Soil Dynamics and Earthquake Engineering 112 (2018) 76–88

Contents lists available at ScienceDirect

Soil Dynamics and Earthquake Engineering

journal homepage: www.elsevier.com/locate/soildyn

Comparison of statistical and machine learning approaches to modeling earthquake damage to water pipelines

Adam Bagriacik[a], Rachel A. Davidson[a,*], Matthew W. Hughes[b], Brendon A. Bradley[c,d], Misko Cubrinovski[c]

[a] Department of Civil and Environmental Engineering, University of Delaware, Newark, DE 19711, USA
[b] Departments of Civil and Natural Resources Engineering and Geological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[c] Department of Civil and Natural Resources Engineering, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[d] Department of Civil and Environmental Engineering, Stanford University, CA, USA

ARTICLE INFO

Keywords:
Buried pipelines
Water pipelines
Logistic regression
Machine learning
Damage analysis
Repair rate
Christchurch
Canterbury earthquake sequence

ABSTRACT

A large dataset of water pipeline damage from the February and June 2011 earthquakes in Christchurch, New Zealand is used to fit and compare four mathematical model types—logistic regression, boosted regression trees (BRT), random forest (RF), and the repair rate (RR) method common in the literature. Cross validation and holdout validation are employed 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.

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 amount and spatial distribution of damage future earthquakes are likely to cause, and what attributes of the pipes, ground motion, and ground conditions 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 (e.g., [1]). In this paper, we add to that body of knowledge by using damage data from the February 22, 2011 (M_w = 6.2) and June 13, 2011 ($M_w = 6.0$) earthquakes in Christchurch, New Zealand to fit and evaluate new mathematical models of earthquake-induced damage to pressurized water pipelines.

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 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 system-wide 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 and trench type, characteristics that might 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 comprehensive dataset from earthquakes that caused extensive damage to a modern water supply system. The dataset includes observations on approximately 84,000 pipes 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 explanatory variables and

---

* Corresponding author.

E-mail addresses: abagria@udel.edu (A. Bagriacik), rdavidso@udel.edu (R.A. Davidson), matthew.hughes@canterbury.ac.nz (M.W. Hughes), brendon.bradley@canterbury.ac.nz (B.A. Bradley), misko.cubrinovski@canterbury.ac.nz (M. Cubrinovski).

https://doi.org/10.1016/j.soildyn.2018.05.010
Received 18 June 2017; Received in revised form 25 April 2018; Accepted 7 May 2018
0267-7261/

Table 1
Summary of hypothesized effects of different factors on earthquake damage to pipes.

Factor	Hypothesized effect	Strength of evidence in the literature
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

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 variables refer more directly to a specified pipe rather than being smoothed over a region. We compare the model types to the simpler repair rate approach conventionally used in the literature. Third, 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 variables. After reviewing the empirical literature on models of earthquake damage to water pipelines in Section 2, we summarize previous findings on influential explanatory variables in Section 3. The data, model types, analyses, and results are described in Sections 4, 5, 6, and 7, respectively.

2. Available models of earthquake damage to water pipelines

Many analyses have been conducted to examine the performance of buried pipelines in earthquakes—Physical experiments (e.g., [2,3]), analytical (e.g., [4]), numerical (e.g., [5]), and empirical or statistical curve fitting (e.g., [6]). References [1,7–9] provide useful reviews of the literature. The focus here is on empirical models, i.e., mathematical relationships fitted to damage data recorded in previous earthquakes (e.g., [10–14]).

Maruyama et al. [15] fitted a logit model to water pipeline damage from the 2011 Tōhoku earthquake. The model used a "water-supply area" as the unit of study and considered four explanatory variables—PGV, length of pipes in the area, vulnerability factor for pipe material, and vulnerability factor for ground condition. In all other examples we found, empirical modeling efforts have 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* (number of repairs, i.e., damage locations, per km. of pipe) to a measure of ground motion, ground deformation, or strain. They often present different curves for different groups of pipe based on their material, diameter, or other characteristic [9]. 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 is computed (total number of repairs/length of pipe), producing a data pair of repair rate and ground motion level. The observations of paired data are plotted and a least squares line is fitted to them.

In this RR approach, each earthquake produces a relatively small number of observations (typically 5–25) [9]. Fitting separate curves for different pipe materials or other subsets of pipes can further reduce the number of observations. Many papers report a coefficient of determination, R^2, as an indication of goodness-of-fit, and they are typically relatively high (e.g., 0.98 in [17]). 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 [18, p. 339] explains, "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.

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 explanatory variables 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 (CV) and holdout validation.

3. Explanatory variables

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 earthquake- and pipe-related explanatory variables are presented in Sections 3.1 and 3.2, respectively, and Table 1 summarizes the resulting hypotheses for our analysis.

3.1. Earthquake-related variables

All previous empirical models have used some measure of earthquake-induced ground motion/deformation—dynamic ground shaking or permanent ground displacement—as the primary explanatory variable and found evidence that more intense ground motion is associated with more damage. Pineda-Porras and Najafi [7] identified at least nine ground motion metrics employed. MMI was used in the 1980s and 1990s (e.g., [19]) 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., [16]), and it is directly related to, but easier to compute than, ground strain, the main cause of damage [7]. More recently, geometric mean peak ground velocity (GMPGV) has been considered to account for the two horizontal components of ground motion [20].

In addition to transient ground motions, pipe damage can be caused by permanent ground deformation due to liquefaction, landslide, fault displacement, or settlement. All of the approximately one dozen studies that have investigated these variables have found evidence that an increase in permanent ground displacement or occurrence of liquefaction