Rain Prediction Portfolio

#### Dylan Vener

#### Statistical Programming Fall 2016

## Introduction

Marinas, to account for the fundamental variability of lake height, have to make many adjustments to their own equipment. In particular, letting in or out the floating dock. They can accurately predict this need with sufficiently accurate forecasts and weather data on precipitation. Because of worry about local weather forecasting covering too broad of an area to be as accurate as necessary for them, they looked for alternative solutions.

The solution three marinas came to was making use of the Community Collaborative Rain, Hail & Snow Network (CoCoRaHs), to more accurately predict the precipitation at their location. The two methods proposed were the K-Nearest-Neighbor algorithm, which takes the average of the nearest k locations, and a radius method, which took the average of all locations within n miles.

The three sites which requested this information were Patona Bay Marina, Culver Marina, and Morse Marina.

## Methodology

The data set provided included five pieces of information about each station, the State, the Station Number, the total precipitation at each station, and the latitude and longitude to five decimal places. Of these, only the last three were of any use for the analysis.

For all calculations which required a distance calculation, I made use of the Vincety formula for an ellipsoid earth, which is the most accurate of the methods for calculating the geodesic, or shortest line connected two points) on the earth. To calculate this I made use of of the geosphere package.

To get started with my analysis I first found accurate Longitude and Latitude data for the three sites to be examined, and load it in.

PatonaBayMarina <- c(-85.78179219999998, 41.3356788 )  
CulverMarina <- c(-86.38940309999998,41.1898399)  
MorseMarina <- c(-86.03378399999997,40.09323699999999)

Once this data was collected, I created functions which implemented the two methods requested.

It was important in the K-Nearest-Neighbors algorithm that I filter out all NaN values, as without the safegaurd, it could suffer potentially catastrophic failure of the algorithm if all of the nearest locations had NaN values, even if there were nearby sites which held useful information.

KNN <- function(target, k){  
 #arrange data by the Vincenty Ellipsoid formula for distance away from target, take the top k of the list, and then average the precipitation at those sites  
 precip <- Precip.df %>%   
 arrange(apply(Precip.df[,c("Longitude","Latitude")],1,distVincentyEllipsoid,p2=target)) %>%  
 filter(Total.Precip>=0) %>%  
 head(k) %>%   
 summarise(Avg.Precip = round(mean(Total.Precip,na.rm=TRUE),6))  
 return(precip)  
}

I ultimately decided that a straight, nonweighted mean, was the optimal approach for calculating the prediction in Nearest Neighbors, as it simplified the code base, and there was no great need to factor distance in, as there were enough data points nearby the three targets to provide suffient resolution for a naive mean to be accurate. I rounded values to three decimal points as the original values only contained two, so I wanted to remain within bounds of that, while providing enough resolution to compare similar results from the two functions.

WithinN <- function(target, n){  
 #Filter by all sites that are within n miles of target, using the Vincenty Ellipsoid formula for geospatial distance, multiplied by a constant to convert form meters to miles, and then mean the precipitation at those points   
 precip <- Precip.df %>% filter(apply(Precip.df[,c("Longitude","Latitude")],1,distVincentyEllipsoid,p2=target)\*0.000621371<=n) %>%   
 summarise(Avg.Precip=round(mean(Total.Precip,na.rm=TRUE),3))  
 return(precip)  
}

## Results

### Nearest 5 Neighbors vs Within 25 Miles

Patona Bay Marina

## [1] "0.102 vs 0.21"

Culver Marina

## [1] "0.07 vs 0.061"

Morse Marina

## [1] "0.336 vs 1.058"

As one can see, the estimates provided by these methods seem to agree, with the exception of Morse Marina, where they differ rather significantly, further exploration showed that in the case of Morse Marina there seemed to be a great divide in the average precipitation, between approximately one inch and close to zero, likely correlated with direction, meaning that in one direction there was rain, and in the other there was none.

This illustrates a significant issue which might arise from Nearest Neighbor, that it might overlook weather in one direction, because there is a cluster of nearby stations in the other direction.

For this reason I would suggest making use of the Within-N algorithm, because of its omnidirectionality, it would overcome this issue, while in the two other stations which did not have this trend, it was similar to Nearest-Neighbor.

## Code for Project

#Must have dplyr, and geosphere installed, tibble is optional. Must have the PrecipData.RData loaded into the home director's download directory, or change the line to point to a different location  
library(dplyr)  
library(tibble)  
library(geosphere)  
load("~/Downloads/PrecipData.RData")  
  
#The three sites to be examined  
PatonaBayMarina <- c(-85.78179219999998, 41.3356788 )  
CulverMarina <- c(-86.38940309999998,41.1898399)  
MorseMarina <- c(-86.03378399999997,40.09323699999999)  
  
#Implementation of K Nearest Neighbors  
KNN <- function(target, k){  
 #arrange data by euclidian distance away from target, take the top k of the list, and then average the precipitation at those sites  
 precip <- Precip.df %>%   
 arrange(apply(Precip.df[,c("Longitude","Latitude")],1,distVincentyEllipsoid,p2=target)) %>%  
 filter(Total.Precip>=0) %>%  
 head(k) %>%   
 summarise(Avg.Precip = round(mean(Total.Precip,na.rm=TRUE),3))  
 return(precip)  
}  
  
WithinN <- function(target, n){  
 #Filter by all sites that are within n miles of target, using the Haversine formula for geospatial distance, and then mean the precipitation at those points   
 precip <- Precip.df %>%  
 filter(apply(Precip.df[,c("Longitude","Latitude")],1,distVincentyEllipsoid,p2=target)\*0.000621371<=n) %>%  
 summarise(Avg.Precip=round(mean(Total.Precip,na.rm=TRUE),3))  
 return(precip)  
}  
  
print("Nearest 5 Neighbors vs Within 25 miles")  
print("Patona Bay Marina")  
print(paste(c(KNN(PatonaBayMarina,5)$Avg.Precip,"vs",WithinN(PatonaBayMarina,25)$Avg.Precip),collapse = " "))  
  
print("Culver Marina")  
print(paste(c(KNN(CulverMarina,5)$Avg.Precip,"vs",WithinN(CulverMarina,25)$Avg.Precip),collapse = " "))  
  
print("Morse Marina")  
print(paste(c(KNN(MorseMarina,5)$Avg.Precip,"vs",WithinN(MorseMarina,25)$Avg.Precip),collapse = " "))