Evaluating Levee Failure Susceptibility on the Lower Mississippi River using Logistic Regression Analysis

Dorian POPOVIC

1 Introduction

The Mississippi River is the longest river of North America and the twentieth longest in the world. As the central river artery of a highly industrialized region, the Mississippi River has been subjected to a remarkable degree of human control and modification. Over the past two centuries, the Mississippi River floodplain has been transformed from forests and prairies to a predominantly agricultural land use, and additional floodplain land has been developed for residential, commercial and industrial expansion.

Levees characterize any low ridge or earthen embankment built along the edges of a stream or river channel to prevent flooding of the adjacent land. They were first used on the Mississippi River in the early 1700s. Although several levee failures that led to major flood incidents demonstrated the limits of "levee-only" protection policies, levees still continue to be an integral part of flood-control efforts today.

As several circumstances might influence levee failure susceptibility, this investigation aims at assessing the relative importance of geologic, geomorphic and other physical factors that led to levee failures in the past century along the *Lower Mississippi River* (LMR). First, the data for this investigation and how they relate to levee failures will be examined through exploratory data analysis. Second, two logistic regression models will be developed based on different significance thresholds. Their efficiencies will finally be assessed and compared to obtain the best logistic regression model for levee failure.

2 Exploratory Data Analysis

The dataset for this study is the `lmr_levee.dat` dataset retrieved from A. Flor et. al.'s study [1]. 82 observations of 13 independent variables provide information about a range of LMR site conditions that might affect the likelihood of levee failure at a given location. Table 1 summarizes the variables. They belong to 4 different types: numerical, ratio, categorical and boolean. The goal is to investigate the relationship between the independent variables and the `failure` dependent binary variable. Before fitting a model, exploratory data analysis needs to be performed in order to gain insight into the structure of the dataset as well as the relative importance of the independent variables in predicting levee failures.

Table 1: Levee failure parameters and variable types.

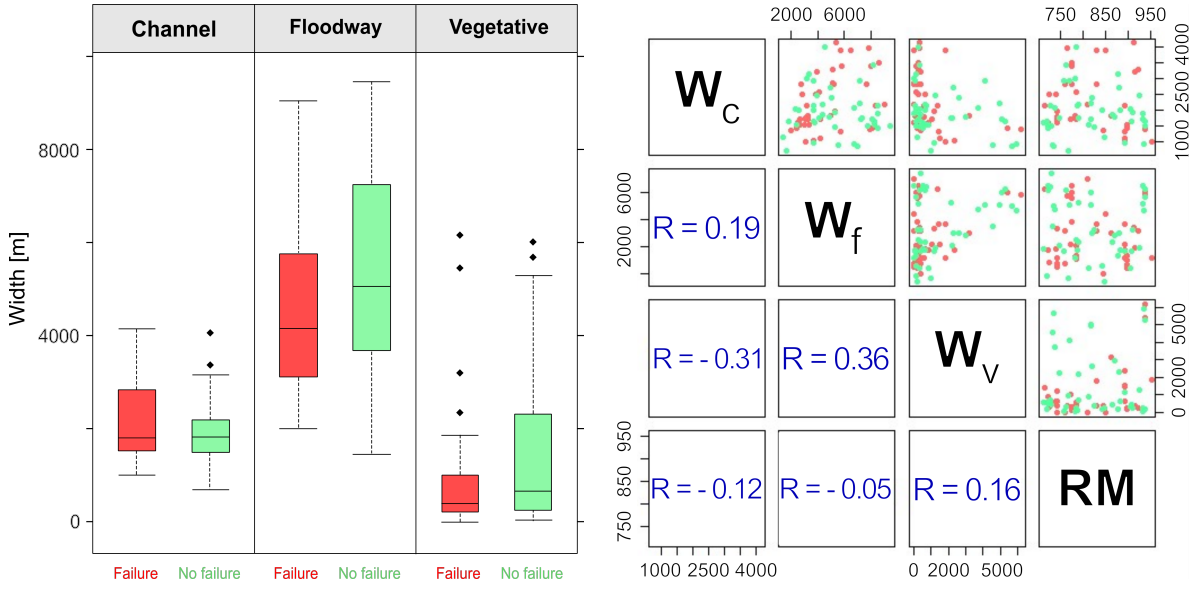
Parameter	*Code*	*Variable type*
Year of levee data	Y	Categorical
River mileage	RM	Numerical
Coarse-grain channel fill	F	Boolean
Borrow pit indicator	BP	Boolean
Meander location	M	Categorical
Channel width	W_c	Numerical
Floodway width	W_f	Numerical
Constriction factor	CF	Ratio
Land cover type	LC	Categorical
Vegetative buffer width	W_v	Numerical
Channel sinuosity	S	Ratio
Dredging intensity	D	Ratio
Bank revetment	R	Boolean

2.1 Numerical Variables

The dataset contains 4 numerical variables: `river mileage`, `channel width`, `floodway width` and `vegetative buffer width`. Figure 1 depicts the main characteristics of those variables.

The distributions of the variables that describe the widths of different levee features are displayed using box plots in Figure 1a. The median values of channel and vegetative buffer widths are similar irrespective of levee failure or not. The median floodway is wider for levees that did not fail. Since in all three instances, none of the median values lie outside the box plot of comparison, there is not likely to be any significant difference between the groups. Floodway widths are more dispersed. There are some outliers for vegetative buffers and channels. Finally, while the channel widths of levees that did not fail are symmetrically distributed, the distributions of all other width variables are positively skewed relative to failure or non-failure.

Figure 1b shows the correlation plots of the numerical variables. There does not appear to be any strong correlation between any of them. The highest positive R^2 is 0.36 between W_v and W_f and the highest negative one is -0.31 between W_c and W_v. Hence multicollinearity between numerical variables is not a concern in this investigation. Additionally there seems to be possible interactions between certain variables regarding levee failure. For example, levees that have low river mileage coupled to a wide vegetative buffer or a wide floodway coupled to a wide vegetative buffer seem to be less prone to failure.



(a) Width box plots of main levee components.

(b) Numerical variables correlation plots.

Figure 1: Exploratory analysis of numerical variables. Failure points are represented in red and non-failure points are represented in green.

2.2 Categorical Variables

There are 3 categorical variables among the independent variables: `data year`, `meander location` and `land cover type`. To assess if levees show differences in any of those categories relative to failure occurrence, Figure 2 shows their comparison using bar plots.

(a) Year categories. **(b)** Meander location categories. **(c)** Land cover categories.

Figure 2: Exploratory analysis of categorical variables distributions.

Figure 2a shows the comparison for the 3 different years at which levee data were collected. The same number of failing and non-failing levees observations was made for each year category. Therefore, this variable will not be relevant for the final levee failure susceptibility predictive model.

A meander is one of a series of regular sinuous curves, loops or turns in the channel of a river. Here, the meander variable captures where on a meander the failure/non-failure point occurred: inside (1), outside (2), chute (3) or straight (4). On Figure 2b it can be observed that failure points are more often located chute or inside a meander while non-failure points happen more often on outside or straight meander locations.

The land cover at each failure/non-failure points was categorized into 4 groups: open water (1), grassy (2), agricultural (3), and forested (4). On Figure 2c it can be observed that failures happen less often when agricultural or forest land covers are present at the levee site while grass and water land covers seem to favor failure points. Failures seem to have happened systematically when the levee was covered by open water, although the number of observations for this specific land cover is low.

2.3 Binary Variables

Among the independent variables, 3 of them are binary: `sediments`, which indicates the presence of a coarse channel fill beneath levee locations, `borrow pit`, which indicates the presence of a borrow pit on the levee and `revetment` which indicates the presence of a bank revetment. Table 2 shows the contingency table for those variables.

The variable sediments does not seem to show any distinctive difference regarding levee failure susceptibility. A very moderate increase in failure points can be observed when there are coarse channel fills beneath levees.

On the other hand, the presence of a borrow pit appears to quite significantly lower the instances of failure points and increase the likelihood of non-failure points.

Finally, among the 82 observations in the dataset, only two indicated the presence of a bank revetment and both were located on a non-failure point. However, although bank revetments seem to systematically avert failure, it might only be due to the lack of observations and this result should be interpreted carefully.

Table 2: Contingency table for binary variables.

	Sediments		**Borrow pit**		**Revetment**	
	No (40)	*Yes* (42)	*No* (44)	*Yes* (38)	*No* (80)	*Yes* (2)
Failure	18 (45%)	23 (55%)	26 (59%)	15 (39%)	41 (51%)	0 (0%)
No failure	22 (55%)	19 (45%)	18 (41%)	23 (61%)	39 (49%)	2 (100%)

2.4 Ratio Variables

The last 3 variables of the dataset are ratio variables, which are interval variables with the added condition that zero measurements indicate that there is none of that variable. `Constriction factor` is a ratio measure of the lateral constriction at the given location. `Channel sinuosity` is calculated as the length along the river divided by the straight-line distance along the river valley. Rivers can have sinuosity ranging from 1 up to 3. Finally, `dredging intensity` measures the removal of sediments and debris from the bottom of the LMR. Figure 3 shows the scatter plots of the ratio variables.

The constriction factor and channel sinuosity are numerical variables while dredging intensity is a categorical ratio variable. There is no important correlation between any of the ratio variables in the dataset. Levees that don't fail seem to have less sinuous channels while a decreased constriction factor seems to favor levee failure. Here again, possible interactions between variables can be observed. For instance, a high channel sinuosity coupled to a high dredging intensity also seems to increase levee failure susceptibility.

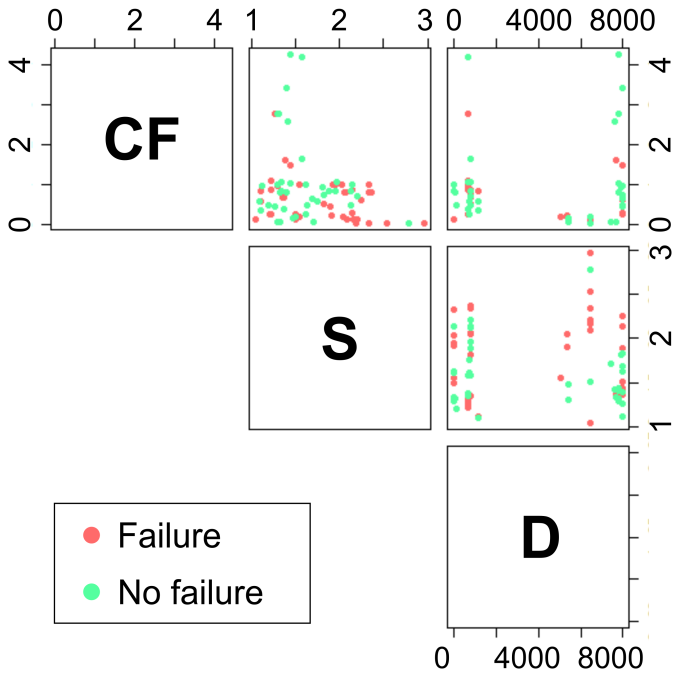


Figure 3: Exploratory analysis of ratio scale variables.

2.5 Conclusion

In conclusion, the dataset for this investigation contains four different variable types. The `year` and `river mileage` variables will most likely be irrelevant for the levee failure prediction model. The `revetment` variable should probably not be included to avoid erroneously constructed models, unless the extremely uneven distribution is dealt with. All other variables seem to show some degree of relevancy in regard to levee failure susceptibility, especially the `meander`, `land cover`, `constriction factor`, `channel sinuosity` and `borrow pit` variables. Because there are several variable types that need to be included in the statistical model, logistic regression is the optimal choice.

3 Model Selection

3.1 Logistic Regression

Logistic regression is a statistical analysis that predicts the group membership of a dependent variable based on a set of independent variables. It is used to predict binary variables and is mathematically defined as:

$$logit(p) = log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{1}$$

Here, the focus is on occurrence probability of levee failure at a given site. The main advantage of a logistic regression model is that the independent variables can take on many types: numerical, categorical, boolean or ratio scale. Additionally, logistic regression allows for the comparison of models with different numbers of predictors. The simplest model would be the one that takes only the constant into account while the most complex one would include all predictors. However, since not every predictor contributes significantly to the outcome, each of the models and variables must be evaluated based on the significance of each predictor. The final model is the one with the highest accuracy and significance and that uses the fewest predictors possible.

3.2 Bivariate Response Models

Before constructing a logistic regression model, the significance of each explanatory variable's correlation with the `failure` variable must be individually assessed. The results are summarized in Table 3.

When possible, categorical and boolean variables were entered in a bivariate response model and evaluated for significance using the Pearson's Chi-squared test with the *null hypothesis* $H_0 : \beta = 0$ assuming a null regression coefficient and the *alternative hypothesis* $H_a : \beta \neq 0$ assuming a regression coefficient different from zero. In this statistical hypothesis test, the sampling distribution of the test statistic is a Chi-square distribution when the null hypothesis is true. Low P-values imply that H_0 can be rejected. Only the `meander` (M) variable has an effect on levee failure susceptibility that is statistically significant at a

Table 3: Individual variables significance results.

Pearson's χ^2	χ^2	*df*	*P-value*
M	9.09	3	0.03 *
F	0.78	1	0.38
BP	3.14	1	0.08 .
LRBM	Std error	*z value*	*Pr(>|z|)*
LC	0.26	-2.10	0.04 *
CF	0.35	-1.99	0.05 *
S	0.55	2.23	0.03 *
D	6.61×10^{-7}	-0.55	0.59
$\mathbf{W}_c$	0.00	1.70	0.09 .
$\mathbf{W}_f$	0.00	-1.22	0.22
$\mathbf{W}_v$	0.00	-1.81	0.07 .

5% level. If the significance level is lowered to $\geq 90\%$, the `borrow pit` (BP) variable yields a statistically significant effect for levee failure susceptibility. `Sediments/coarse-grain channel fill` (F) indicators have no significant effect on levee failure.

Sometimes, the Chi-squared method could not be used because the approximation to the distribution of the test statistic relies on the counts being roughly normally distributed. If many of the expected counts are very small, the approximation may be poor and yield incorrect results. In this case, a simple bivariate logistic regression model (LRBM) response was assessed. The hypotheses were $H_0 : \beta = 0$ and $H_a : \beta \neq 0$. Again, low P-values imply that H_0 can be rejected. The same method was applied to numerical and ratio variables. Three additional variables showed a statistically significant effect on levee failure susceptibility at a 5% level: `land cover` (LC), `constriction factor` (CF) and `channel sinuosity` (S). If the significance level is lowered to $\geq 90\%$, two additional predictors yielded a statistically significant effect on levee failure susceptibility: `channel width` (W_c) and `vegetative buffer width` (W_v). Although they will not be included in the models, the `year`, `river mileage` and `revetment` variables were also tested and as expected they yielded P-values equal to or very close to 1.

3.3 Logistic Regression Models

Two different logistic regression models, *conservative* and *liberal*, based on different significance thresholds, were implemented to investigate levee failure susceptibility.

3.3.1 Conservative Model

The first logistic regression model for predicting levee failure included only the variables that met the $\geq 95\%$ significance level condition (*conservative model*). Four variables met the significance threshold when tested individually, resulting in the following model:

$$\frac{P}{1-P} = e^{(-1.606+0.590M_1-1.079M_2+0.655M_3+17.464LC_1+0.532LC_2-0.422LC_3-0.423CF+1.034S)} \tag{2}$$

The results for the conservative model are given in Table 4. It identified four significant variables to predict levee failure: `meander location`, `land cover type`, `constriction factor` and `channel sinuosity`.

3.3.2 Liberal Model

The second logistic regression model for predicting levee failure lowered the significance threshold to $\geq 90\%$. Although it is not always recommended, this moderate reduction explored the possibility of a broader model with an increased number of predictors (*liberal model*). Three additional variables were included in the following model:

$$\frac{P}{1-P} = e^{(-1.469+0.425M_1-1.086M_2+0.859M_3+17.740LC_1+0.237LC_2-0.312LC_3-0.587CF+0.884S-1.132BP+0.000W_c+0.000W_v)} \tag{3}$$

The results for the liberal model are given in Table 5. On top of the previously identified significant predictors, it included three additional variables to predict levee failure: `borrow pits`, `channel width` and `vegetative buffer width`.

Table 4: Conservative model results.

Variable	*P-value*	*Odds ratio*
(Intercept)	0.19	0.20
$\mathbf{M}_1$	0.40	1.80
$\mathbf{M}_2$	0.20	0.34
$\mathbf{M}_3$	0.38	1.93
$\mathbf{M}_4$	-	-
$\mathbf{LC}_1$	0.99	inf.
$\mathbf{LC}_2$	0.48	1.70
$\mathbf{LC}_3$	0.50	0.66
$\mathbf{LC}_4$	-	-
CF	0.28	0.66
S	0.11	2.81

Table 5: Liberal models results.

Variable	*P-value*	*Odds ratio*
(Intercept)	0.41	0.23
$\mathbf{M}_1$	0.55	1.53
$\mathbf{M}_2$	0.22	0.34
$\mathbf{M}_3$	0.30	2.36
$\mathbf{M}_4$	-	-
$\mathbf{LC}_1$	0.99	inf.
$\mathbf{LC}_2$	0.78	1.27
$\mathbf{LC}_3$	0.64	0.73
$\mathbf{LC}_4$	-	-
CF	0.18	0.56
S	0.21	2.42
BP	0.07 .	0.32
$\mathbf{W}_c$	0.48	1.00
$\mathbf{W}_v$	0.43	1.00

The efficiency of both conservative and liberal models for predicting levee failure now needs to be compared to select the best model.

3.4 Models Assessment Comparison

A comparison of the conservative and liberal models efficiency is first performed through a likelihood ratio test. The deviance between the two models and the null models that include only the intercept is compared. The difference of deviance is subsequently compared to a χ^2 distribution. The results are given in Table 6. The conservative and liberal models yield a P-value of 0.003 and 0.005 respectively. The conservative model has a lower P-value, although the difference is not very important. Both models are significantly better at a 1% level than the null model for predicting failure.

Table 6: Conservative and liberal models analysis of deviance.

Conservative	*Resid. Df*	*Resid. Dev*	*Df*	*Deviance*	*Pr(> Chi)*
Null Model	81	113.68			
Conservative Model	73	90.67	8	23.01	0.003 **
Liberal	*Resid. Df*	*Resid. Dev*	*Df*	*Deviance*	*Pr(> Chi)*
Null Model	81	113.68			
Liberal Model	70	86.86	11	26.82	0.005 **

In order to gain further insight into the differences in efficiency between the two models for predicting levee failure, two different versions of the logistic regression R^2 are computed for each one of them. They evaluate the goodness-of-fit of the models. Logistic regression models are fitted using the method of maximum likelihood - i.e. the parameter estimates are those values which maximize the likelihood of the data which have been observed. McFadden's R^2 measure is defined as $R^2_{\text{McFadden}} = 1 - log(L_c)/log(L_{\text{null}})$ where L_c denotes

the (maximized) likelihood value from the current fitted model, and L_{null} denotes the corresponding value for the null model. Nagelkerke's R^2 is an adjusted version of the Cox and Snell's method that adjusts the scale of the statistic to cover the full range from 0 to 1. The results are given in Table 7. In both cases, the liberal model shows an increase in R^2 values, indicating a better fit.

Table 7: Conservative and liberal models R^2 comparison.

	McFadden's R^2	*Nagelkerke's* R^2
Conservative Model	0.20	0.33
Liberal Model	0.24	0.37

The final step in model efficiency comparison is to assess the conservative and liberal models overall performance. Tables 8 and 9 compare the models performances when predicting levee failure using confusion matrices from which several evaluation measures can be obtained. *Sensitivity* is defined as the number of correct positive predictions divided by the total number of positives $TP/(TP + FN)$. *Specificity* is defined as the number of correct negative predictions divided by the total number of negatives $TN/(TN + FP)$. *Accuracy* is the number of all correct predictions divided by the total number of observations $(TP + TN)/(P + N)$.

The conservative model has a 69.5% accuracy, a 73.2% sensitivity and a 65.9% specificity. The liberal model has an accuracy , a sensitivity and a specificity that are all three equal to 70.7%. Therefore the liberal model has better accuracy and specificity while the conservative model has a better sensitivity.

These results should be interpreted with caution. To have a better estimation of the predictive power of both models, separate training and testing sets should be created, otherwise the results can be biased and are not the best indicator of actual performance.

Table 8: Conservative model confusion matrix.

		Prediction outcome		
		NF	**F**	**total**
actual value	**NF**	TP: **30**	FN: **11**	41
	F	FP: **14**	TN: **27**	41
	total	44	38	

Table 9: Liberal model confusion matrix.

		Prediction outcome		
		NF	**F**	**total**
actual value	**NF**	TP: **29**	FN: **12**	41
	F	FP: **12**	TN: **29**	41
	total	41	41	

4 Discussion

The logistic regression models here were created to identify which site characteristics significantly influenced the occurrence of levee failures. The main constraint in this investigation was data availability. Logistic regression modeling is most robust for large data sets, particularly in complex response systems with a large number of factors. Despite this limitation, both models yielded promising results for predicting levee failure.

In accordance with the exploratory data analysis, the conservative LMR model identified four variables to be significantly associated with levee failure susceptibility: location on a meander, land-cover type, constriction factor and channel sinuosity. The model suggested that levees on the inside of meander bends (M_1) and along chutes (M_3) may fail more often, as the odds ratio in Table 4 are bigger than 1, than on the outside of meander bends (M_2) or straight (M_4). The same was hypothesised during exploratory data analysis (Figure 2b). Open Water (LC_1) and grassy land cover (LC_2) may be more prone to failure than agricultural (LC_3) or forested land cover (LC_4). This was also expected from exploratory data analysis (Figure 2c). Finally, sinuosity increased and constriction decreased the chances of levee failure, as was previously proposed from Figure 3.

By lowering the significance threshold to 90%, the liberal model was created. A relaxed statistical threshold is not generally recommended, but at the least shows how a more complete model using a larger sample size could appear. The liberal model identified three additional significant variables. It suggested that the presence of borrow pits decreased the likelihood of failure ($odds\ ratio = 0.32$), which was exactly observed in column 5 of Table 2. In contrast, channel width and vegetated buffer width had no effect on failure susceptibility ($odds\,ratio = 1$) and did not contribute to the model, as expected when observing Figure 1a.

When comparing both models, the liberal one seems to show better performance when predicting levee failure. However, this should not lead to choosing it over the conservative model for different applications. Indeed, the performance was tested on the same data that were used for training, therefore increased performance when adding predictors is expected. Rather it was developed to see how a more complete model using a larger sample size could be obtained. To avoid over fitting and poor performance for generalized applications, the conservative model should be used when dealing with other data.

5 Conclusion

The question posed in this investigation was: *By looking at site characteristics at locations along the Lower Mississippi River where levees have previously failed, is it possible to predict levees that are more susceptible to fail?*

The null hypothesis H_0 tested was: *the site characteristics do not predict the presence or absence of a levee failure.* The alternative hypothesis H_a was: *the site characteristics do predict levee failure.* Using the data available, the results here quantify the relative importance of factors in past levee failures along the LMR and help predict areas that may be more prone to failure in the future.

The results suggest that location along a meander bend, land cover type, constriction over time and channel sinuosity are significantly associated with the occurrence of levee failures. The final conservative model is:

$$\frac{P}{1-P} = e^{(-1.606+0.590M_1-1.079M_2+0.655M_3+17.464LC_1+0.532LC_2-0.422LC_3-0.423CF+1.034S)} \quad (4)$$

These results could potentially assist engineers and decision-makers to choose more suitable locations and design for levees in the future. Additional work is needed to increase data size to be able to try a different approach such as cross-validation. Finally, given the results obtained from exploratory data analysis, interactions between variables should be assessed in further research to possibly obtain more advanced statistical models.

References

[1] A. Flor, N. Pinter, W.F. Remo (2010). "Evaluating Levee Failure Susceptibility on the Mississippi River Using Logistic Regression Analysis," Engineering Geology, Vol. 116, pp. 139-148.