ten-fold cross validation and holdout validation to assess their out-of-sample predictive power. Cross validation (CV) was conducted by partitioning the dataset into ten randomly sampled folds. For each fold, the 90% of observations not in the fold made up a training set used to fit the models, which were then applied to predict the values for each of the 10% of observations in the fold—the validation set. Note that any sampling to address imbalance (Section 6.1) is done within the cross validation, that is, after a fold has been held out, to ensure observations in a fold are not also in the training set (Altini 2015). To minimize the effect of the fold sampling, we repeated the cross-validation 50 times, each with a different set of randomly generated folds and averaged the resulting 50 estimates of each error metric. In addition, the models were tested using holdout data from the June 13, 2011 earthquake.

The many metrics available to evaluate models of binary response variables can be categorized into three groups—probability, threshold, and ranking (Caruana and Niculescu-Mizil 2004, Fielding and Bell 1997, Hossin and Sulaiman 2015). Probability metrics directly use the predictions and interpret them as probabilities of damage, threshold metrics compare the predictions to a threshold so as to classify each response as positive or negative (damage or no damage in this case), and ranking metrics depend only on the ordering of the predictions, not the actual values. The various metrics measure different aspects of the fit so that a model may perform well on one but not another. We use ten metrics to capture models' ability to achieve the four stated goals—total count, spatial distribution of damaged pipes, classification of individual pipes, and relative importance of covariates (Table 6.1). Absolute and percentage error in the expected total number of damaged pipes for the region (*TE* and *TEP*, respectively) directly address the first goal. Square root of the mean squared error (*RMSE*) and mean absolute error (*MAE*) are also probability-based metrics, but at the individual pipe level, with the former especially penalizing larger errors over multiple small ones. None of the first four metrics distinguishes between false positive and false negative errors. The percentage of damaged pipes correctly classified (sensitivity, *SN*); percentage of undamaged pipes correctly classified (specificity, *SP*) do. To compute them, one sets a threshold; classifies each pipe as damaged if the predicted probability is greater than the threshold and not damaged if it is less than the threshold (we use the damage prevalence in the original dataset, 2.76%); and based on those binary classifications, creates a confusion matrix (Table 6.2).

Like all threshold metrics, *SN* and *SP* are unaffected by how close a prediction is to a threshold, only if it is above or below it. True skill statistic (*TSS*) is a newer threshold metric that unlike *SN* and *SP*, makes full use of the information in the confusion matrix and corrects the overall accuracy by the accuracy expected to occur by chance, and unlike the similar Cohen's kappa, does not depend on prevalence (Allouche et al. 2006). *TSS*, which ranges from -1 to +1 and is also known as the Hanssen-Kuipers discriminant, "compares the number of correct forecasts, minus those attributable to random guessing, to that of a hypothetical set of perfect forecasts," (Allouche et al. 2006, p1226) so that +1 indicates perfect agreement and zero means no better than chance. Threshold-based metrics are necessary if one wants to be able to classify each pipe, for example to create system-wide damage maps to simulate service outages. However, they fail to use all the information in the predictions and obviously are dependent on the choice of threshold.

The area under the receiver operating characteristic (ROC) plot (AUC) is threshold- and prevalence-independent. An ROC plot is a graph of *SN* vs. (1-*SP*), where each point corresponds to a possible threshold value (Fielding and Bell 1997). AUC is a ranking metric so it assesses the correctness of the ordering of the probabilities, but does not distinguish if they range from 0 to 1, for example, or from 0.40 to 0.42. AUC values range from 0.5 to 1 (worthless to perfect). The AUC can also be interpreted as the probability that the model will rank a randomly selected damaged pipe as more likely to be damaged than a randomly selected undamaged pipe. A potential concern in interpreting the AUC is that while it treats all thresholds equally, they may not all be of practical significance. To assess the ability of the models to predict the spatial distribution of damage correctly, we include two metrics

applied at the suburb level. The city of Christchurch is partitioned into 105 suburbs (i.e., neighborhoods). We compute the absolute error and percentage error for each suburb, and report the median value over all suburbs (SE and SEP).

Table 6.1: Evaluation metrics

Goal	Metric	Equation ^a	Type ^b
Prediction of total damage count	Absolute error in total count	$TE = \sum_i^n y_i - \sum_i^n \hat{p}_i$	Probability
	Percentage error in total count	$TEP = TE / \sum_i^n y_i$	
Prediction of individual pipe state	Root mean squared error	$RMSE = \sqrt{\frac{1}{n} \sum_i^n (y_i - \hat{p}_i)^2}$	Probability
	Mean absolute error	$MAE = \frac{1}{n} \sum_i^n y_i - \hat{p}_i $	
	Sensitivity	$SN = tp / (tp + fn)$	Threshold
	Specificity	$SP = tn / (tn + fp)$	
	True skill statistic	$TSS = SN + SP - 1$	
	Area under the ROC plot	AUC =area under ROC plot	Ranking
Prediction of spatial distribution of damage	Median absolute suburb error	Median over all suburbs of: $SE = \sum_{i \in n_S} y_i - \sum_{i \in n_S} \hat{p}_i$	Probability
	Median percentage suburb error	Median over all suburbs of: $SEP = SE / \sum_{i \in n_S} y_i$	

^a y_i = observed response for pipe i (0/1 if undamaged/damaged), \hat{p}_i = estimated probability of damage for pipe i , n =number of pipes, tp =true positive, fp =false positive, tn =true negative, fn =false negative, and n_S =set of pipes in suburb S .

^bLower values are preferred for probability metrics; higher values are preferred for threshold and ranking metrics.

Table 6.2: Example confusion matrix

		Predicted	
		Negative	Positive
Actual	Negative	True negative (<i>tn</i>)	False positive (<i>fp</i>)
	Positive	False negative (<i>fn</i>)	True positive (<i>tp</i>)

Chapter 7

RESULTS

7.1 Balancing Methods

Using the cross validation and the ten metrics defined in Table 6.1, we compare the three balancing methods—undersampling, oversampling, and SMOTE—for both boosted regression trees (BRT), and random forests (RF) (Table 7.1). For the BRT models, the best sampling method depends on which metric is used, but overall, the BRT with the original data performs at least as well as the others. For the random forests models, although the model using the original data has the smallest predictive errors in terms of RMSE and MAE (0.166 and 0.028, respectively), the sensitivity and specificity make it clear that in that case, that model actually predicts no damage for every pipe and the RMSE and MAE are only small because damage is relatively infrequent, i.e., because of the imbalance in the data. All three balancing methods address this anticipated difficulty, with SMOTE performing best across all metrics. Based on this comparison, for the remaining analyses, we use the BRT model with the original data and the RF model with SMOTE.

Table 7.1: Comparison of predictive performance for different sampling methods using cross validation

Model ^a	Total city		Individual pipe						Suburb	
	TE ^b	TPE	RMSE	MAE	SN	SP	TSS	AUC	MASE	MPSE
BRT	42	1.8%	0.153	0.046	0.782	0.813	0.575	0.882	2.3	21%
BRT Under.	-116	5.0%	0.155	0.049	0.815	0.778	0.594	0.877	2.6	31%
BRT Over.	554	24.0%	0.154	0.042	0.752	0.845	0.597	0.886	2.0	29%
BRT Smote	1,568	67.8%	0.160	0.034	0.500	0.939	0.439	0.877	4.3	68%
RF	2,301	99.5%	0.166	0.028	0.000	1.000	0.000	0.577	9.0	99%
RF Under.	-5,448	236%	0.244	0.095	0.871	0.685	0.556	0.857	22.3	210%
RF Over.	-6,891	298%	0.275	0.109	0.877	0.672	0.549	0.853	32.2	249%
RF Smote	-912	39.4%	0.182	0.056	0.697	0.814	0.511	0.831	3.6	55%

^aUnder = undersampling, Over = oversampling

^bTable 6.1 provides the definition for each metric. Lower values are preferred for TE, TPE, RMSE, MAE, MASE, and MPSE; higher values are preferred for SN, SP, TSS, and AUC.

7.2 Model Results

In this section, we compare the four model types—repair rate (RR), logit, boosted regression trees (BRT), and random forests with SMOTE (RF Smote)—considering results from both the cross validation using the February earthquake data, and the holdout validation using the June earthquake data (Table 7.2). Both the cross validation (CV) and holdout validation provide estimates of out-of-sample prediction errors since in both cases, the observations used to fit a model is not used to test it. However, cross validation best estimates the expected prediction error over all training sets, rather than for a specific training set. That is, it is best to compare the methods rather than specific models with specific coefficient values. Holdout validation imagines we use the specific models fitted with the February training set and apply

them for a future earthquake (Hastie et al. 2013, Ch. 7). It can provide a better estimate of prediction error for those specific models, but the estimates can have high variance, so the results depend a lot on the particular future earthquake used as the test dataset. We evaluate, in turn, predictive performance based on (1) errors at the total city level, (2) individual pipe classification, and (3) suburb level, in turn.

Table 7.2: Comparison of predictive performance for four model types using cross validation with the February earthquake data and June earthquake holdout validation

	Model	Total city		Individual pipe						Suburb	
		TE ^a	TPE	RMSE	MAE	SN	SP	TSS	AUC	MASE	MPSE
Feb. CV	RR	654	28.0%	0.162	0.045	0.603	0.782	0.385	0.777	4.0	57%
	Logit	0	0.0%	0.157	0.049	0.777	0.737	0.515	0.836	3.2	35%
	BRT	42	1.8%	0.153	0.046	0.782	0.813	0.575	0.882	2.3	21%
	RF										
	Smote	-912	39.4%	0.182	0.056	0.697	0.814	0.511	0.831	3.6	55%
June hold- out	RR	383	35.7%	0.109	0.020	0.304	0.928	0.231	0.796	2.5	53%
	Logit	-836	78.0%	0.112	0.032	0.787	0.616	0.402	0.780	3.7	72%
	BRT	-471	43.9%	0.112	0.028	0.689	0.777	0.466	0.812	2.0	44%
	RF										
	Smote	-1413	132%	0.129	0.039	0.613	0.800	0.413	0.785	3.8	81%

^aTable 6.1 provides the definition for each metric. Lower values are preferred for TE, TPE, RMSE, MAE, MASE, and MPSE; higher values are preferred for SN, SP, TSS, and AUC.

7.2.1 Total city level prediction

Considering the February CV results, the logit and BRT models produce virtually no error (TPE=0% and 2%, respectively), suggesting that they are better model types in general than RR and RF Smote (Table 7.2). For the June test data in particular, however, the RR does best (underpredicting by 36%), with the BRT producing similar error (overpredicting by 44%), much higher than for the CV. It is difficult to know how representative the June test data is of data from other future

earthquakes the models might be applied to. The ground motions were lower overall than those in the February earthquake (Table 4.1). There also may be some peculiarities of the June earthquake that are not captured in the covariate data, since it followed so closely after two other major events (September 2010 and February 2011).

To the extent that it is representative, the June holdout results suggest the magnitude of prediction error in the total number of damaged pipes for the city. Both RR and BRT suggest on the order of 40% error, which should be acknowledge when applying these models for prediction in the future. RF Smote is substantially worse in both cases, with TPE=39% and 132%, respectively, for CV and June holdout.

7.2.2 Individual pipe classification

If the goal of the analysis is to correctly classify individual pipes as damaged or not, the sensitivity (SN), specificity (SP), and true skill statistics (TSS) suggest that BRT is the best of the four models and RR is the worst (Table 7.2). For the June holdout results, for example, only 30% of damaged pipes would be correctly identified as such by the RR, whereas, 79% or 69% would by the Logit and BRT, respectively. There is a tradeoff between false positives and false negatives. Overall, the TSS is much lower for RR (0.231) than Logit (0.402) or BRT (0.466). While the AUC values are similar (0.78 to 0.881 across all models for the holdout), that is a bit misleading because while the AUC treats all thresholds equally (Section 6.2), in this application, only the low thresholds (say, <0.03) are of practical importance. While the SN and SP across models and analyses are similar above thresholds of 0.2, they are quite different at the lower thresholds of interest. The RR method is not able to differentiate individual pipes as much as the other models which include multiple covariates and the variability of PGV within the contours. As a result, the predicted damage

probabilities from RR for the June dataset only have a range of 0 to 0.49; whereas, those for the other three methods have a range of 0 to 1. The RMSE and MAE are similar across models, but that is largely due to the low prevalence of damaged pipes. In the extreme, an obviously useless model that predicts no damage for any pipe would get an MAE of 0.028 for the February CV and 0.012 for the June holdout. To simulate system-wide damage maps that could be used as input to simulations of customer outages, one needs to classify each pipe as damaged or not. For that type of analysis, the results suggest a model based on individual pipes would be more useful than the RR. In particular, the BRT performs best in terms of classifying individual pipes for both the CV and holdout analyses.

7.2.3 Suburb-level prediction

Between the scale of the entire city and the scale of individual pipes is the suburb-level analysis. The median absolute suburb error (MASE) and median percentage suburb error (MPSE) provide summary metrics of a model's ability to predict the approximate spatial distribution of damage for purposes of identifying neighborhoods likely to experience more or less damage. Based on those metrics, the BRT performs best for both the February CV and June holdout validations (MASE=2.29 and 2.01, and MPSE=21% and 44%, respectively) (Table 7.2). In other words, for the BRT, half of the suburbs have an error in the predicted number of damaged pipes in the June earthquake of 2 or less (MASE=2.01). The number of damaged pipes varies across suburbs from zero to 156 for the February earthquake and 64 for the June earthquake, so in terms of percentage error, the MASE=2.01 corresponds to a median 44% error.

To examine the performance in terms of spatial distribution more carefully, Figures 7.1 and 7.2 show the maps and calibration plots (predicted vs. observed number of damaged pipes) for the February CV results. They show the ability of the logit and BRT in particular to capture the approximate spatial distribution of damage. The RF smote model overpredicts damage in some suburbs by quite a lot. The pattern associated with the RR model follows the PGV contours. It is unable to distinguish the suburbs with the highest number of damaged pipes.

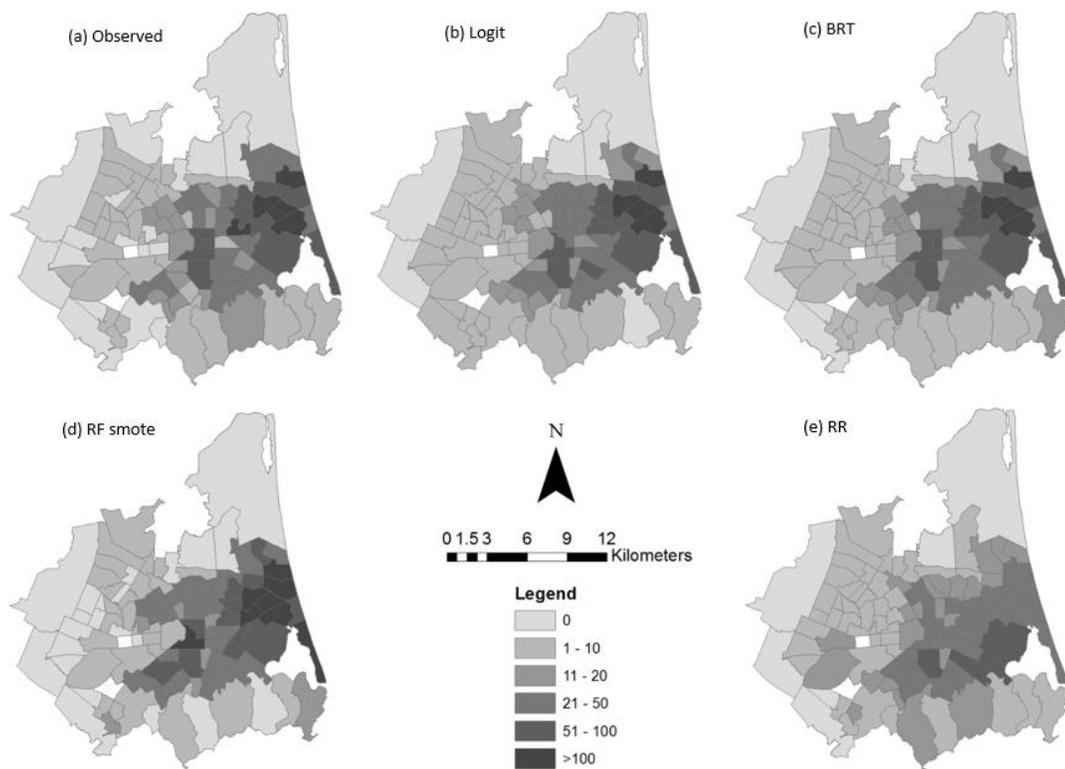


Figure 7.1: Number of damaged pipes in each suburb for February earthquake, (a) Observed and in each model: (b) Logit, (c) BRT, (d) RF smote, and (e) RR

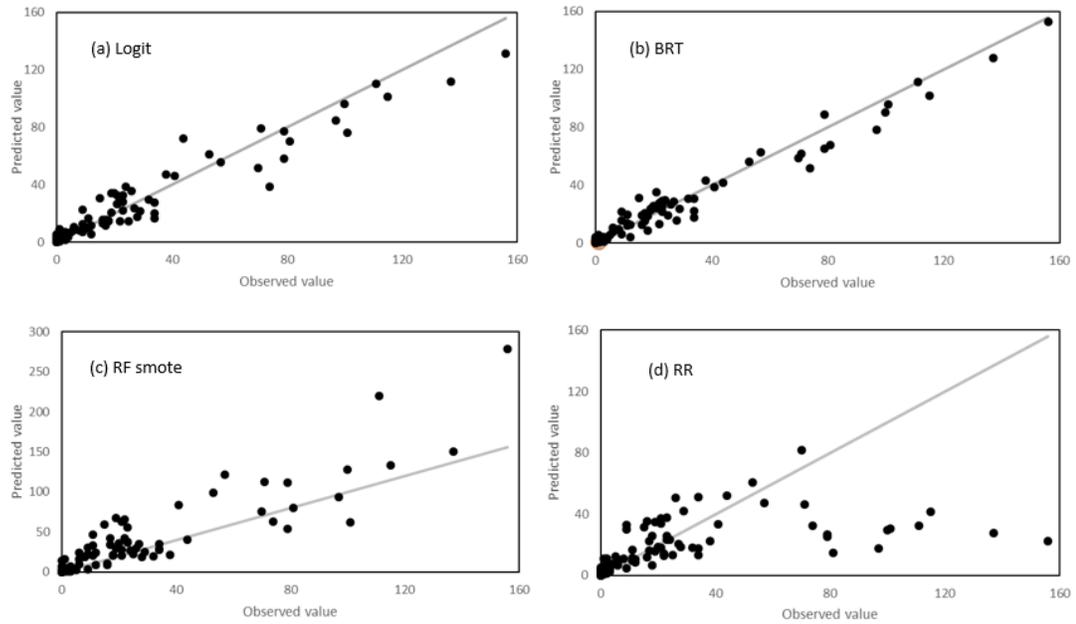


Figure 7.2: Predicted vs. observed number of damaged pipes in each suburb for February earthquake, for each model: (a) Logit, (b) BRT, (c) RF smote (note y-axis range is different), and (d) RR

Figures 7.3 and 7.4 show the maps and calibration plots for the June hold-out predictions. Logit and BRT models predict similar spatial distributions. Both overpredicts damage in Eastern Christchurch. RF smote overpredicts damage to a larger extent than Logit and BRT. RR underestimates the extent of damage. There does not seem to be a trend between observed and predicted damage.

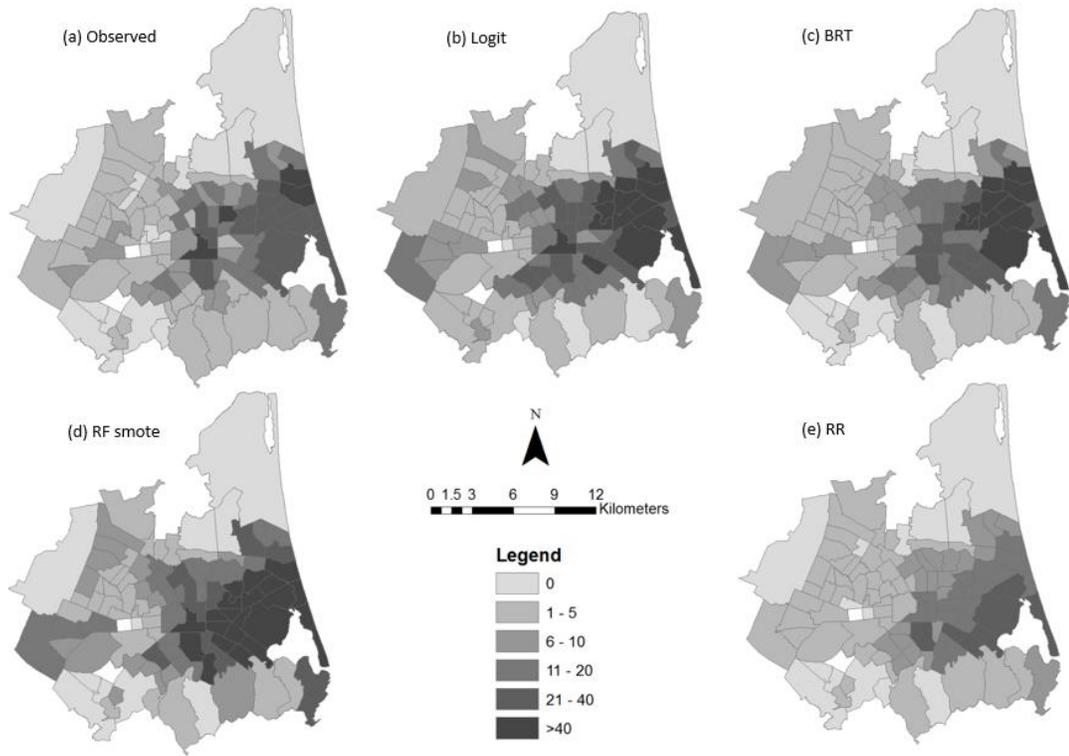


Figure 7.3: Number of damaged pipes in each suburb for June earthquake, (a) Observed and in each model: (b) Logit, (c) BRT, (d) RF smote, and (e) RR

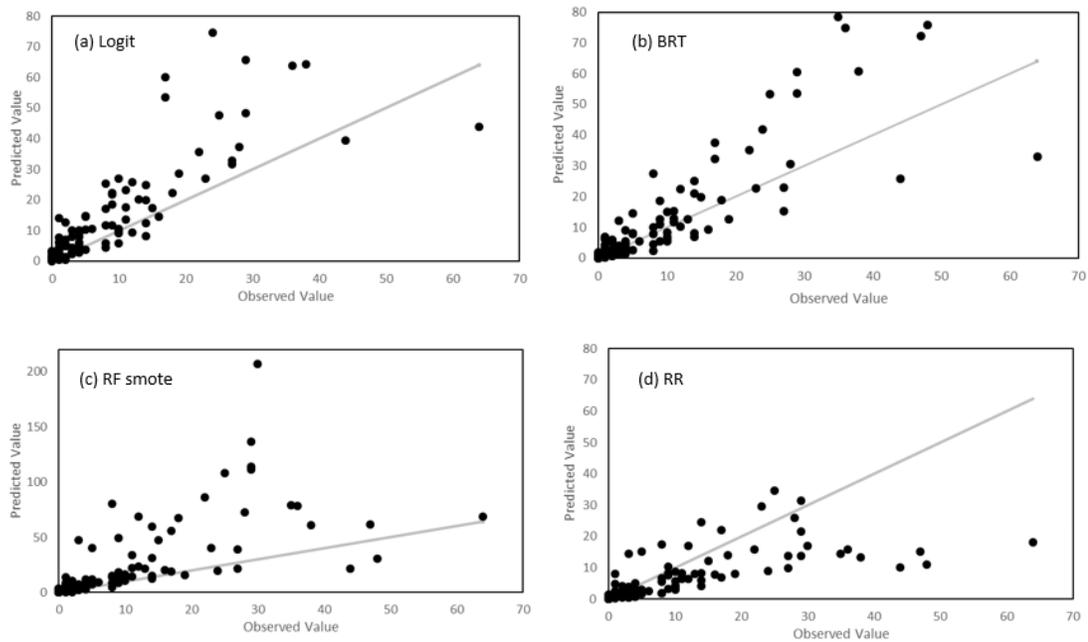


Figure 7.4: Predicted vs. observed number of damaged pipes in each suburb for June earthquake, for each model: (a) Logit, (b) BRT, (c) RF smote (note y-axis range is different), and (d) RR

7.3 Importance of Covariates

In examining the importance of the covariates, we seek to: (1) assess the potential effects of possible mitigation strategies, (2) clarify lingering ambiguity from past research about which pipe attributes are really influential, and (3) examine the tradeoff between predictive power and model simplicity (i.e., number of covariates included). While the RR method cannot provide insight on the relative importance of different covariates, each of the other methods can. This section focuses on the covariate results from the Logit and BRT models. Covariate results from RF smote are

not considered because it did not seem to have potential based on the predictive performance results in Section 7.2. Here we focus on the Logit model results in particular because unlike BRT and RF, they allow examination of the influence of different specific pipe materials, pipe types, and trench types (as opposed to pipe material as a whole).

The coefficients of all covariates in the full Logit model are statistically significant, suggesting that all the attributes of the pipes and ground motion considered contribute to the probability of earthquake damage (Table 7.3, full model). However, changing the pipe material or trench type are the most obvious potential mitigation strategies for reducing damage. For the Logit model, the consequence of making such changes can be determined explicitly by examining the marginal effects. To compute the marginal effect for each categorical covariate, we first compute the change in probability of damage for pipe i resulting from a change from the reference level of x_k to the level of interest of x_k , keeping all other covariates at their original values (Eq. 7.1). We then aggregating over all pipes according to the method of sample enumeration, weighting the marginal effect of each pipe i by the original probability of damage for pipe i (Eq. 7.2). As discussed in Louviere et al. (2000, p.60), this method is preferable to evaluating Eq. 7.1 using the sample average values of each covariate since the model is nonlinear. Marginal effects for the continuous covariates are computed similarly, but by incrementing the covariate instead of changing from the reference level to the covariate level of interest in Eq. 7.1

$$m_{ik} = P_i(y = 1|x, x_k = 1) - P_i(y = 1|x, x_k = 0) \quad (1.3)$$

$$M_k = \frac{\sum p_i m_{ik}}{\sum p_i} \quad (1.4)$$

The rightmost column of Table 7.3 presents the marginal effects for the statistically significant covariates in the full Logit model. It suggests, for example, that changing a pipe from galvanized iron, GI, the reference pipe material, to MPVC reduces the probability of damage by 0.1, all else being equal. Changing the trench type from pre-1984 local, the reference level, to post-2005, AP20 does the same.

Table 7.3: Logit results for full and reduced models

Covariate		Coefficients					Marginal effect ^d
		R1 ^b	R2	R3	R4	Full	Full
	Intercept	-5.22* ^c	-4.70*	-2.39*	-2.26*	8.79 [†]	
	PGV	0.024*	0.021*	0.014*	0.015*	0.013*	1.07(10 ⁻³)
	LRI			-0.677*	-0.662*	-0.696*	55.82(10 ⁻³)
	GWT					-0.062 [‡]	-4.95(10 ⁻³)
	Pipe length	0.012*	0.013*	0.014*	0.015*	0.013*	1.05(10 ⁻³)
	Diameter				-0.008*	-0.005*	-0.38(10 ⁻³)
	Year laid					-0.005 [‡]	-0.41(10 ⁻³)
Pipe material ^a	AC		-0.669*	-0.522*	0.273 [‡]	-0.252	
	CI		-0.885*	-0.885*	-0.123	-0.714 [†]	-0.058
	DI		-1.81*	-1.74*	-0.431	-0.568	
	STEEL		-0.777*	-0.739*	0.316	-0.313	
	CLS		-0.937*	-1.05*	0.263	-0.387	
	UPVC		-2.24*	-2.22*	-1.49*	-1.36*	-0.091
	MPVC		-2.80*	-2.66*	-1.84*	-1.64*	-0.101
	PE80B		-2.01*	-2.12*	-1.91*	-1.19*	-0.084
	PE100		-14.1	-14.1	-13.2	-13.0	
HDPE		-1.49*	-1.48*	-1.40*	-0.868*	-0.067	
LDPE		-13.3	-12.2	-12.2	-12.0		
Pipe type ^a	Trunk					-0.700	
	Submain					-0.072	
	Crossover					-1.34*	-0.075
Trench type ^a	Pre-1984, import					-0.575*	-0.052
	1984-2000, AP40					-0.955*	-0.077
	Post-2005, AP 20					-1.31*	-0.095

^aReference level for pipe material is galvanized iron, for pipe type is main, and for trench type is pre-1984, local.

^bModel R1 includes PGV and pipe length only. Model R2 includes PGV, pipe length, and pipe material. Model R3 includes PGV, pipe length, pipe material, and LRI. Model R4 includes PGV, pipe length, pipe material, LRI, and diameter. Full model includes all covariates.

^{c†} Indicates significance at 0.05, [‡] Indicates significance at 0.01, * Indicates significance at 0.0001

^dMarginal effect on the change in probability of damage due to a unit change, for significant covariates only.

The coefficient signs in the full Logit model (Table 7.3) support the hypotheses in Table 3.1. They suggest that more damage is associated with higher PGV, lower LRI (i.e., increased liquefaction susceptibility), smaller pipe diameter, and older pipes. They also offer new insights related to pipe material, pipe type, and trench type. The results provide evidence that MPVC, UPVC, PE80B, HDPE, and CI were associated with the least damage, and GI with the most. They indicate that crossovers and pipes in the more recent trench types are less likely to be damaged, all else being equal.

As noted in Section 3.3, while the influences of different pipe attributes on damage have been studied previously, in many cases the studies have not controlled for other attributes, making it unclear whether the influence is really due to the attribute of interest or a correlate of it. The statistical significance of the coefficients of all the covariates in the full Logit model suggests that each is important, even when controlling for the others. To further examine their effects, we compare the full model to the four reduced models, which are the same except that they include only a subset of the covariates. Comparing Models R3 and R4, for example, shows that when pipe diameter is added, the covariate for steel changes from being significant with a negative coefficient ($\beta_S = -0.739, p=0.0003$) to not significant ($\beta_S=0.316, p=1452$) (Table 7.3). This suggests that in the model R3, the steel covariate may actually be reflecting the effect of pipe diameter, since steel pipes tend to be large. The average diameter of steel pipes is 190 mm, compared to 83 mm for pipes of all other materials. A model like R2 or R3, therefore, might lead to a conclusion that steel is associated with reduced damage, whereas R3 suggests that when controlling for pipe diameter, there is no evidence that is true. Similarly, comparing Models R2 and R3 indicates that when liquefaction susceptibility is not controlled for, the PGV covariate captures some of the variability it causes. These results highlight the importance of considering what set of covariates are included in a model when interpreting the results for a specific covariate.

Beyond the statistical significance of the covariates, it is worth considering their practical significance since while additional covariates improve prediction, they also increase the data required to apply the model. To investigate this tradeoff between improved prediction and increased complexity and data needs, we examined the

predictive power of the four reduced models to the full models (Table 7.4). These results suggest first, that in terms of error in the total count, the PGV and pipe length alone provide the same level of predictive ability that the full set of covariates does (R1 vs. Full). In terms of prediction at the individual pipe and suburb levels, while the overall performance improves with additional covariates as expected, the marginal improvement declines after the first few most influential covariates. In particular, the R3 models, which include only PGV, pipe length, pipe material, and LRI achieve most of predictive power of the full model. For the Logits, for example, the TSS, AUC, and MASE are 4%, 14%, and 10% worse for the R3 model than the full model. Depending on the intended uses of the model and data availability, therefore, one might choose to use one of the reduced models instead of the full model.

Table 7.4: Predictive performance of reduced models compared to full models

Model ^a	Total city		Individual pipe						Suburb	
	TE ^b	TPE	RMSE	MAE	SN	SP	TSS	AUC	MASE	MPSE
Logit R1	0	0.0%	0.163	0.053	0.621	0.687	0.307	0.685	7.5	65%
Logit R2	0	0.0%	0.162	0.052	0.740	0.687	0.427	0.689	6.6	52%
Logit R3	0	0.0%	0.159	0.050	0.771	0.721	0.492	0.722	3.6	33%
Logit R4	0	0.0%	0.159	0.050	0.765	0.729	0.494	0.730	3.7	39%
Logit Full	0	0.0%	0.157	0.049	0.777	0.737	0.515	0.836	3.2	35%
BRT R1	22	1.0%	0.160	0.051	0.808	0.696	0.504	0.699	5.1	53%
BRT R2	22	0.9%	0.157	0.049	0.788	0.764	0.552	0.765	4.5	44%
BRT R3	20	0.9%	0.155	0.047	0.786	0.789	0.576	0.789	3.0	29%
BRT R4	24	1.0%	0.155	0.047	0.790	0.794	0.583	0.794	3.0	32%
BRT Full	42	1.8%	0.153	0.046	0.782	0.813	0.575	0.882	2.3	21%

^aModel R1 includes PGV and pipe length only. Model R2 includes PGV, pipe length, and pipe material. Model R3 includes PGV, pipe length, pipe material, and LRI. Model R4 includes PGV, pipe length, pipe material, LRI, and diameter. Full model includes all covariates.

^bTable 6.1 provides the definition for each metric. Lower values are preferred for TE, TPE, RMSE, MAE, MASE, and MPSE; higher values are preferred for SN, SP, TSS, and AUC.

Chapter 8

CONCLUSIONS

This paper presents a thorough analysis of earthquake damage to water pipelines using a large database of damage in the February and June 2011 earthquakes in Christchurch, New Zealand. A comparison of four model types using the results of both cross validation and holdout validation suggests first, that the recommended model depends on the intended use and that multiple metrics of predictive power are required to evaluate them fully. Specifically, the results suggest that boosted regression trees (BRT) arguably offer the best overall predictive performance. However, although logit models can overfit the data a bit more than BRT, they are quite good as well, and offer the added benefits of an easily shared closed form model representation and an ability to provide insight into the relative effect of different pipe materials and trench types on damage probabilities. The repair rate (RR) method used almost exclusively in previous studies performs very well in terms of predicting the total number of damaged pipes in the city and is far simpler to apply, but is not as capable of correctly classifying individual pipes as damaged or not, or capturing the spatial distribution of damage. Whichever model type is used, the analysis highlights the need to compute and report the predictive errors of different types and acknowledge them in using the models for subsequent analysis.

Previous studies had identified pipe material and other attributes as being important in determining damage in earthquakes. Due to the type of models, however, questions remained about the relative importance of related attributes. The analysis provides evidence that MPVC, UPVC, PE80B, HDPE, and CI were associated with the least damage, and GI with the most. It further suggests how much the damage is

likely to be reduced if one material is replaced with another. Simulating the replacement of pipes of one material with another and rerunning one of the damage models could provide insight for planning a retrofit program or new construction guidelines. Similarly, the analysis provides evidence that pipes in the more recent trench types are less likely to be damaged, even when controlling for the pipe age. The results also caution against analyses that offer conclusions about the effect of a pipe attribute without controlling for the other attributes. Finally, while all of the pipe and ground attributes examined based on the conceptual framework were statistically significant, a comparison of reduced models suggests that a majority of the predictive power can be obtained by including only the pipe length, PGV, LRI, and pipe material, so if it is necessary to limit the data demands of applying a model for prediction, one might focus on those.

REFERENCES

- Abdoun, T., Ha, D., O'Rourke, M., Symans, M., O'Rourke, T., Palmer, M., and Stewart, H. (2008). "Factors Influencing the Behavior of Buried Pipelines Subjected to Earthquake Faulting", *Soil Dynamics and Earthquake Engineering*, 29, 415-427.
- Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, Wiley, Hoboken, New Jersey, USA.
- Allison, P. (2012). "Logistic Regression for Rare Events", *Statistical Horizons*.
- Allouche, O., Tsoar, A., and Kadmon, R. (2006). "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa, and the True Skill Statistic (TSS)", *Journal of Applied Ecology*, 43, 1223-1232.
- Altini, M. (2015). "Dealing with Imbalanced Data: Undersampling, Oversampling, and Proper Cross-validation".
- American Lifelines Alliance (ALA) (2001). "Seismic Fragility Formulations for Water Systems", *American Society of Civil Engineers (ASCE) and Federal Emergency Management Agency (FEMA)*.
- Anderson, D., Davidson, R. A., Himoto, K., and Scawthorn, C. (2016). "Statistical Modeling of Fire Occurrence Using Data from the Tohoku, Japan Earthquake and Tsunami", *Risk Analysis*, 36(2), 378-395.
- Bouziou, D. and O'Rourke, T. (2015). "Water Distribution System Response to the 22 February 2011 Christchurch Earthquake", *6th International Conference on Earthquake Geotechnical Engineering*.
- Bradley, B. and Hughes, M. (2012). "Conditional Peak Ground Accelerations in the Canterbury Earthquakes for Conventional Liquefaction Assessment," Part 1 and Part 2, *Technical Reports Prepared for the Department of Building and Housing*.
- Breiman, L. (2001). "Random Forests", *Machine Learning*, 45, 5-32.
- Caruana, R., and Niculescu-Mizil, A. "Data mining in metric space: an empirical analysis of supervised learning performance criteria", *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.

- Chawla, N. V. (2010). "Data Mining for Imbalanced Datasets: An Overview", *Data Mining and Knowledge Discovery Handbook*, 2, 875-886.
- Chen, W., Shih, B., Chen, Y., Hung, J., and Hwang, H. (2002). "Seismic Response of Natural Gas and Water Pipelines in the Ji-Ji Earthquake", *Soil Dynamics and Earthquake Engineering*, 22, 1209-1214.
- Cubrinovski, M., Bradley, B., Wotherspoon, L., Green, R.A., Bray, J.D., Wood, C., Pender, M., Allen, J., Bradshaw, A., Rix, G., Taylor, M., Robinson, K., Henderson, D., Giorgini, S., Ma, K., Winkely, A., Zupan, J., O'Rourke, T., DePascale, G., and Wells, D. (2011). "Geotechnical Aspects of the 22 February 2011 Christchurch Earthquake." *Bulletin of the New Zealand Society for Earthquake Engineering*, 44(4), 205-226.
- Cubrinovski, M., Bray, J. D., Taylor, M., Giorgini, S., Bradley, B., Wotherspoon, L., and Zupan, J. (2011). "Soil Liquefaction Effects in the Central Business District during the February 2011 Christchurch Earthquake." *Seismological Research Letters*, 82(6), 893-904.
- Dal Pozzolo, A., Caelen, O., Johnson, R., and Bontempi, G. (2015). "Calibrating Probability with Undersampling for Unbalanced Classification".
- Davis, P., Burn, S., Moglia, M., and Gould, S. (2007). "A Physical Probabilistic Model to Predict Failure Rates in Buried PVC Pipelines", *Reliability Engineering and System Engineering*, 92, 1258-1266.
- Eguchi, R. (1991). "Seismic Hazard Input for Lifeline Systems", *Structural Safety*, 10, 193-198.
- Eidinger, J., and Avila, A., eds. 1999. *Guidelines for the seismic evaluation and upgrade of water transmission facilities, Monograph 15*, Technical Council on Lifeline Earthquake Engineering, American Society of Civil Engineers, Reston, VA.
- Eidinger, J., Maison, B., Lee, D., and Lau, B. (1995). "East Bay Municipal Utility District Water Distribution Damage in 1989 Loma Prieta Earthquake", *Proc., 4th U.S. Conference Lifeline Earthquake Engineering*, 240-247.
- Elith, J., Leathwick, J., and Hastie, T. (2008). "A Working Guide to Boosted Regression Trees", *Journal of Animal Ecology*, 77, 802-813.
- Fielding, A. and Bell, J. (1997). "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models", *Environmental Conservation*, 24(1), 38-49.

- Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.
- He, H. and Garcia, E. (2009). "Learning from Imbalanced Data", *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.
- Hernandez-Orallo, J., Flach, P., and Ferri, C. (2012). "A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss", *Journal of Machine Learning Research*, 13, 2813-2869.
- Heubach, W. (1995). "Seismic Damage Estimation for Buried Pipeline Systems", *Proc., 4th U.S. Conference Lifeline Earthquake Engineering*, 312-319.
- Honneger, D. (1995). "An Approach to Extend Seismic Vulnerability Relationships for Large Diameter Pipelines", *Proc., 4th U.S. Conference Lifeline Earthquake Engineering*, 320-327.
- Hossin, M. and Sulaiman, M. (2015). "A Review on Evaluation Metrics for Data Classification Evaluations", *International Journal of Data Mining and Knowledge Management Process (IJDKP)*, 5(2), 1-11.
- Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*, Wiley, New York, USA.
- Hwang, H. H., and Huo, J. R. (1998). "Seismic fragility analysis of electric substation equipment and structures." *Probabilistic Engineering Mechanics*, 13(2), 107-116.
- Isoyama, R., Ishida, E., Yune, K., and Shirozu, T. (2000). "Seismic Damage Estimation Procedure for Water Supply Pipelines", *Water Supply*, 18(3), 63-68.
- Jeon, S. and O'Rourke T. (2005). "Northridge Earthquake Effects on Pipelines and Residential Buildings", *Bulletin of the Seismological Society of America*, 95(1), 294-318.
- Katayama, T., Kubo, K., and Sato, N. (1975). "Earthquake Damage to Water and Gas Distribution Systems", *Proc., U.S. National Conference on Earthquake Engineering*, 396-405
- Kimishima, K., Maruyama, Y., and Yamakazi, F. (2011). "Damage Ratio of Water Pipes during the 2007 Niigata-Ken Chuetsu-Oki, Japan, Earthquake", *Proc., 12th East Asia-Pacific Conference on Structural Engineering and Construction*, 14, 212-219.

- Lanzano, G., Salzano, E., Santucci de Magistris, F., and Fabbrocino, G. (2014). "Seismic Vulnerability of Gas and Liquid Buried Pipelines", *Journal of Loss Prevention in the Process Industries*, 30, 1361-1376.
- Liaw, A. and Wiener, M. (2002). "Classification and Regression by randomForest", *R News*, 2/3, 18-22.
- Lombardo, L., Cama, M., Conoscenti, C., Marker, M. and Rotigliano, E. (2015). "Binary Logistic Regression Versus Stochastic Gradient Coosted Decision Trees in Assessing Landslide Events: Application to the 2009 Storm Event in Messina (Sicily, Southern Italy)", *Nat Hazards*, 79, 1621-1648.
- Lopez, V., Fernandez, A., Garcia, S., Palade, V., and Herrera, F. (2013). "An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on using Data Intrinsic Characterisctics", *Information Sciences*, 250, 113-141.
- Louviere, J.J., Hensher, D.A. and Swait, J.D. (2000). *Stated Choice Methods: Analysis and Applications*, Cambridge University Press.
- Maruyama, Y., Kimishima, K., and Yamazaki, F. (2011). "Damage Assessment of Buried Pipes due to the 2007 Niigata Chuetsu-oki Earthquake in Japan", *Journal of Earthquake Tsunami*, 5(1), 57-70.
- Maruyama, Y., Nagata, S., and Wakamatsu, K. (2015). "Damage Assessment of Water Distribution Pipelines after the 2011 off the Pacific Coast of Tohoku Earthquake", *Journal of Energy Challenges and Mechanics*, 2(4), 144-149.
- Maruyama, Y. and Yamazaki, F. (2010). "Damage Estimation of Water Distribution Pipes Following Recent Earthquake in Japan", *Proc., 7th International Conference on Urban Earthquake Engineering and 5th International Conference on Earthquake Engineering*, 1487-1491.
- Maruyama, Y., Yamazaki, F., Mizuno, K., Yogai, H., and Tsuchiya, Y. (2008). "Development of Fragility Curves for Highway Embankment Based on Damage Data from Recent Earthquakes in Japan", *Proc., 14th World Conference on Earthquake Engineering*.
- Milashuk, S. and Crane, W. (2012). "Pipeline Damage Caused by Transient Ground Deformation: Case Study of 2010 Darfield and 2011 Christchurch Earthquakes", *Soil Dynamics and Earthquake Engineering*, 41, 84-88.

- Nateghi, R., Guikema, S., and Quiring, S. (2013). "Power Outage Estimation for Tropical Cyclones: Improved Accuracy with Simpler Models", *Risk Analysis*, 1-10.
- O'Rourke, M. and Ayala, G. (1993). "Pipeline Damage Due to Wave Propagation", *Journal of Geotechnical Engineering*, 119(9), 1490-1498.
- O'Rourke, M. and Deyoe, E. (2004). "Seismic Damage to Segmented Buried Pipe", *Earthquake Spectra*, 20(4), 1167-1183.
- O'Rourke, M. and Liu, X. (2012). "Seismic Design of Buried and Offshore Pipelines", *MCEER Monograph MCEER-12-MN04*
- O'Rourke, T. and Jeon, S. (1999). "Factors Affecting the Earthquake Damage of Water Distribution Systems", *Post-Earthquake Lifeline System Reliability*, 16, 379-388.
- O'Rourke, T., Jeon, S., Toprak, S., and Cubrinovski, M. (2012). "Underground Lifeline System Performance during the Canterbury Earthquake Sequence", *Proc., 15th World Conference of Earthquake Engineering*.
- O'Rourke, T. D., Jeon, S. S., Toprak, S., Cubrinovski, M., Hughes, M., van Ballegooy, S., and Bouziou, D. (2014). "Earthquake response of underground pipeline networks in Christchurch, NZ." *Earthquake Spectra*, 30(1), 183-204.
- O'Rourke, T. and Toprak, S. (1997). "GIS Assessment of Water Supply Damage from the Northridge Earthquake", *Soil Dynamics and Earthquake Engineering*, 117-131.
- Pineda, O. and Najafi, M. (2010). "Seismic Damage Estimation for Buried Pipelines: Challenges after Three Decades of Progress", *Journal of Pipeline Systems Engineering and Practice*, 1(1), 19-24.
- Pineda, O. and Ordaz, M. (2003). "Seismic Vulnerability Function for High-Diameter Buried Pipelines: Mexico City's Primary Water System Case", *Pipelines*, 1145-1154.
- Pineda, O. and Ordaz, M. (2007). "A New Seismic Intensity Parameter to Estimate Damage in Buried Pipelines due to Seismic Wave Propagation", *Journal of Earthquake Engineering*, 11, 773-786.

- Pineda, O. and Ordaz, M. (2010). "Seismic Fragility Formulations for Segmented Buried Pipeline Systems Including the impact of Differential Ground Subsidence", *Journal of Pipeline Systems Engineering and Practice*, 1(4), 141-146.
- Prasad, A., Iverson, L., and Liaw, A. (2006). "Newer Classification and Regression Tree Techniques: bagging and Random Forests for Ecological Prediction", *Ecosystems*, 9, 181-199.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <<https://www.R-project.org/>>.
- Ridgeway, G. with contributions from others (2015). Gbm: Generalized Boosted Regression Models. R package version 2.1.1. <<http://CRAN.R-project.org/package=gbm>>
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351-357.
- Shirozu, T. and Isoyama, R. (1998). "Analysis of Damage to Water Mains Caused by the 1995 Hyogoken-nanbu (Kobe) Earthquake", *Water Supply*, 16(1-2), 443-454.
- Terzi, V., Alexoudi, M. and Hatzigogos, T. (2007). "Numerical Assessment of Damage State of Segmented Pipelines due to Permanent Ground Deformation", *Proc., 10th International Conference on Applications of Statistics and Probability in Civil Engineering*, 495-497.
- Tsai, J., Jou, L., and Lin, S. (2000). "Damage to Buried Water Supply Pipelines in the Chi-Chi (Taiwan) Earthquake and a Preliminary Evaluation of Seismic Resistance of Pipe Joints", *Journal of the Chinese Institute of Engineers*, 23(4), 395-408.
- Tromans, I. (2004). "Behaviour of Buried Water Supply Pipelines in Earthquake Zones", *Thesis University of London*.
- Vazouras, P., Karamanos, S., and Dakoulas, P. (2010). "Finite Element Analysis of Buried Steel Pipelines Under Strike-slip Fault Displacements", *Soil Dynamics and Earthquake Engineering*, 30, 1361-1376.
- Weiss, G. (2004). "Mining with Rarity: A Unifying Framework", *Sigkdd Explorations*, 6(1), 7-19.

- Weiss, G. and Provost, F. (2003). "Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction", *Journal of Artificial Intelligence Research*, 19, 315-354.
- Yang, R., Zhang, G., Liu, F., Lu, Y., Yang, F., Yang, F., Yang, M., Zhao, Y., and Li, D. (2016). "Comparison of Boosted Regression Tree and Random Forest Models for Mapping Topsoil Organic Carbon Concentration in an Alpine Ecosystem", *Ecological Indicators*, 60, 870-87.
- Youd, T. L., Member, C., ASCE, Idriss, I. M., Fellow, C.-C., Andrus, R. D., Arango, I., Castro, G., Christian, J. T., Dobry, R., Finn, W. D. L., Harder, Jr., L. F., Hynes, M. E., Ishihara, K., Koester, J. P., Liao, S. S. C., Marcuson. III, W. F., Martin, G. R., Mitchell, J. K., Moriwaki, Y., Power, M. S., Robertson, P. K., Seed, R. B., Stokoe, II, K. H. (2001). "Liquefaction Resistance of Soils: Summary Report from the 1996 NCEER and 1998 NCEER/NSF Workshops on Evaluation of Liquefaction Resistance of Soils", *Journal of Geotechnical and Geoenvironmental Engineering*, 127, 817-833.