Crosswalk Demo

library(tidyverse)  
library(sf)  
library(dataRetrieval) # WQP web service library  
library(XML)  
library(leaflet)  
library(mapview)  
  
knitr::opts\_chunk$set(echo = TRUE, message = FALSE)  
  
NHD\_URL <- 'https://hydro.nationalmap.gov/arcgis/services/nhd/MapServer/WMSServer'  
  
find\_lake\_word <- function(data, search\_term, state\_abbr = NULL) {  
 if (!missing(state\_abbr)) {  
 if (!state\_abbr %in% state.abb) stop("Use a valid 2-letter state abbreviation")  
 }  
 if (missing(state\_abbr)) {  
 filtered\_data <- data[apply(data, 1, function(x) any(grepl(search\_term, x, ignore.case = TRUE))),]  
 }  
 else {  
 filtered\_data <- data[apply(data, 1, function(x) any(grepl(search\_term, x, ignore.case = TRUE))) & data$State == state.abb,]  
 }  
 return\_data <- filtered\_data %>%  
 distinct(lagoslakeid) %>%  
 inner\_join(data, by = "lagoslakeid")  
 return(return\_data)  
   
}

This notebook will show some examples of the complex relationships that can be found in the LAGOS Lake Link (lake identifier crosswalk) as well as examples of problems that could possibly be solved with some more work. For a more detailed introduction to LAGOS Lake Link and a description of the creation process, see the document “LAGOS Lake Link: Creation”

# Definitions in this document

“Lake”: Permanent lake or reservoir.

# What is LAGOS Lake Link?

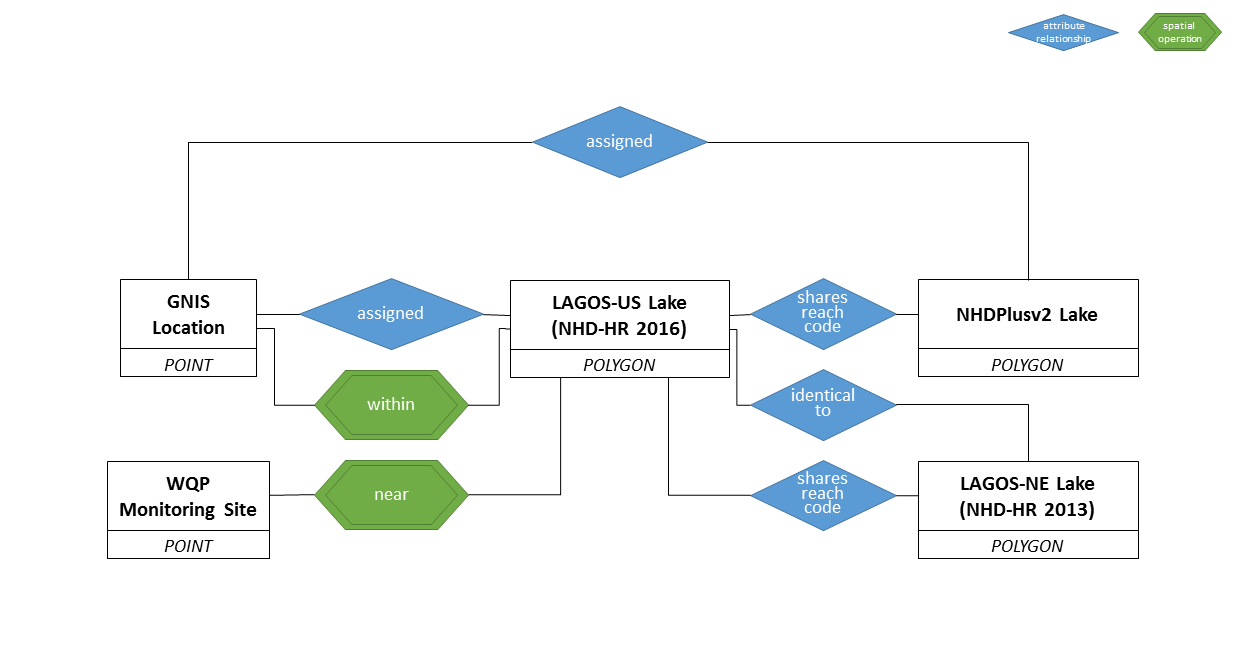
LAGOS Lake Link is a tabular dataset (a crosswalk table) that can be used to connect one lake-related dataset to another for many common lake datasets. Several lake datasets are in common use on their own or as a base for scientific data products and LAGOS Lake Link is intended to make it easier to combine lake-related data between multiple sources. The table can be searched to find identifiers and location for a particular lake, or it can be used in data join operations to convert identifiers en masse.

# Data dictionary

data\_dictionary <- read\_csv("LAGOS\_Lake\_Link\_data\_dictionary.csv")  
data\_dictionary

## # A tibble: 24 x 7  
## `Column Name` `Column Type` `Is Nullable` Definition `Source Dataset`  
## <chr> <chr> <chr> <chr> <chr>   
## 1 lagoslakeid int N Unique lak~ LAGOS-US   
## 2 NHDHR\_Permane~ char N 40-char GU~ NHD-HR   
## 3 NHDHR\_ReachCo~ char Y "Unique id~ NHD-HR   
## 4 NHDHR\_AreaSqKm decimal N Area of ar~ NHD-HR   
## 5 GNIS\_ID int Y Unique ide~ NHD-HR   
## 6 GNIS\_Name char Y Proper nam~ NHD-HR   
## 7 LAGOS\_LakeName char Y Alternate ~ GNIS   
## 8 LAGOS\_CountyN~ <NA> <NA> Current na~ <NA>   
## 9 LAGOS\_CountyF~ <NA> <NA> 5-digit co~ <NA>   
## 10 LAGOS\_Latitud~ decimal N The latitu~ LAGOS-US   
## # ... with 14 more rows, and 2 more variables: `Source Column Name` <chr>,  
## # Example <chr>

# Entity-relationship diagram



Entity-relationship diagram

Preview of the crosswalk table. As a reminder, there are **479950 LAGOS-US lakes**.

glimpse(xwalk, width = 100)

## Observations: 569,329  
## Variables: 26  
## $ lagoslakeid <int> 31198, 31198, 31198, 31198, 31198, 31198, 122, 122, 12...  
## $ NHDHR\_PermanentIdentifier <chr> "120020498", "120020498", "120020498", "120020498", "1...  
## $ NHDHR\_ReachCode <chr> "04010201003330", "04010201003330", "04010201003330", ...  
## $ NHDHR\_AreaSqKm <dbl> 2.745000, 2.745000, 2.745000, 2.745000, 2.745000, 2.74...  
## $ GNIS\_ID <int> 649289, 649289, 649289, 649289, 649289, 649289, 653169...  
## $ GNIS\_Name <chr> "Perch Lake", "Perch Lake", "Perch Lake", "Perch Lake"...  
## $ LAGOS\_LakeName <chr> "Perch Lake", "Perch Lake", "Perch Lake", "Perch Lake"...  
## $ LAGOS\_CountyName <chr> "Carlton County", "Carlton County", "Carlton County", ...  
## $ LAGOS\_CountyFIPS <int> 27017, 27017, 27017, 27017, 27017, 27017, 27017, 27017...  
## $ LAGOS\_LatitudeNAD83 <dbl> 46.68943, 46.68943, 46.68943, 46.68943, 46.68943, 46.6...  
## $ LAGOS\_LongitudeNAD83 <dbl> -92.67066, -92.67066, -92.67066, -92.67066, -92.67066,...  
## $ State <chr> "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN", ...  
## $ WQP\_MonitoringLocationIdentifier <chr> "FONDULAC-114A", "FONDULAC-114B", "FONDULAC\_WQX-114A",...  
## $ WQP\_MonitoringLocationName <chr> "Perch Lake (North Basin)", "Perch Lake (South Basin)"...  
## $ WQP\_ProviderName <chr> "STORET", "STORET", "STORET", "STORET", "STORET", "STO...  
## $ NHDPlusv2\_COMID <chr> "1776072", "1776072", "1776072", "1776072", "1776072",...  
## $ NHDPlusv2\_ReachCode <chr> "04010201003330", "04010201003330", "04010201003330", ...  
## $ NHDPlusv2\_AreaSqKm <dbl> 2.720, 2.720, 2.720, 2.720, 2.720, 2.720, 1.348, 1.348...  
## $ LAGOSNE\_lagoslakeid <int> 31198, 31198, 31198, 31198, 31198, 31198, 122, 122, 12...  
## $ NLA2007\_SITE\_ID <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ NLA2012\_SITE\_ID <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ count\_wqp\_per\_lagos\_id <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 0, 0, 0, 0, 4, 4, ...  
## $ count\_nhdplusv2\_per\_lagos\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 0, 1, 2, 2, ...  
## $ count\_lagos\_per\_nhdplusv2\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, NA, 1, 1, 1,...  
## $ count\_lagosNE\_per\_lagos\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...  
## $ count\_lagos\_per\_lagosNE\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...

# 1:1 relationships, or, “easy ones”

A large proportion of lakes connect easily and don’t split or merge polygons over time. The crosswalk has all lakes greater than or equal to 1 hectare, but some of those are too small to be in the NHDPlusv2, so let’s only check lakes greater than or equal to 4 hectares.

lagos\_lake\_count\_4ha <- xwalk %>%   
 filter(NHDHR\_AreaSqKm >= 0.04) %>%  
 distinct(lagoslakeid) %>%  
 nrow()  
count\_easy <- xwalk %>%  
 filter(NHDHR\_AreaSqKm >= 0.04) %>%  
 filter(count\_lagos\_per\_lagosNE\_id <= 1 & # LAGOS-NE only has partial coverage so 0 is just fine  
 count\_lagos\_per\_nhdplusv2\_id == 1 &  
 count\_lagosNE\_per\_lagos\_id == 1 &  
 count\_nhdplusv2\_per\_lagos\_id == 1) %>%  
 distinct(lagoslakeid) %>%  
 nrow()  
  
pct1 <- round(100\*count\_easy/lagos\_lake\_count\_4ha, 0)

**30% of lakes are very simple!** They may have multiple WQP sites, which is a bit more intuitive than the other 1: many relationships.

Here is an example lake. Bask in its comforting consistency, because things are about to get more complex.

xwalk %>% filter(lagoslakeid == 1)

## # A tibble: 1 x 26  
## lagoslakeid NHDHR\_PermanentIdent~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 1 50524769 02020003000803 1.15 607350  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

Equally simple, a large proportion of lakes has no matches for anything. Here is an example:

xwalk %>% filter(lagoslakeid == 256832)

## # A tibble: 1 x 26  
## lagoslakeid NHDHR\_PermanentIdent~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 256832 {45479AA3-EDAB-4DE4-~ 04030106002915 0.149 NA  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

# 1:many relationship exploration

## LAGOS-US & WQP

(n\_lagos\_wqp <- xwalk %>%  
 mutate(Number\_of\_WQP\_sites\_per\_LAGOS\_lake = case\_when(count\_wqp\_per\_lagos\_id == 0 ~ '0', count\_wqp\_per\_lagos\_id == 1 ~ '1', count\_wqp\_per\_lagos\_id > 1 ~ '>1')) %>%  
 distinct(Number\_of\_WQP\_sites\_per\_LAGOS\_lake, lagoslakeid) %>%  
 count(Number\_of\_WQP\_sites\_per\_LAGOS\_lake)  
)

## # A tibble: 3 x 2  
## Number\_of\_WQP\_sites\_per\_LAGOS\_lake n  
## <chr> <int>  
## 1 >1 14193  
## 2 0 450475  
## 3 1 15282

Out of the lakes that have a WQP sampling site, about half of them have more than one site within. In total, **29475** LAGOS lakes have WQP sites linked.

### Examples of lakes with more than one WQP site

This is what a user will see in the crosswalk that should indicate to them that multiple WQP sites are found within a lake. This lake has 10 sites. The lake information is duplicated for every site (i.e., the crosswalk table is not normalized).

(ex1 <- xwalk %>% filter(lagoslakeid == 3476))

## # A tibble: 10 x 26  
## lagoslakeid NHDHR\_PermanentIden~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 3476 81632411 07010207001341 2.48 640111  
## 2 3476 81632411 07010207001341 2.48 640111  
## 3 3476 81632411 07010207001341 2.48 640111  
## 4 3476 81632411 07010207001341 2.48 640111  
## 5 3476 81632411 07010207001341 2.48 640111  
## 6 3476 81632411 07010207001341 2.48 640111  
## 7 3476 81632411 07010207001341 2.48 640111  
## 8 3476 81632411 07010207001341 2.48 640111  
## 9 3476 81632411 07010207001341 2.48 640111  
## 10 3476 81632411 07010207001341 2.48 640111  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

If the lake also joins to multiple other lakes in either the NHDPlusv2 or LAGOS-NE, the user will see something like this. This site has 7 WQP sites and 2 NHDPlusv2 polygons for a total of 14 rows in the crosswalk. It’s not possible (currently) to tell which WQP sites go with which NHDPlusv2 polygon. Now both lake and WQP information are duplicated.

(ex2 <- xwalk %>% filter(lagoslakeid == 4069))

## # A tibble: 14 x 26  
## lagoslakeid NHDHR\_PermanentIden~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 4069 82817060 04060101008480 0.382 1621364  
## 2 4069 82817060 04060101008480 0.382 1621364  
## 3 4069 82817060 04060101008480 0.382 1621364  
## 4 4069 82817060 04060101008480 0.382 1621364  
## 5 4069 82817060 04060101008480 0.382 1621364  
## 6 4069 82817060 04060101008480 0.382 1621364  
## 7 4069 82817060 04060101008480 0.382 1621364  
## 8 4069 82817060 04060101008480 0.382 1621364  
## 9 4069 82817060 04060101008480 0.382 1621364  
## 10 4069 82817060 04060101008480 0.382 1621364  
## 11 4069 82817060 04060101008480 0.382 1621364  
## 12 4069 82817060 04060101008480 0.382 1621364  
## 13 4069 82817060 04060101008480 0.382 1621364  
## 14 4069 82817060 04060101008480 0.382 1621364  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

### How many WQP sites didn’t link to a LAGOS Lake?

load('./rdata/wqp\_sf.RData')  
wqp\_no\_link <- wqp\_sf %>%  
 anti\_join(distinct(xwalk, WQP\_MonitoringLocationIdentifier), by = c("MonitoringLocationIdentifier" = "WQP\_MonitoringLocationIdentifier"))

28786 out of 131005 sites did not link (22%).

### Map of sites that didn’t link.

As you can see, Great Lakes sites don’t always have the “Great Lake” designation in the WQP and we are fine with these sites not linking.

wqp\_no\_link %>%   
 filter(startsWith(HUCEightDigitCode,'0406')) %>%  
 select(MonitoringLocationIdentifier, MonitoringLocationName, MonitoringLocationDescriptionText, MonitoringLocationTypeName) %>%  
 mapview()

I would like to ask the limnologists: **What do you make of these samples clustered around large lakes?** The point is on the intended site as you can tell by looking at where the indicated township section would be. What are they sampling here? Should it be linked along with the other lake water quality samples?

We could improve some of the WQP matches by doing a “closest” join and enforcing a requirement for shared words.

## LAGOS-US & NHDPlusv2

NHDPlusv2 within LAGOS Lake:

(count\_lagos\_nhd <- xwalk %>%  
 mutate(Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake = case\_when(count\_nhdplusv2\_per\_lagos\_id == 0 ~ '0', count\_nhdplusv2\_per\_lagos\_id == 1 ~ '1', count\_nhdplusv2\_per\_lagos\_id > 1 ~ '>1')) %>%  
 distinct(Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake, lagoslakeid) %>%  
 count(Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake)  
)

## # A tibble: 3 x 2  
## Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake n  
## <chr> <int>  
## 1 >1 1440  
## 2 0 278218  
## 3 1 200292

And vice versa:

(count\_nhd\_lagos <- xwalk %>%  
 mutate(Number\_of\_LAGOS\_Lakes\_per\_NHDPlusv2 = case\_when(count\_lagos\_per\_nhdplusv2\_id == 0 ~ '0', count\_lagos\_per\_nhdplusv2\_id == 1 ~ '1', count\_lagos\_per\_nhdplusv2\_id > 1 ~ '>1')) %>%  
 distinct(Number\_of\_LAGOS\_Lakes\_per\_NHDPlusv2, lagoslakeid) %>%  
 count(Number\_of\_LAGOS\_Lakes\_per\_NHDPlusv2) %>%  
 filter(!is.na(Number\_of\_LAGOS\_Lakes\_per\_NHDPlusv2))  
)

## # A tibble: 2 x 2  
## Number\_of\_LAGOS\_Lakes\_per\_NHDPlusv2 n  
## <chr> <int>  
## 1 >1 5710  
## 2 1 196131

It seems it’s more common for NHDPlusv2 lakes to split into multiple LAGOS-US lakes than the other way around, but both occur. Occurs rarely: less than 1% of lakes.

### Missing LAGOS-NHDPlusv2 connections

A large proportion of LAGOS lakes have no NHDPlusv2 lake connection, but I don’t actually expect that the smallest lakes would be found in the NHD medium-resolution. Check it out using a **size cutoff of 10 hectares**, which is more than large enough to appear in both.

(count\_lagos\_nhd\_4ha <- xwalk %>%  
 filter(NHDHR\_AreaSqKm >= 0.1) %>%  
 mutate(Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake\_over10ha = case\_when(count\_nhdplusv2\_per\_lagos\_id == 0 ~ '0', count\_nhdplusv2\_per\_lagos\_id == 1 ~ '1', count\_nhdplusv2\_per\_lagos\_id > 1 ~ '>1')) %>%  
 distinct(Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake\_over10ha, lagoslakeid) %>%  
 count(Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake\_over10ha)  
)

## # A tibble: 3 x 2  
## Number\_of\_NHDPlusv2\_per\_LAGOS\_Lake\_over10ha n  
## <chr> <int>  
## 1 >1 1114  
## 2 0 16440  
## 3 1 43462

Still, **over 1/4 of 10-hectare lakes have no connection**. This is after we have also searched GNIS\_ID for matches. What are some of these lakes? I originally thought I might need to walk through the NHDReachCrossReference table more times, but I can’t find an instance where that would help in 20 or so lakes I’ve hand-checked.

Instead, I found 3 common situations:

1. New reach code is linked to old reach codes that aren’t in the NHDPlusv2 waterbody layer
2. New reach code is created with no old reach codes indicated
3. There is no equivalent waterbody included in the NHDPlusv2 despite these lakes being more than large enough (this may legimitely happen due to real-world changes over time).

Of these, (3) is a legitimate lack of connection. **(1) and (2) could be linked with a spatial analysis along with a condition that the lake areas are within, say, 25% of each other–I think we could pick up almost 30,000 more lakes this way.**

### Examples of each

This is what a user will see in the crosswalk that should indicate to them that multiple NHDPlusv2 lakes have been condensed to a single LAGOS-US lake. This lake has 4 NHDPlusv2 connections and 2 WQP sites, so 8 rows appear.

xwalk %>% filter(count\_nhdplusv2\_per\_lagos\_id > 3 & count\_nhdplusv2\_per\_lagos\_id <5) %>% sample\_n(10)

## # A tibble: 10 x 26  
## lagoslakeid NHDHR\_PermanentIden~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 4044 155639076 04050001012397 7.62 436929  
## 2 154409 46366370 03030002013128 5.68 996193  
## 3 4044 155639076 04050001012397 7.62 436929  
## 4 457119 60573450 16050302000758 8.89 861754  
## 5 344386 148214878 10140101020437 19.2 1265607  
## 6 3809 152373794 07070002011943 0.120 1560896  
## 7 344386 148214878 10140101020437 19.2 1265607  
## 8 73 120019354 09030003002126 185 662245  
## 9 3809 152373794 07070002011943 0.120 1560896  
## 10 104 120019079 09030001037980 57.3 663188  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

xwalk %>% filter(lagoslakeid == 2107)

## # A tibble: 8 x 26  
## lagoslakeid NHDHR\_PermanentIdent~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 2107 76246872 07050005007101 0.244 1575381  
## 2 2107 76246872 07050005007101 0.244 1575381  
## 3 2107 76246872 07050005007101 0.244 1575381  
## 4 2107 76246872 07050005007101 0.244 1575381  
## 5 2107 76246872 07050005007101 0.244 1575381  
## 6 2107 76246872 07050005007101 0.244 1575381  
## 7 2107 76246872 07050005007101 0.244 1575381  
## 8 2107 76246872 07050005007101 0.244 1575381  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

This is what a user will see in the crosswalk that should indicate to them that 1 NHDPlusv2 lake has been split into multiple LAGOS-US lakes. These 4 lagos lakes used to be 1 NHDPlusv2 lake. 1 of them has 4 WQP sites, while the others have none, resulting in 7 rows appearing. Searching by lagoslakeid alone will not indicate the related lakes, of course.

xwalk %>% filter(NHDPlusv2\_COMID == '18469416')

## # A tibble: 7 x 26  
## lagoslakeid NHDHR\_PermanentIdent~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 32510 90385408 05120202010841 0.0726 437773  
## 2 150 90385400 05120202001140 5.88 437773  
## 3 150 90385400 05120202001140 5.88 437773  
## 4 150 90385400 05120202001140 5.88 437773  
## 5 150 90385400 05120202001140 5.88 437773  
## 6 96986 90385410 05120202010842 0.0333 437773  
## 7 52431 90385406 05120202001139 0.0842 437773  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

## LAGOS-US & LAGOS-NE

The relationships seen here bear a strong resemblence to those seen with the NHDPlusv2, above.

(count\_ne\_per\_us <- xwalk %>%  
 mutate(Number\_of\_NE\_per\_US\_Lake = case\_when(count\_lagosNE\_per\_lagos\_id == 0 ~ '0', count\_lagosNE\_per\_lagos\_id == 1 ~ '1', count\_lagosNE\_per\_lagos\_id > 1 ~ '>1')) %>%  
 distinct(Number\_of\_NE\_per\_US\_Lake, lagoslakeid) %>%  
 count(Number\_of\_NE\_per\_US\_Lake)  
)

## # A tibble: 3 x 2  
## Number\_of\_NE\_per\_US\_Lake n  
## <chr> <int>  
## 1 >1 193  
## 2 0 341083  
## 3 1 138674

And vice versa:

(count\_us\_per\_ne <- xwalk %>%  
 mutate(Number\_of\_US\_per\_NE\_Lake = case\_when(count\_lagos\_per\_lagosNE\_id == 0 ~ '0', count\_lagos\_per\_lagosNE\_id == 1 ~ '1', count\_lagos\_per\_lagosNE\_id > 1 ~ '>1')) %>%  
 distinct(Number\_of\_US\_per\_NE\_Lake, lagoslakeid) %>%  
 count(Number\_of\_US\_per\_NE\_Lake) %>%  
 filter(!is.na(Number\_of\_US\_per\_NE\_Lake))  
)

## # A tibble: 2 x 2  
## Number\_of\_US\_per\_NE\_Lake n  
## <chr> <int>  
## 1 >1 211  
## 2 1 138680

### Example: maximum complexity

Here’s what happens when not only is a LAGOS-US lake associated with multiple LAGOS-NE lake polygons, but also when it has multiples of *everything*. This lake has 2 WQP sites, was once 4 LAGOS-NE lakes, and is associated with 3 NHDPlusv2 lakes. It’s not possibly to untangle which sample sites are where or how the LAGOS-NE lakes might relate to the NHDPlusv2 lakes, currently. 24 rows appear for this lake.

xwalk %>% filter(lagoslakeid == 268427)

## # A tibble: 24 x 26  
## lagoslakeid NHDHR\_PermanentIden~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 2 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 3 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 4 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 5 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 6 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 7 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 8 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 9 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## 10 268427 {28f2e7d4-ede1-46f3~ 05120202063169 0.764 NA  
## # ... with 14 more rows, and 21 more variables: GNIS\_Name <chr>,  
## # LAGOS\_LakeName <chr>, LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

# Searching by name

It is possible to use names to search through all available information in the crosswalk if you build the right sort of function. I have done so in the setup code at the top of this document. Here are the results for the search term ‘Cochituate’. This is a chain of lakes known as Lake Cochituate. Even though the NHD doesn’t identify all of them by name, the linked WQP sites indicate the association.

find\_lake\_word(xwalk, 'Cochituate')

## # A tibble: 8 x 26  
## lagoslakeid NHDHR\_PermanentIdent~ NHDHR\_ReachCode NHDHR\_AreaSqKm GNIS\_ID  
## <int> <chr> <chr> <dbl> <int>  
## 1 12 129738545 01070005000884 0.544 NA  
## 2 12 129738545 01070005000884 0.544 NA  
## 3 50905 129738546 01070005000897 0.970 NA  
## 4 50905 129738546 01070005000897 0.970 NA  
## 5 50905 129738546 01070005000897 0.970 NA  
## 6 14255 129738544 01070005000875 0.792 616989  
## 7 14255 129738544 01070005000875 0.792 616989  
## 8 14255 129738544 01070005000875 0.792 616989  
## # ... with 21 more variables: GNIS\_Name <chr>, LAGOS\_LakeName <chr>,  
## # LAGOS\_CountyName <chr>, LAGOS\_CountyFIPS <int>,  
## # LAGOS\_LatitudeNAD83 <dbl>, LAGOS\_LongitudeNAD83 <dbl>, State <chr>,  
## # WQP\_MonitoringLocationIdentifier <chr>,  
## # WQP\_MonitoringLocationName <chr>, WQP\_ProviderName <chr>,  
## # NHDPlusv2\_COMID <chr>, NHDPlusv2\_ReachCode <chr>,  
## # NHDPlusv2\_AreaSqKm <dbl>, LAGOSNE\_lagoslakeid <int>,  
## # NLA2007\_SITE\_ID <chr>, NLA2012\_SITE\_ID <chr>,  
## # count\_wqp\_per\_lagos\_id <int>, count\_nhdplusv2\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_nhdplusv2\_id <int>, count\_lagosNE\_per\_lagos\_id <int>,  
## # count\_lagos\_per\_lagosNE\_id <int>

# Appendix: Useful bits of code for exploration

This chunk identifies lakes with a missing NHDPlusv2 connection.

missing\_nhdplusV2 <- xwalk %>%  
 filter(NHDHR\_AreaSqKm >= 0.1) %>%  
 filter(count\_nhdplusv2\_per\_lagos\_id < 1) %>%  
 filter(lagoslakeid < 5000) %>%  
 select(-WQP\_MonitoringLocationIdentifier, -WQP\_MonitoringLocationName, -WQP\_ProviderName) %>%  
 distinct()

And with the help of ArcMap to find equivalent lakes visually, this bit of code helps search for the cause of the missing link.

load('./rdata/nhd\_xref.RData')  
# slow but effective  
search\_nhd\_xref <- function(codes, depth = 1) {  
 list\_results <- lapply(codes, function(x) filter(nhd\_xref, OldReachCode == x | NewReachCode == x))  
 df1 <- do.call("rbind", list\_results)  
 if (depth == 2) {  
 codes = c(pull(df1, OldReachCode), pull(df1, NewReachCode))  
 list\_results2 <- lapply(codes, function(x) filter(nhd\_xref, OldReachCode == x | NewReachCode == x))  
 df2 <- do.call("rbind", list\_results)  
 } else {  
 df2 <- df1  
 }  
 return(df2)  
}  
  
(result <- search\_nhd\_xref(c('04050001011691','04050001016492'), depth = 2) %>%  
 mutate(OldReachCode = as.character(OldReachCode), NewReachCode = as.character(NewReachCode)))

## OldReachCode OldReachDate NewReachCode NewReachDate  
## 1 <NA> <NA> 04050001011691 2007-03-23

# need to have nhd\_plus\_orig object loaded  
#nhd\_plus\_orig %>% filter(REACHCODE %in% pull(result, OldReachCode))