Veta Bonnewell – Analysis of Annual Tree root growth from tube scans

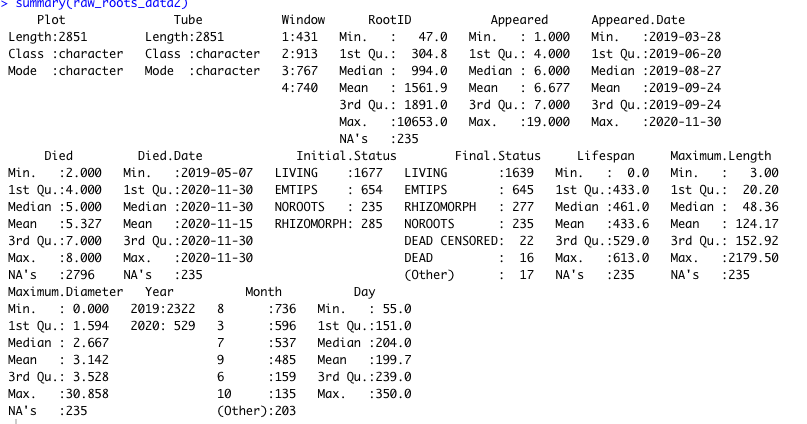
2022-01-13

**Data set**

counts\_roots\_data <- read\_csv("Data/QUBI-E\_counts.csv", na=c("NA", "-", NA))# Converts "NA", "-" to NAs

raw\_roots\_data2 <- read\_csv("Data/QUBI-E\_raw.csv", na=c("NA", "-", NA),…

Summary of raw\_roots\_data2, Note that further investigation shows that the 235 NAs are rows that do not have living roots.



See all the values for factored variables:

> levels(raw\_roots\_data2$Window)

[1] "1" "2" "3" "4"

> levels(raw\_roots\_data2$Initial.Status)

[1] "LIVING" "EMTIPS" "NOROOTS" "RHIZOMORPH"

> levels(raw\_roots\_data2$Final.Status)

[1] "LIVING" "DEAD" "EMTIPS" "NOROOTS" "EMTIPSDEAD" "DEAD CENSORED"

[7] "RHIZOMORPH DEAD" "RHIZOMORPH"

> levels(raw\_roots\_data2$Year)

[1] "2019" "2020"

> levels(raw\_roots\_data2$Month)

[1] "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12"

#The total number of rows in the dataset, and total with rows with the no roots removed

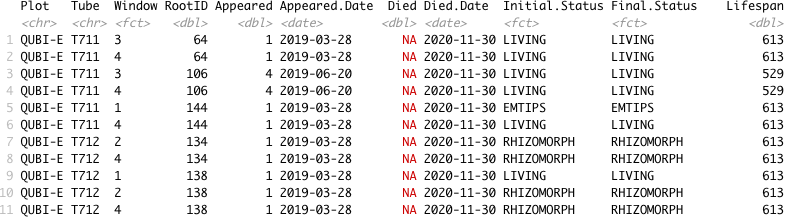
summarise(raw\_roots\_data2,count=n(), rootCountNoNA=sum(!is.na(RootID)))

count r**ootCountNoNA**

<int> <int>

2851 **2616**

#Roots with same id in different windows at same observation time, example



(Assume inconsistency in Died and Died.Date is due to data has not been processed for deaths)

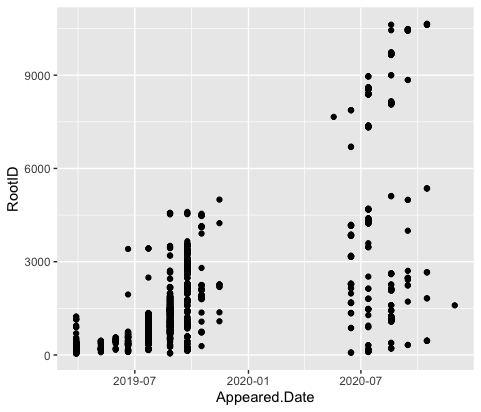
#Further analysis shows that RootID/Tube/Window identifies a specific root across all observations

#. i.e the same RootID in two windows in the same tube on the same date will be counted as 2 roots

#Check the pattern of new root appearance

ggplot(data=raw\_roots\_data2) +

geom\_point(mapping = aes(x =Appeared.Date, y=RootID ))



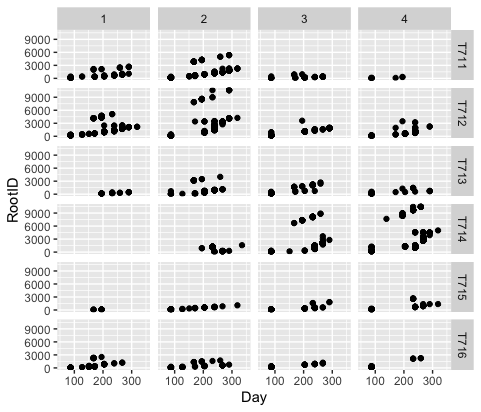
#RootID by tube and window, just looking at pattern, larger RootIDs branches??

# No distinct pattern of more roots in a particular window (depth?)

ggplot(data=raw\_roots\_data2) +

geom\_point(mapping = aes(x =Day, y=RootID ))+

facet\_grid (Tube ~ Window)



**#Plot New roots per date for 2019 and 2020**

All plots with geom\_smooth()` use method = 'loess' and formula 'y ~ x'

New\_RootsByDate<- raw\_roots\_data2 %>%

group\_by(Appeared, Appeared.Date,Year, Month, Day) %>% #These variables should have same value for each Observtion date

summarise(rootCountNA=n(), rootCount=sum(!is.na(RootID)),

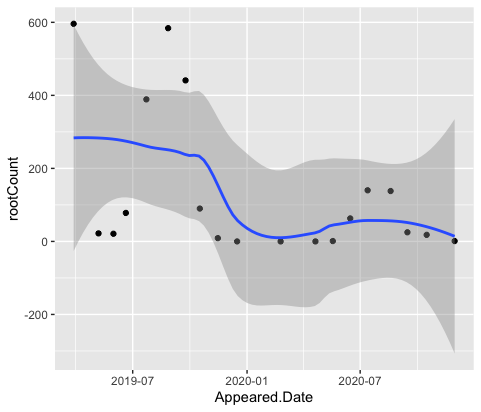
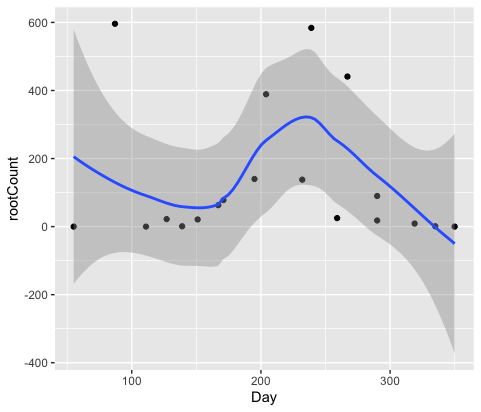
totalLengthmm = sum(Maximum.Length, na.rm=TRUE))

ggplot(data=New\_RootsByDate, mapping =aes(x =Appeared.Date, y=rootCount)) +

#geom\_point(mapping = aes(color=Window )) +

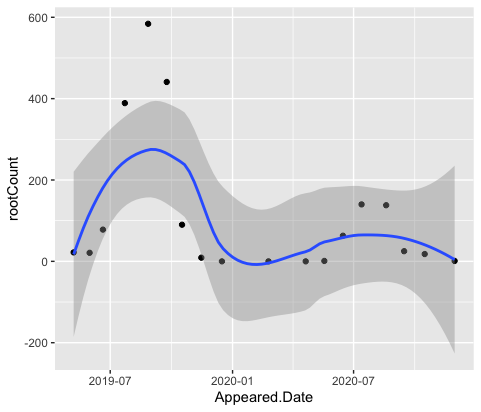
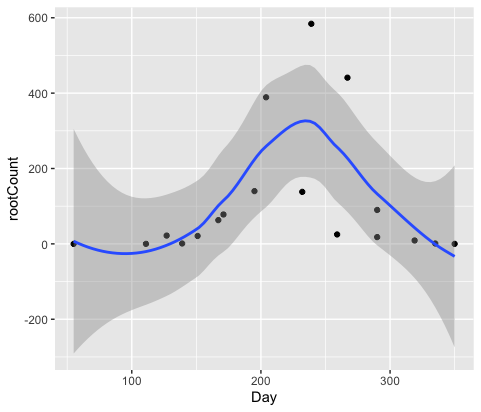
geom\_point() +

geom\_smooth()

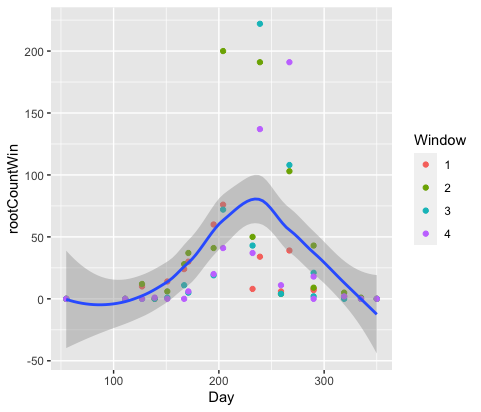


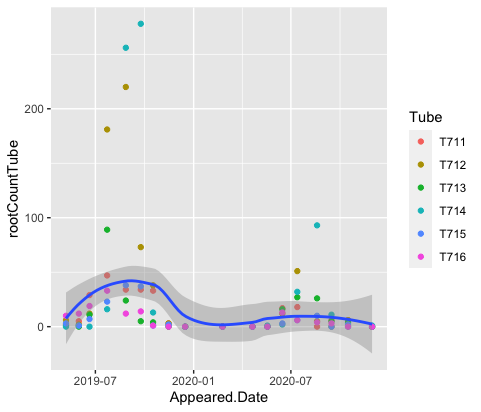
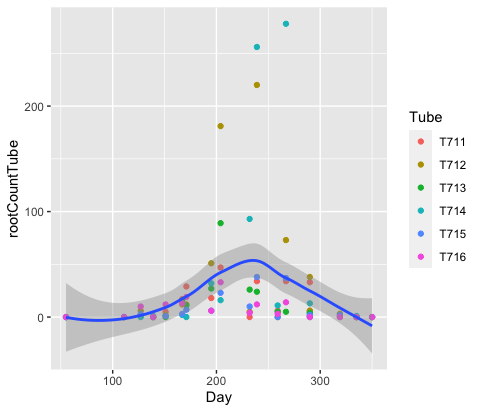
#Fix by removing data for first observation since it is actually comulative from trees past growth

# To increase number of samples, use counts per tube, and counts per window

# Plot of counts per window vs time (all tubes combined)

Note gam is not used since loess is consideredbetter for small datasets (n < 1000)

Further, research into loess vs gam: loess is good for visualizing data and discovering trends.

Does not buil a model that can be used for evaluating curves.

<https://ggplot2.tidyverse.org/reference/geom_smooth.html>

|  |  |
| --- | --- |
| **method** | “Smoothing method (function) to use, accepts either NULL or a character vector, e.g. "lm", "glm", "gam", "loess" or a function, e.g. [MASS::rlm](https://rdrr.io/pkg/MASS/man/rlm.html) or [mgcv::gam](https://rdrr.io/pkg/mgcv/man/gam.html), [stats::lm](https://rdrr.io/r/stats/lm.html), or [stats::loess](https://rdrr.io/r/stats/loess.html). "auto" is also accepted for backwards compatibility. It is equivalent to NULL.  For method = NULL the smoothing method is chosen based on the size of the largest group (across all panels). [**stats::loess()**](https://rdrr.io/r/stats/loess.html) **is used for less than 1,000 observations**; otherwise [mgcv::gam()](https://rdrr.io/pkg/mgcv/man/gam.html) is used with formula = y ~ s(x, bs = "cs") with method = "REML". Somewhat anecdotally, **loess gives a better appearance**, but is *𝑂*(*𝑁*2) in memory, so does not work for larger datasets.  If you have fewer than 1,000 observations but want to use the same gam() model that method = NULL would use, then set method = "gam", formula = y ~ s(x, bs = "cs"). |

ToDo try setting method to gam and see what it generates.

How to generate different models and compare them

Recreate QUBI-E counts?, normalize counts and max lenfth????

Rename raw….2 to remove 2

**2022-01-14. Call with Luke??? Just ask questions without getting bogged down with data.**

?Short version:

What dataset to use? What is target dataset for publishing? 2021 not available yet, so maybe other oaks?

Loess vs gam(). Loess good for showing pattern across years. Assume next step is compare gam to loess results. Get statistical info from gam()

Completed looking at raw data, started generating curves with default routines with breaking the data up in different ways.

Questions for Luke: Any other data to combine to make closer to final analysis.

Show curve with loess overall years (shows variation) then combine years for gam analysis

1. Goal: Generate plots where curves are not subjective, -- goal needs refinement
2. What software used to generate curves previously? Spline-fit, jump
3. Gam not generally recommended for curve fitting, however, alternative loess no stats to compare fit.
4. So assume goal is not just to see curve, but also get some stat measure?
5. Goal: Use gam() to generate curves, investigate parameters of gam() for “best” results

(Jason Fridlay – plant ecologist Syracuse University has some course notes evaluating gam models, presence of birch vs elevation, 1000 obs)

1. Having looked at loess curves for all observations on x axis, day of year (combines years). Can increase number of data points by using counts per window or counts per tube. All give similar shaped curves.
2. Is there any other grouping of data that would simulate the final analysis. For example, data is QUBI-E,

So are there any other QUBI? Assume QUBI is Quercus, are there other oak species if no more of same species.

Luke – Aware of root growing into another window, For now are 2 roots. Before good fake by drawing extra. Current software had previous version that could extend roots but version was buggy so do not want to use it. He and Newton may consider issue again

Luke does want “better” curve fitting, he before varied parameters after running defaults to vary smoothness, but no objective for “best”

Me same true for loess, need stats?

Luke in Future: Uni modal or modality. Was referred to in one of online gam lectures but not how to do

Newton mm?, git?

Tubes had weird data, i.e. vary variable, but interesting to see if including improves stats

Agrees need to compare gam to loess, learn more about gam, if it is useful (ok if no)

Minor points:

1. Confirm RootID are unique for Plot/Tube/Window, i.e. same RootID , same date, in different windows are separate roots.
2. Use git for version control? Sharing? (Me use to share with Arb computer)