CPLN 675 ASSIGNMENT 03

ESTIMATING A FLOOD INUNDATION PROBABILITY MAP

JONATHAN MANURUNG, JUNE JAEGAL

PROJECT PRESENTATION VIDEO:

https://youtu.be/fK_NGajb_p0

INTRODUCTION

In preparation for the high precipitation season and to mitigate more loss, it is imperative to create a probability map to estimate flooding and inundation in a city. Data-driven probability maps will provide inundation-related information to help the city government to make data-based decisions. Inundation probability maps can be generated by creating prediction models, from some related variables that related spatially across space.

Using data from the City of Calgary in Alberta, Canada, this project will build an inundation predictive model to predict areas of flood inundation in the city. The inundation predictive model will be trained borrowing the experience from past inundation in Calgary, associate the past event with related variables, and test the model's validity on predicting the flood inundation event. The generated predictive model will be applied to predict flood inundation for a comparable city, in this case, Denver, Colorado, United States of America. Variable-related data is collected from Calgary and Denver city open data portal and Multi-Resolution Land Cover Consortium (MRLC) website. ArcGIS Pro is used to wrangle those features data, then import the data to be modelled in RStudio.

SIGNIFICANT FEATURES

Four significant features are determined to be the most significant variable in predicting inundation. The first two features are proximity, with the first being distance of a cell in the map from stream, and the second is the distance of a cell in the map from hydrologic features such as lakes and ponds. These first two features were generated from Digital Elevation Model (DEM) of Calgary using Geoprocessing tools in ArcGIS Pro, and the distance calculated using Euclidean Distance tools. Both first two proximity features are considered significant because inundation happens mostly in the area near the water body like streams, lakes, and ponds.

The third feature is elevation data generated from the city DEM. Areas located at lower elevation are more susceptible to flooding and the elevation variation within a landscape can influence the flow of water.

The last significant feature is the National Land Cover Database from MRLC website. This feature explains the class/value of Land Cover in the city. The given Land Cover type then be reclassified to four main class (Water, Medium Residential, Forest, and Agricultural). All these four significant features then joined in a fishnet to be exported to RStudio for prediction modeling. Figures 1, 2, 3, and 4 explain the map of significant features in Calgary city.

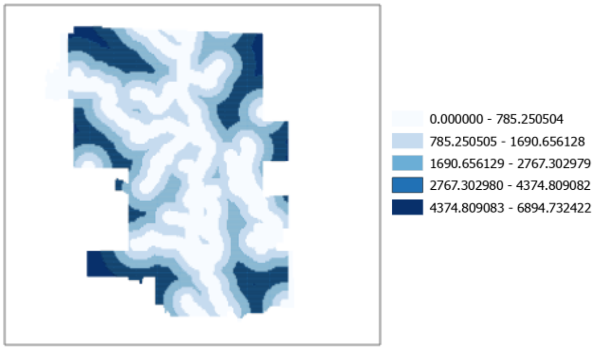


Figure 1. Distance to stream

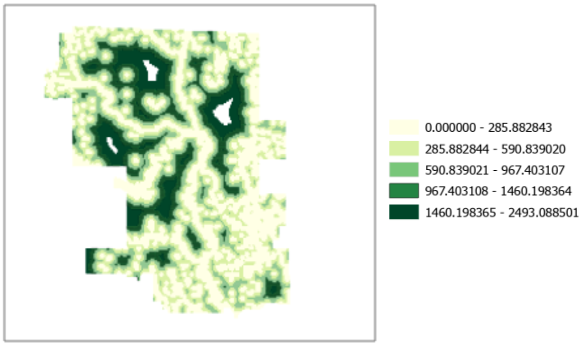


Figure 2. Distance to water bodies

Figure 1 and Figure 2 show the proximity of each fishnet cells to stream, and each fishnet cells to water bodies feature in Calgary city respectively. Lighter blue color in Figure 1 indicating locations close to a stream, meanwhile lighter green color in Figure 2 indicating locations close to a water bodies and indicating high potensial of inundation.

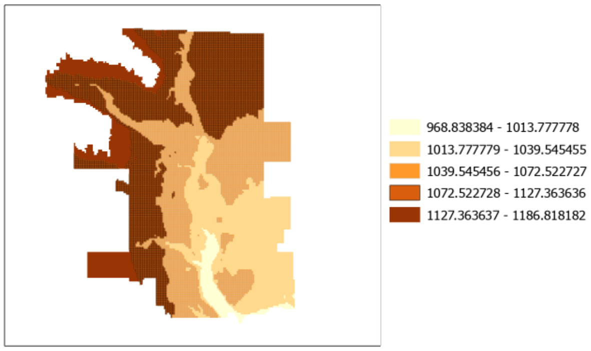


Figure 3. Elevation in Calgary

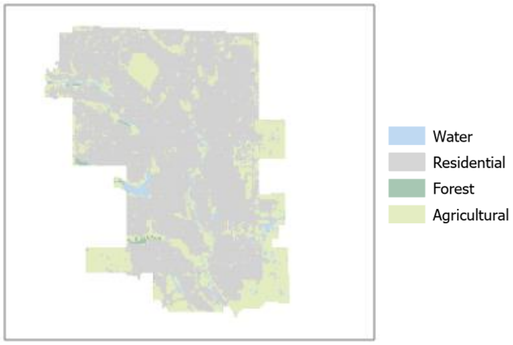
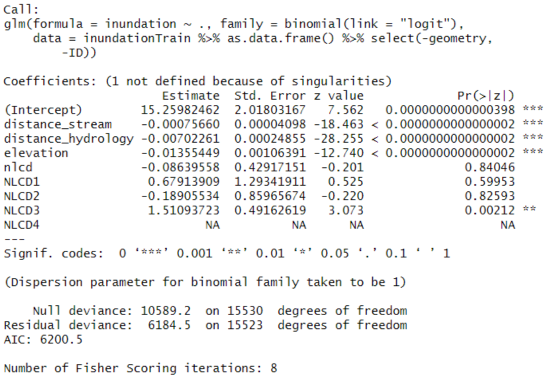


Figure 4. NLCD Classes

Figure 3 shows the value of elevation and Figure 4 show the NLCD class of each fishnet cells in Denver city. Lighter color in Figure 3 indicating locations are in the lower level area and indicating higher inundation potential. Figure 4 describes the classification of land cover in Denver city based on it's type. Different types of land cover affect the ability of the land to absorb water and potentially lead to inundation. These four features then used as data to train and test new binomial logistic model in RStudio.

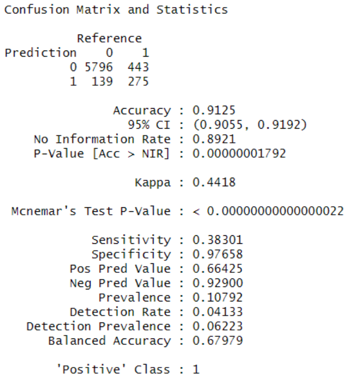
LOGISTIC REGRESSION MODEL SUMMARY

Using ‘glm’ function in RStudio, we can learn the relationship between dependent variable (inundation) and all independent variables shown in the summary of flood inundation in Figure 5. Based on the summary, distance to stream, distance to hydrology, elevation and some NLCD categories are negatively associated with inundation. Figures 6 explain the confusion of matrix for the final model. Based on the sensitivity (0.38301) and specificity value (0.97658), the model excels in identifying negative cases correctly better than identifying positive cases. Final Receiver Operating Characteristic (ROC) Curve of the model is also presented in Figure 7. The area under the curve is 0.9254, indicating the final model can predict inundation at 92.54% accuracy.



```
Call:
glm(formula = inundation ~ ., family = binomial(link = "logit"),
    data = inundationTrain %>% as.data.frame() %>% select(-geometry,
        -ID))

Coefficients: (1 not defined because of singularities)
                     Estimate  Std. Error z value            Pr(>|z|)    
(Intercept)       15.25982462  2.01803167   7.562   0.0000000000000398 ***
distance_stream   -0.00075660  0.00004098 -18.463 < 0.0000000000000002 ***
distance_hydrology -0.00702261  0.00024855 -28.255 < 0.0000000000000002 ***
elevation         -0.01355449  0.00106391 -12.740 < 0.0000000000000002 ***
nlcd              -0.08639558  0.42917151  -0.201              0.84046    
NLCD1              0.67913909  1.29341911   0.525              0.59953    
NLCD2             -0.18905534  0.85965674  -0.220              0.82593    
NLCD3              1.51093723  0.49162619   3.073              0.00212 ** 
NLCD4                      NA          NA      NA                   NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10589.2  on 15530  degrees of freedom
Residual deviance:  6184.5  on 15523  degrees of freedom
AIC: 6200.5

Number of Fisher Scoring iterations: 8
```

Figure 5. Summary of Flood Inundation Model



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5796  443
         1  139  275
                                               
               Accuracy : 0.9125               
                 95% CI : (0.9055, 0.9192)     
    No Information Rate : 0.8921               
    P-Value [Acc > NIR] : 0.00000001792        
                                               
                  Kappa : 0.4418               
                                               
 Mcnemar's Test P-Value : < 0.00000000000000022
                                               
            Sensitivity : 0.38301              
            Specificity : 0.97658              
         Pos Pred Value : 0.66425              
         Neg Pred Value : 0.92900              
             Prevalence : 0.10792              
         Detection Rate : 0.04133              
   Detection Prevalence : 0.06223              
      Balanced Accuracy : 0.67979              
                                               
       'Positive' Class : 1                    
```

Figure 6. Confusion Matrix and Statistics

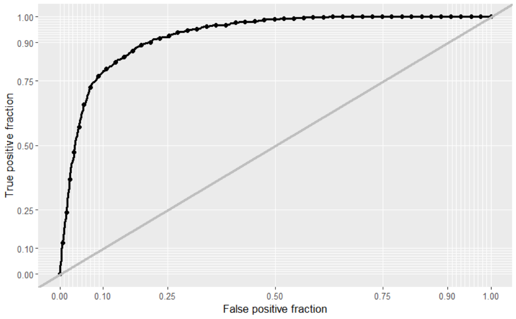


Figure 7. ROC Curve

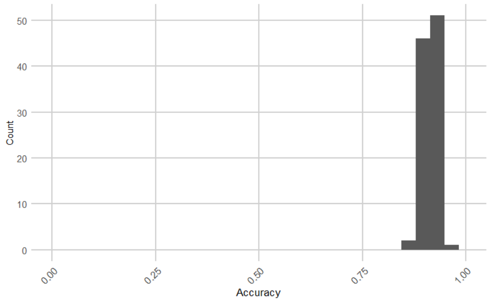


Figure 8. Cross Validation

Figure 8 presents the Cross Validation using k-fold cross-validation methodology, with the average accuracy over 100 folds is in 0.9130936.

PREDICTION RESULT

Figure 9 explains maps of true positives, true negatives, false negatives and false positives for the training set. and Figure 10 shows the final model predicts flood inundation in Calgary would occur.

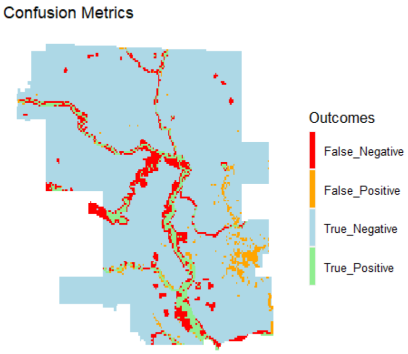


Figure 9. Model Outcomes

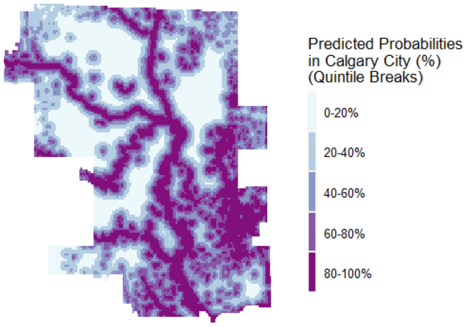


Figure 10. Predicted Inundation in Calgary

Figure 11 presents the final predictions of inundation map in another comparable city (Denver, Colorado), using the model that was trained on Calgary data.

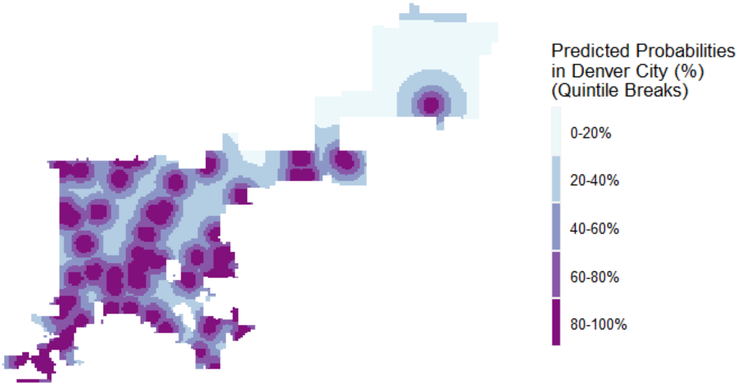


Figure 11. Inundation Prediction Map in Denver, Colorado