Assignment\_5.2

Austin Poole, Zach Hill

September 24, 2019

# Course Project: Milestone 2

As far as science can tell, water is a requirement for life. The human animal itself is comprised mostly of water and with it comes the necessity of an environment capable of providing that water. With an ever-increasing population of humans the water needs of humanity are becoming greater year by year and that burden is felt by the planet more and more. At some point the toll being levied against our groundwater stores will exceed the replacement rate from precipitation in many areas leaving a lack of water not only for humanity but also for the crops required to support the human population and the surrounding flora and fauna.

The United States Geological Survey (USGS) tracks ground water supplies both historically and in real time. This data can be analyzed to help predict if our current groundwater supplies are adequate or if we are in danger of depleting groundwater stores due to overuse. This information can be used to aid in health and safety matters regarding safe and available water as well as in maintaining population control when water thresholds are exceeded.

USGS records groundwater data through the Long-Term Groundwater Data Network which is comprised of " actively-measured periodic, continuous, and/or real-time wells with at least 20 years of measurements" (United States Geological Survey, 2019). The data available provide the option for 20, 30 and 50 year record samples from among over 9510 daily, monthly and annually recorded wells. Using historical data to predict trends in future water usage the federal, state and local governments can work together in forming a plan for the time when water will become limited. Due to the number of records available, the longer term records are generally only available from annual samples. The more recent data include a larger volume of samples in monthly and daily increments which will aid in fine-tuning usage allowances and requirements. Due to the volume of data collected in more recent years, analysis for generalized prediction will be accomplished using the 50 year dataset in annual increments.

The aim of this project is to help identify which areas in the Mid-Atlantic Hydrologic Region are at risk and when areas not already at risk might become so. Through anaylsis there should be insight gained regarding both our present situation, if there is one, and any concerns we might have for the future.

## Preliminary Requirement

Data must first be obtained for the USGS through <https://groundwaterwatch.usgs.gov>. The number of variables available for download is staggering so close attention must be paid when selecting relevant data. For this project, simple water elevations were selected to watch trends in water availability. These elevations are based on the North American Vertical Datum of 1988 (NAVD 88) survey standards. This multi-national survey standard set the height of the tidal benchmark for the continent and helped establish the mean sea level from which most modern measurements are made (National Geodetic Survey, 2018).

# sL <- list(380457075122301,  
# 380457075122302,  
# 380512075125401,  
# 380512075125402,  
# 380730075110401,  
# 380730075110402,  
# 380731075105601,  
# 380731075105602,  
# 380731075111101,  
# 380731075111102,  
# 380734075111301,  
# 380734075111302,  
# 380837075112201,  
# 380837075112202,  
# 381155075091801,  
# 381155075091802,  
# 381157075092901,  
# 381157075092902,  
# 381427075081102,  
# 381428075081401,  
# 381428075081402,  
# 381428075081403,  
# 381450075075401,  
# 381450075075402,  
# 381452075075901,  
# 381452075075902,  
# 381452075080101,  
# 381452075080102,  
# 381804075063401,  
# 381804075063402,  
# 381805075063501,  
# 381805075063502,  
# 381917076035501,  
# 381922076043001,  
# 381922076043002,  
# 381923076043501,  
# 381924076043801,  
# 381924076043901,  
# 381925076043001,  
# 381926076043401,  
# 381929076043001,  
# 381929076043002,  
# 382718076062002,  
# 383100076061101,  
# 383100076061103,  
# 383112076053503,  
# 383122076055701,  
# 383122076055702,  
# 390657076462601,  
# 390657076462602,  
# 390756076464201,  
# 390756076464202,  
# 390826076454801,  
# 390826076454802,  
# 390830076473902,  
# 392521075583301,  
# 392521075583302,  
# 392521075583303,  
# 392537075593001,  
# 392537075593002,  
# 392548075593802,  
# 392556075590001,  
# 392556075590002,  
# 431357073425301)  
  
sL <- list(335335078351901,  
 335629078115406,  
 335629078115407,  
 335631078003604,  
 335631078003605,  
 335631078003606,  
 335849078054301,  
 354302081433201,  
 354302081433202,  
 355031081243202,  
 355031081243203,  
 354616081085101,  
 354616081085102,  
 354855078553201,  
 351121083545002,  
 355359080331701,  
 345051078012109,  
 352012081154301,  
 352012081154302,  
 361011079595401,  
 352315082484401,  
 353827076293001,  
 354126076314201,  
 353135080524201,  
 353135080524202,  
 353135080524203,  
 345809077301408,  
 354133082042201,  
 354133082042203,  
 351126079301401,  
 351132079275301,  
 355944079013401,  
 361829076163201,  
 353219077153801,  
 362231079410801,  
 354057080362601,  
 344520079281001,  
 344544079263701,  
 345812079313401,  
 351709082434101,  
 351808082374302,  
 345609080415102,  
 345609080415103,  
 353509078404102,  
 353833078493301,  
 354144078460102,  
 354315078300101,  
 354315078300103,  
 354359078403104,  
 354649078400701,  
 354649078400702,  
 354748078315901,  
 354818078234101,  
 355457078232701,  
 355635078385101,  
 355657078342601,  
 360352078414401,  
 354216076271201,  
 354418076463601,  
 351849078163901)

### Importing Data to R

After obtaining the site data, the dataRetrieval package in R was used to request the groundwater level data from each site and added to a dataframe for anaylsis.

data <- dataRetrieval::readNWISgwl(sL,   
 startDate = "1900-01-01",   
 endDate = "2019-09-01")  
  
df <- data.frame(data$site\_no, data$lev\_dt, data$lev\_va)  
  
attribs <- attr(data, 'siteInfo')

### Data Structure

Multiple variables were included in the data retrieval process but of primary concern are the lev\_va variable, which is the level above mean sea level (MSL) and the date/time stamps in this analysis. Additionally, site identifiers were used in case further reference is needed, especially for geographical positioning lookup. The site numbers were imported as factors, the lev\_dt variable is imported as Date format, and lev\_va is imported as numeric.

str(df)

## 'data.frame': 5692 obs. of 3 variables:  
## $ data.site\_no: Factor w/ 56 levels "335335078351901",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ data.lev\_dt : Date, format: "2013-11-19" "2013-11-20" ...  
## $ data.lev\_va : num 15.1 15.2 15.4 15.4 14 ...

### Possible Issues

Another hurdle might be consistency of readings as many wells levels were not recorded the full duration of the timespan to be analyzed. This might require a manipulation of the sites using generalized areas where wells are found if geographical coordinates can be obtained.

## Technical Approach

### Analysis

summary(df)

## data.site\_no data.lev\_dt data.lev\_va   
## 335631078003606: 278 Min. :1969-07-07 Min. : -0.920   
## 335631078003604: 226 1st Qu.:2002-10-01 1st Qu.: 6.088   
## 335631078003605: 225 Median :2009-07-29 Median : 15.375   
## 352315082484401: 213 Mean :2007-03-30 Mean : 18.529   
## 335629078115407: 196 3rd Qu.:2014-07-15 3rd Qu.: 26.192   
## 351808082374302: 194 Max. :2019-08-30 Max. :219.400   
## (Other) :4360 NA's :4

While the dataset requested should have contained results at least 50 years prior, the earliest observation with a timestamp was 1975. More recent dates seem to occur more frequently, as expected. One site in particular was very consistent, site number 381427075081102, with over 450 level readings. This comprises around 20% of the total data which will make for reliable analysis in that area but other regions aren’t as consistent.

### Requirement Development

Business and Functional Requirement: Due to the nature of the hypothesis the first step in analysis is finding how the readings at each site relate to the other readings from that same site. This must be done prior to applying analysis to the region.

Technical Requirement: Data was obtained for sixty three wells in the Mid-Atlantic Hydrologic Region. One data set for each well was required for this model.

### Model Deployment

The scope of this project is to build and evaluate the model. Deployment is out of scope, being we don’t have a cloud solution for deployment.

### Testing and Evaluation

The model should be tested and evalated based on the test and train data randomly sampled from the project data set.

### Specific Exclusions from Scope

Depending on the processing performance, we may or may not include groundwater level data prior to 2000. At sixty three sites, multiple scatter plots will be required. Additionally, we may or may not show sites with water levels exceeding thirty feet. The majority of sites have a water level of fifteen feet or less. We’ve also considered removing those wells with less than five to ten years of records.

### Expected Result

Once finished, the scatter plots will tell us if water levels have remained stable or if in fact there are areas at risk. Ideally, we would like to plot the locations of these wells on a map to determine their proximity to one another.

## Management Approach

### Project Plan

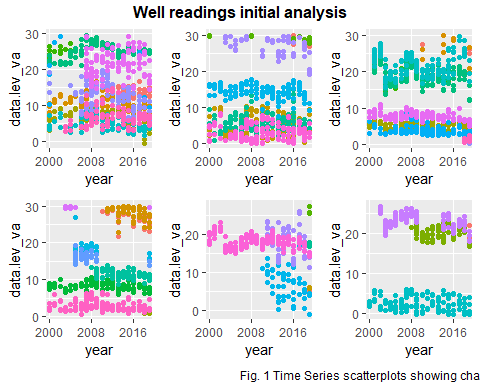
The project will be released in an incremental MVP (minimal viable product) approach as the product is phased as assignments over few weeks. This weekly assignment also governs what will be released each week.

### Project Risk

Since we would be working on this project along with other major assignments, time-line is one risk that we need to mitigate. For the project, would be collecting open data from the Internet, hence data quality might be a risk.

df$year <- substr(df$data.lev\_dt, 1, 4)  
df <- na.omit(df)  
century <- df[df$year >= 2000 & df$data.lev\_va <= 30, ]  
century$site\_num <- as.numeric(century$data.site\_no)  
ts\_df <- dcast(df, data.lev\_dt + data.lev\_va ~ data.site\_no, value.var = 'data.lev\_va', sum)  
  
ten <- century[century$site\_num <= 10, ]  
twenty <- century[century$site\_num <= 20 & century$site\_num >= 11, ]  
thirty <- century[century$site\_num <= 30 & century$site\_num >= 21, ]  
forty <- century[century$site\_num <= 40 & century$site\_num >= 31, ]  
fifty <- century[century$site\_num <= 50 & century$site\_num >= 41, ]  
sixty <- century[century$site\_num <= 60 & century$site\_num >= 51, ]  
seventy <- century[century$site\_num <= 70 & century$site\_num >= 61, ]  
  
p10 <- ggplot(ten, aes(x = year, y = data.lev\_va, group = data.site\_no, color = data.site\_no)) +   
 geom\_point(aes(colour = data.site\_no)) +  
 scale\_x\_discrete(breaks = seq(2000, 2019, 8)) +  
 theme(legend.position = "none")  
   
p20 <- ggplot(twenty, aes(x = year, y = data.lev\_va, group = data.site\_no, color = data.site\_no)) +   
 geom\_point(aes(colour = data.site\_no)) +  
 scale\_x\_discrete(breaks = seq(2000, 2019, 8)) +  
 theme(legend.position = "none")  
   
p30 <- ggplot(thirty, aes(x = year, y = data.lev\_va, group = data.site\_no, color = data.site\_no)) +   
 geom\_point(aes(colour = data.site\_no)) +  
 scale\_x\_discrete(breaks = seq(2000, 2019, 8)) +  
 theme(legend.position = "none")  
   
p40 <- ggplot(forty, aes(x = year, y = data.lev\_va, group = data.site\_no, color = data.site\_no)) +   
 geom\_point(aes(colour = data.site\_no)) +  
 scale\_x\_discrete(breaks = seq(2000, 2019, 8)) +  
 theme(legend.position = "none")  
   
p50 <- ggplot(fifty, aes(x = year, y = data.lev\_va, group = data.site\_no, color = data.site\_no)) +   
 geom\_point(aes(colour = data.site\_no)) +  
 scale\_x\_discrete(breaks = seq(2000, 2019, 8)) +  
 theme(legend.position = "none")  
   
p60 <- ggplot(sixty, aes(x = year, y = data.lev\_va, group = data.site\_no, color = data.site\_no)) +   
 geom\_point(aes(colour = data.site\_no)) +  
 scale\_x\_discrete(breaks = seq(2000, 2019, 8)) +  
 theme(legend.position = "none")  
   
p70 <- ggplot(seventy, aes(x = year, y = data.lev\_va, group = data.site\_no, color = data.site\_no)) +   
 geom\_point(aes(colour = data.site\_no)) +  
 scale\_x\_discrete(breaks = seq(2000, 2019, 8)) +  
 theme(legend.position = "none")

grid.arrange(p10, p20, p30, p40, p50, p60, nrow = 2, ncol = 3,  
 top = textGrob("Well readings initial analysis",  
 gp = gpar(fontface = "bold")),  
 bottom = textGrob("Fig. 1 Time Series scatterplots showing changes in well depth",  
 gp = gpar(fontsize = 9),  
 hjust = 0))



register\_google(key = 'AIzaSyCvvnczEV600J888oFybtq3QrDcHPlM030')  
  
map <- get\_stamenmap(bbox = c(top = 37.2115,  
 right = -75.5614,  
 bottom = 32.9347,  
 left = -86.0628),  
 maptype = 'terrain',  
 zoom = 7)

## Source : http://tile.stamen.com/terrain/7/33/49.png

## Source : http://tile.stamen.com/terrain/7/34/49.png

## Source : http://tile.stamen.com/terrain/7/35/49.png

## Source : http://tile.stamen.com/terrain/7/36/49.png

## Source : http://tile.stamen.com/terrain/7/37/49.png

## Source : http://tile.stamen.com/terrain/7/33/50.png

## Source : http://tile.stamen.com/terrain/7/34/50.png

## Source : http://tile.stamen.com/terrain/7/35/50.png

## Source : http://tile.stamen.com/terrain/7/36/50.png

## Source : http://tile.stamen.com/terrain/7/37/50.png

## Source : http://tile.stamen.com/terrain/7/33/51.png

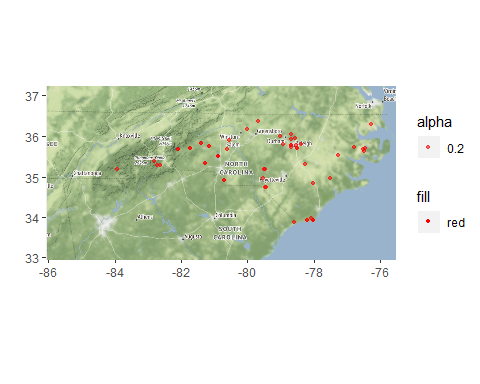
## Source : http://tile.stamen.com/terrain/7/34/51.png

## Source : http://tile.stamen.com/terrain/7/35/51.png

## Source : http://tile.stamen.com/terrain/7/36/51.png

## Source : http://tile.stamen.com/terrain/7/37/51.png

ggpoint <- data.frame(lon = c(-75.20661), lat = c(38.08269))  
  
ggmap(map) +  
 labs(x = "", y = "") +  
 geom\_point(data = attribs, aes(x = dec\_long\_va, y = dec\_lat\_va, fill = "red", alpha = 0.2), size = 1, shape = 19, color = 'red')



# myMap <- get\_stamenmap(bbox = c(left = -105.4969,  
# bottom = 39.8995,  
# right = -104.9579,  
# top = 37.1274),  
# maptype = "terrain",   
# crop = FALSE,  
# zoom = 6)  
# # plot map  
# gage\_location <- data.frame(lon = c(-105.178333), lat = c(40.051667))  
#   
# # create a map with a point location for boulder.  
# ggmap(myMap) + labs(x = "", y = "") +  
# geom\_point(data = gage\_location, aes(x = lon, y = lat, fill = "red", alpha = 0.2), size = 1, shape = 19) +  
# guides(fill = FALSE, alpha = FALSE, size = FALSE)