**ENGO 697 – Final Report**

**Flood Susceptibility Mapping in the Red River Valley,**

**Manitoba, Using Machine Learning**

**Blair William Gerald Scriven**

Introduction:

In the previous decade, Canada has been subjected to major flood events across the country, leading to heavy damages to infrastructure and costing the lives of Canadians. The 2013 Alberta spring flood event resulted 6 billion dollars in damages and killed five people (Global News, 2019). A 2019 flooding event in Eastern Canada caused over 15,000 residents to evacuate in Quebec, New Brunswick, and Ontario, with record breaking water levels being recorded (CTV News, 2019). In 2011, a province-wide flooding event in Manitoba required the Canadian Government to spend $1.2 Billion in disaster relief due to extensive damage of infrastructure and cultivated lands (Government of Manitoba, 2013). One interesting aspect of this case is that the Red River valley, located in-between the city of Winnipeg and the US border, experience less severe damages than other regions of Manitoba due to the development of flood protection infrastructure, like flood barriers. In the 2011 Manitoba flood report, the Government of Manitoba recommended the development of maps to identify the most flood-prone areas of the province to allocated resources for flood protection infrastructure (Government of Manitoba, 2013).

To this end, a flood susceptibility map of Manitoba could be a source of information on spatial flood-proneness in the province. Susceptibility maps are often utilized in earth science and disaster management to estimate the spatial probability of a disaster occurring; they are used not just for predicting flood events, but also landslides and volcanos (Becerril et al., 2013; Tehrany et al, 2015; Di et al., 2019). While some studies have used multivariate statistical models, like logistic regression, to derive these maps, research has found that supervised machine learning algorithms create more accurate maps (Al-Abadi & Al-Najar, 2019; Merghadi et al., 2018). To derive these maps, machine learning methods use a variety of explanatory variables, usually physical and climatic factors, to define their relationships with flood proneness. For flood susceptibility maps in particular, previous research has shown strong results using classification and regression trees (CART) and support vector machine (SVM) models with a radial kernel (Choubin et al., 2019; Al-Abadi & Al-Najar, 2019; Tehrany et al, 2015). Although it has not been used in flood susceptibility mapping before, the gradient boosting machine (GBM) model has performed quite well in landslide susceptibility studies, outperforming other machine learning methods like random forest (RF) and even SVM (Merghadi et al., 2018; Di et al., 2019). It may be interesting to see how well a GBM can derive a flood susceptibility map compared to SVM and CART.

In this study, three flood susceptibility maps of the Red River Valley were created using CART, SVM, and GBM, and then compared for their prediction accuracy. Various physical descriptors of the study site, all publicly available, were incorporated into the models, and the model performance was determined using the receiver operating characteristic (ROC) curve, the area under the ROC curves (AUC), and the Cohen kappa index. The purpose of this study is to examine different strategies for producing flood susceptibility maps in the Red River Valley, which could help develop strategies for flood susceptibility mapping in all of Manitoba and the rest of Canada.

Study Area & Data:

The Red River valley has been subjected to many severe flood events in recorded history besides the 2011 flood, owing to a lack of natural water-storage sites along the watershed. This results in the river being unable to contain higher flows when the spring ice melt occurs, leaving the area vulnerable to significant social and economic impacts if flood infrastructure is not in place (George & Nielsen, 2000; Government of Manitoba, 2013). A flood extent polygon of the 2011 Manitoba flood was used as the dependant variable in this project, and so the study area was defined to capture the bulk of this polygon in the Red River valley (Figure 1). The study cite was centered on the Red River, using a 40 km by 40 km (49° 5' 0.3048"N, 97° 34' 19.3902"W to 49° 27' 7.6566"N, 97° 0' 35.769"W) bounding box. Similar to how Giovannettone et al (2018) prepared the flood data, the flood extent polygon was converted into a binary raster (1 for flood, 0 for non-flood) and a randomized sample of 4,000 points, equal between flood and non-flood cells, were taken in ArcGIS Desktop (Version 10.6) using the Sample tool.

All the explanatory variables chosen for this project were selected because previous flood susceptibility models used these variables in their studies (Table 1). In particular, the elevation and distance to river variables were found to be significant in all the flood susceptibility research studies reviewed for this project (Choubin et al, 2019; Al-Abadi & Al-Najar, 2019; Giovannettone et al., 2018). All the variables are physical descriptors of the Red River valley; climatic descriptors, such as wet days, were not available for the Red River Valley and were not included in this study, despite being significant in other studies (Ji et al., 2013). All the raw data was publicly available from Open Canada data portal (https://open.canada.ca/data /en/dataset?keywords=flood+extent) or from the Manitoba Geological Survey (https://www.manitoba.ca/iem/geo/gis/ surfgeomap.html). The distance to river and the river density datasets were created in ArcGIS using the Euclidean tool and line density tools. Every dataset was resampled into a 30m resolution and then projected into the NAD 1983 UTM Zone 14N projection before being sampled by the Extract Values to Points tool. All samples were converted into a .csv file and then transferred to RStudio.

Methods:

*Data Preparation*

Prior to creating the maps, tests for multicollinearity between the explanatory variables were testing using a Pearson’s coefficient matrix (Figure 2) and the variance inflation factors (VIF). An acceptance threshold of -0.7 to 0.7 for the Pearson’s test was used, and both tests were performed in RStudio using the *ggplot2* and *car* packages (Al-Abadi & Al-Najar, 2019). The Pearson’s matrix showed that there was a strong correlation coefficient (>0.9) for the relationship between the soil type (surficial materials) and river density variables; this was also confirmed by the VIF test. When the soil type variable was removed, both tests passed successfully, and thus the soil type data was removed from the full dataset. The dataset was then randomly split into two subsets, one for training (70%; 2,800 samples) and testing (30%; 1,200 samples); this is a common train/test ratio used in previous studies (Choubin et al., 2019; Giovannettone et al., 2018) and it was performed using the *ISLR* package in RStudio. The training data was utilized to develop the models, while the test data was used to assess the accuracy of the models.

*Classification and Regression Trees (CART)*

The CART method works by recursively partitioning the dataset into various predictor nodes, with each node containing an if/then condition based on values of the explanatory variables; this results in a tree-like diagram (Al-Abadi & Al-Najar, 2019). To avoid overfitting, superfluous branches can be removed (pruning), using a complexity parameter (CP) that balances the goodness of fit and tree size through cross-validation error estimates; the optimum CP value for the model was determined to be 0.01 (Ji et al., 2013). The tree diagram for the Red River Valley flood data is shown in Figure 3; both the original CART model and the pruning were implemented in RStudio using the *rpart* package.

*Support Vector Machine (SVM)*

SVM algorithms work by separating the data space using optimized linear hyperplanes, often requiring a transformation of the dataset to fit the high dimensional feature space (Choubin et al., 2019). The kernel will perform this transform the data, so that every sample can be classified as flood or non-flood; in this study, a radial basis function was used due its strong performance in previous flood susceptibility studies (Choubin et al., 2019; Tehrany et al, 2015). All SVM processing was done in RStudio using the *e1071* package.

*Gradient Boosting Machine (GBM)*

Similar to the CART method, GBM produces decision trees to classify data into flood or non-flood; however, the GBM is iterative, running and sequentially combining a series of trees (weak learners) to minimize the model error (Merghadi et al., 2018). Each iterative tree is added through the shrinkage rate, which reduces the impact of adding trees to the overall structure; a shrinkage rate of 0.05 was used, similar to Di et al (2019). Pruning can be performed by selecting an inflection point (optimal number of trees) where the model performance no longer improves (Singh, 2018). An out of bag sample set (OOB) was used to determine the inflection point, and an optimal number of trees was calculated to be 48 (Figure 4). All GBM processing, including the pruning, was done using the *gbm* package in RStudio.

*Model Prediction Assessment*

A common test for prediction accuracy for binary models and susceptibility maps is the AUC score, which is derived from the ROC curve plot (Choubin et al., 2019). The ROC curve plot compares the rate of false positives and true positives, and the AUC is the integral of this curve, which will determine the overall prediction capacity of the model; a value greater than 0.8 indicates a very good prediction capacity (Al-Abadi & Al-Najar, 2019). A Cohen Kappa index was also performed to assess the model prediction accuracy; any fitted value greater than 0.6 was given a value of 1 (flooded) and a value under 0.6 was given a value of 0 (non-flooded). Once again, a Kappa index of >80% indicates a very good prediction score. All accuracy assessments were performed in RStudio using the *pROC* package.

Results & Discussion:

Every model had an AUC score of above 0.8, indicating a strong prediction capacity; CART had a score of 0.857, SVM had a score of 0.887, and GBM had a score of 0.877. The ROC curve plots are shown in Figure 5. The Cohen Kappa scores show some slightly different results; CART had a score of 73.5%, SVM had a score of 82.0%, and GBM had a score of 81.6%. The results of both assessments indicate that the SVM model had the highest predictive capacity and accuracy for flood susceptibility mapping in the Red River Valley, followed closely by GBM, while the CART slightly under-performed compared to the other models; the Cohen Kappa score of the CART model was low compared to SVM and GBM. These results were similar to the results of previous studies; in the study by Choubin et al (2019), SVM had outperformed CART with prediction accuracy score of 88.0% vs. a score of 85.0%, and in the landslide susceptibility study by Di et al (2019) the GBM and SVM had similar AUC scores, although GBM outperformed with a score of 0.88 vs. a score of 0.86.

Comparing both the CART tree diagram (Figure 3) and the GBM relative influence bar graph (Figure 6), it is evident that the distance-to-river was the most significant factor for flood susceptibility mapping in the Red River Valley. In the CART tree diagram, the primary node was the distance-to-river if/then condition, with the remaining nodes affiliated with slope, river density and elevation. The GBM relative influence had similar results, with the distance-to-river variable having the overwhelmingly highest influence (68.9), followed by elevation (9.12) and river density (7.88).

This strong influence from the distance-to-river variable is also evident in the flood susceptibility prediction maps (Figure 7); all the maps show high flood susceptibility index values within proximity of the Red river, and this value decreases the farther away from the river. Overall, all the maps show a sudden decrease in flood susceptibly at a similar threshold distance from the Red river; going by the CART tree diagram, this threshold distance is probability around 5-6 km off of the river. Thus, this study could indicate that flood protection infrastructure should be present all along the Red River, and that future urban development should be at least 5-6 km away from the river.

Conclusion:

Overall, this study has shown the potential for creating flood susceptibility models using machine learning algorithms, especially for the SVM and GBM models. CART may be too simplistic with the limited number of nodes, something that is expanded on by the GBM model. Overwhelmingly, distance-to-river had the strongest influence on flood proneness, which was expected from previous studies. While these models show good overall results, the influence of the explanatory variables will vary over not only the province but the country. A variety of physical and climatic variables should be considered before any susceptibility map is derived.

Reference:

Al-Abadi, A. M., & Al-Najar, N. A. (2019). Comparative assessment of bivariate, multivariate and machine learning models for mapping flood proneness. *Natural Hazards*, 100, 461–491. doi: 10.1007/s11069-019-03821-y

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., & Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Science of the Total Environment*, 651, 2087-2096. doi: 10.1016/j.scitotenv.2018.10.064

Di, B., Zhang, H., Liu, Y., Li, J., Chen, N., Stamatopoulos, C. A., ... & Zhan, Y. (2019). Assessing Susceptibility of Debris Flow in Southwest China Using Gradient Boosting Machine. *Scientific reports*, *9*(1), 1-12. doi: 10.1038/s41598-019-48986-5

Giovannettone, J., Copenhaver, T., Burns, M., & Choquette, S. (2018). A Statistical Approach to Mapping Flood Susceptibility in the Lower Connecticut River Valley Region. *Water Resources Research*, 54(10), 7603-7618.doi: 10.1029/2018WR023018

Government of Manitoba. (2013). Manitoba 2011 Flood Review Task Force Report. Retrieved from: https://www.gov.mb.ca/asset\_library/en/2011flood/flood\_review\_task\_force\_report.pdf

Ji, Z., Li, N., Xie, W., Wu, J., & Zhou, Y. (2013). Comprehensive assessment of flood risk using the classification and regression tree method. Stochastic environmental research and risk assessment, 27(8), 1815-1828. doi: 10.1007/s00477-013-0716-z

Merghadi, A., Abderrahmane, B., & Tien Bui, D. (2018). Landslide susceptibility assessment at Mila Basin (Algeria): a comparative assessment of prediction capability of advanced machine learning methods. ISPRS International Journal of Geo-Information, 7(7), 268. doi: 10.3390/ijgi7070268

Singh, H. (2018) Understanding Gradient Boosting Machines [Web Article]. Medium. Retrieved from: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types*. Catena*, 125, 91-101. doi: 10.1016/j.catena.2014.10.017

Appendix:

Figures:

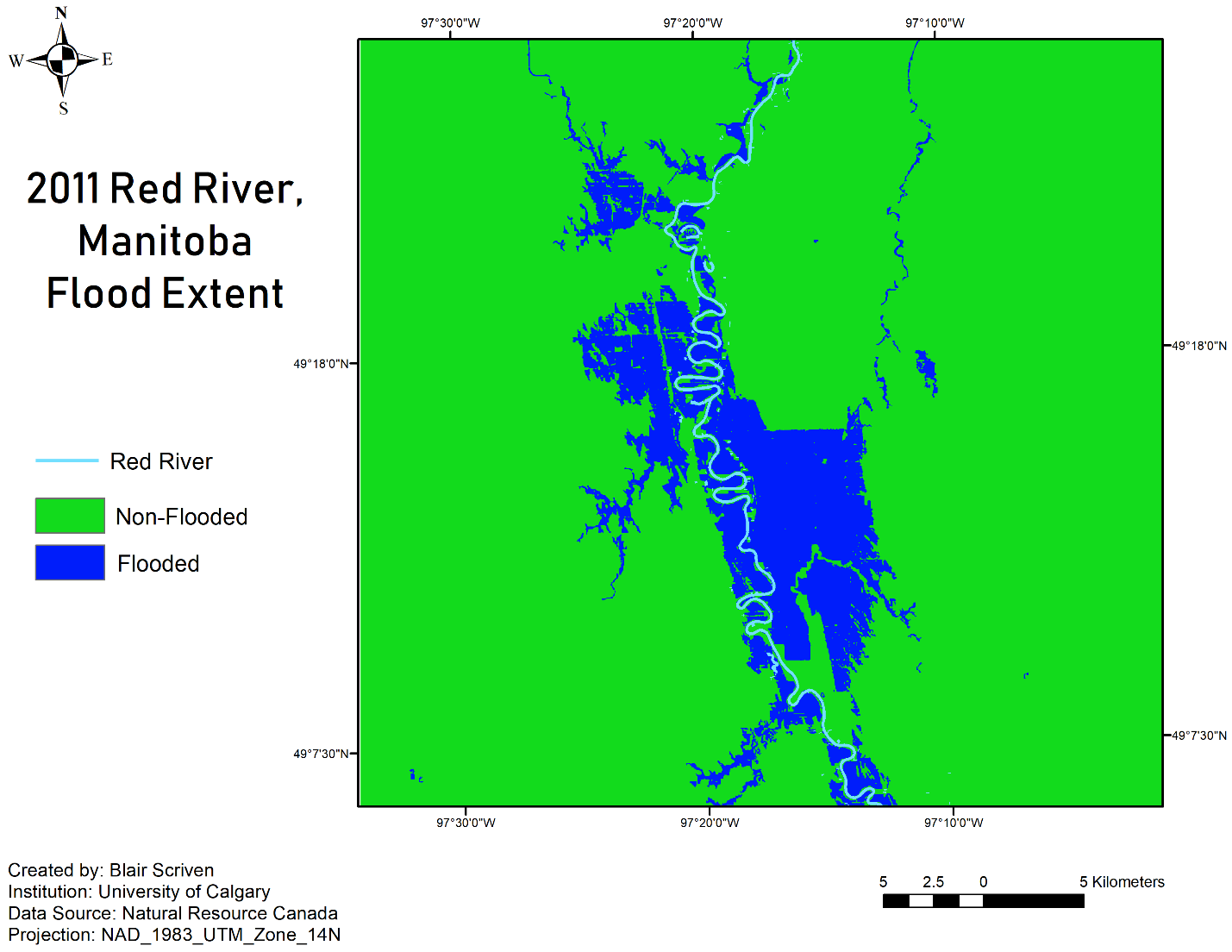


Figure 1. Study site for the Red River Valley, Manitoba

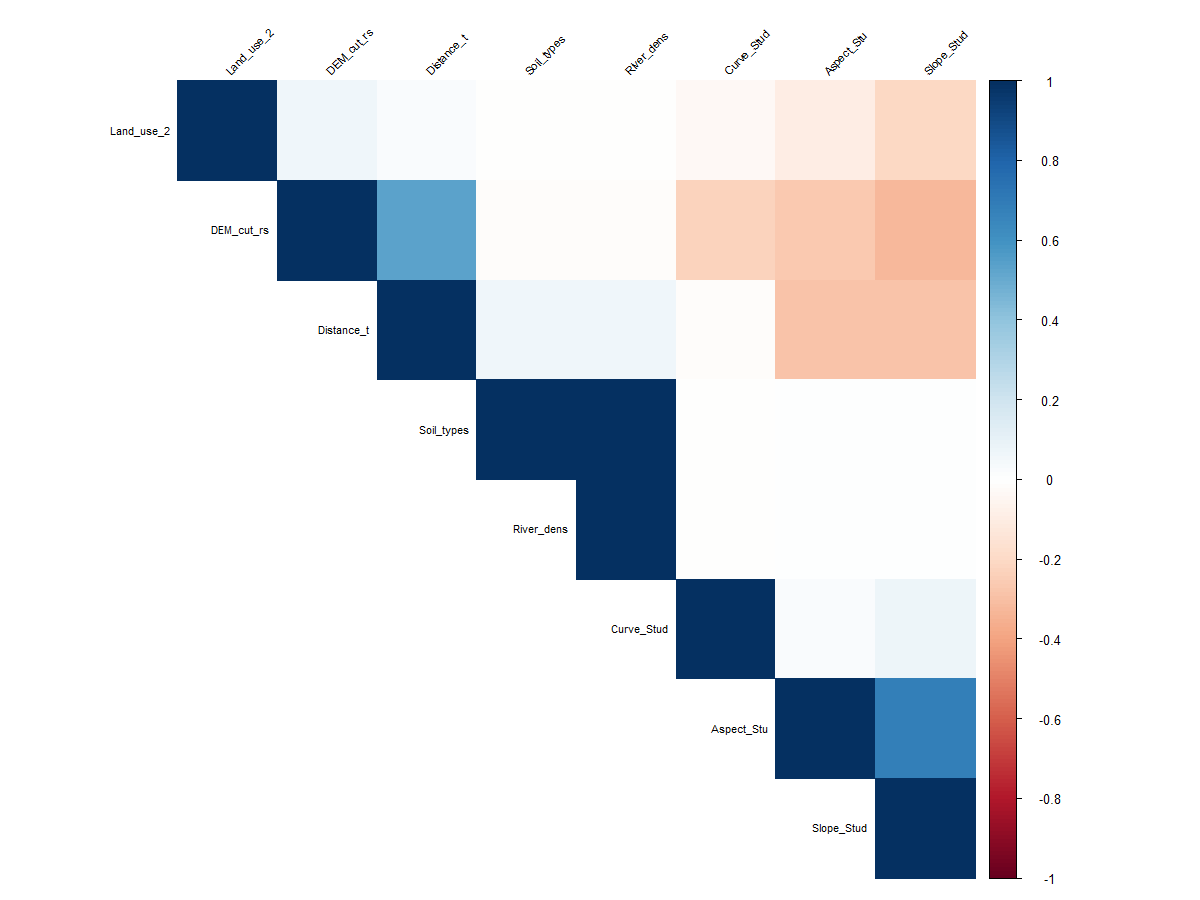
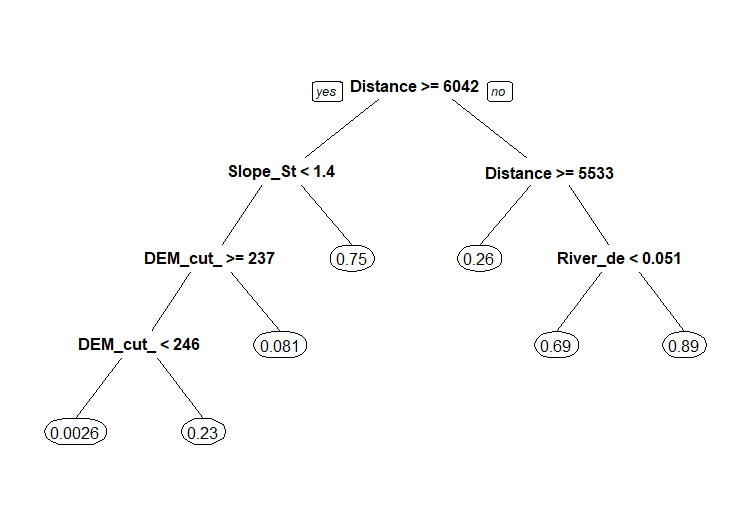


Figure 2. Pearson's Correlation Matric for the explanatory variables



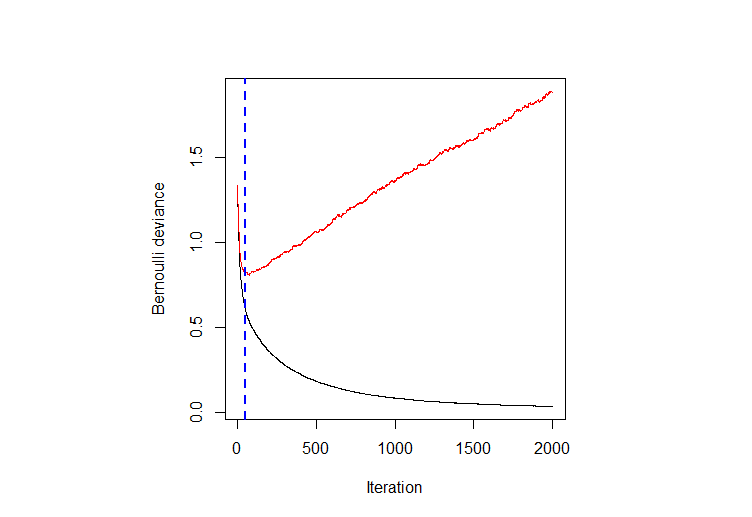
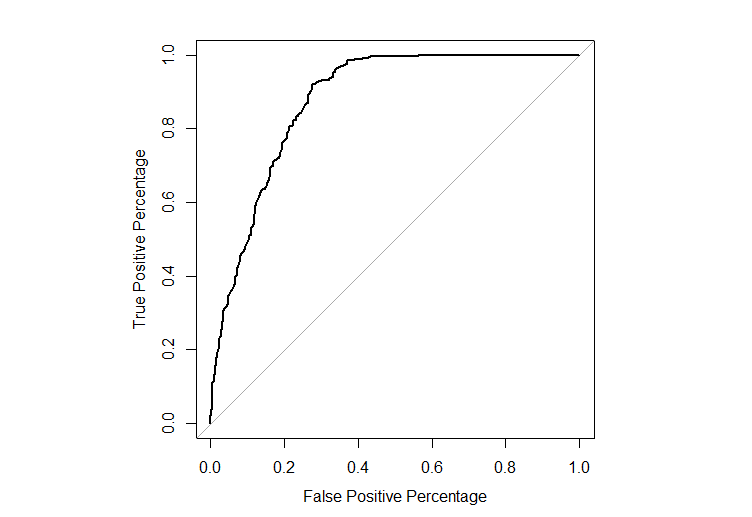
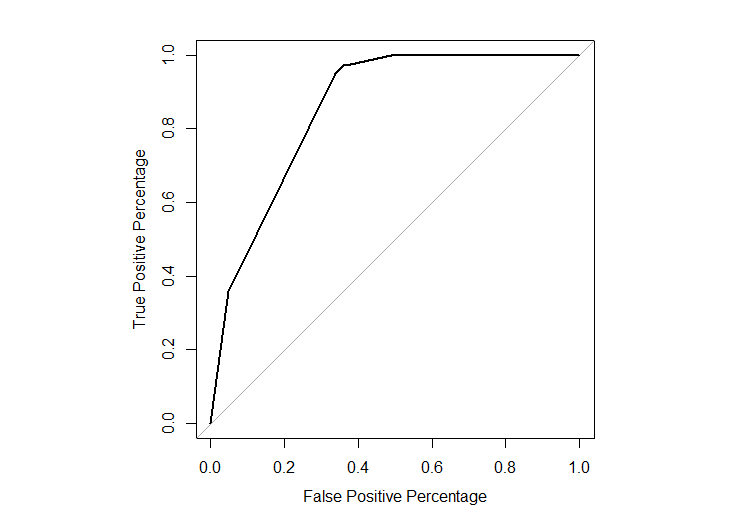


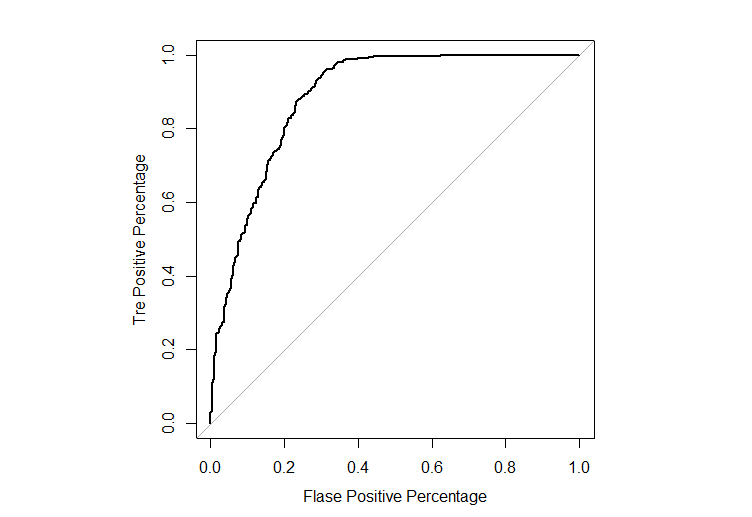
Figure 4. Optimum number of trees using the OOB method; the red line is the error on test, while the black like is the train data set

Figure 3. CART tree diagram

b.

a.





c.

Figure 5. The ROC curve plots of CART (a), GBM (b), and SVM (c)

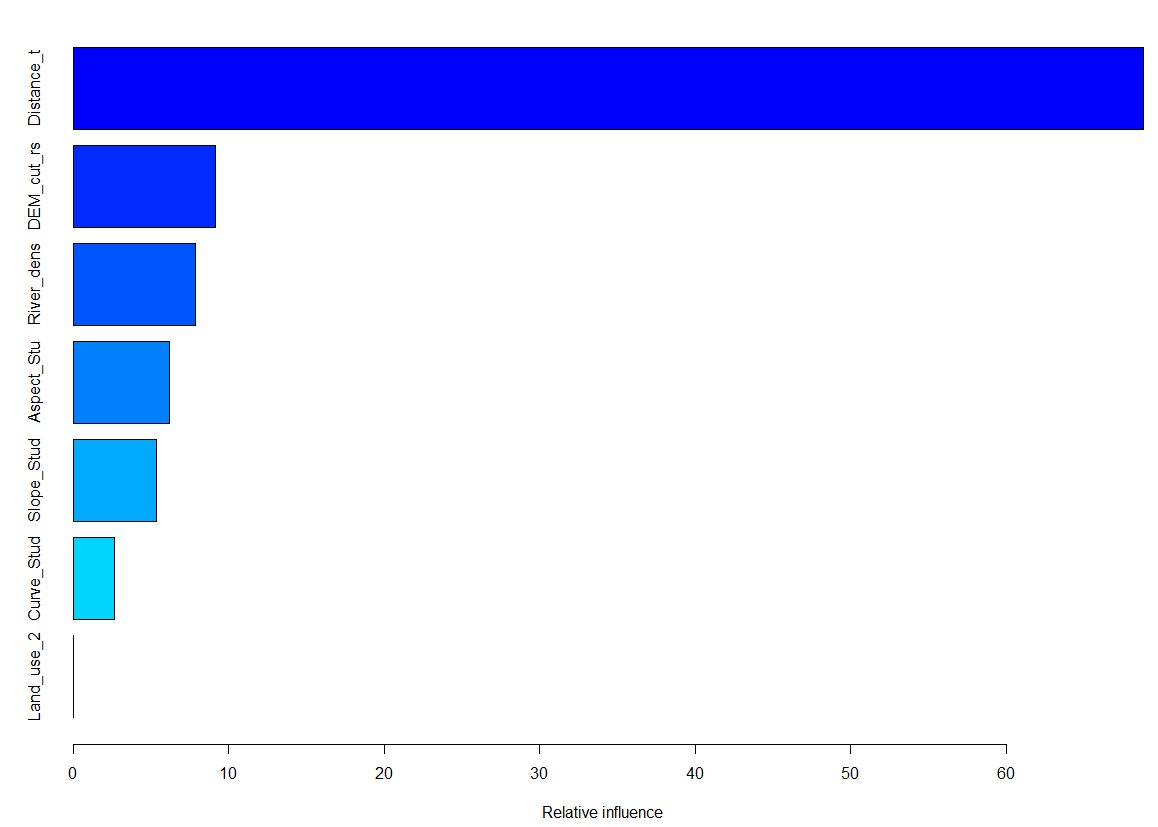
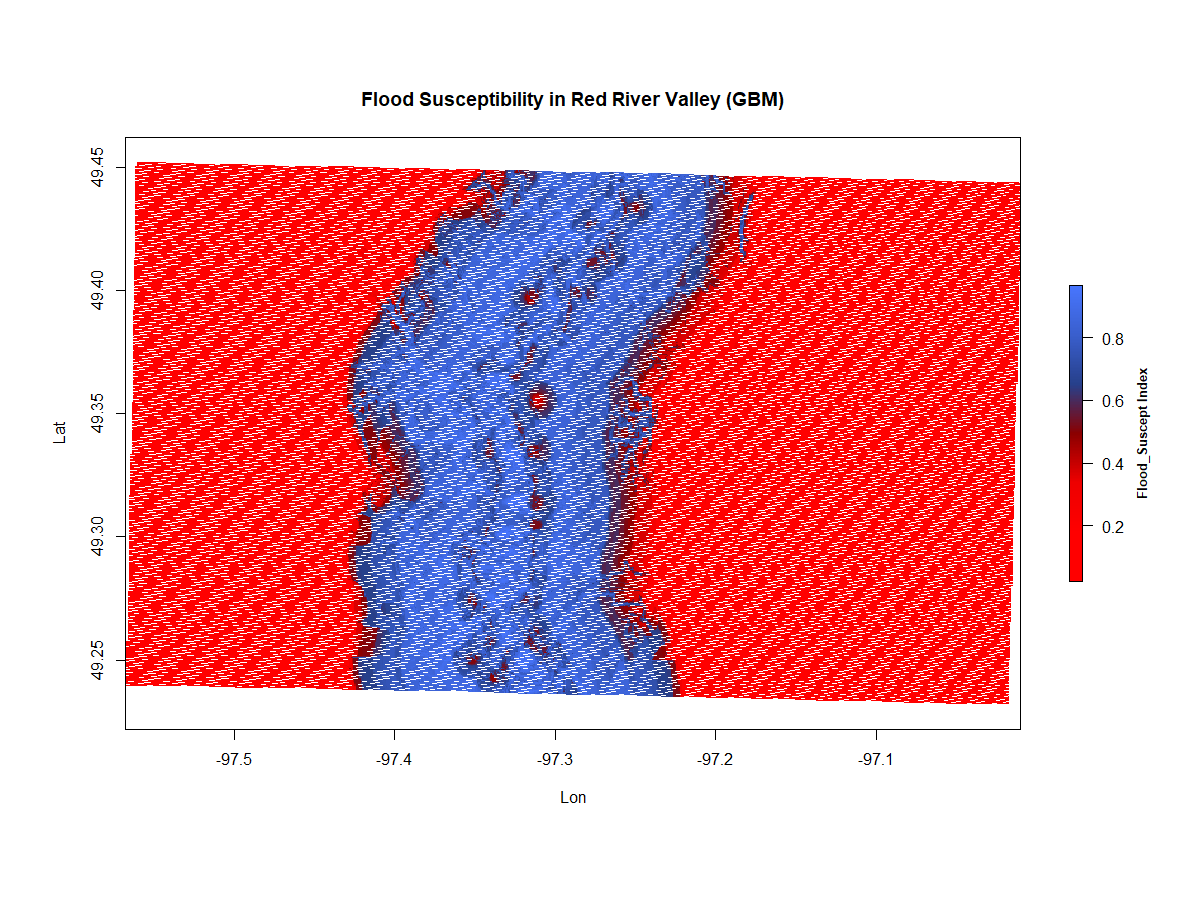
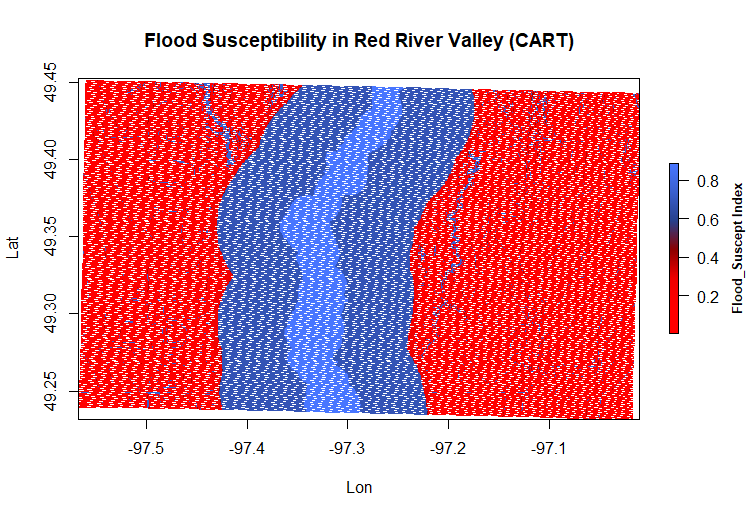
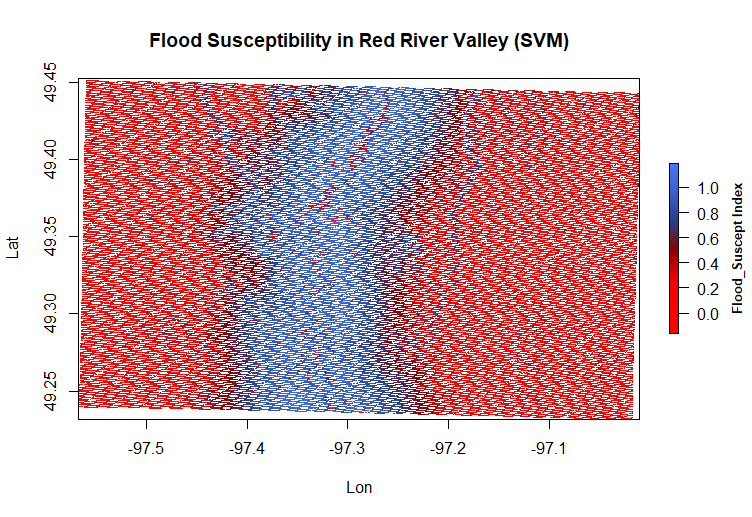


Figure 6. GBM relative influence bar graph; note the overwhelming influence of the distance-to-river variable

b.

a.





c.

Figure 3. Prediction flood susceptibility maps of the CART (a), GBM (b) and SVM (c) models

Tables:

Table 1. Explanatory variables used to create the Red River Valley flood susceptibility maps, including data sources.