8-6-2021

A probabilistic approach to levee overtopping risk assessment

Stefan G. Flynn
stefanflynn22@gmail.com

A probabilistic approach to levee overtopping risk assessment

By

Stefan G. Flynn

Approved by:

Farshid Vahedifard (Major Professor/Graduate Coordinator)
Ghada S. Ellithy
Isaac L. Howard
Jason M. Keith (Dean, Bagley College of Engineering)

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Civil Engineering
in the Richard A. Rula School of Civil and Environmental Engineering

Mississippi State, Mississippi

August 2021

Name: Stefan G. Flynn

Date of Degree: August 6, 2021

Institution: Mississippi State University

Major Field: Civil Engineering

Major Professor: Farshid Vahedifard

Title of Study: A probabilistic approach to levee overtopping risk assessment

Pages in Study: 109

Candidate for Master of Science

The most common mode of levee failure, breach due to overtopping, is generally considered as a function of a complex set of contributing factors. The goal of this research is to enhance the state of the art and practice for performing levee overtopping risk assessment. For this purpose, a dataset of levee overtopping event records within the portfolio of levee systems maintained by the U.S. Army Corps of Engineers (USACE) is presented. The dataset is utilized with logistic regression analysis to develop a probabilistic model to calculate system response probabilities and assess risk related to levee overtopping. The presented dataset can be used for identifying key factors controlling overtopping behavior, validation of model results, and providing new insight into the phenomenon of levee overtopping. The proposed model offers a practical yet robust tool for levee risk analysis and can be readily employed by engineers and other stakeholders.

DEDICATION

To my grandmother, Karen Flynn.

ACKNOWLEDGEMENTS

First, I would like to express my thanks to Dr. Farshid Vahedifard for his continuous support of my study and research, his guidance, and his dedication to the field of geotechnical engineering.

I would like to thank the members of my graduate committee, Dr. Ghada Ellithy and Dr. Isaac Howard for their time and effort related to my studies.

I would like to thank those that have contributed to the publications contained within this thesis, Dave Schaaf, Dr. Soroush Zamanian and Dr. Abdollah Shafieezadeh. Your efforts have been immensely helpful in my educational progress and practical knowledge of statistical analysis.

I thank the U.S. Army Corps of Engineers Rock Island District and the Risk Management Center (RMC) for supporting my studies. My team at Rock Island has been fully supportive of my endeavors while balancing a demanding workload outside of the classroom. The RMC has provided expertise and the data required for this research.

I would like to thank my mother, Shellie Wear, for the example she has set and acknowledge the years of work she has put in to get me to this point. Thank you for being the vision of hard work and perseverance. The world is a better place because of you.

I would like to thank my grandfather, Gary Flynn. I wouldn't be in this profession without you. I'm thankful for the bond we have and your support throughout my life.

Finally, this thesis, and my graduate studies in general, would not be possible without the support of my fiancée, Karrah. You have been my rock throughout this process, supporting and encouraging me every step of the way. It's been a long road, and I will forever be grateful. I appreciate you, and I love you beyond measure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

Understanding and quantifying the risk posed by levees under extreme hydraulic loading is a critical task for engineers and decision makers. Federal agencies such as U.S. Army Corps of Engineers (USACE) and the U.S. Bureau of Reclamation require the assessment of risk associated with flood risk management structures for the purposes of planning, design and construction (Reclamation-USACE 2015). According to the National Levee Database (NLD), levees managed by USACE protect more than 13 million people and over $1.3 trillion in economic assets (USACE 2021). However, this is just a small fraction of levees that exist nationwide. It has been estimated that more than 100,000 miles of levee infrastructure serves to manage flood risk within the United States (CRS 2017). Projected climate models consistently predict patterns of increased frequency and severity of flooding in several regions across the United States (Villarini et al. 2011; Mallakpour and Villarini, 2015; Vahedifard et al. 2016, 2021; USACE 2018; 2021). More frequent extreme weather can directly contribute to increased probability of levee overtopping, which leads to an increase in risk posed to population and critical economic infrastructure existing within leveed areas.

The most common mode of levee failure is breach due to overtopping (Hui et al. 2016; USACE 2018). Breach can be defined as the levee giving way, thus creating an opening through which flood waters can pass and inundate the leveed area (USACE 2018). Factors influencing

overtopping performance are most commonly studied in relation to geometric, geotechnical and hydraulic parameters. Often interdependent, these factors need to also be considered with levee construction history as many levees existing within the United States were built long before engineering standards guided levee design. Understanding what factors have the most significant contribution to the probability of breach is a critical component in assessing levee risk.

Determination of the factors that have the most significant impact on levee overtopping performance can be aided by the study of documented events, for cases of both breach and non-breach. Utilizing this data, statistical models can provide a bridge between observation and analysis. The need for further development of probabilistic methods for assessing flood risk management structures, such as levees, has become a field of interest in recent history (Balistrocchi 2019). Models developed through the use of such methods been employed with documented success in the prediction and assessment of levee performance under various loading conditions, including overtopping (Uno et al., 1987, 1994; Balistrocchi et al. 2019; Isola 2020). However, many models require a complex set of parameters to generate high levels of accuracy.

1.2 Goal and Objectives

The main goal of this research is to enhance the state of the art and practice for performing levee risk amassment against overtopping. Toward this goal, this study is aimed to achieve the following four objectives. The first objective is to introduce a comprehensive dataset of 185 levee performance events and present a refined dataset that specifically considers riverine levees that have experienced overtopping. Second, a data-driven model is developed for determining the probability of levee breach due to overtopping. Next, the model accuracy is tested with additional overtopping event data not available for model creation. Lastly, the results

of known levee risk assessments are compared with model predictions to assess the compatibility between subjective risk elicitation and calculated probabilities. By accomplishing these objectives, a tool for estimating levee overtopping breach probability can be created that is both effective and relatively easy to implement in practice. The proposed model offers a method for levee overtopping risk assessment which can be readily employed in practice through the utilization of a limited number of input variables.

1.3 Scope and Organization of Thesis

This thesis is organized into five chapters. Chapter 1 is the introduction, which provides background on the issue of levee overtopping and applications of statistical methods and analysis, objectives of this research, and an outline of the thesis. Chapters 2 introduces the U.S. Army Corps of Engineers' Levee Loading and Incident Dataset, which includes a wide array of levee performance data contained within the USACE portfolio. Chapters 3 introduces a refined set of the LLID, which focuses on riverine levee overtopping and establishes a proposed logistic regression model for predicting levee breach given overtopping. Chapter 4 applies the model introduced in Chapter 3 to levee risk assessment, comparing the model results to documented risk assessments that have been completed by USACE. Additionally, data for eleven new overtopping incidents are introduced in Chapter 4 to further test the accuracy of the proposed model. Chapter 5 includes culminating conclusions from each chapter, as well as recommendations for future work. This this includes four appendices. Appendix A contains the levee overtopping dataset used to create a logistic regression model for overtopping. Appendix B includes the raw data used for logit model creation prior to data imputation. Appendix C includes the raw data used for logit model creation after data imputation. Appendix D contains the R code script used to create the logit model.

CHAPTER II

A DATASET OF LEVEE OVERTOPPING INCIDENTS

This chapter has been accepted for publication in the proceedings of Geo-Extreme 2021. The paper has been reformatted and replicated herein with minor modifications in order to outfit the purposes of this thesis.

2.1 Introduction

The U.S. Army Corps of Engineers (USACE) maintains a semi-quantitative dataset, which documents reported loading events and historic performance associated with levee segments across the USACE national levee portfolio, known as the USACE Levee Loading and Incident Dataset (LLID). The LLID considers most components that can be part of a flood risk management system (i.e., levee embankment, floodwall, pump station, and closure structures). The dataset contains information on many distress incidents, which are generally tied to commonly-assessed potential failure modes. While the LLID contains information on multiple failure modes, the topic of research presented herein is the dataset failure mode of breach due to levee overtopping. Overtopping is a critical concern for flood risk management systems, and a logical dataset to consider first, given that breach due to overtopping is the most common mode of levee failure (Hui et al. 2016; USACE 2018). Assessing levee overtopping requires a complex combination of hydrological and geotechnical parameters. As such, it is imperative that datasets such as the LLID continue to be refined and calibrated. Validation of this data will allow for

future creation of models to achieve better insight into the uncertainty of the overtopping phenomenon.

The overtopping data subset is semi-quantitative in that it contains both qualitative and quantitative variables. The qualitative data for each system documented includes, but is not limited to, location, operational and maintenance responsibilities, generalized material classification, construction entity, protective structure type (i.e. levee or floodwall), evacuation considerations and anecdotal evidence. Quantitative and semi-quantitative data regarding overtopping events includes date of event, breach locations, number of breaches, loading conditions, embankment geometry, breach dimensions, and evacuation time range data. Quantitative data is often presented in ranges due to a lack of precise measurements. Currently, the dataset includes 230 levee overtopping incidents that range in date from 1948 to 2019. Data is derived primarily from an exhaustive research effort of USACE construction documentation, flood fight reports, and post event repair documentation provided by individual districts. Additional overtopping information comes from dated aerial videos and photographs showing locations of overtopping and breaches. Additionally, general aerial images from software such as Google Earth and Bing Aerial also provide information relative to breach locations after the floodwaters have receded. Overtopping data is compiled from the USACE districts across the nation including Buffalo, Louisville, Pittsburgh, New Orleans, St. Paul, Rock Island, St. Louis, Baltimore, Philadelphia, Kansas City, Portland, Seattle, Walla Walla, Omaha, Jacksonville, Mobile, San Francisco, Fort Worth, and Tulsa. Fig. 2.1 shows a map of all USACE district boundaries for reference.

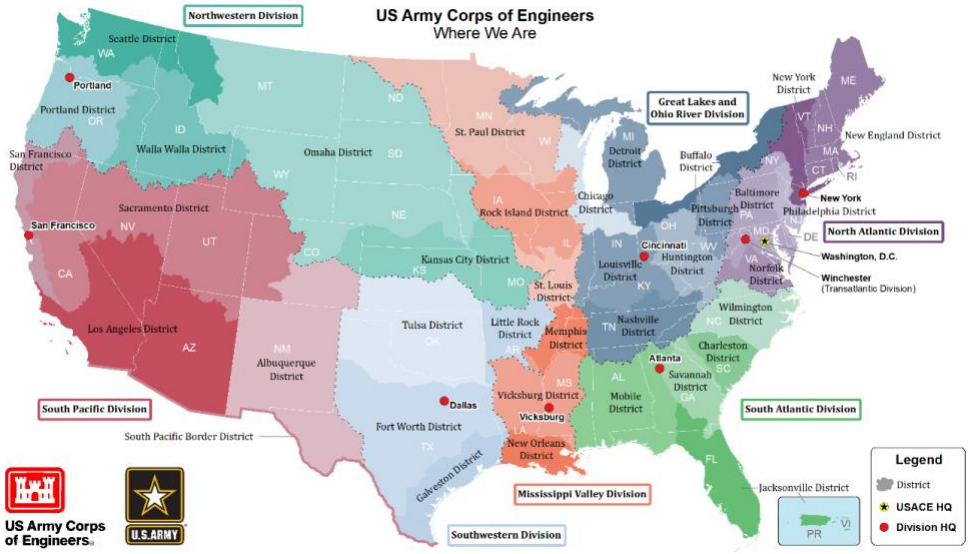

Figure 2.1 Map of USACE Districts (Source: U.S. Army Corps of Engineers, Headquarters Website)

While a great deal of information has been collected and compiled regarding parameterized compositional and spatial data throughout the USACE levee portfolio, it is critical to further that effort to include refined performance data for a more complete assessment of these levees. The overarching goal of gathering and analyzing this data is to inform decision making with regard to design and evaluation of flood risk management systems in an effort to better inform guidance, policy and risk assessment. As more data is collected and assessed, it is the intent of the author to expand efforts to further assess other areas of the LLID related to additional failure modes.

The dataset discussed in this study is expected to add to the collective field of failure analysis. Several others (Gui et al. 1998; Isola et al. 2020; Kamalzare et al. 2013; Ozer et al. 2020) have made attempts to create databases and models to better understand levee overtopping failures. Ozer et al. (2020) presented a review of various flood risk databases in dam and levee

safety. Gui et al (1998) detailed an earlier attempt of creating reliability models for riverine levee segments. Several models have since been created, including a bivariate methodology which looks at hydrological characterization of levee overtopping (Isola et al. 2020). Overtopping based on surface erosion has been investigated Kamalzare et al. (2013), where erosion parameters are modeled in a controlled setting and applied to computational analysis. In addition to those discussed, several other research endeavors have considered database analysis, field and scaled testing, and parametric study. Common to all of these models is the need for data, which the LLID presents in a manner than can be applied to various statistical models.

2.2 USACE Levee Overtopping Dataset

A The focus of current research is to evaluate the USACE dataset specifically related to overtopping events. Levee breach caused by overtopping is a common failure mode and is considered in most, if not all, risk assessments of USACE levee systems. Levee overtopping occurs when flood water elevation exceeds the height of a levee at any given point along its alignment. Therefore, it is necessary to understand two critical inputs when assessing the likelihood that overtopping is going to occur for a given system. These two inputs are the elevation of the levee and the probability of flood water exceeding this elevation.

Levees within the USACE portfolio are typically surveyed on a periodic basis, with elevation data stored within USACE district offices and the National Levee Database (USACE 2020). When assessing overtopping probability, it is important to locate extended areas along the alignment that are likely to overtop for a given hydrological event such that efforts to raise isolated low points no longer is practical. This is often referred to as the incipient overtopping location(s). These locations can occur anywhere along a levee system, can change over time, and may be present due to a number of reasons including, but not limited to, settlement, rutting from

vehicular traffic, manmade crossings, rodent borrows and distress during a previous flood event. For the purposes of understanding the LLID, overtopping events are considered where widespread overtopping occurred that could not be contained by flood fight measures.

Flood loading is considered in terms of the frequency of occurrence of a flood event and is often described as the annual chance of exceedance (ACE) or annual exceedance probability (AEP). The ACE, or AEP, is given a probability value based on a hydrological interpretation of the likelihood of occurrence. For example, a 1% event correlates to an event that is expected to have a 1% chance of occurring each year. More colloquially, this might be described as a 100-year flood event, as it statistically has a 1-in-100 chance of occurring in a given year (USACE 2018).

The probability of overtopping as a finite frequency is provided to risk assessment teams as part of the general background information of a levee system. This frequency of occurrence is the starting point for levee risk assessment when considering overtopping leading to breach. The distinction between breach and non-breach overtopping is critical in that consequences of overtopping without breach are generally lower, in terms of both life and financial loss, than the same event leading to overtopping with breach. Understanding the contributing factors that act as a cumulative tipping point between the two scenarios are critical to what is being investigated in this assessment of the LLID.

What comes next in the assessment is the evaluation of resiliency of the levee when subjected to that given overtopping event. Along with technical assessment, the effects of related floods at corresponding elevations are discussed with the local entity or group responsible for flood fighting the levee system to assess the levees resiliency. Resiliency, in this sense, is the levee's ability to withstand overtopping loading and subsequent breach. Considerations in

assessing levee resiliency when overtopping include embankment material type, duration of overtopping, depth of overtopping, embankment slope protection, embankment height, and steepness of the embankment slope. When evaluating resiliency, subjectivity becomes critical in risk assessment and design evaluation, and it is a goal of this study to better understand how levees perform when subjected to overtopping loading as supported by empirical evidence.

Finally, some data has been inferred or elicited from district or levee personnel, therefore portions of the qualitative data rely on human recollection and judgment. As a result of documenting events from decades ago, information gaps exist in many of the earlier noted overtopping events. It should also be noted that some breach locations were reported as "multiple", or without having a real sense for exact number as this information wasn't available in the post flood repair reports. For the purposes of this analysis, these recorded events were considered as a single event.

2.3 Categorization of Variables

To practically analyze the overtopping data set, incidents are organized based upon physical differences. The data set is broken down for the purposes of initial research based on a flood load source, generalized erosion resistance classification, and construction and maintenance responsibility. General trends are observed related to these general categorizations of data and future efforts in refining the dataset will serve to better correlate and utilize additional available data.

Flood load source refers to the type of loading the levee system experiences, i.e. riverine, canal, or coastal loading events. Riverine levee loading refers to any inland water source loading and is generally considered in terms of steady-state (static) or transient (dynamic) loading. In regard to the evaluation of overtopping events, transient loading is the more critical consideration

because duration of overtopping is related to the dynamic process of the water rising over the crest and receding below the crest. Canal levee segments are typically highly regulated, therefore dynamics involved with canal loading differ from riverine loading. Canal loading makes up a very small percentage of the LLID, with only one documented event which was included in the coastal dataset. As more canal loading events are documented, this information will likely be distributed. Coastal levee loading refers to any loading related to surge or tidal action. These events are always dynamic, and often related to tropical storms and hurricanes. Of the 230 levee embankment overtopping events documented, 214 are riverine events (93%) and 16 are coastal events (7%). Of the coastal levee events, 14 of 16 (87.5%) are directly related to 2005 gulf coast hurricane events, which have been heavily evaluated (Briaud et al 2008; Seed et al 2005, 2008; Sills et al 2008; Ubilla et al 2008). This distribution agrees well with the overall USACE levee portfolio which breaks includes approximately 95% percent riverine levees and 5% coastal levees (USACE 2018).

Each embankment contained within the LLID is categorized by erosion resistance related to descriptions found within available documents held by individual districts. Descriptions of levee system materials where breaches occur vary from broad material type descriptions to laboratory classification and are subject to engineering interpretation. Erosion resistance categories within the LLID are defined as “low”, “moderate”, or “high” relative erosion resistance with embankments having little to no associated material data being classified as “other” or “no classification”. Material descriptions included within the low relative erosion resistance category include sand, silty sand, silty sand with gravel, sand/silt mix, sand/gravel mix, sand/gravel mix with silt, sandy silt, and sandy gravel. Material descriptions included within the moderate relative erosion resistance category include silt, clayey silt, silt with

sand/clay, silt/clay mix with sand, silty loam, silty/clayey loam, sand/silt mix with clay, sand with silty/clay, and clayey sand. Material descriptions included within the high relative erosion resistance category include clay, clay/silt mix, clay with sand/silt, zoned embankment with impervious cover, and clay enlargement of an existing sand levee. Fig. 2.2 shows a breakdown of all embankments classified within the LLID overtopping dataset where 25% of levees are classified as low relative erosion resistance, 33% are classified as moderate relative erosion resistance, 38% are classified as high relative erosion resistance and 4% are classified as other or not classified.

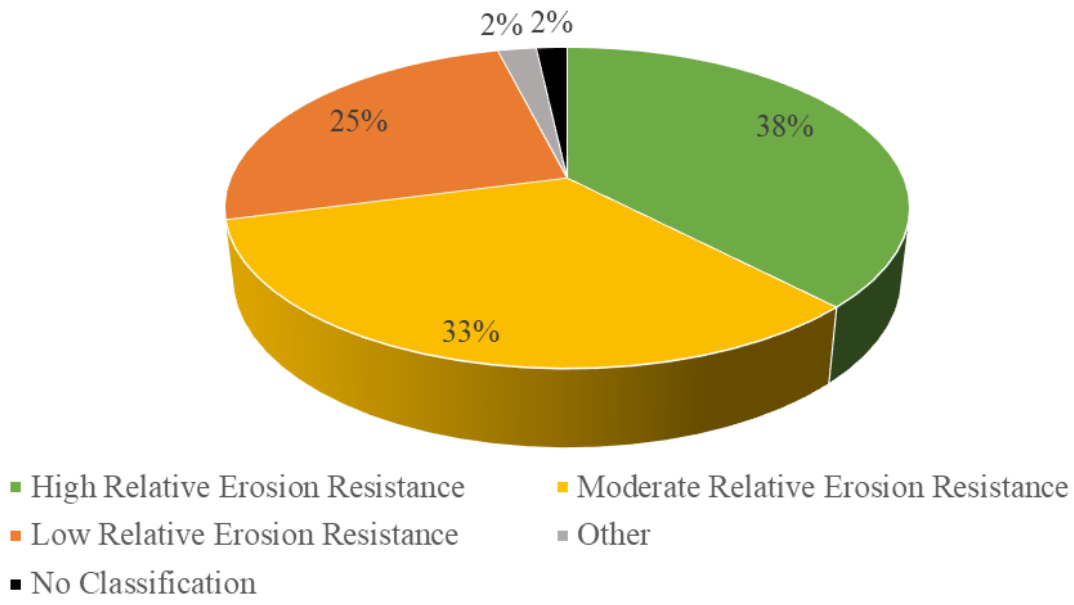

Figure 2.2 Summary of Erosion Resistance Classifications

Levee systems are also categorized by quality of construction and maintenance associated with the levee embankments. This distinction is separated into two categories, “locally constructed/maintained and re-classified federal levees” and “federally constructed/improved

levees". The differences in these two designations are centered on construction authorization, quality of original design/construction, available data and observed maintenance actions. A "re-classified" federal levee is one which has known design/construction or widespread historical maintenance deficiencies. Of all embankments classified within the LLID overtopping dataset, 55.7% were classified as locally constructed/maintained and re-classified federal levees and 44.3% were classified as federally constructed/improved levees. Given that several levee systems throughout the country were constructed long before federal construction authorizations and appropriations for levees existed, the difference in distribution is considered reasonable.

2.4 Levee Overtopping Performance

The remainder of presented analysis will focus on overtopping events relating the aforementioned categorizations of relative erosion resistance and construction and maintenance designation. First, overall breach rates are considered for overtopping events. Breach rate (R_b) in this analysis is simply the ratio of the number of overtopping events resulting in breach (N_b) to the total number of overtopping events in a single category (N). The total number of overtopping events is equal to the sum of breach events (N_b) and non-breach (N_n) events per given category. This can be shown mathematically as:

$$R_b = \frac{N_b}{N} \times 100\% = \frac{Nb}{N_b + N_n} \times 100\% \qquad (2.1)$$

Fig. 2.3 shows the distribution of overtopping events that resulted in breach versus non-breach for each construction and maintenance designation. As can be interpreted from the histogram, breach rates of levees when overtopped are significantly lower when comparing federally constructed/improved levees (46%) to locally constructed/maintained re-classified

federal levee segments (80%). The cumulative breach rate of all systems in the levee overtopping data set is 65% given that overtopping of the embankment occurs, regardless of classification.

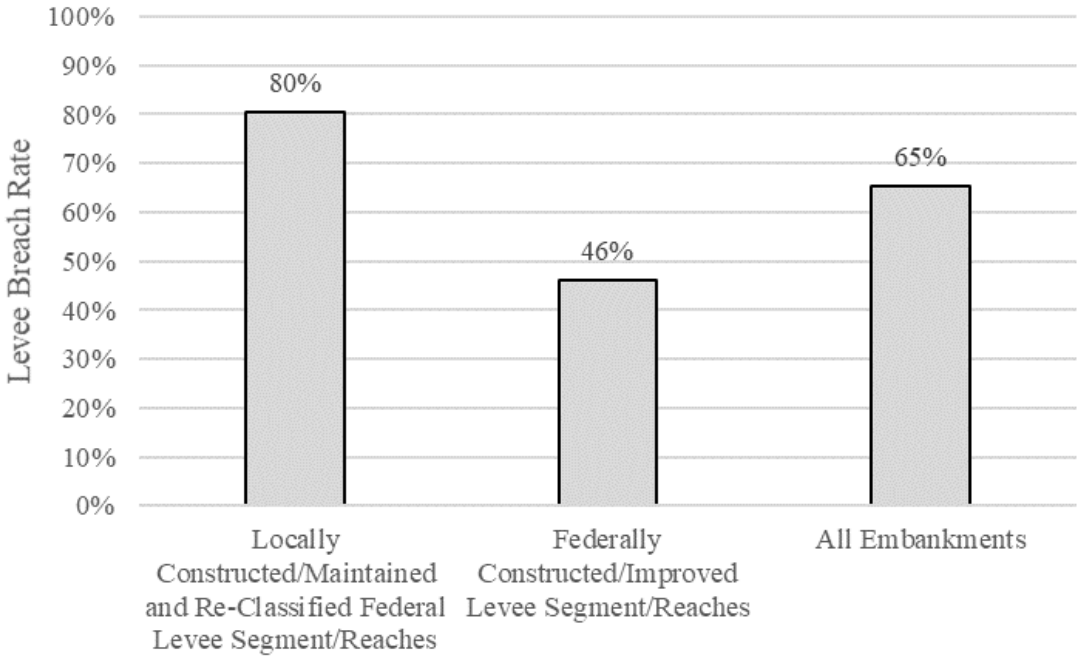

Figure 2.3 Summary of Levee Overtopping Breach Rates by Construction and Maintenance Designation

Fig. 2.4 further refines the breach rate analysis, where overtopping rates are shown for both construction and maintenance designation and relative erosion resistance. As shown in the figure, likelihood of breach due to overtopping increases as relative erosion resistance decreases, which is expected. Considering only the relative erosion resistance of the embankment material, levees with low relative erosion resistance breach at a rate of 84% when overtopped, this figure decreases to 74% for levees with moderate relative erosion resistance, and to 45% for levees with high relative erosion resistance. Significantly, it is noted that as relative erosion resistance categorization of the levee improves, construction and maintenance designation play a major role

in breach rates. While all embankments with low relative erosion resistance breach at a rate between 83-85%, levees with moderate relative erosion resistance show a breach rate disparity of 30%, and those with high erosion resistance show a disparity of 43%, when considering locally constructed/maintained and re-classified federal levees versus federally constructed/improved levees. This is a strong indicator that federally constructed and maintained levees are typically more resilient than those that are locally constructed and maintained. When considering riverine versus coastal levees, a similar relationship is observed. Fig. 2.5 shows that, when coastal events are excluded from the data set, relative erosion resistance has a slightly increased effect on breach rate for federally constructed and maintained levees. Low relative erosion resistance levee breach rate is unchanged while moderate and high relative erosion resistance levees breach at a rate 2% and 6% lower, respectively. Table 2.1 includes a summary of all levee overtopping breach and non-breach data.

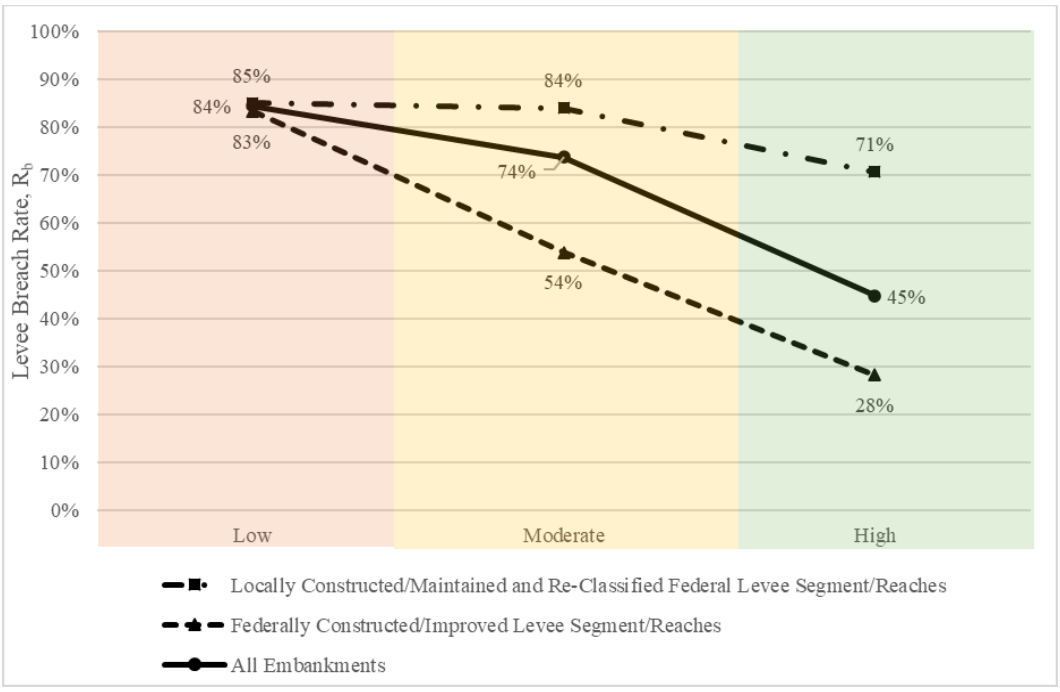

Figure 2.4 Summary of LLID Levee Breach Rates by Relative Erosion Resistance

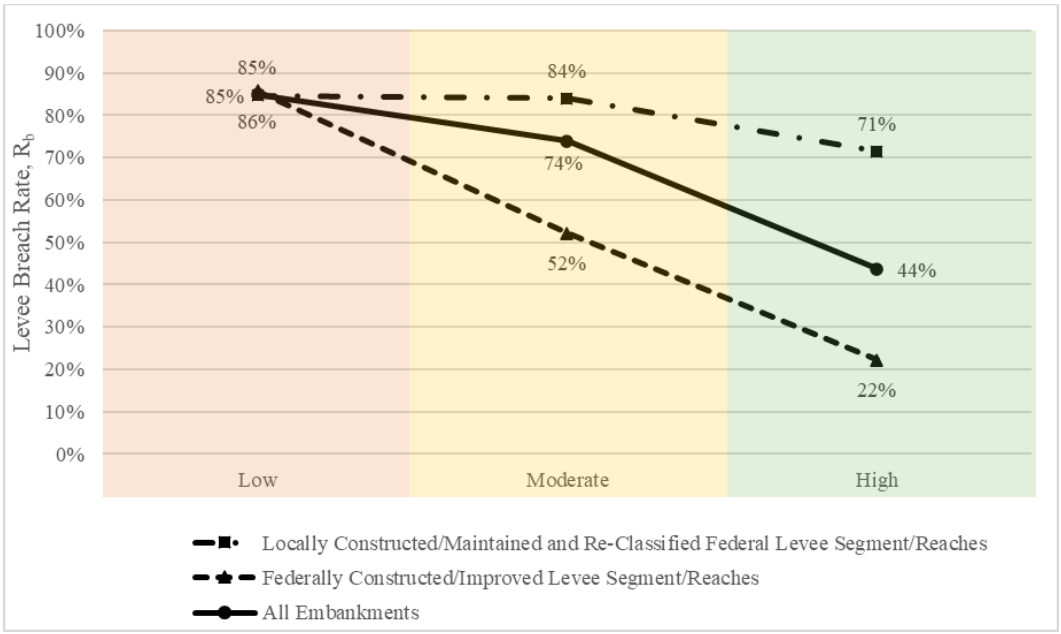

Figure 2.5 Summary of LLID Riverine Levee Breach Rates by Relative Erosion Resistance

Table 2.1 Summary of LLID Overtopping Data

Erosion Classification	Construction Categorization	Total Number of Overtopping Events	Overtopping Events w/ Breach	Overtopping Events w/o Breach
All	Local and Re-Classified Federal Segment/Reaches	128	103	25
	Federal Constructed, Well Maintained	102	47	55
	All Embankment Overtopping Events	*230*	*150*	*80*
High	Local and Re-Classified Federal Segment/Reaches	34	24	10
	Federal Constructed, Well Maintained	53	15	38
	Higher Erosion Resistance Embank OT Events	*87*	*39*	*48*
Moderate	Local and Re-Classified Federal Segment/Reaches	50	42	8
	Federal Constructed, Well Maintained	26	14	12
	Moderate Erosion Resistance Embank OT Events	*76*	*56*	*20*
Low	Local and Re-Classified Federal Segment/Reaches	40	34	6
	Federal Constructed, Well Maintained	18	15	3
	Low Erosion Resistance Embank OT Events	*58*	*49*	*9*
Other	Local and Re-Classified Federal Segment/Reaches	0	0	0
	Federal Constructed, Well Maintained	5	3	2
	Other Erosion Resistance Embank OT Events	*5*	*3*	*2*
No Classification	Local and Re-Classified Federal Segment/Reaches	4	3	1
	Federal Constructed, Well Maintained	0	0	0
	No Erosion Resistance Embank OT Events	*4*	*3*	*1*

In addition to analysis of breach rates of embankments, current efforts are investigating physical properties collected for each breach event. One example of this is breach top width, or

the length of the levee that has breached during for a particular event. This set of analysis considers all levees in the LLID with a defined number of breaches given an overtopping event. Breach top width is assessed for each segment given that a breach occurs. Analysis of this dataset does not include levees which had an undefined or unknown number of breaches for a given overtopping event. The number indicated on the horizontal axis is the top end of the range, whereas the value to the left of a given horizontal axis values is the bottom end of the range. For example, if 30 events occur with a Breach Width of 250 feet, these 30 events are between 200 feet and 250 feet in length. Fig. 2.6 shows a summary of the frequency of all overtopping event breach widths contained within the LLID dataset. Breach width data indicates a heavy grouping of collected measurements less than 400 feet, with individual breach widths greater than 400 feet being less common. As part of future research and analysis of the LLID, physical performance parameters such as breach width will be further investigated.

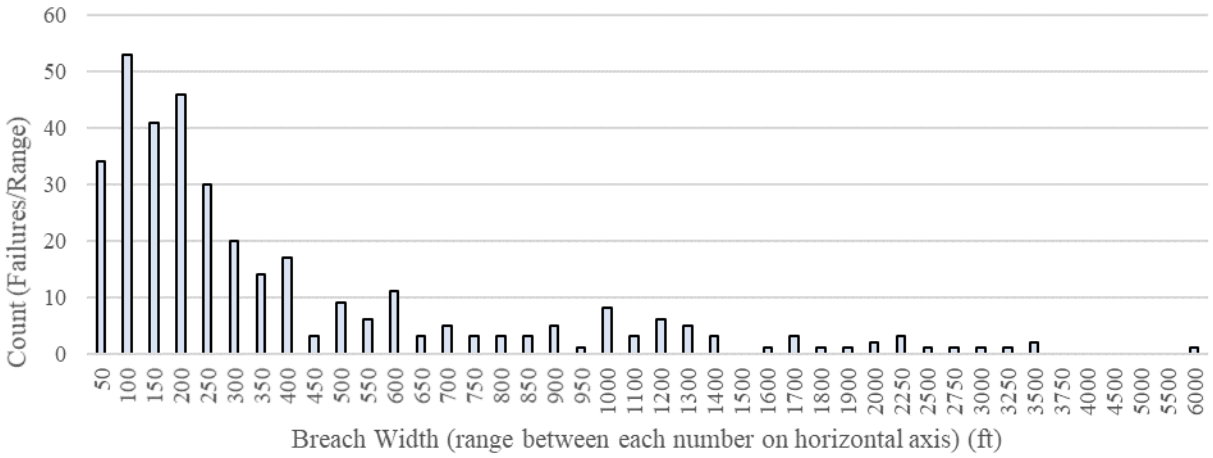

Figure 2.6 Summary of LLID Breach Width Data.

CHAPTER III

DATA-DRIVEN MODEL FOR PROBABILITY OF LEVEE BREACH DUE TO OVERTOPPING

This chapter has been submitted for review and possible publication in a scholarly journal. The paper is currently under peer review process while this thesis has been written. This chapter has been reformatted and replicated herein with minor modifications in order to outfit the purposes of this dissertation.

3.1 Introduction

Earthen levees are a critical component of flood risk management in the United States. Over 160,000 km (~100,000 miles) of levees protect the safety and economy of flood-prone areas across the United States (CRS 2017). The U.S. Army Corps of Engineers (USACE) levee safety portfolio includes over 24,000 km (~15,000 miles) of documented levee systems as communicated by the National Levee Database (NLD). More than 14% of USACE levee systems are classified as having very high, high, or moderate risk (USACE 2021). Often, high risk levees are associated with high economic and potential life loss consequences. Within the USACE levee portfolio, over 86% of the population at risk resides behind just 7% of the all levee systems (ASCE 2021). With over 2,000 systems contained within the portfolio, the USACE maintained portion of the national levee inventory protects a population of nearly 13 million and property value exceeding $1.3 trillion alone (USACE 2021). According to the American Society of Civil

Engineers (ASCE) Report Card released in 2021, identified damages from the recent 2019 flood event in the Midwest exceeded $20 billion dollars with more than 1,100 km (700 miles) of levee requiring repair. Historic records and projected future models consistently show exacerbating patterns in the frequency and severity of floods in several regions (Villarini et al. 2011; Mallakpour and Villarini, 2015; Vahedifard et al. 2016, 2021; USACE 2018; 2021), which can increase the probability of levee overtopping and, subsequently, the risk posed to population and critical economic infrastructure existing within leveed areas. It is evident that levee performance is of critical interest to engineers, municipalities, and policy makers alike.

Levee overtopping is a critical concern for flood risk management systems, with breach due to overtopping being the most common mode of levee failure (Hui et al. 2016; USACE 2018). Breach occurs when a levee gives way, allowing flood water to pass through the barrier and inundate the leveed area (USACE 2018). When a levee gives way after being overtopped, the flow through the breach can be substantial, as a river attempts to drain through a relatively small opening. For a levee to breach due to overtopping, sustained flow over the embankment must first occur long enough to initiate erosion. Once levee erosion has initiated, the continuation of erosion unravels the embankment leading to significant material loss, resulting in a breach. However, a levee can overtop without breaching. A levee breach may be avoided if the embankment is resilient enough to substantially resist erosion caused by overtopping flows. A non-breach overtopping event occurs when the levee is not substantially degraded. Figs. 3.1 and 3.2 show examples of levee overtopping leading to breach and non-breach, respectively. When breach does not occur, the consequences related to life safety and economic loss are typically reduced.

Figure 3.1 Levee overtopping leading to breach (Source: USACE Rock Island District 2019 Mississippi River Flood Fight digital photo library)

Figure 3.2 Levee overtopping leading to non-breach (Source: USACE Rock Island District 2019 Mississippi River Flood Fight digital photo library)

As the quantity of levee performance data increases, so too should the number and quality of models for describing the risk of levee breach due to overtopping. Working towards a combined levee performance data repository is a worthy goal and has been proposed by many with the intent to increase understanding of the phenomenon of levee overtopping. Ozer et al. (2020) provides just one example of such an effort, which has introduced the International Levee Performance Database (ILPD). This work seeks to create a global data source for levee breach

analysis. Expanding upon the information held within this database and others like it, such as the NLD, is just one step of many towards creating reliable levee performance models both for overtopping breach and many other failure modes.

The primary objective of this study is to develop a data-driven model using logistic regression for determining the probability of levee breach due to overtopping. Additionally, a comprehensive dataset documenting levee overtopping event performance is presented. A logistic regression model was trained utilizing applicable data from the dataset. The model allows for the calculation of breach probability based on a selected number of independent variables. The proposed model has been validated using k-fold cross validation and a random test dataset. The resulting product of this study is a reasonably accurate and efficient predictive model which utilizes a relatively minimal number of levee overtopping parameters to create a screening level understanding of the probability of levee breach occurrence given that overtopping is occurring.

3.2 Background

The assessment of risk associated with levee overtopping requires an understanding of geometric, geotechnical, and hydraulic parameters that contribute to the probability of breach. Geometric factors such as levee height, width, and slope steepness inherently affect the initiation and progression of embankment erosion when overtopped. Geometric factors can be influenced by spatial availability, geotechnical properties of available embankment material and potential flood magnitude. Spatial availability refers to the area needed to construct a levee, which may be constrained by factors such as floodway impacts, real estate rights, environmental and/or cultural impacts. Geotechnical factors that affect levee geometry, and subsequently the probability of overtopping breach, are directly related to soil properties when properly designed.

Geometric and geotechnical design constraints are intended to control the performance of an earthen levee when subjected to hydraulic loading. Factors such as the shear strength of embankment material impacts slope stability, thus defining the allowable steepness of slopes. Erosion resistance properties vary significantly by soil type. These properties have been shown to have a significant impact on the breach potential of an overtopped levee (Briaud et al. 2008) and are generally considered when establishing the cross-sectional dimensions of a levee embankment using modern standards of care. Therefore, geotechnical and geometric properties are often interdependent.

However, this is not always the case as many levees existing within the United States were built long before the utilization of geotechnical design standards were standard practice, which is why it is important to consider levee design and construction history where possible in assessing performance. Geometric constraints placed on levee design by hydraulic conditions may have some of the most significant impact, as these constraints typically dictate the height of the levee. Hydraulic parameters are also significant in that they are constantly evolving as more data is collected and statistically updated. Thus, hydraulic impacts must be continuously updated to assess the ability of levees to withstand the most up-to-date maximum projected flood.

With a proper knowledge of how these parameters affect levee performance, in conjunction with known historical design and construction conditions, this information can be used to assist in understanding levee overtopping performance. Employing statistical models can be valuable in predicting the probability of levee breach as a function of controlling geometric, geotechnical, and hydraulic parameters. Statistical models as predictive tools have become an integral component in developing risk assessment frameworks for various structures and infrastructure systems due to advances in computational efficiency and robust predictive

capability (Rahimi et al., 2019; Zamanian et al., 2020; Dehghani et al., 2021). This progress continues to expand the toolbox of the practicing engineer and improve the profession's collective ability to perform informed empirical analyses which informs decision making when dealing with large scale performance data. Logistic regression models have been widely used for data analysis in which the outcome variable is binary or dichotomous and follows a Bernoulli distribution. Several studies have employed logistic regression to assess a wide array of geotechnical failure mechanisms such as soil liquefaction and slope instability (Das et al. 2010; Gandomi et al. 2013; Zhang et al. 2013; Vahedifard et al. 2017). Logistic regression has been also deployed specifically to evaluate the performance of levees (Uno et al. 1987; Flor et al. 2010; Heyer et al. 2010; Danka and Zhang 2015), including considerations of both the rate of breach and overall breach propagation. A summary of previous levee performance logit models is well documented by Heyer et al. (2013) in an effort that considers the benefits and limitations of the use of logit models in assessing levee failure.

3.3 Levee Overtopping Dataset

The dataset presented in this study is a subset of the Levee Loading and Incident Database (LLID), which consists of a collection of both quantitative and qualitative information that documents past performance of USACE levees under flood loading (Flynn et al. 2021a). The cumulative database considers a wide range of flood risk management system components including levees, floodwalls, pump stations and closure structures. The LLID takes a risk-based approach to the organization of data, with the data subdivided into categories based not only on structure type, but also the distress mechanism, i.e. overtopping, internal erosion, stability, etc. Levee data included within the database comprises event information dating back several decades, with levee performance data ranging from 1948 to 2019.

The compilation of this data was initiated concurrent with the establishment of the USACE Levee Safety Program in 2006, following the widespread levee breach events that occurred during the Hurricane Katrina disaster. The result of both past and on-going research and data synthesis, the LLID was created by a team of risk experts using project design, construction, inspection, and flood response records with the initial goal of creating a simplified risk screening tool. Additionally, data compilation and assessment had the goal of establishing base failure rates based on historic performance. To date, the LLID includes performance information on 22% of the overall USACE levee portfolio, focused initially on levees with known major loading performance history. The information contained within the LLID reflects 23,889 years of cumulative performance data.

This study focuses on riverine levee overtopping events due to the fact that, while there are some coastal events included within the overall dataset, the physics of overtopping are materially different. At the time of this study, the LLID contains information on 214 riverine levee overtopping events. Of the more than 30 variables included to describe the full range of data in the database, only physically representative data were used in this study. The levees considered in this study vary significantly both compositionally and in terms of how they are loaded hydraulically. Levees range from agricultural dikes that have been kept in place for decades to federally authorized and designed systems protecting large populations at risk. Flood sources, ranging from minor tributaries to larger rivers, are captured in the data and represent a wide range of loading, both in terms of magnitude and duration. Data for overtopping events is a culmination of both measurement and correlation. Strong emphasis is placed on the fact that a significant portion of this data is based on human estimation and recollection, which leads to data being reported in terms of range, rather than as an exact value in many cases.

The levee overtopping data used in this study include levee systems over a wide geographical range in the continental United States, ranging from New York to Washington, east to west, and northern Minnesota to Central Texas, north to south. The total combined length of the levee systems included in this study is approximately 1348.6 km (848 miles), or about 6% of the NLD inventory in terms of total levee system length (USACE 2020). The levees evaluated in this study represent the range of unique factors that differentiate flood risk management systems in the United States, reflective of the overall LLID. Of note, the study area includes a high density of events that have occurred in the Mississippi, Missouri, and Ohio Rivers, which are part of three of the largest watersheds in the Midwest and eastern United States. These rivers account for 9.6%, 8.9%, and 4.3% of all LLID events, respectively.

Of the 214 levee overtopping events contained within the LLID overtopping data set, 185 were selected for developing a logistic regression model (see Appendix A for the complete dataset used in this study). In the initial screening of data, event records with significant gaps were excluded since these data could not be relied upon to provide valuable information. Once the final dataset for model development was established, the data was assigned to various categorical levels which are described in following sections.

3.4 Selection of Model Variables

While the LLID considers a wide range of parameters and situational conditions to describe each overtopping event, only the variables that are intuitively deemed to have a physical impact on the performance of levee systems were selected for this study. Seven total variables were considered for model inclusion based on a review of database metrics. The variables assessed in this study include those relating to the physical composition of the levee in terms of geometry, material type and consideration of construction quality. Additionally, external loading

on the levee is considered in the form of hydraulic loading of the levee both prior to and during overtopping. In summary, these factors are limited in scope to geometrical, hydraulic, and geotechnical categorization.

The consideration of the number of input variables in this study is driven by the desire to create a screening level approach that informs risk by utilizing readily available data that can be categorized to fit a simplistic model. In addition to numerical variables, some variables need to be estimated or ascertained by grouping them into ranges. These data are considered categorical, which are data whose inputs are grouped into levels such that the model does not require a precise numerical value. Categorical variables are also appropriate for descriptive variables that are non-quantitative, and address the qualitative information contained within the study dataset. A summary of model variables and the associated categorical levels is shown in Table 3.1, with a representation of variables in Fig. 3.3. Fig. 3.4 shows a distribution of the model variables. See APPENDIX A for the complete list of variables for all 185 overtopping events used as a basis for this study.

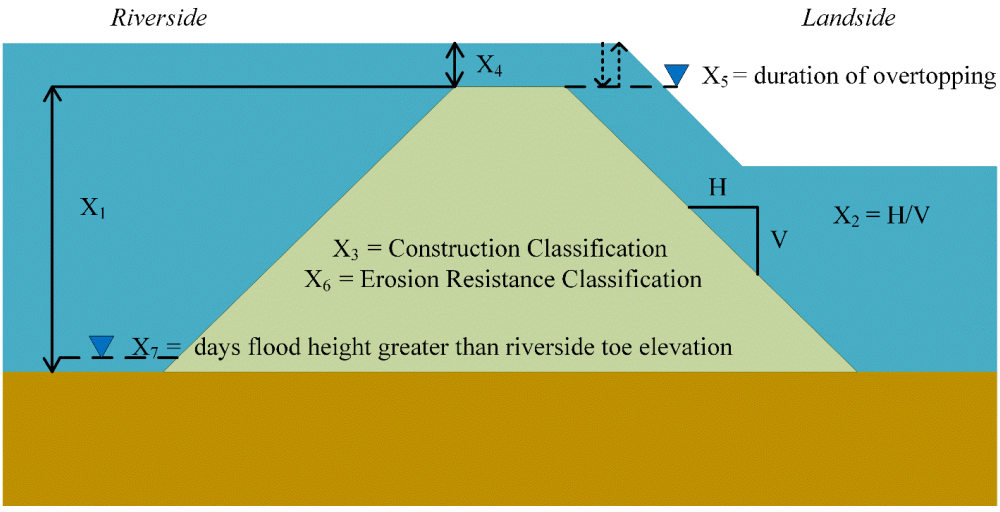

Figure 3.3 Representation of model variables

Table 3.1 Summary of Model Variables

Code	Variable	Type	Level Code	Level Description
X_1	Levee Height	Numerical	-	(meters)
X_2	Slope Geometry	Categorical	1	<3H:1V
			2	≥3H:1V
X_3	Levee Construction Entity	Categorical	1	Local
			2	Federal
X_4	Water Depth Over Levee	Categorical	1	< 0.152 m (< 0.5 ft)
			2	0.152 m - 0.305 m (0.5 ft - 1 ft)
			3	> 0.305 m (> 1 ft)
X_5	Duration of Overtopping Flow Prior to Breach	Categorical	1	<6 hours
			2	6-24 hours
			3	>24 hours
X_6	Erosion Resistance Classification	Categorical	1	Low
			2	Moderate
			3	High
X_7	Duration of Levee Loading Prior to Overtopping	Categorical	1	<3 day
			2	3-14 day
			3	>14 day

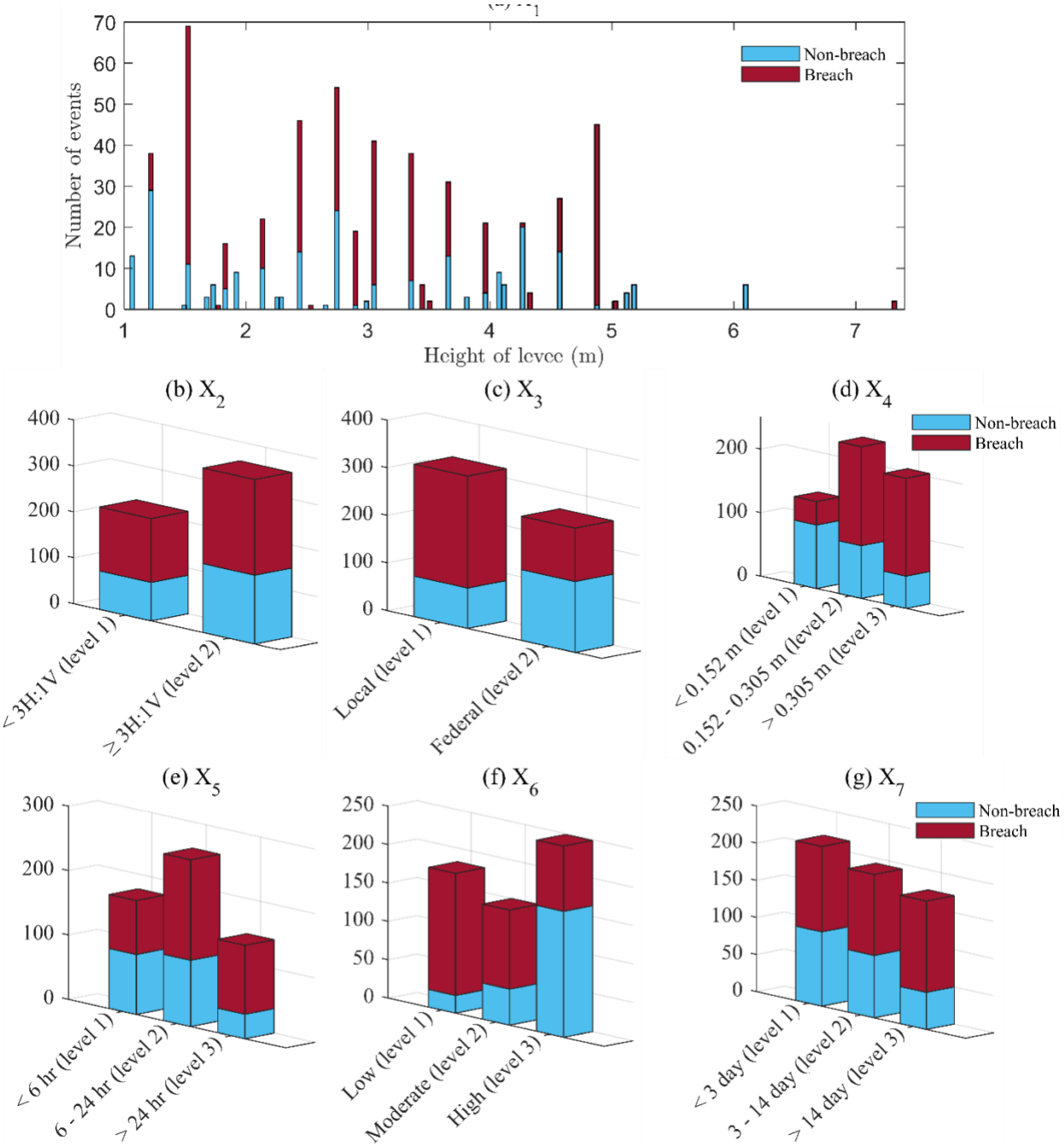

Figure 3.4 Distribution of variables included in the dataset (a) X_1: Levee height; (b) X_2: Slope geometry, (c) X_3: Levee construction entity; (d) X_4: Water depth over levee; (e) X_5: Duration of overtopping flow prior to breach; (f) X_6: Erosion resistance classification; (g) X_7: Duration of levee loading prior to overtopping.

3.4.2 Levee Height (X_1)

Levee height is a critical component of levee geometry that is often considered in overtopping analyses and associated risk. Levee height as a contributing metric has been included when considering levee performance models by others (Gui et al. 1998; Heyer et al. 2010; Danka and Zhang 2015). Levee height in this study refers to the average height of the overtopped embankment. Levee height data is incorporated into analysis as a continuous numerical variable, using the exact number presented in the LLID. The loading height for all levees ranges from 0.91 m (3 feet) to 7.32 m (24 feet).

3.4.3 Slope Geometry (X_2)

Slope geometry in this study refers to the steepness of the landside levee slope. The landside slope steepness for all levees ranges from 1H:1V to 5H:1V. Slope geometry was considered as a categorical variable in this study with two levels. Level 1 corresponds to a landside slope less than 3H:1V, and level 2 corresponds to a landside slope greater than or equal to 3H:1V. The physical meaning in selecting the delineator between the two levels of X_2 is that modernly constructed levees are specified to have a minimum slope of 3H:1V to allow for ease of maintenance (USACE 2000). By using this logic, the authors acknowledge that a partial relationship inherently exists between X_2 and X_3, Levee Construction Classification, which will be described in the following section.

3.4.4 Levee Construction Classification (X_3)

Levee systems are also categorized by quality of construction and maintenance associated with the levee embankments. This categorical variable is classified into two levels, level 1 includes “locally constructed/maintained and re-classified federal levees” and level 2 includes

"federally constructed/improved levees". The differences in these two designations consider construction authorization, quality of original design and construction, available data, and observed maintenance actions. A "re-classified" federal levee is one which has known design, construction, or historical maintenance deficiencies. Given that many levees throughout the country were constructed long before federal construction authorizations and appropriations for levees existed, the difference in this distribution is considered reasonable. Although levee construction classification is easily applied to the dataset described within this study for levees within the USACE portfolio, the categorization can also be applied to any levee with known construction history. The general difference of level of care should be considered when classifying levees from different data sources. Level 1 should be used in any case where a high level of quality control was not undertaken during the design and construction process.

3.4.5 Overtopping Depth (X_4)

Overtopping depth refers to the height of water over the levee while being overtopped. The hydraulic load of overtopping depth has been widely considered in both numerical (Sharp et al. 2011) and statistical models for different types of flood loading including canal loading (Lendering et al. 2012), riverine loading (Amabile et al. 2016; Isola et al. 2020), and compound riverine and coastal loading (Jasim et al. 2020). Data for this variable is based on either physical measurement or using nearby stream gage data to approximate depth at the location of breach. Overtopping depth is a categorical variable in this model with three levels. Level 1 includes overtopping depths less than 0.152 m (0.5 feet), level 2 represents overtopping depths between 0.152 m and 0.305 m (0.5 ft and 1.0 ft), and level 3 denotes overtopping depths greater than 0.305 m (1 ft). These three levels can be considered as "minor", "moderate", and "major" overtopping, respectively. Assigning categorical levels for this data was based on an assessment

of relative data distribution. It was noted during initial review of the dataset that a large percentage of overtopping events, both breach and non-breach, had associated overtopping depths of less than 0.305 m (1 ft). The physical relevance of this observation is that often hydraulic load leading to overtopping is constrained by other systemic factors that restrict the ability of the river to rise significantly above the crest of the levee. Some of these factors might be nearby diversion structures, lower levels of protection across the flood source channel, or breach at the site of interest. It is critical to note that overtopping depth in this dataset is recorded for both breach and non-breach events. The depth considered for breach events assumes that the levee breached in the range represented by the categorical level, therefore was not able to increase in height after breach.

3.4.6 Overtopping Duration (X_5)

While overtopping duration takes on two distinct meanings within the dataset, it can be treated as a single variable. In the event that overtopping leads to breach, the overtopping duration is a measure of the duration that the levee crest elevation is first exceeded until the breach occurs. For events where overtopping does not result in breach, the overtopping duration is simply the measure of time between the levee crest elevation being exceeded by flood water and the flood water returning to an elevation below the levee crest. In either case, the duration represents the total load on the levee, so the variable is treated proportionally for each case. Data for this variable is approximated based on either physical measurement or the use of stream gage data. Overtopping duration is a categorical variable in this model with three levels. Level 1 corresponds to overtopping that occurred for less than 6 hours, level 2 considers overtopping that took place for 6 to 24 hours, and level 3 corresponds to overtopping events that occurred for more than 24 hours. In terms of physical relevance, level 1 overtopping duration can be

correlated to flashy, or short-term flood events. Level 2 corresponds to moderate term flood events, or those typically experienced on riverine levees. Level 3 is the long-term case in which a levee is overtopped for more than 1 day prior to breaching or does not breach at all.

3.4.7 Erosion Resistance Classification (X_6)

Each levee embankment used in this study is categorized by erosion resistance, which is determined by the material descriptions contained within the LLID. Erosion resistance refers to the general ability of the levee to resist degradation when subject to overtopping load. Surface erosion is a field of study in its own with much work having been done to understand how various material combinations generally resist hydraulic stresses. Erosion specific to levees has been studied by many who have looked at contributing factors such as soil type, shear stress from the hydraulic load, soil shear strength, etc. (Briaud et al. 2008; Kamalzare et al. 2013; Ellithy et al. 2017; Osouli et al. 2018). Additionally, many of the previously referenced levee performance logistic regression models considered erosion resistance as input variable (Flor et al. 2010; Heyer et al. 2010; Danka and Zhang 2015). Erosion resistance classification is a categorical variable in this model with levels 1, 2 and 3 defined as “low”, “moderate”, or “high” relative erosion resistance, respectively. For this study, erosion resistance typically followed the logic that coarser grained embankment materials are more susceptible to erosion than finer grained embankment materials. Forensic analysis of wide scale levee breach due to overtopping documented after the Hurricane Katrina event indicated that roller compacted clay levees performed much better than silt and sand levees when overtopped (Sills et al. 2008). This logic is based on a general trend, and materials encountered within levee embankments that are not listed below will need to be considered accordingly. Erosion resistance classification does not consider any landside slope armoring or the effects of vegetive cover. Material descriptions corresponding

to the categorical levels of erosion resistance classification are shown below in Table 3.2. Generalized erosion resistance levels are based upon a general review of surface and embankment erosion literature.

Table 3.2 Material Description for Erosion Resistance Classification

Low Erosion Resistance	Moderate Erosion Resistance	High Erosion Resistance
• sand • silty sand • silty sand with gravel • sand/silt mix • sand/gravel mix • sand/gravel mix with silt • sandy silt • sandy gravel	• silt • clayey silt • silt with sand/clay • silt/clay mix with sand • silty loam • silty/clayey loam • sand/silt mix with clay • clayey sand	• clay • clay/silt mix • clay with sand/silt • zoned embankment with impervious cover • clay enlargement of an existing sand levee

3.4.8 Duration of Levee Loading Prior to Overtopping (X_7)

Duration of levee loading prior to overtopping refers to the length of time the embankment was subjected to hydraulic load above the riverside toe prior to overtopping. This variable is used to represent the effect of saturation of the levee prior to overtopping. Physically, when the saturation of the levee increases, porewater pressures within the levee increase, and the overall strength and stability of the levee decreases. Thus, breach during overtopping can be assumed to be generally more likely. Given how the multiple material types and varying levee geometries included in this study vary physically, this factor does not behave in a perfectly linear manner from a statistical perspective. Levee loading duration is a categorical variable in this model, divided into three levels. Level 1 corresponds to flood water exceeding the riverside toe 3 days prior to overtopping, level 2 duration is 3 to 14 days, level 3 duration represents a duration

greater than 14 days. The physical relationship between these levels corresponds to the general hydrograph of a given river during flood loading. Level 1 can be considered as flashy loading, level 2 corresponds to a moderately rising river and level 3 corresponds to a slow rising river. Similar to the overtopping duration (X_5), the duration of levee loading data is approximated based on either physical measurement or using stream gage data to determine river levels at the breach area of interest.

3.5 Data Cleaning and Processing

3.5.1 Cumulative Effects of Hydraulic Variables

In As previously discussed, the data were evaluated based on representation of physical processes. Given this consideration, the cumulative effects of each variable also needed to be incorporated into the model. Most notably, variables related to hydraulic loading (X_4, X_5, and X_7) require the consideration of cumulative effects of increased or decreased loading, given breach or non-breach results, respectively. That is, the data needed to be extrapolated such that if a levee experienced breach at a low level of loading, it should be assumed to fail at the higher levels of loading if all other variables remain constant. Conversely, if the levee did not breach at the highest level, it would not fail at lower levels given all other variables remain constant. This is a critical assumption in the model that leads to a controlled approach in estimating physical factors. All other variables not directly related to hydraulic loading were considered independent of cumulative or ordinal effects. Data expansion to account for cumulative hydraulic effects had a minimal effect on the base breach rate of the entire data set. Prior to data expansion, breach events accounted for 63.8% of all events. After expansion, 59.7% of the dataset events resulted in breach. This is because expansion of events was not uniform. Some events were expanded to account for more additional scenarios than others given the precedent condition.

3.5.2 Data Imputation

Data for at least one variable was not recorded for 104 of the total 185 levee overtopping events. Variables with missing information include overtopping depth (X_4), overtopping duration (X_5) and duration of levee loading (X_7). To account for missing data, the k-Nearest Neighbor (kNN) imputation algorithm through the 'VIM' package for R (Templ et al. 2020) was used. The kNN imputation algorithm takes advantage of the association between the variable of interest that contains missing data and the auxiliary variables that are fully populated (Beretta et al. 2016). kNN imputation has the distinct advantage of being able to work with multivariate data to fill one or more data gaps using pattern recognition. Depending on types of the variable (e.g. categorical and numerical), an aggregation of the k-values of the nearest neighbors is employed as imputed value (Kowarik and Templ 2016). In this study, an extension of the Gower distance (Gower 1971) as the most popular distance for mixed-type variables which enables the handling of distance for binary, categorical, ordered, numerical, and semi-numerical variables is used for the purpose of kNN imputation (Kowarik & Templ, 2016; D'Orazio, 2021).

The kNN is derived for a mix of numerical and categorical variables, and the distance between i^{th} and j^{th} observations is the weighted mean of the contributions of each variable. The weight represents the importance of the variable and is selected based on the importance of variables (Kowarik & Templ, 2016; Templ, et al. 2020). The distance between i^{th} and j^{th} observations can be determined as follows:

$$d_{i,j} = \frac{\sum_{m=1}^{p} w_m \delta_{i,j,m}}{\sum_{m=1}^{p} w_m} \tag{3.1}$$

where w_m denotes the weight $\delta_{i,j,m}$ represents the contribution of the m^{th} variable. Shortcomings of this method are predominantly centered on the effects of imputation on data distribution and representation. Additionally, the selection of k-value is not predetermined based on a set mathematical relationship. Rather, k-value can be selected through an iterative process that establishes the best correlation to the distribution of each imputed variable in the original data.

A simplified visual representation of k-value selection for imputation in this study is shown in Fig. 3.5. In this figure, the unweighted $d_{i,j}$ which corresponds to setting $w_m = 1$ for all the variables is applied. That is to say, the imputation is solely performed based on contribution of the variable ($\delta_{i,j}$) in this figure. Random data points are shown based on training data for kNN imputation, with example kNN groupings. Each data point represents the tendency of the imputation process to either assign a missing categorical variable to level 1, level 2 or level 3. In this scenario, a k-value equal to 3 predicts a level 1 response for the missing data because the level 1 data outweigh the other two levels data. If the k-value is increased to 8, then the missing categorical variable is imputed as level 3.

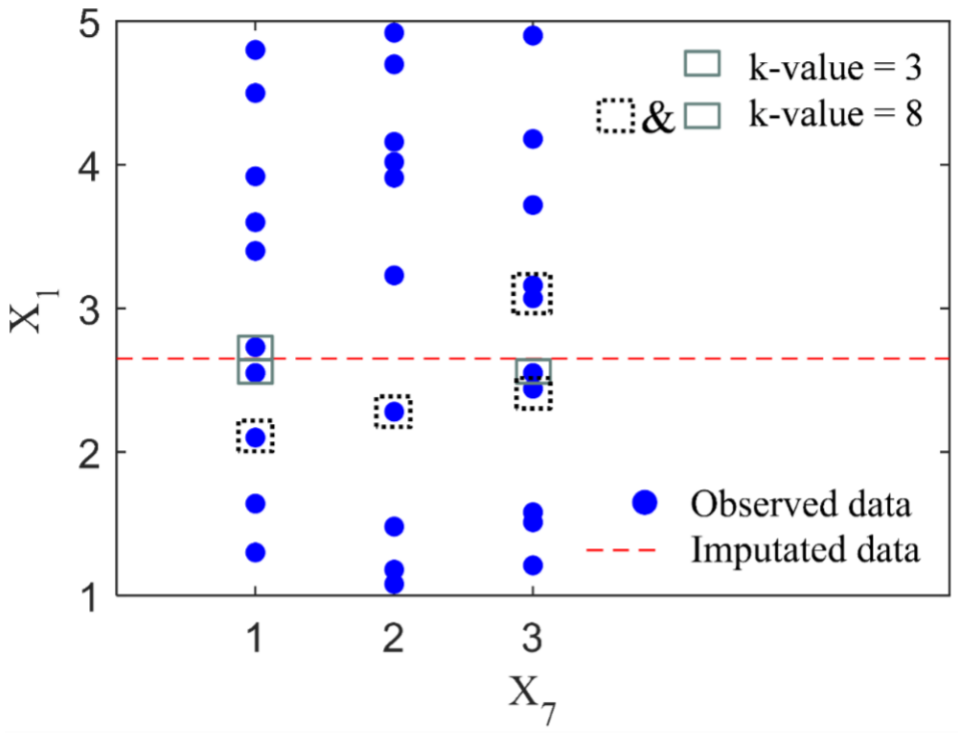

Figure 3.5 Two-dimensional kNN imputation visualization

A k-value of 8 was iteratively selected for this model based on both accuracy metrics analysis and individual variable distribution. The accuracy metrics considered were error rates when validating the data using (1) k-fold cross validation and (2) test data using a random set of real data that were excluded from the training set. Both accuracy metrics will be described in further detail in subsequent sections. As shown in Fig. 3.6, volatility is high for low k-values and attenuates as the k-value increases. In addition to assessing the model error convergence, distribution of imputed variables needed to be simultaneously considered such that the physics of the model relationship remained unchanged. As k-values were increased beyond 8 within the model, distribution smoothing to the mean was generally observed, leading to greater deviation from the original data distribution. While change in original data distribution was less when

selecting k-values less than 8, the offset in error rate was not seen to justify the minor improvements in distribution changes.

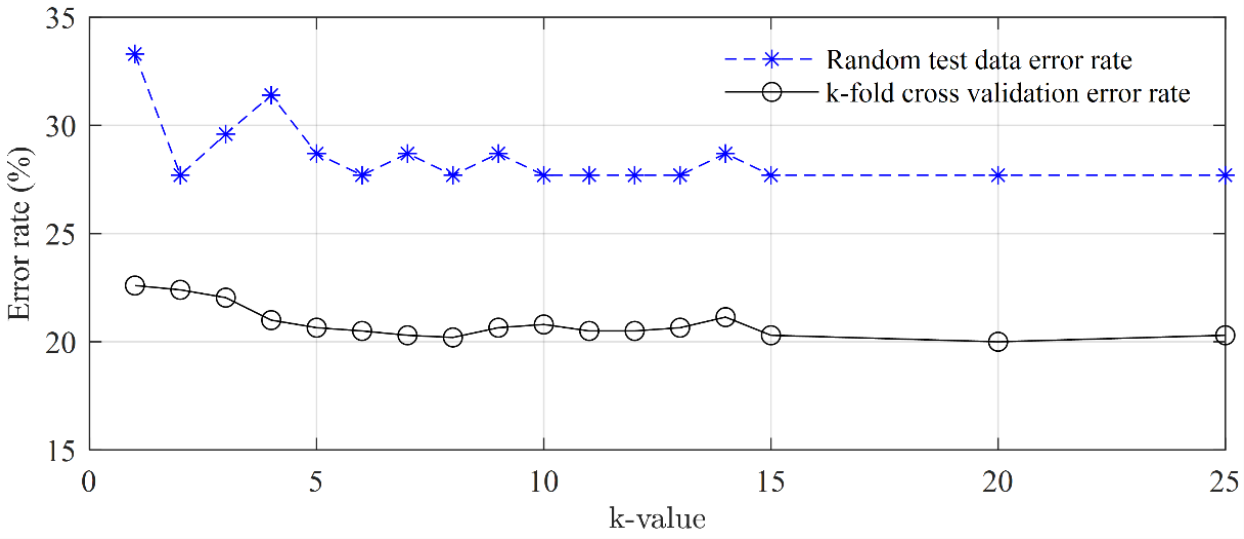

Figure 3.6 kNN error rate sensitivity

Fig. 3.7 shows the effect of imputation on the data distribution for each level, with results summarized in Table 3.3. Using a k-value of 8, the imputation of X_4 (overtopping depth) and X_7 (duration of levee loading prior to breach) had maximum distribution changes of 3.8%, and 2.0%, respectively, with increase in distribution towards level 2. Distribution changes of this magnitude are considered relatively insignificant. Once data was expanded, X_4 had 10.2% of missing measurement, and X_7 had 2.3% of missing measurement. The variable with the most missing data was X_5 (overtopping duration), with 24.1% of data unavailable. Thus, imputation had the greatest effect on this variable, with the maximum distribution change being 8.5%. The result of imputation was to force more of the data to the level 2, given three levels, which in this instance is preferrable due to the fact that many of the missing data were correlated to larger

river events, i.e. the Mississippi and Missouri Rivers. Level 2 of X_5 correlates to moderate duration overtopping (6-24 hours). This result aligns well with the data when comparing to similar events, therefore it is not considered to be a gross misrepresentation of data or trend. With these checks, it is ensured that the applied data imputation does not alter the key characteristics of the levee data. Appendices B and C contain the raw data used for model creation before and after data imputation, respectively.

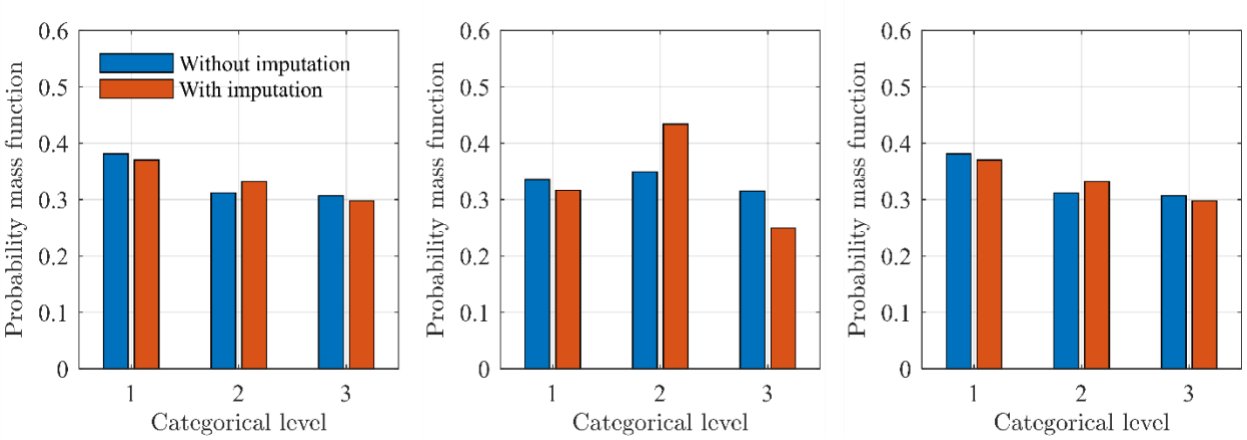

Figure 3.7 Probability change due to imputation with k = 8 for (a) X_4, (b) X_5, and (c) X_7

Table 3.3 Change in probability of X_4, X_5 and X_7 after kNN imputation for k = 8

Level	X_4	X_5	X_7
1	-2.4%	-1.9%	-1.1%
2	3.8%	8.5%	2.0%
3	-1.5%	-6.6%	-0.9%

3.6 Development of Logistic Regression Model

Logistic regression is a statistical method which allows for the utilization of both numerical and categorical independent input parameters to predict an outcome. Often models are used to predict relationships which have binary response (Kleinbaum 1994; Hilbe 2011). While logistic regression models use a linear relationship similar to multilinear regression, a logit transformation, which is replaced with other link functions, is applied to the dependent variable. Since the nature of the studied data consist of both numerical and categorical independent variables, and the dependent variables are reported in two levels, binary logistic regression was used in this study. This method of statistical analysis for levee overtopping is appropriate given that the phenomenon of levee overtopping is generally well understood in terms of what factors generally contribute, and that the result of overtopping is binary in terms of breach or non-breach. The logit model serves to assess of the degree of importance of each parameter that contributes to overtopping.

The logistic regression model developed for this study attempts to predict the probability levee breach given a set of one numerical and six categorical variables related to construction, geotechnical, hydraulic, and geometrical factors. These variables are denoted as X_1 through X_7. In this study, levee breach is defined as Y=1, and non-breach as Y=0. The probability of breach, $P(Y = 1)$, is defined by the following general form:

$$P(Y = 1) = \frac{e^Z}{1 + e^Z} \tag{3.2}$$

where

$$Z = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{3.3}$$

and X_i is the observation or predictor, with β_0 and β_i being the coefficients estimated by the regression model. When considering categorical input variables, the output of the proposed model considers each categorical level as a predictor. So, rather than an X_i input for the categorical variable, a coefficient is calculated for each level of the categorical input variable relative to the base condition (categorical level 1). In the event that a non-base condition is met, the contribution of the event occurring at that categorical level to the model is $\beta_i \times 1$. In the event that the only the base categorical level is met, the contribution is $\beta_i \times 0$. The probability of breach is not calculated as a binary output, but rather a probability of occurrence between 0 and 1. Therefore, a threshold is established to determine if the individual event is likely to result in breach. Breach is considered likely to occur if $P(Y = 1)$ is greater than 0.5, and non-breach is considered likely to occur if $P(Y = 1)$ is less than 0.5.

The first step in evaluating a logistic regression model is to assess the statistical significance of each variable, which is accomplished in the proposed model using the p-value. It is a long-established practice in logistic regression analysis that a variable is significant if its p-value is less than 0.05, which reflects 95% confidence (Fisher 1925). However, there have others who have proposed confidence levels as high as 99.5%, or p-value of 0.005 depending on the application (Di Leo et al. 2020). In this study, p-value computation is done using a base general logistic regression model, which considers the additive properties of each variable in succession, and is represented by:

$$\begin{aligned} Z = \beta_0 + \beta_1.X_1 + \beta_2.X_{X_2=2} + \beta_3.X_{X_3=2} + \beta_4.X_{X_4=2} + \beta_5.X_{X_4=3} \\ + \beta_6.X_{X_5=2} + \beta_7.X_{X_5=3} + \beta_8.X_{X_6=2} + \beta_9.X_{X_6=3} \\ + \beta_{10}.X_{X_7=2} + \beta_{11}.X_{X_7=3} \end{aligned} \quad (3.4)$$

The p-values calculated for each level of the base model are shown in Table 3.4. It should be noted that if any level of a categorical variable was found to be significant, all levels had to be included to maintain completeness of the model. Significance test results indicate that all variables are significant based on the significance threshold of 0.05, except for X_1 (levee height). This result for X_1 is intuitive when looking at the distribution of breach and non-breach data relative to an increasing levee height, as shown in Fig. 3.4. Of the variables with greater significance, variable X_2 was reviewed by considering the raw data breach rates to verify its significance as this value was close to the threshold. The p-value of 0.037 was found to have a loose relationship with the base events in the overtopping dataset which indicated that X_2 level 1 was only 4% more likely to breach than X_2 level 2 when evaluating the pre-imputation data. This observation can be explained by the inverse relationship between levee erosion resistance and standard levee design, which increases the slope ratio when material with lower erosion resistance is utilized.

Table 3.4 Model variable significance

Description	Variable	Level	p-value
Levee Height	X_1	N/A	0.566
Slope Geometry	X_2	2	0.037
Levee Construction Classification	X_3	2	1.79×10^{-5}
Overtopping Depth	X_4	2	4.07×10^{-5}
		3	1.07×10^{-9}
Overtopping Duration	X_5	2	0.089
		3	1.12×10^{-7}
Erosion Resistance Classification	X_6	2	4.00×10^{-4}
		3	$< 2 \times 10^{-16}$
Duration of Levee Loading Prior to Overtopping	X_7	2	0.022
		3	0.0123

Note: All variable significance references level 1.

Stepwise regression is a commonly used method to determine if the addition or exclusion of an individual variable, or a combination of independent variables, can improve accuracy using covariables (Steyerberg et al. 1999). Stepwise regression was used for the logistic regression model to determine if covariables in the form of two-way interactions of variables and quadratic terms could improve the model accuracy while maintaining the simplicity in the developed model. The stepwise logistic regression model was evaluated based on Akaike's Information Criterion (AIC) accuracy metric for each combination. When using AIC to compare model fitting, the model with the lowest AIC value represents the best performance. AIC scores competing models by reducing the value if information loss is minimalized and increasing the value if the model contains unnecessary complexity (Wagenmakers 2007).

The odds that an event occurs, in this case breach, is represented by the odds ratio. For the presented data, odds ratio represents the incremental change in breach probability for a given categorical level of an individual variable. Odds ratios and regression coefficients for each model term are presented in Table 3.5. In each case, the odds ratio is relative to the base case, or level 1. When the odds ratio is less than 1, breach is less likely as the categorical level increases. When the odds ratio is greater than 1, the probability of breach occurrence is more likely as the categorical level increases. For example, X_6 at level 3 has an odds ratio of 0.02. This implies that breach is 50 (1/0.02) times more likely to occur when overtopped for a low erosion resistant levee than a high erosion resistant levee. Conversely, when the X_5 variable is at level 3, the odds ratio is 6.37. This implies that overtopping duration greater than 24 hours leads to a breach probability 6.37 times higher than an overtopping event duration less than 6 hours. Where two levels are specified for the odds ratio, both cases reference level 1 for each variable. To demonstrate, when $X_3 = 2$ and $X_6 = 3$, the probability of breach is 1.67 (1/0.60) times less likely

compared to the base condition of $X_3 = 1$ and $X_6 = 1$. Conversely, when $X_4 = 3$ and $X_7 = 3$, breach is 9.89 times more likely than the base condition of $X_4 = 1$ and $X_7 = 1$.

Table 3.5 Variable Odds Ratio and Coefficient

Variable	Coefficient (β)	Odds ratio
Intercept	$\beta_0 = 0.93$	2.53
Levee Construction Entity (X_3) = 2	$\beta_3 = -1.13$	0.32
Water Depth Over Levee (X_4) = 2	$\beta_4 = 0.87$	2.38
Water Depth Over Levee (X_4) = 3	$\beta_5 = 1.27$	3.55
Duration of Overtopping Flow Prior to Breach (X_5) = 2	$\beta_6 = 0.08$	1.08
Duration of Overtopping Flow Prior to Breach (X_5) = 3	$\beta_7 = 1.85$	6.37
Erosion Resistance Classification (X_6) = 2	$\beta_8 = -1.74$	0.18
Erosion Resistance Classification (X_6) = 3	$\beta_9 = -3.93$	0.02
Duration of Levee Loading Prior to Overtopping (X_7) = 2	$\beta_{10} = -0.42$	0.66
Duration of Levee Loading Prior to Overtopping (X_7) = 3	$\beta_{11} = 0.17$	1.18
$X_4 = 2$ and $X_7 = 2$	$\beta_{12} = 2.18$	8.82
$X_4 = 3$ and $X_7 = 2$	$\beta_{13} = 2.65$	14.08
$X_4 = 2$ and $X_7 = 3$	$\beta_{14} = 0.60$	1.83
$X_4 = 3$ and $X_7 = 3$	$\beta_{15} = 2.29$	9.89
$X_3 = 2$ and $X_6 = 2$	$\beta_{16} = 1.35$	3.84
$X_3 = 2$ and $X_6 = 3$	$\beta_{17} = -0.51$	0.60

Note: β_1 and β_2 were not used in the proposed model.

After the stepwise regression was assessed, X_1 was removed due to lack of significance as observed in the base model. X_2 was considered in the stepwise model but was not included in the final model based on the calculated AIC. As previously discussed, the rejection of this variable in the final model agrees with the assessment of base data, which shows that slope geometry does not have a significant statistical impact on the base rate of breach.

The proposed logistic regression model for the LLID overtopping dataset is presented as follows:

$$\begin{aligned} Z = \beta_0 &+ \beta_3.X_{X_3=2} + \beta_4.X_{X_4=2} + \beta_5.X_{X_4=3} + \beta_6.X_{X_5=2} + \beta_7.X_{X_5=3} \\ &+ \beta_8.X_{X_6=2} + \beta_9.X_{X_6=3} + \beta_{10}.X_{X_7=2} + \beta_{11}.X_{X_7=3} \\ &+ \beta_{12}.X_{X_4=2,X_7=2} + \beta_{13}.X_{X_4=3,X_7=2} + \beta_{14}.X_{X_4=2,X_7=3} \\ &+ \beta_{15}.X_{X_4=3,X_7=3} + \beta_{16}.X_{X_3=2,X_6=2} + \beta_{17}.X_{X_3=2,X_6=3} \end{aligned} \quad (3.5)$$

The proposed model considers the base interactions of X_3, X_4, X_5, X_6 and X_7 and two-way interactions of X_4 with X_7 and X_3 with X_6. The code used for model creation and calculations can be found in Appendix D.

3.7 Model Validation

The proposed model was validated using cross validation as well as a test dataset where a portion of the base data was set aside to be used to evaluate the fitted model. Cross validation was conducted with k-fold cross validation where the dataset is divided into 'k' evenly distributed groups and subsequently compared against the selected logistic regression model to determine the model accuracy (Valavi et al. 2020). Each time the data is redistributed to validate, a fold is created. For this study, a k-value of 5 was selected to test 20% of the dataset against the remainder of the dataset in each fold. In this process, each event for validation is selected

randomly in each fold. To evaluate the accuracy of the fitted model in the k-fold cross validation, the confusion matrix and Cohen's Kappa value are used. Cohen's Kappa value, which ranges between -1 to +1, assesses relative model agreement for multiclass data with an unbalanced response (Delgado et al. 2019). Results of cross validation indicated model accuracy of 80.7% with a standard deviation of 3.4% and Cohen's Kappa value of 0.592. While there is no standardized scale for Kappa score, a value between 0.41-0.60 has been considered as moderate agreement, with values of 0.61-0.80 as being substantial agreement (Landis and Koch 1971).

When checking model accuracy using test dataset, 20% of the expanded model data was set aside that contained fully defined events before imputation. Using the confusion matrix, the accuracy of the fitted model based on the randomly selected test dataset was 73.3%, which is comparable to the k-fold cross validation results. Table 3.6 shows the results of the test data accuracy in the form of a contingency table. Test data indicates that there is a slight tendency of the model to predict a false positive, as opposed to a false negative. The probability of false positive and false negative predictions when using the test data are 14.7% and 12.1%, respectively.

Fig. 3.8 presents the calculated breach probability of all real base events used to create the proposed model with respect to the incident number, as identified in the presented dataset. Finally, the overall range of breach probability values were assessed using the proposed breach probability model which indicated a minimum probability of Y=1 equal 1.05% and a maximum probability of Y=1 equal to 99.12%.

Table 3.6 Test Data Accuracy

	Predicted Value		
Actual Value	Y=0	Y=1	Σ
Y=0	26	17	43
Y=1	14	59	73
Σ	40	76	116

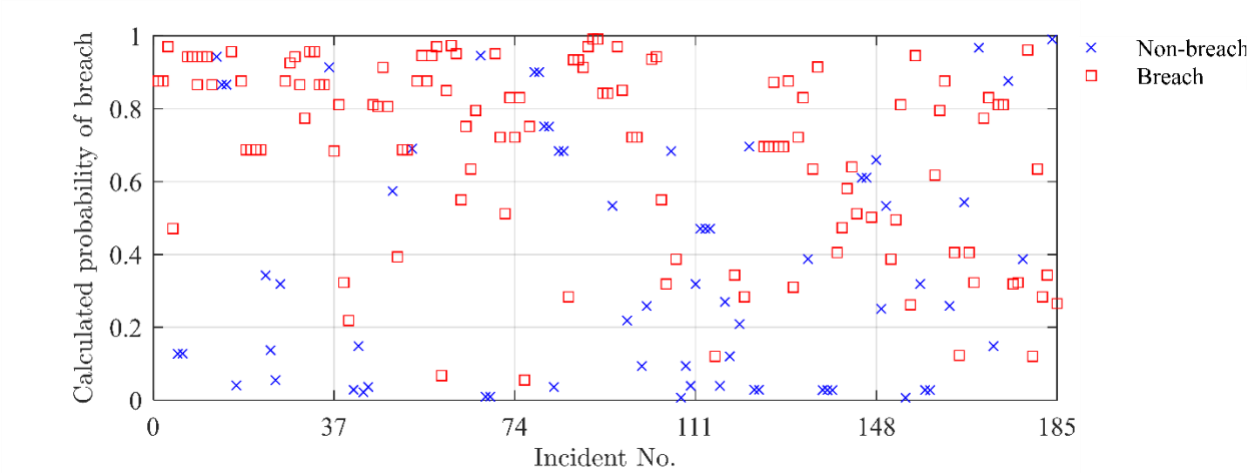

Figure 3.8 Calculated probability of breach for 185 overtopping incidents included in the dataset

3.8 Discussion

The LLID overtopping dataset can be used for many other purposes which support both USACE and other parties interested in levee performance. For instance, the information in the overtopping dataset can be used to highlight areas of typical poor performance of past overtopping locations for various systems. While just a portion of the all levee performance maintained and regularly updated, the LLID overtopping dataset can serve to provide valuable

information for researchers looking to develop models related to overtopping reliability or similar topics such as breach propagation. Additionally, the presented model can be modified such that only specific variables of interest, such as hydraulic loading or erosion resistance, are considered relative to each system to inform more specific research or programmatic interests. Potential studies could be considered using the overtopping dataset that focus on regional performance specific to watershed, river, or other specific geographical locations.

While the assessment of levee breach risk is generally thought to require a complex set of hydrological and geotechnical parameters, one valuable trait of the proposed model is that the input parameters are relatively easy to ascertain or estimate. Variables such as levee construction classification can be categorized based on known history of the structure. Erosion resistance classification can be determined by visual inspection or with minimal sampling. Often, if engineering records exist for a levee, this information is located within design documentation. The hydraulic variables related to overtopping depth, overtopping duration, and duration of levee loading prior to breach can generally be estimated using known river gage information for levees within the United States. Furthermore, visual inspection prior to, or during, the event can often provide enough insight into the event that magnitude of these variables can be estimated. Geometric factors such as levee height and slope steepness were found to be minimally impactful to the overall model, which was not an expected result. A key benefit of the model presented within this study, is that a relatively small number of variables are required. This reduces the complexity of interactions within the model and eliminates some potential factors that may have highly variable effects. With that being said, further studies are suggested to examine whether the inclusion of additional variables could be beneficial in potential future model updating where the variables are easily reasonably attained and do not

lead to unnecessary model complexity. For example, the proposed model does not explicitly consider the effect of slope vegetation or overtopping velocities, which are shown affect the mechanism of soil external erosion and levee overtopping breach. Considering such variables, if available, can potentially improve the model accuracy but would need to be accurately assessed in future studies.

In addition to establishing a repository of levee performance data and establishing baseline failure rates, the LLID effort was initiated to create statistical inferences based on the data using such tools as the proposed model. Levee overtopping breach probability is typically considered through the event tree analysis which defines the process as being the product of subjective probabilities elicited for each individual node occurring on the event tree. For overtopping, the nodes on the event tree typically are a combination of (1) a hydraulic event occurring, (2) initiation of embankment erosion, (3) propagation of embankment erosion progressing beyond a critical state, and (4) breach. The proposed model represents steps (2) through (4) of this process. So, the application of model results should be treated as a factor which, when multiplied by the statistical probability of a flood event causing overtopping occurring in a given year, yields an annual probability of breach. The results of risk assessments are typically documented for each system that has undergone screening or assessment within that organization's portfolio. Overtopping breach probabilities can be used in the field of risk assessment portfolio management to review previous risk estimates based on the results of this model and additional known performance information.

CHAPTER IV

RISK ASSESSMENT OF LEVEE OVERTOPPING BREACH RISK USING A LOGIT MODEL

This chapter has been submitted for publication in the proceedings of Geo-Congress 2022. The paper has been reformatted and replicated herein with minor modifications in order to outfit the purposes of this thesis.

4.1 Introduction

More than 100,000 miles of levee protect the inhabitants and economies located within floodplains across the United States (CRS 2017). Breach due to overtopping has been, and continues to be, the most common mode of failure for earthen levees (Hui et al. 2016; USACE 2018). Breach can be defined as the levee failing to restrict the passage of water, allowing flow to pass through the embankment and inundate the leveed area (USACE 2018). When a levee breaches, the result can include significant life safety and economic consequences. However, when breach does not occur, the consequences related to life safety and economic loss are typically reduced. Thus, it is evident that understanding the incremental difference between these types of overtopping events, in an effort to mitigate against an ever-present risk, is a worthwhile endeavor.

With an increased understanding of what factors contribute to breach when a levee is overtopped, levee risk can be better managed. Statistical models have the potential to be

invaluable in assisting engineers in predicting the probability of levee breach more reliably as a function of controlling parameters. Statistical models as predictive tools have become a key component in developing the risk assessment framework for numerous infrastructure systems as a result of advances in computational efficiency and predictive capability (Rahimi et al. 2019; Zamanian et al. 2020; Dehghani et al. 2021). Predictive models have been employed with success to evaluate the performance of levees under various loading conditions, including overtopping (Uno et al. 1987, 1994; Heyer et al. 2010; Heyer and Stamm 2013; Vahedifard et al. 2017, 2020; Balistrocchi et al. 2019). This progress, combined with on-going efforts to implement probabilistic methods in geotechnical engineering, continues to expand the toolbox of the practicing engineer by allowing for the assessment of levee performance and flood risk through informed data-driven analyses which support improved decision making.

As levee performance data become more readily available, the quality of models for describing the risk of levee breach due to overtopping should increase. The primary objective of this study is to apply a recently developed probabilistic model for overtopping breach in support of levee risk assessment. The probabilistic model is introduced and compared against the results of 8 documented levee risk assessments based on the elicitation of expert opinion. The model is then further validated using a set of eleven levee overtopping breach events not previously included in model development.

4.2 Levee Overtopping Performance Logistic Regression Model

A logistic regression model, referred to henceforth as the “logit model”, was recently developed by Flynn et al. (2021b) with the purpose of estimating the probability of levee breach given overtopping based on a performance dataset of 185 historical riverine levee overtopping events. For completeness, the key features of the logit model proposed by Flynn et al. (2021b)

are encapsulated in this section. The logistic regression model serves to assess the degree of importance of each parameter that contributes to the outcome, culminating in a relationship that predicts the behavior of a multivariate system. In this model, levee breach is defined as Y=1, and non-breach as Y=0. The probability of breach, P(Y=1), is defined by the following general form:

$$P(Y = 1) = \frac{e^Z}{1 + e^Z} \tag{4.1}$$

where

$$Z = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{4.2}$$

and X_i is the observation which, when utilizing the logit model, is the categorical level of an input variable. β_0 and β_i are the coefficients estimated by the regression model.

This method of statistical analysis is appropriate for levee overtopping since the contributing parameters are generally intuitive and the result of interest in an overtopping event is binary, specifically in terms of breach versus non-breach. The probability of breach as determined by this model utilizes a threshold to determine if the individual event is likely to result in breach since the model output is in terms of probability, i.e. between 0.0 and 1.0. Breach is assumed to occur if $P(Y = 1) > 0.5$, with non-breach occurring if $P(Y = 1) < 0.5$.

The logit model attempts to predict the probability of levee breach given a set five categorical variables related to construction history and geotechnical and hydraulic factors. These variables are denoted as X_3 through X_7, each of which having either two or three categorical levels as shown in Table 4.1. Table 4.2 shows the material description for erosion

resistance classification. In initial iterations of the model, two other variables, X_1 and X_2, were considered, however it was determined that these variables were not significant in predicting the model outcome, thus were excluded. The logit model relationship is shown in Eq. 3, which considers the base interactions of X_3, X_4, X_5, X_6 and X_7 and two-way interactions of X_4 with X_7 and X_3 with X_6. Coefficients for the given relationship are presented in Table 4.3, along with the odds ratio for each input parameter.

$$\begin{aligned} Z = \beta_0 &+ \beta_3.X_{X_3=2} + \beta_4.X_{X_4=2} + \beta_5.X_{X_4=3} + \beta_6.X_{X_5=2} + \beta_7.X_{X_5=3} \\ &+ \beta_8.X_{X_6=2} + \beta_9.X_{X_6=3} + \beta_{10}.X_{X_7=2} + \beta_{11}.X_{X_7=3} \\ &+ \beta_{12}.X_{X_4=2,X_7=2} + \beta_{13}.X_{X_4=3,X_7=2} + \beta_{14}.X_{X_4=2,X_7=3} \\ &+ \beta_{15}.X_{X_4=3,X_7=3} + \beta_{16}.X_{X_3=2,X_6=2} + \beta_{17}.X_{X_3=2,X_6=3} \end{aligned} \quad (4.3)$$

Table 4.1 Summary of Logit Model Variables

Code	Variable	Type	Level Code	Level Description
X_3	Levee Construction Entity	Categorical	1 2	Local Federal
X_4	Water Depth Over Levee	Categorical	1 2 3	< 0.5 ft 0.5 ft - 1 ft > 1 ft
X_5	Duration of Overtopping Flow Prior to Breach	Categorical	1 2 3	<6 hours 6-24 hours >24 hours
X_6[1]	Erosion Resistance Classification	Categorical	1 2 3	Low Moderate High
X_7	Duration of Levee Loading Prior to Overtopping	Categorical	1 2 3	<3 day 3-14 day >14 day

[1]See **Table 4.2** for description of material descriptions used for erosion resistance classification.

Table 4.2 Material Description for Erosion Resistance Classification

Low Erosion Resistance	Moderate Erosion Resistance	High Erosion Resistance
• sand • silty sand • silty sand with gravel • sand/silt mix • sand/gravel mix • sand/gravel mix with silt • sandy silt • sandy gravel	• silt • clayey silt • silt with sand/clay • silt/clay mix with sand • silty loam • silty/clayey loam • sand/silt mix with clay • clayey sand	• clay • clay/silt mix • clay with sand/silt • zoned embankment with impervious cover • clay enlargement of an existing sand levee

Table 4.3 Variable Odds Ratio and Coefficients of Logit Model

Variable	Coefficient (β)	Odds ratio
Intercept	$\beta_0 = 0.93$	2.53
Levee Construction Entity (X_3) = 2	$\beta_3 = -1.13$	0.32
Water Depth Over Levee (X_4) = 2	$\beta_4 = 0.87$	2.38
Water Depth Over Levee (X_4) = 3	$\beta_5 = 1.27$	3.55
Duration of Overtopping Flow Prior to Breach (X_5) = 2	$\beta_6 = 0.08$	1.08
Duration of Overtopping Flow Prior to Breach (X_5) = 3	$\beta_7 = 1.85$	6.37
Erosion Resistance Classification (X_6) = 2	$\beta_8 = -1.74$	0.18
Erosion Resistance Classification (X_6) = 3	$\beta_9 = -3.93$	0.02
Duration of Levee Loading Prior to Overtopping (X_7) = 2	$\beta_{10} = -0.42$	0.66
Duration of Levee Loading Prior to Overtopping (X_7) = 3	$\beta_{11} = 0.17$	1.18
X_4 = 2 and X_7 = 2	$\beta_{12} = 2.18$	8.82
X_4 = 3 and X_7= 2	$\beta_{13} = 2.65$	14.08
X_4 = 2 and X_7 = 3	$\beta_{14} = 0.60$	1.83
X_4 = 3 and X_7= 3	$\beta_{15} = 2.29$	9.89
X_3 = 2 and X_6 = 2	$\beta_{16} = 1.35$	3.84
X_3 = 2 and X_6 = 3	$\beta_{17} = -0.51$	0.60

4.3 Risk Assessment of Levee Breach Due to Overtopping

Risk to levee systems is defined as a function of hazard, performance, and consequences (USACE 2018). In the framework of levee risk assessment, the hazard is the environmental load on the levee system such as flood or earthquake conditions. Performance refers specifically to the resiliency of the levee against the hazard, or how the hazard will affect a levee's ability to contain flooding. Finally, consequences can be described as potential loss in terms of population and economic assets that result from combined effects of hazards and unsatisfactory levee performance.

Levee risk assessments are commonly conducted utilizing a method referred to as semi-quantitative risk assessment (SQRA) within engineering practice. An SQRA is defined as a risk assessment which utilizes both numerical estimates and qualitative descriptions that result in order of magnitude risk estimates (USACE, 2019). This method of risk assessment allows for the overall asset management of levee infrastructure portfolios, as well as in-depth evaluation of individual levee systems.

When conducting a levee SQRA, the first step is to establish potential failure modes. Each potential failure mode is analyzed using an event tree which considers the successive probabilities of individual events, or nodes, that must occur sequentially such that the result is failure of the embankment. Within the SQRA framework, the first node is typically the hydrologic or seismic event that leads to levee distress. In the case of hydrologic loading, this first node is represented quantitatively as the annual exceedance probability (AEP) of the hydraulic load on the levee. The AEP establishes the risk baseline for a failure mode, as it is a fixed probability of flood frequency determined through statistical analysis and hydrologic modeling. Subsequent nodes on the event tree are assigned subjective probabilities elicited from

a risk team based on circumstantial evidence and engineering judgement (O'Leary 2018). These nodes account for initiation and progression of the failure mode to the point of failure. The product of the nodes on an event tree yields the estimated annual probability of failure (APF), which is typically represented as an order of magnitude estimate to account for uncertainty in the expert elicitation of nodal probability values. This relationship can be represented as:

$$\text{APF} = AEP \times \Pi_{i=1}^{n} P(i) \tag{4.4}$$

where P(i) = probability of individual node occurring.

When calculating the annual probability of failure of levee breach due to overtopping, the nodes on the event tree are generally a combination of (1) a hydraulic event occurring which leads to the depth of overtopping being considered, (2) initiation of embankment erosion, (3) progression of embankment erosion beyond a critical state, and (4) widespread levee breach, where node one is the AEP and subsequent nodes describe the initiation and continuation of failure. This process is demonstrated in Fig. 4.1, with an example event tree shown in Fig. 4.2.

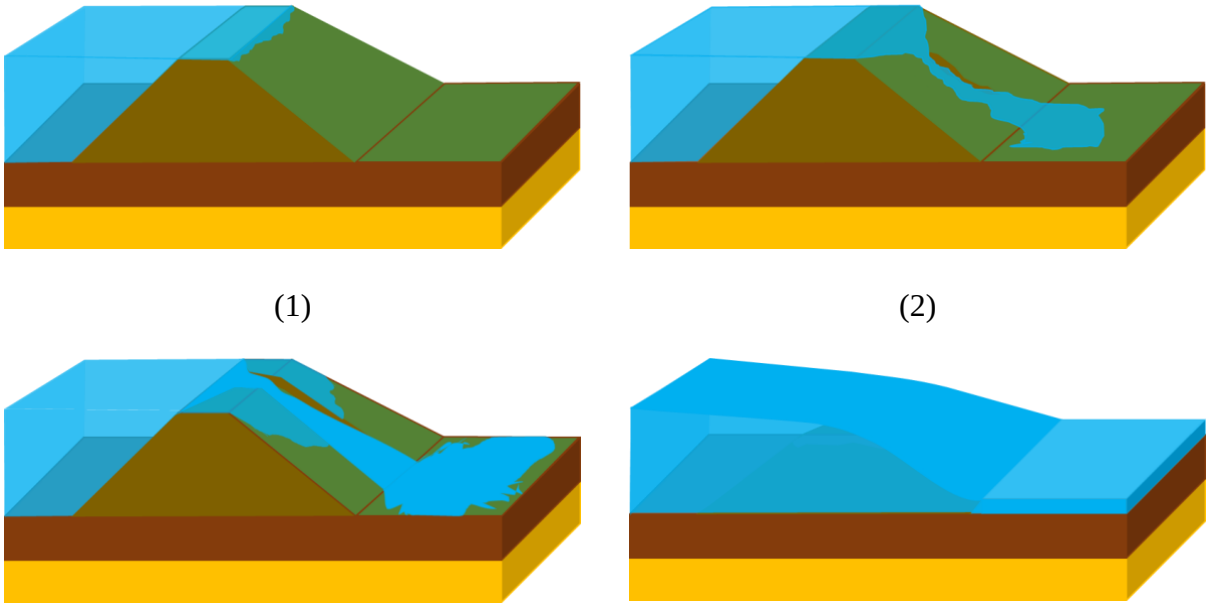

Figure 4.1 Levee Overtopping Breach Development

The proposed logit model provides a data driven tool for assisting in the assessment of risk related to levee overtopping and offers a viable alternative to the subjective, judgment-based component of the levee SQRA process. Implementation of the logit model should not seek to replace the SQRA framework, but rather serve as a supplementary tool which merits consideration when addressing risk related to levee overtopping. Utilization of the logit model does not account for a total assessment of risk, as consequences are not a component of the model. Rather, the model serves as a method for calculating the annual probability of failure, when combined with annual exceedance probability of overtopping by a given depth, as further expanded upon in subsequent discussion.

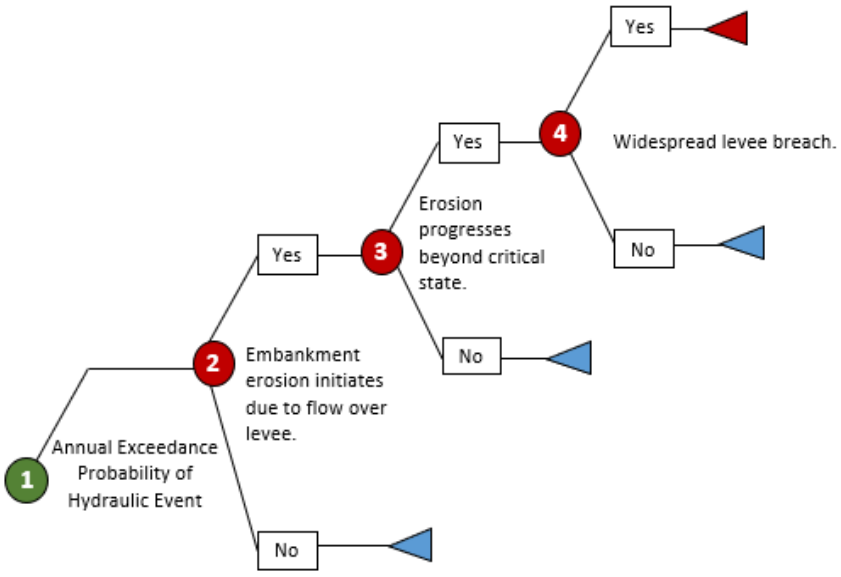

Figure 4.2 Event Tree for Levee Overtopping

4.4 Comparison of Logit Model Versus Semi-Quantitative Methods

The logit model represents steps (2) through (4) of the process described in Fig. 4.2, which can also be described as the system response probability (SRP) of breach. The SRP in this study is equal to $P(Y=1)_{Logit}$ where defined by the logit model, and $P(Y=1)_{SQRA}$ where back-calculated using SQRA results. The results of eight SQRAs, which considered riverine levee breach due to overtopping were evaluated for coherence with the logit model, with data and assumptions from the SQRA re-binned to fit the categorical levels of the logit model variables. Specific identifying system information, such as name and location were withheld from the presented data as this information may be considered sensitive.

The first step in evaluating the SQRA data was to feed the data from the risk assessment into the logit model in order to generate the calculated probability of breach, $P(Y=1)_{Logit}$. The input data and logit model SRP estimates are shown in Table 4.4. As can be

seen in Table 4.4, the SQRA data includes a well distributed range of potential model conditions, with each variable category represented.

Table 4.4 Logit Model Calculations Utilizing SQRA Data

Risk Assessment ID	Logit Model Inputs[1]					$P(Y=1)_{Logit}$[2]
	X_3	X_4	X_5	X_6	X_7	
1	Federal	> 1 ft	>24 hr	High	>14 days	0.421
2	Federal	>1 ft	6-24hr	High	3-14 days	0.366
3	Federal	<0.5 ft	<6 hr	Low	3-14 days	0.350
4	Federal	0.5-1 ft	<6 hr	High	<3 days	0.023
5	Local	<0.5 ft	6-24 hr	High	3-14 days	0.034
6	Local	>1 ft	>24 hr	High	3-14 days	0.913
7	Local	>1 ft	>24 hr	High	3-14 days	0.913
8	Local	>1 ft	>24 hr	Moderate	3-14 days	0.989

[1]Logit model inputs described in "LOGISTIC REGRESSION MODEL FOR LEVEE OVERTOPPING"
[2]Probability of breach given overtopping, as calculated by the logit model.

The next step in evaluating the SQRA results was to back-calculate the system response probability of breach estimated by the SQRA, or $P(Y=1)_{SQRA}$. To compare the logit model to SQRA results, two components of the SQRA risk estimate need to be known, the annual exceedance probability and annual probability of failure. Since the system response probability of breach calculated by the logit model represents the nodes on the event tree that occur after the AEP is established, the calculated probability of breach can be compared to the SQRA estimated value using:

$$P(Y=1)_{SQRA} = SRP = \frac{APF}{AEP} \quad (4.5)$$

Results of the system response probability comparisons between the logit model and SQRA results are presented in Table 4.5 and Fig. 4.3. While the logit model is a quantitative point estimate of APF, the SQRA APF shown in Table 4.5 is the geometric mean of the range

presented in the risk assessment. Therefore, the SQRA APF could be assumed to have an uncertainty bound one half order of magnitude greater than, or less than, the value shown. For two of the cases, this led to a $P(Y=1)_{SQRA}$ value of 1.000 for Risk Assessments 1 and 3. For Risk Assessments 6, 7 and 8, the $P(Y=1)_{SQRA}$ was back-calculated to be 1.5, which is does not fit the logic of the logit model. Given that the AEP for these events was 2.0×10^{-3}, the APF as a point estimate could not exceed 2.0×10^{-3}. This is due to the fact that the risk estimates are presented as the geometric mean of the order of magnitude range. It is assumed that the SRP for each of these events ranged between 0.989 and 1.000 based on the uncertainty bounds of the order of magnitude range. To account for this, the $P(Y=1)_{SQRA}$ is presented as 1.000 for these cases, which is the theoretical maximum when treating the APF as a point estimate. For most cases, the two methods had reasonable agreement. However, with the maximum order of magnitude difference being up to 0.910 in one case, interpretation of the risk could result in significantly different action.

Table 4.5 Comparison of Results Using Logit Model Versus SQRA

Risk Assessment ID	AEP of Hydraulic Event[1]	$P(Y=1)_{Logit}$	Logit Model APF	$P(Y=1)_{SQRA}$[2]	SQRA APF[3]	Order of Magnitude Difference[4,5]
1	1.0×10^{-4}	0.421	8.4×10^{-5}	1.000	1.0×10^{-4}	0.076
2	6.7×10^{-4}	0.366	2.4×10^{-4}	0.045	3.0×10^{-5}	-0.910
3	2.0×10^{-4}	0.350	7.0×10^{-5}	1.000	2.0×10^{-4}	0.456
4	3.3×10^{-4}	0.023	7.5×10^{-5}	0.030	1.0×10^{-4}	0.125
5	2.0×10^{-3}	0.034	6.8×10^{-5}	0.015	3.0×10^{-5}	-0.358
6	2.0×10^{-3}	0.913	1.8×10^{-3}	*1.000*	3.0×10^{-3}	0.222
7	2.0×10^{-3}	0.913	1.8×10^{-3}	*1.000*	3.0×10^{-3}	0.222
8	2.0×10^{-3}	0.989	2.0×10^{-3}	*1.000*	3.0×10^{-3}	0.180

[1]AEP = Annual Exceedance Probability
[2]P(Y=1) calculated based on given AEP and APF presented in SQRA.
[3]Annual Probability of Failure as estimated through expert elicitation.
[4]Order of magnitude calculated by comparing base 10 logarithm differences between APF estimates.
[5]A negative order of magnitude difference indicates that the logit model produced a higher APF than the SQRA.

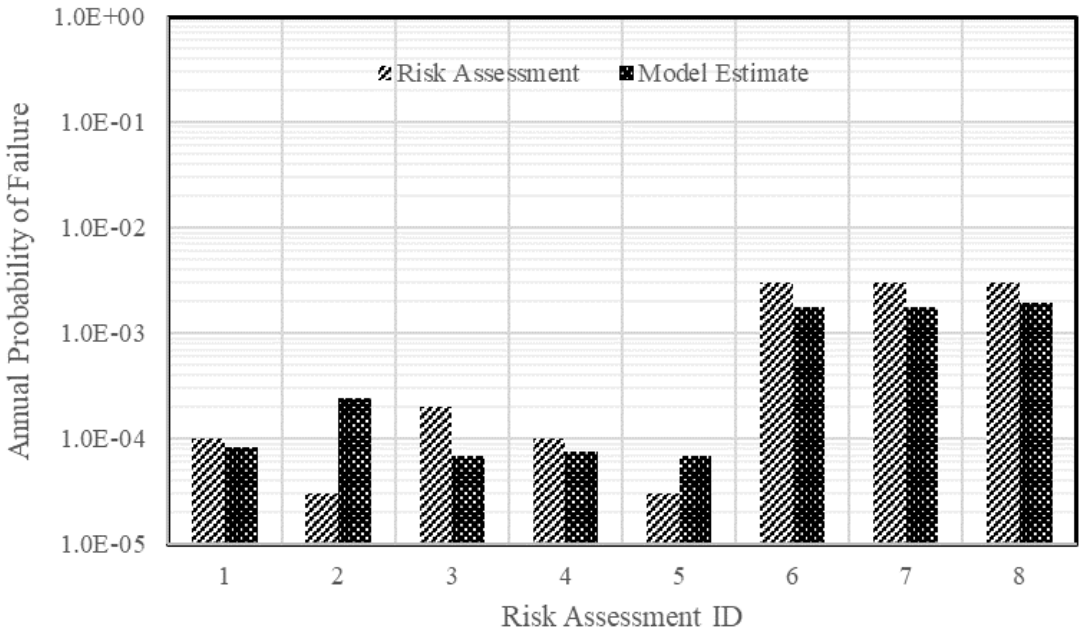

Figure 4.3 Comparison of Annual Probability of Failure from Logit Model and SQRA

4.5 Validation of Logit Model Using New Data

Semi- Eleven new events were established for testing of the logit model. These events include data received after the logit model was developed, as well as events previously excluded due to lack of information that have since been refined. The new data are presented in Table 4.6, with data described in the categorical ranges required by the logit model. New data account for variables that describe each categorical level and include 5 breach events and 6 non-breach events. Logit model predictions for the breach probability of new data range from 0.016 to 0.992, therefore representing nearly the entire range of breach probabilities that can be calculated by the model.

Table 4.6 New Overtopping Event Data

Event Number	Levee Segment	Year	Logit Model Inputs[1]					
			X_3	X_4	X_5	X_6	X_7	Y^2
1	Cherry Valley	2007	Local	0.5-1 ft	6-24 hr	Low	<3 days	1
2	Ste. Genevieve	2015	Local	0.5-1 ft	<6 hr	Moderate	3-14 days	1
3	Eel River	1964	Local	0.5-1 ft	6-24 hr	Low	<3 days	1
4	Bean Lake	2007	Local	< 0.5 ft	<6 hr	Moderate	<3 days	0
5	Plowboy Bend	1995	Local	> 1 ft	>24 hr	Moderate	3-14 days	0
6	Plowboy Bend	2019	Local	> 1 ft	>24 hr	Moderate	>14 days	0
7	MRLS L-575 B-W	2019	Federal	0.5-1 ft	6-24 hr	Moderate	3-14 days	1
8	Hunt-Lima (Bear Creek)	2019	Federal	< 0.5 ft	<6 hr	High	>14 days	0
9	NSA Big Creek	2010	Federal	< 0.5 ft	<6 hr	High	<3 days	0
10	Henry Pohl	2011	Local	< 0.5 ft	6-24 hr	Moderate	>14 days	0
11	Henry Pohl	2019	Local	0.5-1 ft	6-24 hr	Moderate	3-14 days	1

[1]Logit model inputs described in "LOGISTIC REGRESSION MODEL FOR LEVEE OVERTOPPING"
[2]A value of 1 refers to breach and 0 refers to non-breach.

The predictive accuracy of the logit model was 81.8% when computing the expected results of the 11 new data points. This result is comparable to the k-fold cross validation accuracy of 80.7% and exceeds the test data accuracy of 73.3% previously presented by Flynn et al. (2021b). Fig. 4.4 presents the calculated breach probability of all new events when assessed using the logit model with respect to the event number. The results of the newly implemented test data show that there is a tendency of the model to predict a false positive, which is a conservative result. The probability of false positive for the new data was 18.2%, accounting for all incorrectly estimated data. The model correctly predicted the observed result for all documented breach events.

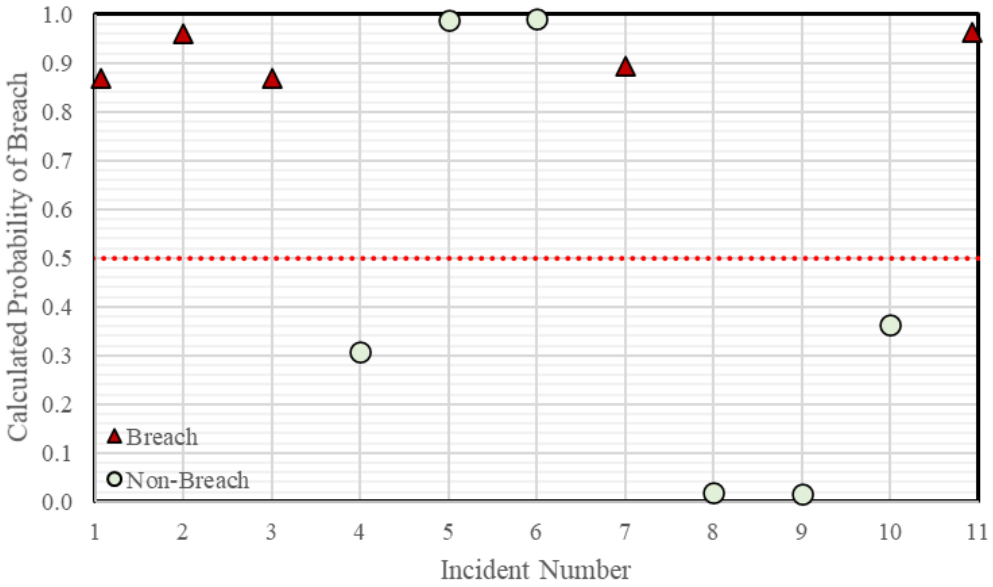

Figure 4.4 Calculated probability of breach for 11 new overtopping incidents

Events 5 and 6 are of particular interest in this dataset. The intuitive reaction to seeing the conditions of these events would lead an experienced engineer to assume these events would likely result in breach given substantial overtopping depths and durations observed. Following this logic, the logit model assumes a probability of breach for each of these events to exceed 0.988. However, these events did not result in breach. This is just one example of the variability of the levee overtopping that leads to difficulty in predicting performance with absolute certainty.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

5.1 Conclusions from Chapter II

The Levee Loading and Incident Dataset (LLID)'s overtopping subset can best serve to inform risk assessment of overtopping failure modes for levees both in the design and post-construction phases of a project. Contained within the presented analysis are just two general trends derived from the overtopping subset Levee Loading and Incident Dataset which serve as insight into valuable correlations that can be derived from future efforts to analyze the LLID.

Overtopping breach probability based on construction and maintenance designation and relative erosion resistance for a limited dataset is presented as an example of data that can be used when assigning likelihoods of breach occurrence given levee overtopping.

Results presented in this study synthesize a portion data contained within the overtopping data set of the LLID only, and would benefit from additional data collection to further define the existing overtopping events within the dataset, as well as to expand the dataset to include additional, previously undocumented overtopping events. With further data collection and refinement, greater confidence can be placed in the various subsets within the LLID, including the overtopping dataset analysis, which will allow for more guided use in the levee design and risk assessment processes.

5.2 Conclusions from Chapter III

Flood risk within the United States poses an existing and increasing threat due to the increased frequency of extreme weather events and expanding development within floodplains. It is imperative that risk associated with levees are thoroughly evaluated and regularly assessed. Utilizing available tools and evolving methods, levee data collection should be prioritized such that levee performance can be better understood through both statistical and deterministic means.

A robust dataset is presented within this study that documents the performance of known overtopping events within the United States. The dataset, along with any future data additions, has the potential to allow for further refinement or expansion of numerical values and categorical levels such that variables could be considered more acutely.

The proposed model uses known ranges of performance information in an effort to help calibrate the minds of those who elicit risk for levee systems. One significant advantage of the proposed model is that it can be updated as additional data is collected and refined. The proposed model can be used by many with a basic understanding of the limited range of inputs that the proposed model requires.

The current study yields promising results when utilizing the proposed model as a screening measure given then extensive work that has gone into data collection. An accuracy of 80.7% indicates quality model fitting of data for a process that is inherently variable. Therefore, the method presented within this study can be used to create screening tools which help guide the assessment of levee overtopping breach risk.

Continued data collection efforts should consider data which furthers the understanding of breach due to overtopping in a way that not only can predict its occurrence but also its magnitude. It is highly recommended that unified data collection efforts be considered that

document levee performance, not only of USACE levees but also of those within the United States and abroad which are maintained by other entities. In doing so, models such as the one presented in this study can gain efficiency. An increased understanding of levee performance during flood events serves to benefit those who maintain, design, and create policy for flood risk management projects.

5.3 Conclusions from Chapter IV

An evaluation of the applicability of the logit model to levee risk assessment has been presented along with model validation utilizing new overtopping event data. Comparison of model results with semi-quantitative risk assessments showed a correlation in expected results in most cases. However, there were instances where differences were on the order of one-half to one full order magnitude. One order of magnitude in a risk assessment could potentially lead to differentiating plans of addressing risk. So, differences in assumptions should be considered when evaluating risk for the cases where the methods result in large differences. In risk assessments where the subjective elicitation of breach probability due to overtopping is simplified, or grossly estimated, the logit model is viewed as an improved method that more carefully considers a range of factors.

The evaluated data indicated that accuracy of the logit model in predicting system response is maintained, with the new dataset showing slightly increased accuracy. The logit model correctly predicted the results of new overtopping data at a rate of 81.8%. Therefore, the logit model should be considered a reliable tool for risk assessment of levee breach due to overtopping, as well as calibrating subjective risk elicitation.

It is recommended that continued efforts be made in the areas of data collection, with the goal of a unified repository for levee performance information. As predictive models continue to

develop and improve based on new information and expanded capability, it is critical that they are periodically tested and adjusted where necessary. It has been previously stated that geotechnical engineering is necessarily Bayesian (Christian, 2004). This is particularly true for the given logit model in that, even though it is created with traditional statistical methods, the model can be updated and improved with new data in an effort to better understand the phenomenon of levee overtopping.

5.4 Recommendations for future work

Levee performance has been a popular topic of research, primarily over the past two decades. Many theoretical models have been introduced that consider both component-based and overall system performance. However, the application of these studies has not taken hold in practice on a widespread scale. The work presented in this study is a just one step towards making probabilistic assessment of levee performance and risk an endeavor that can be made by the practicing engineer. The model presented is an effective screening tool that should be continued to be developed as more data becomes available and additional research is made regarding critical levee performance factors. Working to build on this effort is critical to the progress of risk based levee assessment. Some recommendations for further research are listed below:

- Revisit the dataset used to create the levee overtopping logit model and update any missing hydraulic and hydrologic data where possible.
- Consider the addition of relevant variables such as overtopping velocity and slope vegetation
- Apply the model methodology to similar, non-USACE databases that document levee performance.

- Apply the model methodology to additional failure modes contained within the Levee Loading and Incident Database to include internal erosion and embankment instability.

- Create a levee overtopping model specific to coastal and canal levee events.

- Undertake of an effort to collect and compile data from all known levee overtopping events globally by merging known, available datasets.

REFERENCES

Amabile, A., Cordao-Neto, M., De Polo, F., Taratino, A., (2016). "Reliability analysis of flood embankments taking into account a stochastic distribution of hydraulic loading." *E-UNSAT 2016*. (9). DOI: 10.1051/e3sconf/20160919005

American Society of Civil Engineers (ASCE), (2021). "A Comprehensive Assessment of America's Infrastructure." 75-82. https://infrastructurereportcard.org/wp-content/uploads/2020/12/National_IRC_2021-report.pdf

Balistrocchi, M., Moretti, G., Orlandini, S., and Ranzi, R. (2019). "Copula-based modeling of earthen levee breach due to overtopping." *Advances in Water Resources*, 134. 10.1016/j.advwatres.2019.103433

Beretta, L., Santaniello, A. (2016). "Nearest neighbor imputation algorithms: a critical evaluation." *BMC Medical Informatics and Decision Making*. 16 (Suppl 3).

Briaud, J. L., Chen H. C., Govindasamy, A. V., and Storesund, R. (2008). "Levee erosion by overtopping in New Orleans during the Katrina Hurricane." *Journal of Geotechnical and Geoenvironmental Engineering, ASCE*, 134: 618-632.

Christian, J. (2004). "Geotechnical Engineering Reliability: How Well Do We Know What We Are Doing?" *Journal of Geotechnical and Geoenvironmental Engineering,* 130(10), 985-1003, DOI: 10.1061/(ASCE)1090-0241(2004)130:10(985)

Congressional Research Service (CRS). (2017). "Levee Safety and Risk: Status and Considerations." Version 3. IF10788. https://crsreports.congress.gov/

Danka, J., and L. Zhang. (2015). "Dike Failure Mechanisms and Breaching Parameters." *Journal of Geotechnical and Geoenvironmental Engineering*. 141 (9): 04015039.

Das, I., Sahoo, S., van Westen, C., Stein, A., Hack, R. (2010). "Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classification system, along a road section in the northern Himalayas (India)." *Geomorphology*. 114(4). 627-637.

Dehghani, N. L., Jeddi, A. B., & Shafieezadeh, A. (2021). Intelligent hurricane resilience enhancement of power distribution systems via deep reinforcement learning. *Applied Energy*, *285*, 116355.

Delgado R., Tibau X-A. (2019). "Why Cohen's *Kappa* should be avoided as performance measure in classification." *PLoS ONE* 14(9): e0222916. https://doi.org/10.1371/journal.pone.0222916

Di Leo, G., Sardanelli, F. (2020). "Statistical significance: *p* value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach." *European Radiology Experimental* 18(4) https://doi.org/10.1186/s41747-020-0145-y

D'Orazio, M., (2021). Distances with mixed type variables some modified Gower's coefficients. *arXiv* preprint arXiv:2101.02481.

Ellithy, S., Savant, G., Wibowo, J. (2017). "Effect of Soil Mix on Overtopping Erosion." *World Environmental and Water Resources Congress 2017*. 34-49.

Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd. ISBN 0-050-02170-2.

Flynn, S., Vahedifard, F., Schaaf, D. (2021a). "A Dataset of Levee Overtopping Events" *Accepted by ASCE GeoExtreme 2021* (pending publication).

Flynn, S., Zamanian, S., Vahedifard, F., Shafieezadeh, A., Schaaf, D. (2021b). "Data-Driven Model for Probability of Levee Breach Due to Overtopping" *Journal of Geotechnical and Geoenvironmental Engineering, ASCE.* (In Review).

Gandomi, A., Fridline, M., Roke, D. (2013). "Decision Tree Approach for Soil Liquefaction Assessment." *The Scientific World Journal*. 2013 Dec 30. DOI: 10.1155/2013/346285

Gui, S., Zhang, R., Xue, X. (1998). "Overtopping Reliability Models for River Levee" *Journal of Hydraulic Engineering, ASCE*, 124(12): 1227-1234.

Heyer, T., and J. Stamm. (2013). "Levee Reliability Analysis Using Logistic Regression Models, Abilities, Limitations and Practical Considerations." *Georisk* 7 (2): 77–87. DOI:10.1080/17499518.2013.790734.

Heyer, T., H. B. Horlacher, and J. Stamm. (2010). "Multicriteria Stability Analysis of River Embankments based on Past Experience." In Proceedings, First European IAHR Congress, Edinburgh, UK.

Heyer, T. (2016). "Reliability assessment of levees based on failure investigations" *Vodohospodářské technicko-ekonomické informace*, **58**(3), 28– 33.

Hilbe, J. (2011). "Logistic Regression" *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg. 755-758.

Hui, R., Jachens, E., Lund, J,. 2016. "Risk-based planning analysis for a single levee". *Water Resources Research,* 52(4), 2513-2528.

Isola, M., Caporali, E., Garrote, L. (2020). "River Levee Overtopping: A Bivariate Methodology for Hydrological Characterization of Overtopping Failure". *Journal of Hydrological Engineering, ASCE,* 25(6).

Jasim, F. H., Vahedifard, F., Alborzi, A., Moftakhari, H., & Aghakouchak, A. (2020). Effect of compound flooding on performance of earthen levees. In *Geo-Congress 2020: Engineering, Monitoring, and Management of Geotechnical Infrastructure* (pp. 707-716). Reston, VA: American Society of Civil Engineers.

Lendering, K., Schweckendiek, T., Kok, M. (2018). "Quantifying the failure probability of a canal levee" *Georisk*. 12(3), 203-217. DOI: 10.1080/17499518.2018.1426865

Kamalzare, M., Han, T. S., McMullan, M., Stuetzle, C., Zimmie, T. F., Cutler, B., & Franklin, W. R. (2013). Computer simulation of levee erosion and overtopping. In *Geo-Congress 2013: Stability and Performance of Slopes and Embankments III* (pp. 1851-1860).

Kleinbaum,D. G. (1994). Logistic Regression: A Self-Learning Text. New York: Springer-Verlag.

Kowarik, A. and Templ, M., (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(7), pp.1-16.

Mallakpour, I. and Villarini, G. (2015). "The changing nature of flooding across the central United States" *Nature Climate Change Letters*, Macmillan, Vol. 5, 250-254

O'Leary, T. (2018). "SQRA Calculation Methodology, RMC-TN-2018-01." United States Army Corps of Engineers, Institute for Water Resources, Risk Management Center, Lakewood, Colorado.

Osouli, A., Bahri, P. (2018). "Erosion Rate Prediction Model for Levee-Floodwall Overtopping Applications in Fine-Grained Soils" *Geotechnical and Geological Engineering*. Springer. 36. 2823-2838.

Ozer, I.E., van Damme, M., Jonkman, N., (2020). "Towards an International Levee Performance Database (ILPD) and Its Use for Macro-Scale Analysis of Levee Breaches and Failures" *Water 2020. 12., 19.*

Rahimi, M., Wang, Z., Shafieezadeh, A., Wood, D. and Kubatko, E.J., (2020). "Exploring passive and active metamodeling-based reliability analysis methods for soil slopes: a new approach to active training." *International Journal of Geomechanics*, *20*(3), p.04020009.Ripley, B.,

Reclamation-USACE. (2015). "Best practices in dam and levee safety risk analysis. U.S. Bureau of Reclamation, Denver, CO. http://www.usbr.gov/ssle/damsafety/risk/methodology.html

Seed, R.B., et al. (2008). "New Orleans and Hurricane Katrina. I: Introduction, Overview, and the East Flank." *Journal of Geotechnical and Geoenvironmental Engineering, ASCE*, 134: 701-717.

Seed, R. B., et al. (2005). “Preliminary report on the performance of the New Orleans levee system in Hurricane Katrina on August 29, 2005.” *Rep. No. UCB/CITRIS-05/01*, National Science Foundation.

Sills, G. L., Vroman, N. D., Wahl, R. E., and Shwanz, N. T. (2008). "Overview of New Orleans Levee Failures: Lessons Learned and Their Impact on National Levee Design and Assessment." *Journal of Geotechnical and Geoenvironmental Engineering, ASCE*, 134(5): 556-565.

Sharp, J., McAnally, W. (2011). “Numerical Modeling of surge overtopping of a levee.” *Applied Mathematical Modelling*. 36. 1359-1370. https://doi.org/10.1016/S0895-4356(99)00103-1.

Steyerberg, E., Eijkemans, M., Habbema, F. (1999). “Stepwise Selection in Small Data Sets: A Simulation Study of Bias in Logistic Regression Analysis.” *Journal of Clinical Epidemiology* 51(10). 935-942.

Templ, M., Alfons, A., Kowarik, A., Prantner, B. and Templ, M.M., (2020). Package ‘VIM’.

Ubilla, J., Abdoun, T., Sasanakul, I., Sharp, M., Steedman, S., Vanadit-Ellis, W., Zimmie, T., (2008). “New Orleans Levee System Performance during Hurricane Katrina: London Avenue and Orleans Canal South.” *Journal of Geotechnical and Geoenvironmental Engineering, ASCE*, 134: 668-680.

Uno, T., H. Morisugi,T. Sugii,and K. Ohashi. (1987). “Application of a Logit Model to Stability Evaluation of River Levees.” Journal of Natural Disaster Science 9(1): 61_77.

Uno, T., Sugii, T., Hayashi, M., (1994). “Logit model for river levee stability evaluation considering the flood return period” *Structural Safety*. 14, 81-102.

U.S. Army Corps of Engineers. (2021). Climate Hydrology Assessment Tool v1.0. United States Army Corps of Engineers, Web. <https://corpsmapz.usace.army.mil/apex/f?p=313:2:0::NO>

U.S. Army Corps of Engineers. (2000). “Design and Construction of Levees.” United States Army Corps of Engineers, Engineer Manual, EM 1110-2-1913.

U.S. Army Corps of Engineers. (2019). “Interim Approach for Risk-Informed Designs for Dam and Levee Projects.” United States Army Corps of Engineers, Engineering and Construction Bulletin, ECB No. 2019-15

U.S. Army Corps of Engineers. (2018). “Levee Portfolio Report” Levee Safety Program, Headquarters, Washington, D.C.

U.S. Army Corps of Engineers. (2021). “National Levee Database” Website. Headquarters, Washington, D.C. < https://levees.sec.usace.army.mil/#/>

U.S. Army Corps of Engineers. (2020) “Where We Are” Website. Headquarters, Washington, D.C. https://www.usace.army.mil/locations.aspx

US Global Change Research Program (USGCRP). (2018). “Fourth national climate assessment.” Accessed October 8, 2019. <https://nca2018 .globalchange.gov>

Vahedifard, F., AghaKouchak, A., Jafari. N. H. (2016). “Compound hazards yield Louisiana flood.” *Science*, 353(6306), 1374, DOI: 10.1126/science.aai8579.

Vahedifard, F., Sehat, S., Aanstoos, J. (2017). “Effects of rainfall, geomorphological and geometrical variables on vulnerability of the lower Mississippi River levee system to slump slides” *GeoRisk*., 11(3), 257-271.

Vahedifard, F., Jasim, F. H., Tracy, F. T., Abdollahi, M., Alborzi, A., AghaKouchak, A. (2020). “Levee Fragility Behavior under Projected Future Flooding in a Warming Climate.” *Journal of Geotechnical and Geoenvironmental Engineering*, 146(12), 04020139, DOI: 10.1061/(ASCE)GT.1943-5606.0002399.

Valavi, R, Elith, J, Lahoz-Monfort, JJ, Guillera-Arroita, G. (2019). “blockCV: An r package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models.” *Methods in Ecology and Evolution. British Ecological Society*. 10. 225– 232. https://doi.org/10.1111/2041-210X.13107.

Venables, W., & Ripley, M. B. (2015). Package ‘class’. *The Comprehensive R Archive Network*.

Villarini, G., Smith, J., Baeck, M., Witold, F. (2011). “Examining Flood Frequency Distributions in the Midwest U.S.” *Journal of the American Water Resources Association*, 462(3): 447-463.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779-804.

Zamanian, S., Hur, J. and Shafieezadeh, A., (2020). Significant variables for leakage and collapse of buried concrete sewer pipes: A global sensitivity analysis via Bayesian additive regression trees and Sobol’indices. *Structure and infrastructure engineering*, 1-13.

Zhang, W., Goh, A. (2013). “Multivariate adaptive regression splines for analysis of geotechnical engineering systems.” *Computers and Geotechnics*. 48. 82-95.

APPENDIX A

LEVEE OVERTOPPING DATASET

Table A.1 Levee Loading and Incident Database

LEVEE LOADING & INCIDENT DATABASE - SUMMARY OF PORTFOLIO OVERTOPPING BREACH DATA

Reference: Flynn, S., Zamanian, S., Vahedifard, F., Shafieezadeh, A., Schaaf, D. (2021) "Data-Driven Model for Probability of Levee Overtopping Due to Breach", Submitted to *Journal of Geotechnical and Geoenvironmental Engineering*, ASCE.

DISCLAIMER: The data contained within this dataset has been compiled by the U.S. Army Corps of Engineers since 2006 and includes information mined from inspection reports, daily flood fight reports, historic flood reports, project rehabilitation reports, USGS and NOAA daily gage data, project modeling data, high water profiles, as-built plans, dated photography, media releases, and discussions with USACE personnel and local levee sponsors. The information contained within this dataset represents an extensive effort to characterize known overtopping events based on both measurements and estimated ranges of data and should be considered accordingly.

LEVEE SEGMENT	EVENT NUMBER	YEAR	DID LEVEE BREACH?	LEVEE CLASS[1]	APPROX. WATER DEPTH OVER LEVEE (ft)[2]	DURATION OF FLOW PRIOR TO BREACH (hours)[3]	LEVEE HEIGHT (ft)[4]	LANDSIDE SLOPE RATIO[5]	MATERIAL DESCRIPTION	EROSION RESISTANCE CLASSIFICATION	DURATION OF LEVEE LOADING PRIOR TO OVERTOPPING (days)[6]
ALWARD SEGMENT 2	1	2006	Y	Local-Local	unknown	unknown	5.8	1.5	silty gravel with sands	Low	1-2 days
AMBRAW	2	2008	Y	Local-Local	< 1	< 6 hours	12.0	2.0	clay/silt mix	High	3-6 days
AMES DIKING DISTRICT	3	2019	Y	Local-Local	unknown	unknown	6.0	2.0	sand/silt mix	Low	1-2 days
	4	1978	Y	Local-Local	unknown	unknown	6.0	2.0		Low	3-6 days
AUGUSTA, KS	5	1998	N	Local-Local	unknown	unknown	8.0	2.5	clay	High	< 1 day
BIG PAPIO RB - L STREET to THOMPSON CREEK	6	1997	N	Local-Local	unknown	< 2 hours	7.0	3.0	clay/silt mix	High	< 1 day
	7	1999	Y	Local-Local	unknown	< 2 hours	7.0	3.0		High	< 1 day
BLOCKSOM & JENCKES	8	1985	Y	Local-Local	2.0 - 3.0	unknown	9.0	2.0	silt/clay mix with sands	Moderate	3-6 days
	9	1991	Y	Local-Local	2.0 - 3.0	unknown	9.0	2.0		Moderate	7-14 days
	10	1994	Y	Local-Local	< 1	unknown	9.0	2.0		Moderate	3-6 days
	11	2003	Y	Local-Local	1.0 - 2.0	unknown	9.0	2.0		Moderate	3-6 days
	12	2008	Y	Local-Local	1.0 - 2.0	unknown	9.0	2.0		Moderate	3-6 days
	13	2013	N	Local-Local	< 1	unknown	9.0	2.0		Moderate	7-14 days
	14	2005	N	Local-Local	3.0 - 4.0	unknown	9.0	2.0		Moderate	-
	15	2008	N	Local-Local	< 1	unknown	9.0	2.0		Moderate	-
	16	2011	Y	Local-Local	< 1	unknown	9.0	2.0		Moderate	-
BLUFFDALE	17	1993	N	Local-Local	> 4	unknown	12.0	3.0	silty loam	Moderate	> 30 days

Table A.1 (continued)

	18	2019	Y	Local-Local	< 1	> 48 hours	12.0	3.0		Moderate	> 30 days
BOWMAN	19	1997	Y	Local-Local	unknown	unknown	4.0	2.0	sandy gravel with cobbles	Low	1-2 days
BREVATOR	20	2015	Y	Local-Local	unknown	unknown	8.0	2.5	clay/silt mix	High	-
	21	1982	Y	Local-Local	unknown	unknown	8.0	2.5		High	-
	22	1979	Y	Local-Local	unknown	unknown	8.0	2.5		High	-
	23	1960	N	Local-Local	unknown	unknown	8.0	2.5		High	-
	24	1993	N	Local-Local	> 4	unknown	8.0	2.5		High	15-30 days
	25	2008	N	Local-Local	2.0 - 3.0	72 hours	8.0	2.5		High	7-14 days
	26	2013	N	Local-Local	< 0.5	12 - 24 hours	8.0	2.5		High	3-6 days
	27	2019	Y	Local-Local	< 1	< 24 hours	8.0	2.5		High	> 30 days
BROOK'S ADDITION	28	2011	Y	Fed-Local	> 4	unknown	6.0	3.0	clay	High	> 30 days
CARPENTER	29	1997	Y	Local-Local	unknown	unknown	4.0	2.0	sand/gravel mix	Low	1-2 days
CHEHALIS AIRPORT	30	2007	Y	Local-Local	unknown	< 2 hours	9.0	2.0	sandy clay	Moderate	3-6 days
	31	1996	Y	Local-Local	unknown	unknown	9.0	2.0		Moderate	3-6 days
	32	1990	Y	Local-Local	< 1	unknown	9.0	2.0		Moderate	3-6 days
CHOUTEAU ISLAND	33	1973	Y	Local-Local	< 1	unknown	11.0	3.0	silty/clayey loam	Moderate	> 30 days
	34	1993	Y	Local-Local	1.0 - 2.0	unknown	11.0	3.0		Moderate	> 30 days
	35	1995	Y	Local-Local	1.0 - 2.0	unknown	11.0	3.0		Moderate	15-30 days
	36	2013	N	Local-Local	< 1	unknown	11.0	3.0		Moderate	3-6 days
	37	2015	Y	Local-Local	< 1	unknown	11.0	3.0		Moderate	7-14 days
	38	2017	Y	Local-Local	< 1	2-3 days	11.0	3.0		Moderate	3-6 days
	39	2019	Y	Local-Local	< 1	6-12 hours	11.0	3.0		Moderate	> 30 days
COFFEYVILLE	40	2007	Y	Local-Local	3.0 - 4.0	24-48 hours	13.4	3.0	clay	High	3-6 days
COUNTRY CLUB ESTATES	41	2011	N	Fed-Local	1.0 - 2.0	24-48 hours	5.0	3.0	clay	High	> 30 days
COWLITZ NO. 2 - LEWIS	42	1948	N	Fed-Local	< 1	< 2 hours	13.0	3.0	silty sand	Low	15-30 days
DALLAS LID 1a	43	1990	N	Local-Local	unknown	unknown	16.0	1.0	highly plastic clay	High	15-30 days

Table A.1 (continued)

DARDANELLE	44	2019	N	Fed-Local	< 0.5	< 2 hours	8.0	3.0	silt with sand/clay	Moderate	3-6 days
DARST BOTTOMS	45	1995	Y	Local-Local	< 1	2-6 hours	11.0	3.5	silty/clayey loam	Moderate	7-14 days
DEKALB LB	46	1983	Y	Fed-Local	1.0 - 2.0	24-48 hours	3.5	3.0	clay	High	1-2 days
	47	2007	Y	Fed-Local	< 1	12-24 hours	3.5	3.0		High	1-2 days
DEKALB RB	48	1983	Y	Fed-Local	2.0 - 3.0	24-48 hours	4.0	3.0	clay	High	1-2 days
	49	1996	N	Fed-Local	< 1	6-12 hours	4.0	3.0		High	< 1 day
	50	2007	Y	Fed-Local	1.0 - 2.0	12-24 hours	4.0	3.0		High	< 1 day
DES MOINES & MISS LEVEE 1	51	1993	Y	Fed-Local	< 1	<6 hours	11.3	4.0	sand fill over clay embankment	Low	> 30 days
EFFLAND	52	1974	Y	Local-Local	unknown	unknown	10.0	3.0	silty clay	High	3-6 days
	53	1993	N	Local-Local	1.0 - 2.0	unknown	10.0	3.0		High	3-6 days
	54	2013	Y	Local-Local	unknown	unknown	10.0	3.0		High	3-6 days
ELKADER	55	2008	Y	Local-Local	1.0 - 2.0	2 - 6 hours	7.4	3.0	clayey sand	Moderate	1-2 days
ELSBERRY	56	1973	Y	Local-Local	< 0.5	24-48 hours	10.0	3.0	clay/silt mix	High	> 30 days
	57	1993	Y	Local-Local	< 1	24-48 hours	10.0	3.0		High	> 30 days
	58	2019	Y	Local-Local	< 1	24-48 hours	10.0	3.0		High	> 30 days
FABIUS RIVER DD	59	1993	Y	Fed-Local	< 1	< 6 hours	16.0	5.0	sand fill over clay embankment	Low	> 30 days
FULTON LB CHARTIERS CREEK	60	2004	Y	Fed-Local	> 4	unknown	5.5	2.0	sand/gravel mix with silt	Low	< 1 day
GERLACH	61	1997	Y	Local-Local	unknown	unknown	4.0	2.0	sand/gravel mix	Low	1-2 days
GRAPE-BOLLIN-SCHWARTZ	62	1993	Y	Local-Local	unknown	unknown	6.0	3.0	sand/silt mix	Low	15-30 days
	63	2007	Y	Local-Local	< 1	unknown	6.0	3.0		Low	1-2 days
	64	2011	Y	Local-Local	unknown	unknown	6.0	3.0		Low	> 30 days
	65	2019	Y	Local-Local	unknown	unknown	6.0	3.0		Low	7-14 days
GREGORY	66	1993	Y	Fed-Local	<1	6-12 hours	16.0	4.0	zoned - impervious R/S, pervious L/S	High	> 30 days
	67	2008	N	Fed-Local	< 1	< 2 hours	16.0	4.0	sand fill over clay embankment	Low	> 30 days

Table A.1 (continued)

	68	2019	N	Fed-Local	< 1	< 2 hours	16.0	4.0	sand fill of previous breach	Low	> 30 days
GREEN BAY NO. 2	69	1993	N	Fed-Local	unknown	<6 HR	14.0	5.0	sand fill over clay embankment	Low	> 30 days
HARTWELL	70	1993	Y	Fed-Local	unknown	unknown	13.0	3.0	silt/clay mix with sand	Moderate	> 30 days
HEISE-ROBERTS LB	71	1976	Y	Fed-Local	unknown	unknown	7.0	2.0	sand/silt mix with clay	Moderate	< 1 day
	72	1997	Y	Fed-Local	unknown	unknown	7.0	2.0		Moderate	> 30 days
HEISE-ROBERTS RB UPPER	73	1976	Y	Fed-Local	unknown	unknown	6.0	2.0	sand/silt mix	Low	< 1 day
HENRIETTA CROOKED SEC 1	74	1993	Y	Local-Local	< 0.5	unknown	13.0	3.0	sand/silt mix	Low	> 30 days
	75	2019	Y	Local-Local	< 1	12-24 hours	13.0	3.0		Low	> 30 days
HERRIED	76	1987	Y	Fed-Local	< 0.5	unknown	5.0	2.5	clay	High	1-2 days
	77	2009	Y	Fed-Local	< 0.5	unknown	5.0	2.5		High	1-2 days
HILLVIEW	78	1993	N	Fed-Local	unknown	unknown	16.0	3.0	silt/clay mix with sand	Moderate	> 30 days
HOLT COUNTY NO. 10	79	1993	N	Local-Local	unknown	unknown	7.0	3.5	sand with silt/clay	Moderate	15-30 days
	80	2007	N	Local-Local	< 1	unknown	7.0	3.5		Moderate	1-2 days
	81	2010	N	Local-Local	unknown	unknown	7.0	3.5		Moderate	7-14 days
	82	2011	N	Local-Local	unknown	unknown	7.0	3.5		Moderate	> 30 days
	83	2019	N	Local-Local	unknown	unknown	7.0	3.5		Moderate	3-6 days
HONEY CREEK	84	2013	N	Local-Local	< 0.5	unknown	11.0	2.0	silt/clay mix	High	7-14 days
HOVANDER PARK	85	1975	Y	Local-Local	< 0.5	12-24 hours	4.0	2.0	silty sand	Low	< 1 day
	86	1989	Y	Local-Local	1.0 - 2.0	unknown	4.0	2.0		Low	1-2 days
	87	Early Nov 1990	Y	Local-Local	1.0 - 2.0	unknown	4.0	2.0		Low	< 1 day
	88	Late Nov 1990	Y	Local-Local	< 0.5	unknown	4.0	2.0		Low	< 1 day
	89	1995	Y	Local-Local	< 0.5	unknown	4.0	2.0		Low	1-2 days
	90	2009	Y	Local-Local	< 1	12-24 hours	4.0	2.0		Low	< 1 day
JOHNSON CITY	91	2013	Y	Fed-Local	1.0 - 2.0	12-24 hours	14.0	2.5	silt/clay mix	High	1-2 days

Table A.1 (continued)

JOHNSONS ADDITION	92	2011	Y	Fed-Local	3.0 - 4.0	> 48 hours	6.3	3.0	silt/clay mix	High	> 30 days
KASKASKIA ISLAND	93	1973	Y	Fed-Local	1.0 - 2.0	6-12 hours	24.0	3.0	clay with silt	High	> 30 days
KEACH	94	1993	N	Fed-Local	3.0 - 4.0	> 48 hours	15.0	4.0	impervious fill	High	> 30 days
KINGSTON-TO-EXETER	95	1972	Y	Fed-Local	unknown	unknown	16.0	2.5	sand/silt mix with clay	Moderate	1-2 days
KISSINGER	96	1993	Y	Local-Local	unknown	unknown	13.0	3.0	silt/clay mix	High	> 30 days
	97	2008	N	Local-Local	unknown	unknown	13.0	3.0		High	> 30 days
	98	2013	Y	Local-Local	unknown	24-48 hours	13.0	3.0		High	7-14 days
	99	2019	Y	Local-Local	< 1	12-24 hours	13.0	3.0		High	> 30 days
KS DEPT OF CORRECTIONS	100	2019	N	Local-Local	< 1	unknown	10.0	2.3	sand/silt mix	Low	7-14 days
	101	2011	N	Local-Local	unknown	unknown	10.0	2.3		Low	> 30 days
	102	1993	Y	Local-Local	unknown	unknown	10.0	2.3		Low	15-30 days
LEACH ROAD	103	2006	Y	Local-Local	< 1	< 6 hours	5.0	2.0	silty gravel with sands	Low	1-2 days
	104	2009	Y	Local-Local	< 1	< 6 hours	5.0	2.0		Low	< 1 day
LEONARD PARK	105	2007	Y	Local-Local	> 4	unknown	7.5	2.0	clay	High	< 1 day
LETHA BRIDGE	106	1997	N	Local-Local	unknown	unknown	8.0	2.0	sandy gravel	Low	3-6 days
LEVEE UNIT NO. 8	107	2008	Y	Fed-Local	< 1	unknown	11.0	3.0	silt/clay mix with sands	Moderate	7-14 days
LONG ROAD	108	2007	N	Fed-Local	unknown	unknown	6.0	2.0	sand/silt mix with clay	Moderate	3-6 days
LOUISA COUNTY LD 11	109	1993	N	Local-Local	< 1	<6hr	12.0	4.0	sand/clay mix	Moderate	> 30 days
	110	2008	N	Local-Local	<1	<6 hr	12.0	5.0		Moderate	> 30 days
LYFORD	111	2005	N	Fed-Local	< 1	24-48 hours	14.0	3.0	clay with sand/silt	High	7-14 days
	112	2013	N	Fed-Local	< 1	2-6 hours	14.0	3.0		High	7-14 days
McLEAN BOTTOM LEVEE & PS	113	2019	N	Fed-Fed	1.0 - 2.0	12-24 hours	20.0	3.0	silt/clay mix	High	7-14 days
McLEAN BOTTOM NO. 3	114	2019	N	Fed-Local	1.0 - 2.0	12-24 hours	15.0	3.0	silt with sand/clay	Moderate	3-6 days
McGINNIS	115	2008	Y	Local-Local	unknown	unknown	8.0	2.5	silt/clay mix with sands	Moderate	3-6 days
MILES POINT	116	2007	N	Local-Local	< 1	< 2 hours	8.0	3.0	sand/silt mix	Low	7-14 days

Table A.1 (continued)

MONTEZUMA	117	1994	N	Fed-Local	unknown	unknown	15.0	2.3	silt/clay mix with sands	Moderate	1-2 days
MRLS 471-460 R	118	1993	N	Fed-Local	1.0 - 2.0	12 -24 hours	11.5	3.0	hydraulic/random fill with impervious cover over R/S face and crest	High	15-30 days
MRLS 500-R	119	1993	Y	Fed-Local	unknown	unknown	11.0	3.0	hydraulic/random fill with impervious cover over R/S face and crest	High	> 30 days
	120	2019	N	Fed-Local	< 1	24-48 hours	11.0	3.0		High	7-14 days
MRLS R-548 MISSOURI RB/BROWNVILLE LD #2	121	1993	Y	Fed-Local	< 0.5	unknown	14.0	3.0	hydraulic/random fill with impervious cover over R/S face and crest	High	-
	122	2011	N	Fed-Local	< 1	unknown	14.0	3.0		High	-
	123	2019	N	Fed-Local	< 0.5	24-48 hours	14.0	3.0		High	-
MRLS L-550	124	1993	N	Fed-Local	1.0 - 2.0	unknown	15.0	3.0	hydraulic/random fill with impervious cover over R/S face and crest	High	> 30 days
	125	2019	Y	Fed-Local	< 1	12- 24 hours	15.0	3.0		High	7-14 days
MRLS R-573	126	2019	Y	Fed-Local	< 1	24-48 hours	13.5	3.0	silt/clay mix	High	3-6 days
MRLS L-575 EAST	127	1993	Y	Fed-Local	unknown	unknown	10.0	3.0	silt/clay mix with sand	Moderate	> 30 days
	128	1998	Y	Fed-Local	unknown	unknown	10.0	3.0		Moderate	> 30 days
	129	2007	Y	Fed-Local	unknown	unknown	10.0	3.0		Moderate	> 30 days
MRLS L 611-614	130	2019	Y	Fed-Local	< 1	12-24 hours	14.2	3.0	hydraulic/random fill with impervious cover over R/S face and crest	High	3-6 days
MRLS R-616	131	2019	Y	Fed-Local	< 0.5	24-48 hours	12.5	3.0	hydraulic/random fill with impervious cover over R/S face and crest	High	3-6 days
NORTH ADDISON	132	1972	Y	Fed-Local	< 0.5	4-8 hours	9.8	2.5	silt with sand/clay	Moderate	3-6 days
NORTH ELMIRA	133	1975	Y	Fed-Local	< 1	< 6 hours	15.0	2.5	sandy silt	Low	1-2 days

Table A.1 (continued)

OSAWATOMIE	134	2007	N	Fed-Local	< 1	N/A	15.0	3.0	lean clay	High	3-6 days
PAJARO RIVER LB	135	1995	Y	Fed-Local	< 0.5	< 2 hours	9.0	2.0	sand/silt mix with clay	Moderate	< 1 day
PAJARO RIVER RB - D/S	136	1998	Y	Fed-Local	< 0.5	< 24 hours	9.0	2.0	sand/silt mix with clay	Moderate	1-2 days
PENNY SLOUGH	137	1997	N	Fed-Local	unknown	N/A	8.7	3.0	clay	High	7-14 days
PLYMOUTH	138	1972	N	Fed-Local	unknown	unknown	16.0	2.5	clayey silt	Moderate	1-2 days
PORTVILLE SOUTH DODGE CREEK	139	1972	N	Fed-Local	< 1	12-24 hours	16.8	2.5	sand/silt mix with gravel	Low	1-2 days
PUNXSUTAWNEY LB	140	1996	Y	Fed-Local	< 1	6-12 hours	7.0	2.0	silt/clay mix	High	< 1 day
PUNXSUTAWNEY RB	141	1996	Y	Fed-Local	< 1	6-12 hours	7.0	2.0	silt/clay mix	High	< 1 day
RAINBOW SLOUGH	142	1975	Y	Local-Local	< 0.5	<6hr	5.0	2.5	silty sand	Low	1-2 days
	143	1989	Y	Local-Local	< 0.5	< 6 hr	5.0	2.5		Low	1-2 days
	144	1990	Y	Local-Local	1.0 - 2.0	< 6 hr	5.0	2.5		Low	< 1 day
	145	1995	N	Local-Local	< 0.5	< 6 hr	5.0	2.5		Low	1-2 days
	146	2009	N	Local-Local	< 0.5	< 6 hr	5.0	2.5		Low	1-2 days
RAYHORST	147	1990	Y	Local-Local	< 1	unknown	6.0	8.0	silty sand	Low	< 1 day
RENZ ITEM #36	148	1990	N	Local-Local	< 0.5	unknown	8.0	4.5	sand/silt mix with clay	Moderate	1-2 days
	149	2019	N	Local-Local	< 1	unknown	8.0	4.5		Moderate	15-30 days
	150	1993	N	Local-Local	unknown	unknown	8.0	4.5		Moderate	7-14 days
ROSEAU RIVER - DUXBY	151	2002	Y	Fed-Local	< 1	24-48 hours	5.7	3.0	silt/clay mix	High	3-6 days
RUNNING WATER	152	2011	Y	Fed-Local	< 1	2-6 hours	5.0	3.0	silty sand, *silty clay for one breach	Low	1-2 days
	153	2017	Y	Fed-Local	< 1	6-12 hours	7.0	3.0	silty sand	Low	3-6 days
RUSHFORD AB	154	2007	N	Fed-Local	unknown	< 4 hours	9.0	3.0	hydraulic/random fill with impervious cover	High	< 1 day

Table A.1 (continued)

RUSHFORD C	155	2008	Y	Fed-Local	unknown	< 4 hours	9.0	3.0	hydraulic/random fill with impervious cover	High	< 1 day
RUSHFORD DEF	156	2007	Y	Fed-Local	unknown	< 4 hours	9.5	3.0	hydraulic/random fill with impervious cover	High	< 1 day
RUSSELL & ALLISON	157	2008	N	Local-Local	< 1	< 2 hours	10.0	3.0	clay	High	3-6 days
	158	2011	N	Local-Local	< 1	6-12 hours	10.0	3.0		High	7-14 days
SAINTE MARIE	159	1950	N	Local-Local	< 1	unknown	10.0	2.5	silt/clay mix with sands	Moderate	1-2 days
	160	1957	Y	Local-Local	1.0 - 2.0	unknown	10.0	2.5		Moderate	1-2 days
	161	2008	Y	Local-Local	< 1	< 2 hours	10.0	2.5		Moderate	1-2 days
SALAMANCA LB	162	1972	Y	Fed-Local	2.0 - 3.0	unknown	9.0	2.5	sand/clay mix	Moderate	1-2 days
SALAMANCA RB	163	1972	N	Fed-Local	2.0 - 3.0	unknown	10.0	2.5	sand/clay mix	Moderate	1-2 days
SANDY CREEK	164	1993	Y	Local-Local	> 4	> 48 hours	9.0	3.0	silty/clayey loam	Moderate	> 30 days
	165	2008	Y	Local-Local	unknown	unknown	9.0	3.0		Moderate	> 30 days
	166	2013	N	Local-Local	< 0.5	24-48 hours	9.0	3.0		Moderate	7-14 days
SAWYER WEST	167	2011	Y	Fed-Local	unknown	unknown	4.9	3.0	silt/clay mix	High	> 30 days
SAYRE	168	2011	Y	Local-Local	1.0 - 2.0	24-48 hours	12.0	4.0	impervious - silt/clay, pervious - sand/silt	High	< 1 day
SIX MILE DIVERSION	169	2019	N	Fed-Fed	< 1	unknown	16.5	3.0	clay	High	7-14 days
SOUTH ELMIRA	170	1972	Y	Fed-Local	unknown	unknown	10.0	2.5	sand/silt mix	Low	1-2 days
SOUTH RIVER DD	171	1993	Y	Fed-Local	<1	<6hr	11.0	5.0	sand fill over clay embankment	Low	> 30 days
SUNBURY	172	1972	N	Fed-Local	< 0.5	2-6 hours	16.0	2.5	silt/clay mix	High	3-6 days
TALBOTT'S NURSERY	173	2011	Y	Fed-Local	1.0 - 2.0	unknown	8.3	3.0	silt/clay mix	High	> 30 days
TETESEAU BEND	174	1993	Y	Local-Local	unknown	unknown	10.0	3.0	sand/silt mix	Low	> 30 days
TIERRECITA VALLEJO	175	2011	N	Fed-Local	unknown	unknown	6.0	3.0	clay	High	> 30 days

Table A.1 (continued)

TULSA-WEST TULSA A	176	1984	Y	Fed-Local	1.0 - 2.0	2-6 hours	5.0	2.5	zoned - impervious R/S, pervious L/S	High	< 1 day
TULSA-WEST TULSA B	177	1984	Y	Fed-Local	1.0 - 2.0	2-6 hours	5.0	2.5	zoned - impervious R/S, pervious L/S	High	< 1 day
UNION TOWNSHIP	178	2008	N	Local-Local	< 0.5	< 2 hours	16.0	4.0	sand fill over clay embankment, *one section silt/clay mix with sands	Low	7-14 days
	179	2019	Y	Local-Local	< 0.5	< 2 hours	16.0	4.0	sand fill over clay embankment	Low	> 30 days
VANDERPOL	180	1975	Y	Local-Local	unknown	unknown	6.0	2.0	silty sand	Low	1-2 days
VESTAL-TWIN ORCHARDS	181	2011	Y	Fed-Local	1.0 - 2.0	12-24 hours	17.0	2.5	silt/clay mix	High	3-6 days
WINFIELD PIN OAKS	182	1993	Y	Local-Local	< 1	< 6 hours	9.5	3.0	clay	High	> 30 days
	183	2015	Y	Local-Local	< 1	< 6 hours	9.5	3.0		High	7-14 days
	184	2019	N	Local-Local	< 0.5	12-24 hours	9.5	3.0		High	> 30 days
YORK E DOWNTOWN	185	1972	Y	Fed-Fed	3.0 - 4.0	< 4 hours	5.0	2.5	compact sand with silt/clay	Moderate	< 1 day

[1]First descriptor indicates construction entity, second description indicates maintaining entity.
[2]Represents the approximate depth (feet) of water flowing over the levee at the time of the breach or maximum depth of water over levee during non-breach overtopping event.
[3]Represents the duration (hours) of widespread overtopping before breach occurred OR duration of overtopping without breach before river receded below levee embankment.

[4]Average cross sectional height of levee in breached segment.
[5]Represents ratio of horizontal length to vertical height of slope. (i.e. 3H:1V = 3)

[6]Days a flood load was on the riverside levee slope prior to overtopping.

APPENDIX B

EXPANDED LEVEE OVERTOPPING DATASET WITHOUT IMPUTED DATA

Table B.1 Expanded Dataset without Imputation

X1	X2	X3	X4	X5	X6	X7	Y
1.76784	1	1			1	1	1
1.8288	1	1			1	1	1
1.8288	1	1			1	2	1
2.4384	1	1			3	1	1
2.1336	2	1		1	3	1	0
2.1336	2	1		1	3	1	0
2.7432	1	1	3		2	2	1
2.7432	1	1	3		2	2	1
2.7432	1	1	2		2	2	1
2.7432	1	1	3		2	2	1
2.7432	1	1	3		2	2	1
2.7432	1	1	3		2	2	1
2.7432	1	1	2		2	2	1
2.7432	1	1	3		2	2	1
2.7432	1	1	3		2		0
2.7432	1	1	2		2		0
2.7432	1	1	1		2		0
2.7432	1	1	2		2		0
2.7432	1	1	1		2		0
2.7432	1	1	2		2		0
2.7432	1	1	1		2		0
3.6576	2	1	3		2	3	1
3.6576	2	1	2	3	2	3	0
3.6576	2	1	1	3	2	3	0
3.6576	2	1	2	2	2	3	0
3.6576	2	1	2	1	2	3	0
1.2192	1	1			1	1	1
2.4384	1	1			3		1
2.4384	1	1			3		1
2.4384	1	1			3		1
2.4384	1	1			3		1
2.4384	1	1	3		3	3	0
2.4384	1	1	2		3	3	0
2.4384	1	1	1		3	3	0
2.4384	1	1	3	3	3	2	0
2.4384	1	1	2	3	3	2	0
2.4384	1	1	1	3	3	2	0
2.4384	1	1	3	2	3	2	0
2.4384	1	1	3	1	3	2	0
2.4384	1	1	1	1	3	2	0
2.4384	1	1	1	2	3	2	0
2.4384	1	1	2	1	3	2	0
2.4384	1	1	2	2	3	2	0
2.4384	1	1	1	2	3	2	0
2.4384	1	1	1	1	3	2	0
1.8288	2	2	3		3	3	0
1.8288	2	2	2		3	3	0
1.8288	2	2	1		3	3	0
1.2192	1	1			1	1	1
2.7432	1	1		1	2	2	1
2.7432	1	1		2	2	2	1
2.7432	1	1		3	2	2	1
2.7432	1	1			2	2	1
2.7432	1	1	2		2	2	1
2.7432	1	1	3		2	2	1
3.3528	2	1	2		2	3	1
3.3528	2	1	3		2	3	1
3.3528	2	1	3		2	3	1

Table B.1 (continued)

3.3528	2	1	3		2	3	1
3.3528	2	1	2		2	2	1
3.3528	2	1	3		2	2	1
3.3528	2	1	2		2	2	1
3.3528	2	1	3		2	2	1
4.08432	2	1	3	3	3	2	0
4.08432	2	1	3	2	3	2	0
4.08432	2	1	3	1	3	2	0
4.08432	2	1	2	3	3	2	0
4.08432	2	1	2	2	3	2	0
4.08432	2	1	2	1	3	2	0
4.08432	2	1	1	3	3	2	0
4.08432	2	1	1	2	3	2	0
4.08432	2	1	1	1	3	2	0
1.524	2	2	3	3	3	3	1
3.9624	2	2	2	1	1	3	1
3.9624	2	2	3	1	1	3	1
3.9624	2	2	3	2	1	3	1
3.9624	2	2	3	3	1	3	1
3.9624	2	2	2	2	1	3	1
3.9624	2	2	2	3	1	3	1
4.8768	1	1			3	3	1
2.4384	2	2	1	1	2	2	1
2.4384	2	2	1	2	2	2	1
2.4384	2	2	1	3	2	2	1
2.4384	2	2	2	1	2	2	1
2.4384	2	2	2	2	2	2	1
2.4384	2	2	2	3	2	2	1
2.4384	2	2	3	1	2	2	1
2.4384	2	2	3	2	2	2	1
2.4384	2	2	3	3	2	2	1
1.0668	2	2	2	2	3	1	0
1.0668	2	2	1	1	3	1	0
1.0668	2	2	1	2	3	1	0
1.0668	2	2	2	1	3	1	0
1.2192	2	2	3	3	3	1	0
1.2192	2	2	3	2	3	1	0
1.2192	2	2	3	1	3	1	0
1.2192	2	2	2	3	3	1	0
1.2192	2	2	2	2	3	1	0
1.2192	2	2	2	1	3	1	0
1.2192	2	2	1	1	3	1	0
1.2192	2	2	1	2	3	1	0
1.2192	2	2	1	3	3	1	0
1.2192	2	2	2	1	3	1	0
1.2192	2	2	1	1	3	1	0
1.2192	2	2	3	2	3	1	0
1.2192	2	2	2	2	3	1	0
1.2192	2	2	1	2	3	1	0
1.2192	2	2	3	1	3	1	0
1.2192	2	2	2	1	3	1	0
1.2192	2	2	1	1	3	1	0
3.44424	2	2	2	1	1	3	1
3.44424	2	2	3	2	1	3	1
3.44424	2	2	3	3	1	3	1
3.44424	2	2	2	2	1	3	1
3.44424	2	2	2	3	1	3	1
3.44424	2	2	3	1	1	3	1
3.048	2	1			3	2	1

Table B.1 Expanded Dataset without Imputation

3.048	2	1	3		3	2	1
3.048	2	1			3	2	1
2.25552	2	1	3	1	2	1	0
2.25552	2	1	2	1	2	1	0
2.25552	2	1	1	1	2	1	0
3.048	2	1	1	3	3	3	1
3.048	2	1	2	3	3	3	1
3.048	2	1	3	3	3	3	1
3.048	2	1	2	3	3	3	1
3.048	2	1	3	3	3	3	1
3.048	2	1	2	3	3	3	1
3.048	2	1	3	3	3	3	1
1.6764	1	2	3		1	1	0
1.6764	1	2	2		1	1	0
1.6764	1	2	1		1	1	0
1.2192	1	1			1	1	1
1.8288	2	1			1	3	1
1.8288	2	1	2		1	1	1
1.8288	2	1	3		1	1	1
1.8288	2	1			1	3	1
1.8288	2	1			1	2	1
4.8768	2	2	2	2	3	3	1
4.8768	2	2	2	3	3	3	1
4.8768	2	2	3	2	3	3	1
4.8768	2	2	3	3	3	3	1
4.8768	2	2	2	1	1	3	1
4.8768	2	2	2	2	1	3	1
4.8768	2	2	2	3	1	3	1
4.8768	2	2	3	1	1	3	1
4.8768	2	2	3	2	1	3	1
4.8768	2	2	3	3	1	3	1
4.2672	2	2			1	3	1
3.9624	2	2			2	3	1
2.1336	1	2			2	1	1
2.1336	1	2			2	3	1
1.8288	1	2			1	1	1
3.9624	2	1	1		1	3	1
3.9624	2	1	2		1	3	1
3.9624	2	1	3		1	3	1
3.9624	2	1	2	2	1	3	0
3.9624	2	1	1	2	1	3	0
3.9624	2	1	2	1	1	3	0
3.9624	2	1	1	1	1	3	0
1.524	1	2	1		3	1	0
1.524	1	2	1		3	1	0
4.8768	2	2			2	3	1
2.1336	2	1			2	3	1
2.1336	2	1	2		2	1	1
2.1336	2	1	3		2	1	1
2.1336	2	1			2	2	1
2.1336	2	1			2	3	1
2.1336	2	1			2	2	1
3.3528	1	1	1		3	2	1
3.3528	1	1	2		3	2	1
3.3528	1	1	3		3	2	1
1.2192	1	1	1	2	1	1	1
1.2192	1	1	1	3	1	1	1
1.2192	1	1	2	2	1	1	1
1.2192	1	1	2	3	1	1	1

Table B.1 Expanded Dataset without Imputation

1.2192	1	1	3	2	1	1	1
1.2192	1	1	3	3	1	1	1
1.2192	1	1	3		1	1	0
1.2192	1	1	2		1	1	0
1.2192	1	1	1		1	1	0
1.2192	1	1	3		1	1	0
1.2192	1	1	2		1	1	0
1.2192	1	1	1		1	1	0
1.2192	1	1	1		1	1	0
1.2192	1	1	1		1	1	0
4.2672	1	2	3	2	3	1	0
4.2672	1	2	3	1	3	1	0
4.2672	1	2	2	2	3	1	0
4.2672	1	2	2	1	3	1	0
4.2672	1	2	1	2	3	1	0
4.2672	1	2	1	1	3	1	0
1.92024	2	2	3	3	3	3	0
1.92024	2	2	3	2	3	3	0
1.92024	2	2	3	1	3	3	0
1.92024	2	2	2	3	3	3	0
1.92024	2	2	2	2	3	3	0
1.92024	2	2	2	1	3	3	0
1.92024	2	2	1	1	3	3	0
1.92024	2	2	1	2	3	3	0
1.92024	2	2	1	3	3	3	0
4.572	2	2	3	3	3	3	0
4.572	2	2	3	2	3	3	0
4.572	2	2	3	1	3	3	0
4.572	2	2	2	3	3	3	0
4.572	2	2	2	2	3	3	0
4.572	2	2	2	1	3	3	0
4.572	2	2	1	1	3	3	0
4.572	2	2	1	2	3	3	0
4.572	2	2	1	3	3	3	0
4.8768	1	2			2	1	1
3.9624	2	1			3	3	1
3.9624	2	1			3	3	1
3.9624	2	1		3	3	2	1
3.048	1	1	2		1	2	1
3.048	1	1	3		1	2	1
3.048	1	1			1	3	1
3.048	1	1			1	3	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
2.286	1	1	3		3	1	0
2.286	1	1	2		3	1	0
2.286	1	1	1		3	1	0
2.4384	1	1			1	2	1
3.3528	2	2	2		2	2	1

Table B.1 Expanded Dataset without Imputation

3.3528	2	2	3		2	2	1
1.8288	1	2			2	2	0
3.6576	2	1	2	1	2	3	1
3.6576	2	1	2	2	2	3	1
3.6576	2	1	2	3	2	3	1
3.6576	2	1	3	1	2	3	1
3.6576	2	1	3	2	2	3	1
3.6576	2	1	3	3	2	3	1
3.6576	2	1	2	1	2	3	1
3.6576	2	1	2	2	2	3	1
3.6576	2	1	2	3	2	3	1
3.6576	2	1	3	1	2	3	1
3.6576	2	1	3	2	2	3	1
3.6576	2	1	3	3	2	3	1
4.2672	2	2	2	1	3	2	0
4.2672	2	2	1	1	3	2	0
6.096	2	2	3	2	3	2	0
6.096	2	2	3	1	3	2	0
6.096	2	2	2	2	3	2	0
6.096	2	2	2	1	3	2	0
6.096	2	2	1	2	3	2	0
6.096	2	2	1	1	3	2	0
4.572	2	2	3	2	2	2	1
4.572	2	2	3	3	2	2	1
2.4384	1	1			2	2	1
4.572	1	2			2	1	1
3.5052	2	2	3	2	3	3	1
3.5052	2	2	3	3	3	3	1
3.3528	2	2			3	3	0
3.3528	2	2	2	3	3	2	1
3.3528	2	2	3	3	3	2	1
4.2672	2	2	1		3		0
4.2672	2	2	2		3		0
4.2672	2	2	1		3		0
4.2672	2	2	1	3	3		0
4.2672	2	2	1	2	3		0
4.2672	2	2	1	1	3		0
4.572	2	2	3		3	3	0
4.572	2	2	2		3	3	0
4.572	2	2	1		3	3	0
3.048	2	2			2	3	0
3.048	2	2			2	3	0
3.048	2	2			2	3	0
4.32816	2	2	2	2	3	2	1
4.32816	2	2	2	3	3	2	1
4.32816	2	2	3	2	3	2	1
4.32816	2	2	3	3	3	2	1
3.81	2	2	1	3	3	2	0
3.81	2	2	1	2	3	2	0
3.81	2	2	1	1	3	2	0
2.98704	1	2	1	2	2	2	0
2.98704	1	2	1	1	2	2	0
4.572	2	2	2		3	2	0
4.572	2	2	1		3	2	0
2.7432	1	2	1	2	2	1	1
2.7432	1	2	1	3	2	1	1
2.7432	1	2	2	2	2	1	1
2.7432	1	2	2	3	2	1	1
2.7432	1	2	3	2	2	1	1

Table B.1 Expanded Dataset without Imputation

2.7432	1	2	3	3	2	1	1
2.65176	2	2			3	2	0
4.8768	1	2			2	1	1
5.12064	1	2	2	2	1	1	0
5.12064	1	2	2	1	1	1	0
5.12064	1	2	1	2	1	1	0
5.12064	1	2	1	1	1	1	0
2.1336	1	2	2	2	3	1	0
2.1336	1	2	2	1	3	1	0
2.1336	1	2	1	2	3	1	0
2.1336	1	2	1	1	3	1	0
2.1336	1	2	2	2	3	1	0
2.1336	1	2	2	1	3	1	0
2.1336	1	2	1	2	3	1	0
2.1336	1	2	1	1	3	1	0
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.8288	2	1	2		1	1	1
1.8288	2	1	3		1	1	1
2.4384	2	1	1		2	1	1
2.4384	2	1	2		2	1	1
2.4384	2	1	3		2	1	1

Table B.1 Expanded Dataset without Imputation

2.4384	2	1	2		2	3	1
2.4384	2	1	3		2	3	1
2.4384	2	1			2	2	1
1.73736	2	2	2	3	3	2	0
1.73736	2	2	2	2	3	2	0
1.73736	2	2	2	1	3	2	0
1.73736	2	2	1	3	3	2	0
1.73736	2	2	1	2	3	2	0
1.73736	2	2	1	1	3	2	0
1.524	2	2	2	1	1	1	1
1.524	2	2	2	2	1	1	1
1.524	2	2	2	3	1	1	1
1.524	2	2	3	1	1	1	1
1.524	2	2	3	2	1	1	1
1.524	2	2	3	3	1	1	1
2.1336	2	2	2	2	1	2	1
2.1336	2	2	2	3	1	2	1
2.1336	2	2	3	2	1	2	1
2.1336	2	2	3	3	1	2	1
2.7432	2	2		1	3	1	0
2.7432	2	2		1	3	1	0
2.8956	2	2		1	3	1	0
3.048	2	1	2	1	3	2	1
3.048	2	1	2	2	3	2	1
3.048	2	1	2	3	3	2	1
3.048	2	1	3	1	3	2	1
3.048	2	1	3	2	3	2	1
3.048	2	1	3	3	3	2	1
3.048	2	1	2	2	3	2	1
3.048	2	1	2	3	3	2	1
3.048	2	1	3	2	3	2	1
3.048	2	1	3	3	3	2	1
3.048	1	1	2		2	1	1
3.048	1	1	3		2	1	1
3.048	1	1	3		2	1	1
3.048	1	1	2	1	2	1	1
3.048	1	1	2	2	2	1	1
3.048	1	1	2	3	2	1	1
3.048	1	1	3	1	2	1	1
3.048	1	1	3	2	2	1	1
3.048	1	1	3	3	2	1	1
2.7432	1	2	3		2	1	0
2.7432	1	2	2		2	1	0
2.7432	1	2	1		2	1	0
3.048	1	2	3		2	1	0
3.048	1	2	2		2	1	0
3.048	1	2	1		2	1	0
2.7432	2	1			2	3	1
2.7432	2	1	1	3	2	2	0
2.7432	2	1	1	2	2	2	0
2.7432	2	1	1	1	2	2	0
1.49352	2	2			3	3	0
3.6576	2	1	3	3	3	1	0
3.6576	2	1	3	2	3	1	0
3.6576	2	1	3	1	3	1	0
3.6576	2	1	2	3	3	1	0
3.6576	2	1	2	2	3	1	0
3.6576	2	1	2	1	3	1	0
3.6576	2	1	1	3	3	1	0

Table B.1 Expanded Dataset without Imputation

3.6576	2	1	1	2	3	1	0
3.6576	2	1	1	1	3	1	0
5.0292	2	2	2		3	2	1
5.0292	2	2	3		3	2	1
3.048	1	2			1	1	1
3.3528	2	2	2	1	1	3	1
3.3528	2	2	2	2	1	3	1
3.3528	2	2	2	3	1	3	1
3.3528	2	2	3	1	1	3	1
3.3528	2	2	3	2	1	3	1
3.3528	2	2	3	3	1	3	1
4.8768	1	2	1	1	3	2	0
2.52984	2	2	3		3	3	1
3.048	2	1			1	3	1
1.8288	2	2			3	3	0
1.524	1	2	3	1	3	1	0
1.524	1	2	2	1	3	1	0
1.524	1	2	1	1	3	1	0
1.524	1	2	3	1	3	1	0
1.524	1	2	2	1	3	1	0
1.524	1	2	1	1	3	1	0
4.8768	2	1	1	1	1	2	1
4.8768	2	1	1	2	1	2	1
4.8768	2	1	1	3	1	2	1
4.8768	2	1	2	1	1	2	1
4.8768	2	1	2	2	1	2	1
4.8768	2	1	2	3	1	2	1
4.8768	2	1	3	1	1	2	1
4.8768	2	1	3	2	1	2	1
4.8768	2	1	3	3	1	2	1
4.8768	2	1	1	1	1	3	1
4.8768	2	1	1	2	1	3	1
4.8768	2	1	1	3	1	3	1
4.8768	2	1	2	1	1	3	1
4.8768	2	1	2	2	1	3	1
4.8768	2	1	2	3	1	3	1
4.8768	2	1	3	1	1	3	1
4.8768	2	1	3	2	1	3	1
4.8768	2	1	3	3	1	3	1
1.8288	1	1			1	1	1
5.1816	1	2	3	2	3	2	0
5.1816	1	2	3	1	3	2	0
5.1816	1	2	2	2	3	2	0
5.1816	1	2	2	1	3	2	0
5.1816	1	2	1	2	3	2	0
5.1816	1	2	1	1	3	2	0
2.8956	2	1	2	1	3	2	1
2.8956	2	1	2	2	3	2	1
2.8956	2	1	2	3	3	2	1
2.8956	2	1	3	1	3	2	1
2.8956	2	1	3	2	3	2	1
2.8956	2	1	3	3	3	2	1
2.8956	2	1	1	2	3	3	1
2.8956	2	1	2	2	3	3	1
2.8956	2	1	3	2	3	3	1
2.8956	2	1	3	2	3	3	1
2.8956	2	1	3	3	3	3	1
2.8956	2	1	1	3	3	3	1
1.524	1	2	3	1	2	1	0

Table B.1 Expanded Dataset without Imputation

1.524	1	2	2	1	2	1	0
1.524	1	2	1	1	2	1	0
3.6576	1	1	2	1	3	2	1
3.6576	1	1	3	1	3	2	1
3.6576	1	1	2	2	3	2	1
3.6576	1	1	2	3	3	2	1
3.6576	1	1	3	3	3	2	1
2.4384	1	1	2	2	3	3	1
2.4384	1	1	3	2	3	3	1
2.4384	1	1	2	3	3	3	1
2.4384	1	1	3	3	3	3	1
3.3528	2	1	2	3	2	2	0
3.3528	2	1	1	3	2	2	0
3.3528	2	1	2	2	2	2	0
3.3528	2	1	2	1	2	2	0
3.3528	2	1	1	2	2	2	0
3.3528	2	1	1	1	2	2	0
3.3528	2	1	2	2	2	3	1
3.3528	2	1	3	2	2	3	1
3.3528	2	1	2	3	2	3	1
3.3528	2	1	3	3	2	3	1
3.3528	2	1	2	1	2	2	1
3.3528	2	1	3	2	2	2	1
3.3528	2	1	3	3	2	2	1
3.3528	2	1	2	2	2	2	1
3.3528	2	1	2	3	2	2	1
3.3528	2	1	3	1	2	2	1
1.0668	2	2	3	3	3	1	0
1.0668	2	2	3	2	3	1	0
1.0668	2	2	3	1	3	1	0
1.0668	2	2	2	3	3	1	0
1.0668	2	2	2	2	3	1	0
1.0668	2	2	2	1	3	1	0
1.0668	2	2	1	1	3	1	0
1.0668	2	2	1	2	3	1	0
1.0668	2	2	1	3	3	1	0
4.8768	2	2	2	1	1	3	1
4.8768	2	2	2	2	1	3	1
4.8768	2	2	2	3	1	3	1
4.8768	2	2	3	1	1	3	1
4.8768	2	2	3	2	1	3	1
4.8768	2	2	3	3	1	3	1
4.8768	2	2	2	1	1	3	1
4.8768	2	2	2	2	1	3	1
4.8768	2	2	2	3	1	3	1
4.8768	2	2	3	1	1	3	1
4.8768	2	2	3	2	1	3	1
4.8768	2	2	3	3	1	3	1
1.2192	1	1	2	2	1	1	0
1.2192	1	1	1	2	1	1	0
1.2192	1	1	2	1	1	1	0
1.2192	1	1	1	1	1	1	0
7.3152	2	2	3	2	3	3	1
7.3152	2	2	3	3	3	3	1
3.9624	2	1	2	2	3	3	1
3.9624	2	1	2	3	3	3	1
3.9624	2	1	3	2	3	3	1
3.9624	2	1	3	3	3	3	1
4.2672	2	2	2	3	3	2	0

Table B.1 Expanded Dataset without Imputation

4.2672	2	2	2	2	3	2	0
4.2672	2	2	2	1	3	2	0
4.2672	2	2	1	3	3	2	0
4.2672	2	2	1	2	3	2	0
4.2672	2	2	1	1	3	2	0
2.4384	2	1	2	1	1	2	1
2.4384	2	1	2	2	1	2	1
2.4384	2	1	2	3	1	2	1
2.4384	2	1	3	1	1	2	1
2.4384	2	1	3	2	1	2	1
2.4384	2	1	3	3	1	2	1
4.572	2	2	2	2	3	2	1
4.572	2	2	3	2	3	2	1
4.572	2	2	2	3	3	2	1
4.572	2	2	3	3	3	2	1
4.1148	2	2	2	3	3	2	0
4.1148	2	2	2	2	3	2	0
4.1148	2	2	2	1	3	2	0
4.1148	2	2	1	3	3	2	0
4.1148	2	2	1	2	3	2	0
4.1148	2	2	1	1	3	2	0
4.572	1	2	2	1	1	1	1
4.572	1	2	2	2	1	1	1
4.572	1	2	2	3	1	1	1
4.572	1	2	3	1	1	1	1
4.572	1	2	3	2	1	1	1
4.572	1	2	3	3	1	1	1
2.7432	1	2	1	1	2	1	1
2.7432	1	2	1	2	2	1	1
2.7432	1	2	1	3	2	1	1
2.7432	1	2	2	1	2	1	1
2.7432	1	2	2	2	2	1	1
2.7432	1	2	2	3	2	1	1
2.7432	1	2	3	1	2	1	1
2.7432	1	2	3	2	2	1	1
2.7432	1	2	3	3	2	1	1
2.7432	2	1	3	3	2	3	0
2.7432	2	1	3	2	2	3	0
2.7432	2	1	3	1	2	3	0
2.7432	2	1	2	3	2	3	0
2.7432	2	1	2	2	2	3	0
2.7432	2	1	2	1	2	3	0
2.7432	2	1	1	3	2	3	0
2.7432	2	1	1	2	2	3	0
2.7432	2	1	1	1	2	3	0
2.8956	2	1	2	1	3	3	1
2.8956	2	1	2	2	3	3	1
2.8956	2	1	2	3	3	3	1
2.8956	2	1	3	1	3	3	1
2.8956	2	1	3	2	3	3	1
2.8956	2	1	3	3	3	3	1

APPENDIX C

EXPANDED LEVEE OVERTOPPING DATASET WITH IMPUTED DATA

Table C.1 Expanded Dataset with Imputation

X1	X2	X3	X4	X5	X6	X7	Y
1.76784	1	1	2	2	1	1	1
1.8288	1	1	2	2	1	1	1
1.8288	1	1	2	2	1	2	1
2.4384	1	1	2	2	3	1	1
2.1336	2	1	2	1	3	1	0
2.1336	2	1	2	1	3	1	0
2.7432	1	1	3	2	2	2	1
2.7432	1	1	3	2	2	2	1
2.7432	1	1	2	2	2	2	1
2.7432	1	1	3	2	2	2	1
2.7432	1	1	3	2	2	2	1
2.7432	1	1	3	2	2	2	1
2.7432	1	1	2	2	2	2	1
2.7432	1	1	3	2	2	2	1
2.7432	1	1	3	2	2	2	0
2.7432	1	1	2	2	2	2	0
2.7432	1	1	1	2	2	2	0
2.7432	1	1	2	2	2	2	0
2.7432	1	1	1	2	2	2	0
2.7432	1	1	2	2	2	2	0
2.7432	1	1	1	2	2	2	0
3.6576	2	1	3	2	2	3	1
3.6576	2	1	2	3	2	3	0
3.6576	2	1	1	3	2	3	0
3.6576	2	1	2	2	2	3	0
3.6576	2	1	2	1	2	3	0
1.2192	1	1	2	2	1	1	1
2.4384	1	1	2	2	3	2	1
2.4384	1	1	2	2	3	2	1
2.4384	1	1	2	2	3	2	1
2.4384	1	1	2	2	3	2	1
2.4384	1	1	3	2	3	3	0
2.4384	1	1	2	2	3	3	0
2.4384	1	1	1	2	3	3	0
2.4384	1	1	3	3	3	2	0
2.4384	1	1	2	3	3	2	0
2.4384	1	1	1	3	3	2	0
2.4384	1	1	3	2	3	2	0
2.4384	1	1	3	1	3	2	0
2.4384	1	1	1	1	3	2	0
2.4384	1	1	1	2	3	2	0
2.4384	1	1	2	1	3	2	0
2.4384	1	1	2	2	3	2	0
2.4384	1	1	1	2	3	2	0
2.4384	1	1	1	1	3	2	0
1.8288	2	2	3	2	3	3	0
1.8288	2	2	2	2	3	3	0
1.8288	2	2	1	2	3	3	0
1.2192	1	1	2	2	1	1	1
2.7432	1	1	3	1	2	2	1
2.7432	1	1	3	2	2	2	1
2.7432	1	1	3	3	2	2	1
2.7432	1	1	3	2	2	2	1
2.7432	1	1	2	2	2	2	1
2.7432	1	1	3	2	2	2	1
3.3528	2	1	2	2	2	3	1
3.3528	2	1	3	2	2	3	1
3.3528	2	1	3	2	2	3	1

Table C.1 (continued)

3.3528	2	1	3	2	2	3	1
3.3528	2	1	2	2	2	2	1
3.3528	2	1	3	2	2	2	1
3.3528	2	1	2	2	2	2	1
3.3528	2	1	3	2	2	2	1
4.08432	2	1	3	3	3	2	0
4.08432	2	1	3	2	3	2	0
4.08432	2	1	3	1	3	2	0
4.08432	2	1	2	3	3	2	0
4.08432	2	1	2	2	3	2	0
4.08432	2	1	2	1	3	2	0
4.08432	2	1	1	3	3	2	0
4.08432	2	1	1	2	3	2	0
4.08432	2	1	1	1	3	2	0
1.524	2	2	3	3	3	3	1
3.9624	2	2	2	1	1	3	1
3.9624	2	2	3	1	1	3	1
3.9624	2	2	3	2	1	3	1
3.9624	2	2	3	3	1	3	1
3.9624	2	2	2	2	1	3	1
3.9624	2	2	2	3	1	3	1
4.8768	1	1	2	1	3	3	1
2.4384	2	2	1	1	2	2	1
2.4384	2	2	1	2	2	2	1
2.4384	2	2	1	3	2	2	1
2.4384	2	2	2	1	2	2	1
2.4384	2	2	2	2	2	2	1
2.4384	2	2	2	3	2	2	1
2.4384	2	2	3	1	2	2	1
2.4384	2	2	3	2	2	2	1
2.4384	2	2	3	3	2	2	1
1.0668	2	2	2	2	3	1	0
1.0668	2	2	1	1	3	1	0
1.0668	2	2	1	2	3	1	0
1.0668	2	2	2	1	3	1	0
1.2192	2	2	3	3	3	1	0
1.2192	2	2	3	2	3	1	0
1.2192	2	2	3	1	3	1	0
1.2192	2	2	2	3	3	1	0
1.2192	2	2	2	2	3	1	0
1.2192	2	2	2	1	3	1	0
1.2192	2	2	1	1	3	1	0
1.2192	2	2	1	2	3	1	0
1.2192	2	2	1	3	3	1	0
1.2192	2	2	2	1	3	1	0
1.2192	2	2	1	1	3	1	0
1.2192	2	2	3	2	3	1	0
1.2192	2	2	2	2	3	1	0
1.2192	2	2	1	2	3	1	0
1.2192	2	2	3	1	3	1	0
1.2192	2	2	2	1	3	1	0
1.2192	2	2	1	1	3	1	0
3.44424	2	2	2	1	1	3	1
3.44424	2	2	3	2	1	3	1
3.44424	2	2	3	3	1	3	1
3.44424	2	2	2	2	1	3	1
3.44424	2	2	2	3	1	3	1
3.44424	2	2	3	1	1	3	1
3.048	2	1	2	3	3	2	1

Table C.1 (continued)

3.048	2	1	3	3	3	2	1
3.048	2	1	2	3	3	2	1
2.25552	2	1	3	1	2	1	0
2.25552	2	1	2	1	2	1	0
2.25552	2	1	1	1	2	1	0
3.048	2	1	1	3	3	3	1
3.048	2	1	2	3	3	3	1
3.048	2	1	3	3	3	3	1
3.048	2	1	2	3	3	3	1
3.048	2	1	3	3	3	3	1
3.048	2	1	2	3	3	3	1
3.048	2	1	3	3	3	3	1
1.6764	1	2	3	1	1	1	0
1.6764	1	2	2	1	1	1	0
1.6764	1	2	1	1	1	1	0
1.2192	1	1	2	2	1	1	1
1.8288	2	1	2	1	1	3	1
1.8288	2	1	2	1	1	1	1
1.8288	2	1	3	1	1	1	1
1.8288	2	1	2	1	1	3	1
1.8288	2	1	2	1	1	2	1
4.8768	2	2	2	2	3	3	1
4.8768	2	2	2	3	3	3	1
4.8768	2	2	3	2	3	3	1
4.8768	2	2	3	3	3	3	1
4.8768	2	2	2	1	1	3	1
4.8768	2	2	2	2	1	3	1
4.8768	2	2	2	3	1	3	1
4.8768	2	2	3	1	1	3	1
4.8768	2	2	3	2	1	3	1
4.8768	2	2	3	3	1	3	1
4.2672	2	2	2	2	1	3	1
3.9624	2	2	3	2	2	3	1
2.1336	1	2	2	2	2	1	1
2.1336	1	2	2	2	2	3	1
1.8288	1	2	2	1	1	1	1
3.9624	2	1	1	1	1	3	1
3.9624	2	1	2	1	1	3	1
3.9624	2	1	3	1	1	3	1
3.9624	2	1	2	2	1	3	0
3.9624	2	1	1	2	1	3	0
3.9624	2	1	2	1	1	3	0
3.9624	2	1	1	1	1	3	0
1.524	1	2	1	1	3	1	0
1.524	1	2	1	1	3	1	0
4.8768	2	2	2	2	2	3	1
2.1336	2	1	2	1	2	3	1
2.1336	2	1	2	1	2	1	1
2.1336	2	1	3	1	2	1	1
2.1336	2	1	2	1	2	2	1
2.1336	2	1	2	1	2	3	1
2.1336	2	1	2	1	2	2	1
3.3528	1	1	1	3	3	2	1
3.3528	1	1	2	3	3	2	1
3.3528	1	1	3	3	3	2	1
1.2192	1	1	1	2	1	1	1
1.2192	1	1	1	3	1	1	1
1.2192	1	1	2	2	1	1	1
1.2192	1	1	2	3	1	1	1

Table C.1 (continued)

1.2192	1	1	3	2	1	1	1
1.2192	1	1	3	3	1	1	1
1.2192	1	1	3	2	1	1	0
1.2192	1	1	2	2	1	1	0
1.2192	1	1	1	2	1	1	0
1.2192	1	1	3	2	1	1	0
1.2192	1	1	2	2	1	1	0
1.2192	1	1	1	2	1	1	0
1.2192	1	1	1	2	1	1	0
1.2192	1	1	1	2	1	1	0
4.2672	1	2	3	2	3	1	0
4.2672	1	2	3	1	3	1	0
4.2672	1	2	2	2	3	1	0
4.2672	1	2	2	1	3	1	0
4.2672	1	2	1	2	3	1	0
4.2672	1	2	1	1	3	1	0
1.92024	2	2	3	3	3	3	0
1.92024	2	2	3	2	3	3	0
1.92024	2	2	3	1	3	3	0
1.92024	2	2	2	3	3	3	0
1.92024	2	2	2	2	3	3	0
1.92024	2	2	2	1	3	3	0
1.92024	2	2	1	1	3	3	0
1.92024	2	2	1	2	3	3	0
1.92024	2	2	1	3	3	3	0
4.572	2	2	3	3	3	3	0
4.572	2	2	3	2	3	3	0
4.572	2	2	3	1	3	3	0
4.572	2	2	2	3	3	3	0
4.572	2	2	2	2	3	3	0
4.572	2	2	2	1	3	3	0
4.572	2	2	1	1	3	3	0
4.572	2	2	1	2	3	3	0
4.572	2	2	1	3	3	3	0
4.8768	1	2	2	1	2	1	1
3.9624	2	1	3	2	3	3	1
3.9624	2	1	3	2	3	3	1
3.9624	2	1	3	3	3	2	1
3.048	1	1	2	2	1	2	1
3.048	1	1	3	2	1	2	1
3.048	1	1	2	2	1	3	1
3.048	1	1	2	2	1	3	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
2.286	1	1	3	2	3	1	0
2.286	1	1	2	2	3	1	0
2.286	1	1	1	2	3	1	0
2.4384	1	1	2	2	1	2	1
3.3528	2	2	2	2	2	2	1

Table C.1 (continued)

3.3528	2	2	3	2	2	2	1
1.8288	1	2	2	2	2	2	0
3.6576	2	1	2	1	2	3	1
3.6576	2	1	2	2	2	3	1
3.6576	2	1	2	3	2	3	1
3.6576	2	1	3	1	2	3	1
3.6576	2	1	3	2	2	3	1
3.6576	2	1	3	3	2	3	1
3.6576	2	1	2	1	2	3	1
3.6576	2	1	2	2	2	3	1
3.6576	2	1	2	3	2	3	1
3.6576	2	1	3	1	2	3	1
3.6576	2	1	3	2	2	3	1
3.6576	2	1	3	3	2	3	1
4.2672	2	2	2	1	3	2	0
4.2672	2	2	1	1	3	2	0
6.096	2	2	3	2	3	2	0
6.096	2	2	3	1	3	2	0
6.096	2	2	2	2	3	2	0
6.096	2	2	2	1	3	2	0
6.096	2	2	1	2	3	2	0
6.096	2	2	1	1	3	2	0
4.572	2	2	3	2	2	2	1
4.572	2	2	3	3	2	2	1
2.4384	1	1	3	2	2	2	1
4.572	1	2	3	2	2	1	1
3.5052	2	2	3	2	3	3	1
3.5052	2	2	3	3	3	3	1
3.3528	2	2	2	2	3	3	0
3.3528	2	2	2	3	3	2	1
3.3528	2	2	3	3	3	2	1
4.2672	2	2	1	1	3	2	0
4.2672	2	2	2	1	3	2	0
4.2672	2	2	1	1	3	2	0
4.2672	2	2	1	3	3	2	0
4.2672	2	2	1	2	3	2	0
4.2672	2	2	1	1	3	2	0
4.572	2	2	3	2	3	3	0
4.572	2	2	2	2	3	3	0
4.572	2	2	1	2	3	3	0
3.048	2	2	2	2	2	3	0
3.048	2	2	2	2	2	3	0
3.048	2	2	2	2	2	3	0
4.32816	2	2	2	2	3	2	1
4.32816	2	2	2	3	3	2	1
4.32816	2	2	3	2	3	2	1
4.32816	2	2	3	3	3	2	1
3.81	2	2	1	3	3	2	0
3.81	2	2	1	2	3	2	0
3.81	2	2	1	1	3	2	0
2.98704	1	2	1	2	2	2	0
2.98704	1	2	1	1	2	2	0
4.572	2	2	2	2	3	2	0
4.572	2	2	1	2	3	2	0
2.7432	1	2	1	2	2	1	1
2.7432	1	2	1	3	2	1	1
2.7432	1	2	2	2	2	1	1
2.7432	1	2	2	3	2	1	1
2.7432	1	2	3	2	2	1	1

Table C.1 (continued)

2.7432	1	2	3	3	2	1	1
2.65176	2	2	3	1	3	2	0
4.8768	1	2	2	1	2	1	1
5.12064	1	2	2	2	1	1	0
5.12064	1	2	2	1	1	1	0
5.12064	1	2	1	2	1	1	0
5.12064	1	2	1	1	1	1	0
2.1336	1	2	2	2	3	1	0
2.1336	1	2	2	1	3	1	0
2.1336	1	2	1	2	3	1	0
2.1336	1	2	1	1	3	1	0
2.1336	1	2	2	2	3	1	0
2.1336	1	2	2	1	3	1	0
2.1336	1	2	1	2	3	1	0
2.1336	1	2	1	1	3	1	0
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.524	1	1	1	1	1	1	1
1.524	1	1	1	2	1	1	1
1.524	1	1	1	3	1	1	1
1.524	1	1	2	1	1	1	1
1.524	1	1	2	2	1	1	1
1.524	1	1	2	3	1	1	1
1.524	1	1	3	1	1	1	1
1.524	1	1	3	2	1	1	1
1.524	1	1	3	3	1	1	1
1.8288	2	1	2	1	1	1	1
1.8288	2	1	3	1	1	1	1
2.4384	2	1	1	1	2	1	1
2.4384	2	1	2	1	2	1	1
2.4384	2	1	3	1	2	1	1

Table C.1 (continued)

2.4384	2	1	2	1	2	3	1
2.4384	2	1	3	1	2	3	1
2.4384	2	1	2	1	2	2	1
1.73736	2	2	2	3	3	2	0
1.73736	2	2	2	2	3	2	0
1.73736	2	2	2	1	3	2	0
1.73736	2	2	1	3	3	2	0
1.73736	2	2	1	2	3	2	0
1.73736	2	2	1	1	3	2	0
1.524	2	2	2	1	1	1	1
1.524	2	2	2	2	1	1	1
1.524	2	2	2	3	1	1	1
1.524	2	2	3	1	1	1	1
1.524	2	2	3	2	1	1	1
1.524	2	2	3	3	1	1	1
2.1336	2	2	2	2	1	2	1
2.1336	2	2	2	3	1	2	1
2.1336	2	2	3	2	1	2	1
2.1336	2	2	3	3	1	2	1
2.7432	2	2	3	1	3	1	0
2.7432	2	2	3	1	3	1	0
2.8956	2	2	2	1	3	1	0
3.048	2	1	2	1	3	2	1
3.048	2	1	2	2	3	2	1
3.048	2	1	2	3	3	2	1
3.048	2	1	3	1	3	2	1
3.048	2	1	3	2	3	2	1
3.048	2	1	3	3	3	2	1
3.048	2	1	2	2	3	2	1
3.048	2	1	2	3	3	2	1
3.048	2	1	3	2	3	2	1
3.048	2	1	3	3	3	2	1
3.048	1	1	2	2	2	1	1
3.048	1	1	3	2	2	1	1
3.048	1	1	3	2	2	1	1
3.048	1	1	2	1	2	1	1
3.048	1	1	2	2	2	1	1
3.048	1	1	2	3	2	1	1
3.048	1	1	3	1	2	1	1
3.048	1	1	3	2	2	1	1
3.048	1	1	3	3	2	1	1
2.7432	1	2	3	2	2	1	0
2.7432	1	2	2	2	2	1	0
2.7432	1	2	1	2	2	1	0
3.048	1	2	3	2	2	1	0
3.048	1	2	2	2	2	1	0
3.048	1	2	1	2	2	1	0
2.7432	2	1	2	2	2	3	1
2.7432	2	1	1	3	2	2	0
2.7432	2	1	1	2	2	2	0
2.7432	2	1	1	1	2	2	0
1.49352	2	2	2	2	3	3	0
3.6576	2	1	3	3	3	1	0
3.6576	2	1	3	2	3	1	0
3.6576	2	1	3	1	3	1	0
3.6576	2	1	2	3	3	1	0
3.6576	2	1	2	2	3	1	0
3.6576	2	1	2	1	3	1	0
3.6576	2	1	1	3	3	1	0

Table C.1 (continued)

3.6576	2	1	1	2	3	1	0
3.6576	2	1	1	1	3	1	0
5.0292	2	2	2	2	3	2	1
5.0292	2	2	3	2	3	2	1
3.048	1	2	1	2	1	1	1
3.3528	2	2	2	1	1	3	1
3.3528	2	2	2	2	1	3	1
3.3528	2	2	2	3	1	3	1
3.3528	2	2	3	1	1	3	1
3.3528	2	2	3	2	1	3	1
3.3528	2	2	3	3	1	3	1
4.8768	1	2	1	1	3	2	0
2.52984	2	2	3	1	3	3	1
3.048	2	1	2	2	1	3	1
1.8288	2	2	2	2	3	3	0
1.524	1	2	3	1	3	1	0
1.524	1	2	2	1	3	1	0
1.524	1	2	1	1	3	1	0
1.524	1	2	3	1	3	1	0
1.524	1	2	2	1	3	1	0
1.524	1	2	1	1	3	1	0
4.8768	2	1	1	1	1	2	1
4.8768	2	1	1	2	1	2	1
4.8768	2	1	1	3	1	2	1
4.8768	2	1	2	1	1	2	1
4.8768	2	1	2	2	1	2	1
4.8768	2	1	2	3	1	2	1
4.8768	2	1	3	1	1	2	1
4.8768	2	1	3	2	1	2	1
4.8768	2	1	3	3	1	2	1
4.8768	2	1	1	1	1	3	1
4.8768	2	1	1	2	1	3	1
4.8768	2	1	1	3	1	3	1
4.8768	2	1	2	1	1	3	1
4.8768	2	1	2	2	1	3	1
4.8768	2	1	2	3	1	3	1
4.8768	2	1	3	1	1	3	1
4.8768	2	1	3	2	1	3	1
4.8768	2	1	3	3	1	3	1
1.8288	1	1	2	2	1	1	1
5.1816	1	2	3	2	3	2	0
5.1816	1	2	3	1	3	2	0
5.1816	1	2	2	2	3	2	0
5.1816	1	2	2	1	3	2	0
5.1816	1	2	1	2	3	2	0
5.1816	1	2	1	1	3	2	0
2.8956	2	1	2	1	3	2	1
2.8956	2	1	2	2	3	2	1
2.8956	2	1	2	3	3	2	1
2.8956	2	1	3	1	3	2	1
2.8956	2	1	3	2	3	2	1
2.8956	2	1	3	3	3	2	1
2.8956	2	1	1	2	3	3	1
2.8956	2	1	2	2	3	3	1
2.8956	2	1	3	2	3	3	1
2.8956	2	1	3	2	3	3	1
2.8956	2	1	3	3	3	3	1
2.8956	2	1	1	3	3	3	1
1.524	1	2	3	1	2	1	0

Table C.1 (continued)

1.524	1	2	2	1	2	1	0
1.524	1	2	1	1	2	1	0
3.6576	1	1	2	1	3	2	1
3.6576	1	1	3	1	3	2	1
3.6576	1	1	2	2	3	2	1
3.6576	1	1	2	3	3	2	1
3.6576	1	1	3	3	3	2	1
2.4384	1	1	2	2	3	3	1
2.4384	1	1	3	2	3	3	1
2.4384	1	1	2	3	3	3	1
2.4384	1	1	3	3	3	3	1
3.3528	2	1	2	3	2	2	0
3.3528	2	1	1	3	2	2	0
3.3528	2	1	2	2	2	2	0
3.3528	2	1	2	1	2	2	0
3.3528	2	1	1	2	2	2	0
3.3528	2	1	1	1	2	2	0
3.3528	2	1	2	2	2	3	1
3.3528	2	1	3	2	2	3	1
3.3528	2	1	2	3	2	3	1
3.3528	2	1	3	3	2	3	1
3.3528	2	1	2	1	2	2	1
3.3528	2	1	3	2	2	2	1
3.3528	2	1	3	3	2	2	1
3.3528	2	1	2	2	2	2	1
3.3528	2	1	2	3	2	2	1
3.3528	2	1	3	1	2	2	1
1.0668	2	2	3	3	3	1	0
1.0668	2	2	3	2	3	1	0
1.0668	2	2	3	1	3	1	0
1.0668	2	2	2	3	3	1	0
1.0668	2	2	2	2	3	1	0
1.0668	2	2	2	1	3	1	0
1.0668	2	2	1	1	3	1	0
1.0668	2	2	1	2	3	1	0
1.0668	2	2	1	3	3	1	0
4.8768	2	2	2	1	1	3	1
4.8768	2	2	2	2	1	3	1
4.8768	2	2	2	3	1	3	1
4.8768	2	2	3	1	1	3	1
4.8768	2	2	3	2	1	3	1
4.8768	2	2	3	3	1	3	1
4.8768	2	2	2	1	1	3	1
4.8768	2	2	2	2	1	3	1
4.8768	2	2	2	3	1	3	1
4.8768	2	2	3	1	1	3	1
4.8768	2	2	3	2	1	3	1
4.8768	2	2	3	3	1	3	1
1.2192	1	1	2	2	1	1	0
1.2192	1	1	1	2	1	1	0
1.2192	1	1	2	1	1	1	0
1.2192	1	1	1	1	1	1	0
7.3152	2	2	3	2	3	3	1
7.3152	2	2	3	3	3	3	1
3.9624	2	1	2	2	3	3	1
3.9624	2	1	2	3	3	3	1
3.9624	2	1	3	2	3	3	1
3.9624	2	1	3	3	3	3	1
4.2672	2	2	2	3	3	2	0

Table C.1 (continued)

4.2672	2	2	2	2	3	2	0
4.2672	2	2	2	1	3	2	0
4.2672	2	2	1	3	3	2	0
4.2672	2	2	1	2	3	2	0
4.2672	2	2	1	1	3	2	0
2.4384	2	1	2	1	1	2	1
2.4384	2	1	2	2	1	2	1
2.4384	2	1	2	3	1	2	1
2.4384	2	1	3	1	1	2	1
2.4384	2	1	3	2	1	2	1
2.4384	2	1	3	3	1	2	1
4.572	2	2	2	2	3	2	1
4.572	2	2	3	2	3	2	1
4.572	2	2	2	3	3	2	1
4.572	2	2	3	3	3	2	1
4.1148	2	2	2	3	3	2	0
4.1148	2	2	2	2	3	2	0
4.1148	2	2	2	1	3	2	0
4.1148	2	2	1	3	3	2	0
4.1148	2	2	1	2	3	2	0
4.1148	2	2	1	1	3	2	0
4.572	1	2	2	1	1	1	1
4.572	1	2	2	2	1	1	1
4.572	1	2	2	3	1	1	1
4.572	1	2	3	1	1	1	1
4.572	1	2	3	2	1	1	1
4.572	1	2	3	3	1	1	1
2.7432	1	2	1	1	2	1	1
2.7432	1	2	1	2	2	1	1
2.7432	1	2	1	3	2	1	1
2.7432	1	2	2	1	2	1	1
2.7432	1	2	2	2	2	1	1
2.7432	1	2	2	3	2	1	1
2.7432	1	2	3	1	2	1	1
2.7432	1	2	3	2	2	1	1
2.7432	1	2	3	3	2	1	1
2.7432	2	1	3	3	2	3	0
2.7432	2	1	3	2	2	3	0
2.7432	2	1	3	1	2	3	0
2.7432	2	1	2	3	2	3	0
2.7432	2	1	2	2	2	3	0
2.7432	2	1	2	1	2	3	0
2.7432	2	1	1	3	2	3	0
2.7432	2	1	1	2	2	3	0
2.7432	2	1	1	1	2	3	0
2.8956	2	1	2	1	3	3	1
2.8956	2	1	2	2	3	3	1
2.8956	2	1	2	3	3	3	1
2.8956	2	1	3	1	3	3	1
2.8956	2	1	3	2	3	3	1
2.8956	2	1	3	3	3	3	1

APPENDIX D

LOGISTIC REGRESSION MODEL CODE (R)

```
#Set File Path
#setwd("\\Users\\b5ecgsgf\\Desktop\\Temp_telework\\Mississippi
State\\Logistic Regression")

################################  Inputs ################################
##---------Data processing-------##
# normalize the input points in the range [0,1]
input.normalize<-function(X,xmax,xmin){
  for(i in 1:nrow(X)){
    X[i,] = (X[i,]-xmin)/(xmax-xmin)
  }
  return(X)
}

###------------Define training data with N observations

filename = 'LO_Working Data_20pct Test_Feb.16.csv'
library(VIM)

N = 581 # Number of Samples
df = read.csv(filename,header = TRUE)
str(df) # Determine which variables are integers.
summary(df)
aggr(x=df, numbers=TRUE)
df$X2 <- factor(df$X2)
df$X3 <- factor(df$X3)
df$X4 <- factor(df$X4)
df$X5 <- factor(df$X5)
df$X6 <- factor(df$X6)
df$X7 <- factor(df$X7)
df$Y <- factor(df$Y)
adf <- kNN(df,
variable = c('X4','X5','X7'),
dist_var = c('X1','X2','X3','X6'),
weights = "auto",
catFun = maxCat,
k = 8)

summary(adf)
aggr(adf, delimiter="_imp", numbers=TRUE)
aggr(adf, delimiter="_imp")

# Insert kNN Imputation
str(adf)
write.csv(adf, file="adf.csv",row.names = FALSE)
colnames(adf) = c('X1','X2','X3','X4','X5','X6','X7','Y')
colnames(adf) =
c('Load_H','Slope','Constr.','Depth','Duration','Erosion','Days','Breach')

N = 465
X = adf[1:N,1:7] # inputs
Y = adf[1:N,8]   # response
data_train=data.frame(X,Y)
str(data_train)
###--------------test data
```

```
Xpred = adf[(N+1):581,1:7]
Ypred_true = adf[(N+1):581,8]
data_test=data.frame(Xpred,Ypred_true)
str(data_test)
##------------Fit the Logistic model-------##--

### 1) Base LR Model utilizing all variables

Logistic_1<- glm(Y ~.,data = data_train, family="binomial")

#(Logistic_1)
summary(Logistic_1)
anova(Logistic_1, test = "Chisq")

res2 <-predict(Logistic_1,data_test,type='response')
res <-predict(Logistic_1,data_train,type='response')

#validate the model _Confusion matrix
confmatrix <- table(Actual_value=data_train$Y, predicted_value=res>0.5)
confmatrix

#validate the model _Confusion matrix for test data
confmatrix2 <- table(Actual_value=data_test$Y, predicted_value=res2>0.5)
confmatrix2

#Accuracy train
(confmatrix[[1,1]]+confmatrix[[2,2]]) / sum(confmatrix)

#Accuracy test
(confmatrix2[[1,1]]+confmatrix2[[2,2]]) / sum(confmatrix2)

library(MASS)
library(magrittr)
N = 465
X = adf[1:N,1:7] # inputs
Y = adf[1:N,8]   # response
data_train=data.frame(X,Y)
model <- glm(Y ~
Constr.+Depth+Duration+Erosion+(Depth*Days)+(Constr.*Erosion)
             ,data = data_train, family = "binomial") %>%
  stepAIC(trace = TRUE)

summary(model)
anova(model, test = "Chisq")
coef(model)
model_check=predict(model,data_train,type='response')

#odds ratio
exp(coef(model))

#Accuracy_Confusion matrix_Test data
res_test <-predict(model,data_test,type='response')
confmatrix_model <- table(Actual_value=data_test$Y,
predicted_value=res_test>0.5)
confmatrix_model
```

```
#Accuracy test
(confmatrix_model[[1,1]]+confmatrix_model[[2,2]]) / sum(confmatrix_model)

###-------------k cross fold validation
#install.packages("caret")
library(caret)
set.seed(100)
#colnames(adf) =
c('Load_H','Slope','Constr.','Depth','Duration','Errosion','Days','Breach')
N = 581
X = adf[1:N,1:7] # inputs
Y = adf[1:N,8]   # response
data_train_CV=data.frame(X,Y)
str(data_train_CV)
train_control <- trainControl(method="cv", number=5)

model_val <- train(Y
~Constr.+Depth+Duration+Erosion+Days+(Depth*Days)+(Constr.*Erosion),
data=data_train_CV
                   , trControl=train_control
                   , method='glm',family=binomial()
                   )
print(model_val)
model_val$results
summary(model_val)
exp(coef(model))
```