1 - Data Analysis.R

* Opens and cleans salinity and stations data
* Separates between Shallow, Deep and NA depth dataframes
  + Trends look similar for Shallow (surface) and Deep (bottom) water measurements, although average salinity for deep measurements are higher.
* Plots the trends for each station in separate PDF files
* Plots the trends for delta salinity (salinity(month) – salinity(month-1))

2 – Data Analysis.R

* Makes distances dataframe
  + From the coordinates of each station and each river/lock
    - Cano Quebrado, Gatun, Ciri Grande, Trinidad, Atlantic and Pacific
    - Makes min\_dist -> minimum distance to a lock, be Atlantic or Pacific (by selecting the smallest of the two distances)
* Checks avg salinity levels before and after neopanamax (just for curiosity)
  + Average before neopanamax
  + Average after neopanamax
  + Average while selecting only stations that have had data collected before and after the opening of neopanamax (some stations only have recent data)
  + Average while only selecting stations that have coordinates (some of them don’t)
* Subsets distances dataframe between Lake, Rivers and Alajuela stations, so that only Lake stations are used to build the lake model.
* Makes traffic dataframe
  + Uses monthly\_traffic.RDS (came from Dat, it’s old data) and monthly\_traffic\_2018.csv (came from ACP’s website, it’s from 2018 to 2020).
  + Extrapolates neo information for the 2018 year (first 12 months), since it wasn’t given, but we know it was on average 15% of total traffic.
  + Possible sources of error are a potential incompleteness of ACP data – I’m unsure if they’ve included all small traffic, which may not have been economically significant for them.
* Estimates delay – makes delay dataframe
  + Sees how long in days did each station take to show a positive delta in salinity after neopanamax’s opening date.
* Plots El Niño and freshwater flow data

9/07/2020

7/07/2020

1. How am I going to incorporate depth?
   1. The original model had averaged surface and bottom values over the same point in the lake.
   2. Also, I could add point depth as a predictor in the model. Probably the best idea here.
2. How am I going to incorporate delay?
   1. One way is what I was originally planning – number of days until a spike in salinity is seen after Neopanamax. However, this has flaws – how can we really guarantee a spike observed is due to the increase in traffic? What if magnitudes vary?
   2. Another way, which I believe is best, is looking at the passive diffusion model and seeing when it expects each station to receive an increase in salinity. We can then see if this model gives us values that are similar to the distribution of stations that receive salinity spikes sooner or later.

* Reconstructing all\_data (with depth)
* Improving the model file
* Working on the passive diffusion model to tell the delays.

6/07/2020

The main issues right now are the unreliable values for traffic before Neopanamax. However, we may get reliable numbers soon from the engineer?

Indeed, separating the river variables makes a huge difference. Indeed, the original model performs very poorly with only post-2016 data, but the one with the separate variables and the fraction does much much better.

With full data:

Model 1: joint terms, river+neo+main – 64%

Model 2: all separate variables – 55%

Model 3: river flow/river dist, joint neo+main – 67.5% (I don’t know why? Trying to find that out)

Model 4: river flow/river dist, frac/dist – 56%

Using only post Neopanamax data:

Model 1: joint terms, river+neo+main – 70%

Model 2: all separate variables – 60%

Model 3: river flow/river dist, joint neo+main – 73%

Model 4: river flow/river dist, frac/dist – 56% (obviously the same)

I believe this is as good as we can get the shipping information going – the percentage of neopanamax as sole predictor, at least until we have full main shipping information from the engineer or another source to include the previous years as well.

In relation with the model building, I believe the best I can do to improve it for now is to think of the delay. It really seems that the percentage, for now, is the best we can do to represent traffic, and the river variables are doing good for the four rivers we have.

For optim, things work fine for the function x. Morgane told me a bit more about how the function works, so I do get how it’s so much more flexible, but it’s still raising me an error.

What I’m planning on doing tomorrow:

* Reconstruct all\_data.csv dataframe in a clearer way
* Look into depth as a term
* See if it’d be possible to improve the distance term (from linear distance to the diffusion model’s output)