Surface Elevation Table- Marker Horizon (SET-MH) Data Handling and Analaysis in R

The Nature Conservancy - Long Island Chapter

Adam Starke

## Warning: package 'ProjectTemplate' was built under R version 3.3.1

## Warning in .check.version(config): Your configuration is compatible with version 0.6 of the ProjectTemplate package.  
## Please run ProjectTemplate::migrate.project() to migrate to the installed version 0.7.

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector  
  
## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector

## Warning in left\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factor and character vector, coercing into character vector  
  
## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factor and character vector, coercing into character vector

## Warning: attributes are not identical across measure variables; they will  
## be dropped

## Warning: Too few values at 18180 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,  
## 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...

## Current Data and Analysis File Structure

This document is intended to document the process used for database storage and data analysis of our SET-MH data. Our current workflow uses Access2010 to store the data and RStudio (running the R language) for data handling. R is a powerful, though often times overwhelming tool. It is hoped that this document will aid anyone unfamiliar with R through our current data processing methods.

### Data Storage

The SET-MH method of marsh monitoring generates a large set of data that needs to be carefully maintained and catalogued. Fortunately, there has been a substaintial effort put forth by the National Park Service's National Capital Region Network Inventory and Monitoring Program to create a user-friendly Access database (DB). This DB allows for easy field data entry, data browsing and the creation of summary plots. There are also methods for exporting data to other software if desired. Managing SET-MH data in such as system is critical.  
Although the statistical analyses needed to calculate trends in marsh elevation are somewhat basic the restructuring of the data to allow for analysis can be somewhat complex. With this as a characteristic of the data we decided to use the statistical programming language R. R is one of the most widely used programming languages and contains an enormous amount of tools and scripts (known as packages in R) that make these restructuring of data possible. The ability to create 'scripts' allow for quick updates to analysis after data entry. A number of other features also exist that make R a good choice, particulary the graphics capabilities and the support for creation of 'data-driven documents' which allows for reproducible research. This document is actually written with R code embedded within it. This allows one to actually run the code that performs the analysis along side the documentation of the code. This allows the user to describe and show the exact steps taken for data organization, restructuring and analysis. Actual R code run in this document will be highlighted and have a slight grey background. This code can be cut and pasted or run as a whole document.

As mentioned the data are currently stored in an Access database located at:

[1] "T:/Coastal/Marsh-Wetlands/SET-MH\_project/SET\_Data/SET\_Monitoring\_Database/Database\_storage/SET\_DB\_BE\_ver\_2.94\_TNC\_Master.mdb"

This file is referred to as the 'backend database' which holds all the raw SET-MH data.The 'front-end' of the database contains the user interface (UI) and is linked to the backend allowing multiple 'front ends' all pointing to one dataset. It's suggested that the front-end UI be used for quick checks on data and it is where **ALL** data entry takes place.

### Accessing Our Data Through R

Using the power of R's scripting language capabilities the quickest, most direct, way to update SET-MH analyses is to simply navigate to the folder containing the RStudio project file (SET-MH\_Analysis.Rproj) currently located at: C:/Users/astarke/Desktop/Box Sync/R\_Code/SET-MH\_Analysis.

Opening this file in RStudio will load all the code associated with this project. Run the script 'MarshProject.R' by opening it and clicking the *'Source'* button in the top-right of the window. This will pull all the code from the open file and run it. You will see activity in the "Console" window confirming that it is running.

The main feature in this script is the *ProjectTemplate* package command: load.project() which automatically runs (sources) scripts contained in the '/data' folder, then cleans it up using scripts in the '/munge' folder. Munging simply means to remove irrelevant or missing values, restructure the data, transforming variable names etc. The scripts in each of these folders will be explained below. There are then three other lines of code that source code that performs the actual analysis and creates summary tables and plots. This will also be explained in detail below.

## Data Processing Steps

This section will walk through each of our data processing steps; from loading data into R to exporting plots and summary tables. Most of these code blocks have been split to allow for commentary on what the code is doing. Again, any text that is bounded and highlighted is actual R code that is run in the software. If needed this text can be copied and pasted into an active R console and run.

#### Data Loading

The first step is to connect to the data. To do so R needs to communicate with the Access database through an *"ODBC- protocol"* or translator. For this we use the package **RODBC** which allows R to talk with Access.

###  
# Data importing from SET\_Monitoring\_Database 'Backend Data Storage File' currently SET\_DB\_BE\_ver\_2.92\_TNC\_Master  
# ProjectTemplate automatically runs this script, importing the dataframes below into the R workspace  
#  
###  
  
library(RODBC)  
  
SET.DB.path <- "T:/Coastal/Marsh-Wetlands/SET-MH\_project/SET\_Data/SET\_Monitoring\_Database/Database\_storage/SET\_DB\_BE\_ver\_2.94\_TNC\_Master.mdb"  
  
  
SET.DB <- odbcConnectAccess2007(SET.DB.path)

This code block creates a pathway to the Access database that is then used to communicate with the database. Once the connection is established several tables will be pulled into R as 'dataframes'; R's equivelent to a spreadsheet.

### Pull in the tables of interest containing data needed for analysis ====  
#Location data- i.e. study sites and stations within sites  
Sites <- sqlFetch(SET.DB, sqtable= "tbl\_Sites")  
Locations <- sqlFetch(SET.DB, sqtable= "tbl\_Locations")  
  
# Sites and Locations will be joined in munge by 'site\_ID'  
# Use capwords function to standardize caps and revert Stratafication to as a factor  
Locations$Stratafication <- as.factor(capwords((as.character(Locations$Stratafication))))  
  
# Sampling events  
Events <- sqlFetch(SET.DB, sqtable= "tbl\_Events")  
  
# Contacts/Field Personel  
EventContacts <- sqlFetch(SET.DB, sqtable = "xref\_Event\_Contacts")  
Contacts <- sqlFetch(SET.DB, sqtable = "tlu\_Contacts")  
  
# SET Rod data  
Positions <- sqlFetch(SET.DB, sqtable= "tbl\_SET\_Position")  
SETdata<- sqlFetch(SET.DB, sqtable= "tbl\_SET\_Data")  
  
# Surface Accretion data  
SAccret <- sqlFetch(SET.DB, sqtable= "tbl\_Accretion\_Data")   
SA\_Layers <- sqlFetch(SET.DB, sqtable= "tbl\_Feldspar\_Layers")  
  
close(SET.DB) #close the connection to the Access SET database.

This code block will load 7 tables into R containing:  
\* Sites- data associated with each study site, i.e. general location, name, date of establishment  
\* Locations- data associated with each station within each study site  
\* Events- dates that samples were made  
\* Positions- ID's for each SET arm position  
\* SETdata- SET readings raw data  
\* SAccret- Marker Horizon (MH) surface accretion raw data  
\* SA\_Layers- MH layer ID information

#### Munging

Data munging (aka data wrangling) is the process of formatting and rearranging the 'shape' of the data to allow for analysis. As mentioned earlier the actual analysis of SET-MH data is not that complex. The primary crux of the analysis is getting the data into a form that makes the analysis programatically easy. The basic steps are to merge the tables brought in from the DB, retaining all relevant data from each table and then reshaping the data to allow for the analysis to run quickly.

#SQL type joins to flatten the tables from the database. ----  
  
#Study Sites with Locations data --Location\_ID is a numeric key analagous to Plot\_Name   
StudyStations <- left\_join(Sites, Locations, by="Site\_ID") # Change to a left join --@^\*  
StudyStationLocations <- StudyStations %>% select(Location\_ID, Site\_Name, Stratafication, Plot\_Name, X\_Coord, Y\_Coord, Coord\_System, UTM\_Zone, Datum, SET\_Type)  
  
#Surface Accretion data   
SA <- inner\_join(SA\_Layers, SAccret, by="Layer\_ID")

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector

#SET Rod data  
SET <- inner\_join(SETdata, Positions, by="Position\_ID")

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector

# SET readers- Gather up Event\_ID, SET Reader and Contact\_ID -- Then match up Contact ID with Contacts Last name-  
Contacts %<>% select(Contact\_ID, Last\_Name, First\_Name, Organization) %>% mutate(FullName = paste(First\_Name, Last\_Name, sep = " "))  
  
SET\_readers <- EventContacts %>%   
 left\_join(Contacts, by = "Contact\_ID") %>%   
 select(Event\_ID, Contact\_ID, Contact\_Role, Last\_Name, First\_Name, Organization, FullName)%>%   
 mutate(ID = paste(Event\_ID, Contact\_ID, sep = "\_")) %>%   
 spread(key = Contact\_Role, value = FullName) %>%   
 dplyr::rename(SET\_Reader = `SET Reader`) %>%   
 select(Event\_ID, Last\_Name, First\_Name, Organization, SET\_Reader) %>%   
 filter(complete.cases(.))

## Warning in left\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector

# Sampling events and sites  
# join site-location data with sampling 'events'  
Samplings <- inner\_join(StudyStations, Events, by= "Location\_ID")   
Samplings <- inner\_join(Samplings, SET\_readers, by = "Event\_ID")

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factors with different levels, coercing to character vector

# Complete datasets with ALL variables  
SA.data <- inner\_join(SA, Samplings, by="Event\_ID")

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factor and character vector, coercing into character vector

SET.data <- inner\_join(SET, Samplings, by="Event\_ID")

## Warning in inner\_join\_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining  
## factor and character vector, coercing into character vector

# Remove excess dataframes leaving only SA and SET data ----  
rm(Events,  
 Contacts,  
 EventContacts,  
 SET\_readers,  
 Positions,  
 SA,  
 SA\_Layers,  
 SAccret,  
 SET,  
 SETdata,   
 Samplings,  
 Sites,  
 Locations  
 )

The next step is to clean up some of the variables in the tables.

We now have two complete datasets that contain all our relevant SET-MH data. One set contains SET data with location, date and pin height data. One has Surface Accretion data with location, date and accretion data.  
We need to continue cleaning up the variable names and reshape the way the data is stored in the dataframes.

# list of Notes regarding issues with pin placement   
# removes <- c("deer trail", 'hole', 'Hole', )  
  
  
# Complete SET data in a WIDE format ----  
# Trim excess columns and clean charcter strings.  
SET.data %<>% tbl\_df() %>%   
 select(Pin1:Pin9\_Notes, Arm\_Direction, Site\_Name, SET\_Type, Stratafication:Plot\_Name, Location\_ID.x, Position\_ID, Start\_Date, Organization, SET\_Reader) %>%   
 mutate(Stratafication = capwords(as.character(Stratafication)), Start\_Date = as.Date(Start\_Date)) # Eventually add a filter that will filter out only 'clean' readings  
   
attr(SET.data, 'Datainfo') <-"Full SET dataset including all measures in a WIDE format" # give dataframe some metadata attributes  
  
# Complete SET dataset in a LONG format ----  
SET.data.long <- SET.data %>%  
 select(Site\_Name, Stratafication, Plot\_Name, SET\_Type, Pin1:Pin9\_Notes, Arm\_Direction, Location\_ID.x, Position\_ID, Start\_Date, SET\_Reader)%>%   
 group\_by(Position\_ID, Start\_Date) %>%   
 gather(pin, measure, Pin1:Pin9\_Notes) %>% filter(!is.na(measure)) %>% # Remvoe NA from PinX\_Notes   
 separate(pin, c('name', 'note'), "\_", remove = TRUE) %>%   
 separate(name, c('name', 'Pin\_number'), 3, remove = TRUE) %>%   
 mutate(key = ifelse(is.na(note),yes = "Raw", no = note)) %>%   
 select(-note, -name) %>%   
 spread(key, measure) %>%   
 mutate(pin\_ID = paste(Position\_ID, Pin\_number, sep = "\_")) %>% # Above all transposing and repositioning dataframe.  
 ungroup() %>% # Below- adding columns, renaming variables, and reordering rows.  
 dplyr::rename(Date = Start\_Date, Location\_ID = Location\_ID.x) %>% # rename SET reading date  
 group\_by(pin\_ID) %>% # group by pinID to   
 mutate(EstDate = min(Date)) %>% # create a column identifying the EstDate (date of the first SET-MH station reading)  
 arrange(Date) %>% # arrange by date  
 mutate(Change = as.numeric(Raw) - as.numeric(Raw[1]),   
 incrementalChange = c(NA, diff(Change))) %>% # add column of incremental change  
 ungroup() %>%   
 mutate(DecYear = round((((as.numeric(difftime(.$Date, .$EstDate, units = "days"))))/365),3))

## Warning: attributes are not identical across measure variables; they will  
## be dropped

## Warning: Too few values at 18180 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,  
## 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...

attr(SET.data.long, 'Datainfo') <-"Full SET dataset including all measures in a LONG format" # give dataframe some metadata attributes  
  
  
# Clean outliers and 'issues' at time of reading ----  
# Two strategies- 1) drop complete time series for a pin that has an 'issue' at any one point in the series,   
# vs 2) drop only that data point that has an issue.  
# Or a merger of both stratagies....   
# For holes: drop the whole timeseries  
# For others: drop only that value  
  
# Create a list of pins that have a note regarding an issue ----  
troublePins <- SET.data.long %>% ungroup() %>%   
 select(Notes, pin\_ID) %>%   
 filter(complete.cases(.)) %>%   
 `attr<-`("Datainfo", "List of pins that have reported issues (holes, etc)")  
  
pinlistClean <- unique(troublePins$pin\_ID)  
  
# SET data cleaned of any pins that have an 'issue' in the timeseries- Most restrictive dataset strategy 1 from above. ----  
  
   
SET.data.cleanV1 <- SET.data.long %>%   
 filter(!pin\_ID %in% troublePins$pin\_ID) %>%   
 droplevels()  
  
  
attr(SET.data.cleanV1, 'Datainfo') <- "Any pin with a 'history of an issue' has been dropped" # edit metadata  
  
# SET data cleaned of any measures of a pin that had an issue ----  
SET.data.cleanV2 <- SET.data.long %>%  
 filter(is.na(Notes))%>%   
 droplevels()  
  
attr(SET.data.cleanV2, 'Datainfo') <- "Any individual pin reading with a 'an issue' has been dropped" # edit metadata  
  
# SET data cleaned of any pins that were in a hole or on a mussel.  
bigIssues <- c("Hole", "hole", "mussel", "Holr", "Shell", "Mussel", "edge of hole", "hole next to mussel")  
bigIssuePins <- troublePins %>% filter(Notes %in% bigIssues)  
  
  
SET.data.cleanV3 <- SET.data.long %>%   
 filter(!pin\_ID %in% bigIssuePins$pin\_ID)%>% # Filter OUT the pins that have been recorded on a shell/hole/mussel  
 droplevels()  
  
attr(SET.data.cleanV3, 'Datainfo') <- "Any individual pin reading taken atop a hole or mussle has been dropped" # edit metadata  
  
# Identifiers used for reshape ----  
iders <- c("SET\_Type",   
 "Position\_ID",   
 "Location\_ID",   
 "Site\_Name",   
 "Stratafication",   
 "Plot\_Name",   
 "Start\_Date",  
 "Arm\_Direction",   
 "Position\_Name",   
 "X\_Coord",  
 "Y\_Coord",  
 "Coord\_Units",  
 "Coord\_System",  
 "UTM\_Zone",  
 "Datum")

By *'melting'* the SET and SA dataframes we will transpose or pivot the measurement data to allow for calculations of incremental change in accretion and elevation, the change in date and the cumulative change that has been measured. This technique relies on a split-apply-combine strategy common in data sciences. The process first groups the data by some variables and value(s). Each grouping of data is then processed through a set of functions or formulas and the results are then grouped back into the dataset to rebuild the original dataset. The grouping on the SET data is a combination of Study site, location within that study site, the direction of the SET reading arm and finally each pin.  
Be sure to read the comments embedded in the R code blocks as they provide some additional commentary on what's taking place in the code.

The same basic steps are taken for the Surface Accretion data.

###  
# Surface Accretion -----  
#  
#Clean up SA.data dataframe and reshape; same as above but working on surface accretion data  
# similar steps as SET.data  
  
SA.data.long <- SA.data %>% gather('measure', 'Accretion', Measure\_1:Measure\_6) %>%   
 filter(!is.na(Accretion)) %>%   
 select(Layer\_ID, Layer\_Label, Location\_ID.x, Estab\_Date, Start\_Date, Accretion, Core\_Type, Site\_ID, Plot\_Name, Organization) %>%   
 mutate(DecYear = round((((as.numeric(difftime(Start\_Date, Estab\_Date, units = "days"))))/365),3)) %>%   
 dplyr::rename(Date = Start\_Date)  
  
  
  
  
  
### All clear to hear-- 07Mar2014- AFS -----

We now have two additional datasets that contain the SA and SET data in a transposed or *'long'* format