LAGOS Package Documentation

Farzan Masrour

October 16, 2015

This document is the first draft of documentation for using LAGOS package in **R**.

1 preparation

Using this code is quite easy. First, you need to set your working directory to be the main folder of package Version1.054.1 that contains the *"LAGOS1.054.1.R"* file. Second, in order to use the data and methods in this code you need to call the *"LAGOS1.054.1.R"* file using source command as follow.

```
> #First set the working directory
> setwd("/Users/farzan/Desktop/Dropbox/
+       Summer\ 2015/Job/Lagos\ Package/LAGOS\ Package")
> #Second loading Data and functions
> source("LAGOS1.054.1.R")
```

If you running the above code in *RStudio* you might be able to see the added data frames and functions in your Environment.

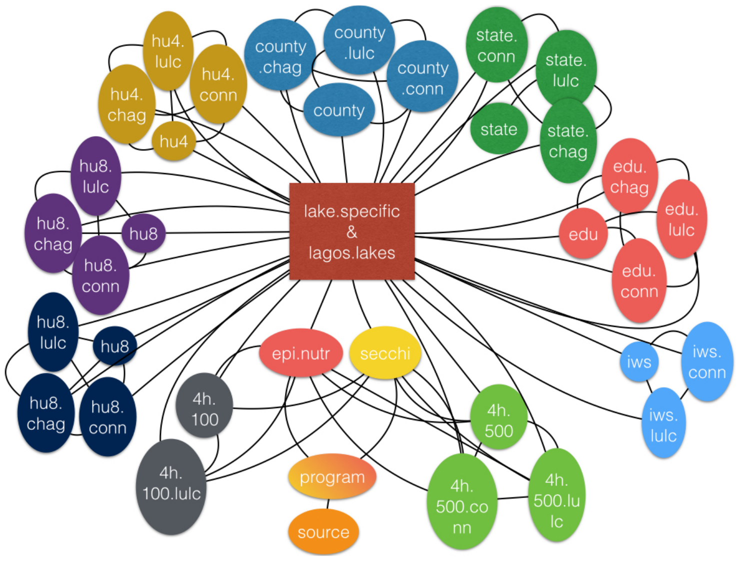


Figure 1: The network of data frames. There is a connection between two data frame if and only if they have common identifiers.

2 Data Structure and Preprocessing

All the csv files of LAGOS are imported to the *LGOS* package and ready to use. This package contains 38 data sets where all saved in a single *".RData"* file. You can find the last version of the data file in the main folder of the package. By calling the *"LAGOS.R"* all the data sets loaded as 38 data frames Table 1. The package also contains an *iforamtionTable* data frame that contains the details of all the data sets, and two *lists* limno and geo, that separate *geo* and *limno* data frames.

You can get access to more detail information about each data frame by using commands like *summary()* , *names()*, *head()*, *etc.* For example to see what are the columns of secchi data frame,

```
> names(secchi)
```

```
 [1] "eventidc10541"     "lagoslakeid"       "programname"
 [4] "programtype"       "lagosversion"      "sampledate"
 [7] "secchi"            "secchi_censorcode" "secchi_qual"
[10] "secchi_methodinfo" "greatlakes"        "sampleyear"
[13] "samplemonth"
```

Or to see the statistical details of secchi data

```
> summary(secchi)
```

```
 eventidc10541      lagoslakeid                      programname
 Min.   : 901626   Min.   :     1   MN_MPCA_SECCHI        :261856
 1st Qu.:1127028   1st Qu.:  2318   MN_MPCA_CHEM_1999_2012:232995
 Median :1352409   Median :  4387   WI_DNR_NUTRIENT       :125987
 Mean   :1352438   Mean   : 20669   MI_CORPS_CHEM         : 85900
 3rd Qu.:1577869   3rd Qu.:  8428   ME_DEP_CHEM           : 84332
 Max.   :1803250   Max.   :141326   VT_DWQ_NUTRIENT       : 24295
                                    (Other)               : 86160
                                              programtype      lagosversion
 State Agency/Citizen Monitoring Program          :652999   1.054.1:901525
 State Agency/University/Citizen Monitoring Program:198270
 State Agency                                     : 37504
 Non-Profit Agency                                :  4798
 LTER                                             :  4042
 University                                       :  2689
 (Other)                                          :  1223
      sampledate         secchi       secchi_censorcode  secchi_qual
 2009-07-19:   838   Min.   : 0.000   GT:  8734         NO     :113755
 2010-06-20:   753   1st Qu.: 1.680   LT:     5         2      : 43311
 2007-07-15:   751   Median : 2.900   NC:892786         4      : 37249
 2009-06-14:   727   Mean   : 3.224                     YES    :  3759
 2008-07-20:   708   3rd Qu.: 4.400                     5      :  2277
 2008-08-24:   627   Max.   :26.822                     (Other):  3336
 (Other)   :897121                                      NA's   :697838
          secchi_methodinfo    greatlakes   sampleyear     samplemonth
 SECCHI_VIEW        :  3382   Min.   :0    Min.   :1937   Min.   : 1.000
 SECCHI_VIEW_UNKNOWN: 11073   1st Qu.:0    1st Qu.:1995   1st Qu.: 6.000
 NA's               :887070   Median :0    Median :2002   Median : 7.000
                              Mean   :0    Mean   :2000   Mean   : 7.184
                              3rd Qu.:0    3rd Qu.:2007   3rd Qu.: 8.000
                              Max.   :0    Max.   :2013   Max.   :12.000
```

2.1 Merging

In order to build your own data frame you have two options. The first option is using the R function *merge()*. For example, if you want to have a data frame that contains the "secchi" value and "sampleyear" from *secchi* data frame, and "lakes4ha_buffer500m_streamdensity_streams_sum_lengthm" from *lakes4ha.buffer500m.conn* data frame. First, you have to find the column number of each value

```
> names(lakes4ha.buffer500m.conn)
```

```
 [1] "lakes4ha_buffer500m_nhdid"
 [2] "lakes4ha_buffer500m_canalditchdensity_sum_lengthm"
 [3] "lakes4ha_buffer500m_canalditchdensity_density_mperha"
 [4] "lakes4ha_buffer500m_streamdensity_streams_sum_lengthm"
 [5] "lakes4ha_buffer500m_streamdensity_streams_density_mperha"
 [6] "lakes4ha_buffer500m_streamdensity_headwaters_sum_lengthm"
 [7] "lakes4ha_buffer500m_streamdensity_headwaters_density_mperha"
 [8] "lakes4ha_buffer500m_streamdensity_midreaches_sum_lengthm"
 [9] "lakes4ha_buffer500m_streamdensity_midreaches_density_mperha"
[10] "lakes4ha_buffer500m_streamdensity_rivers_sum_lengthm"
[11] "lakes4ha_buffer500m_streamdensity_rivers_density_mperha"
[12] "lagoslakeid"
```

```
> names(secchi)
```

```
 [1] "eventidc10541"     "lagoslakeid"       "programname"
 [4] "programtype"       "lagosversion"      "sampledate"
 [7] "secchi"            "secchi_censorcode" "secchi_qual"
[10] "secchi_methodinfo" "greatlakes"        "sampleyear"
[13] "samplemonth"
```

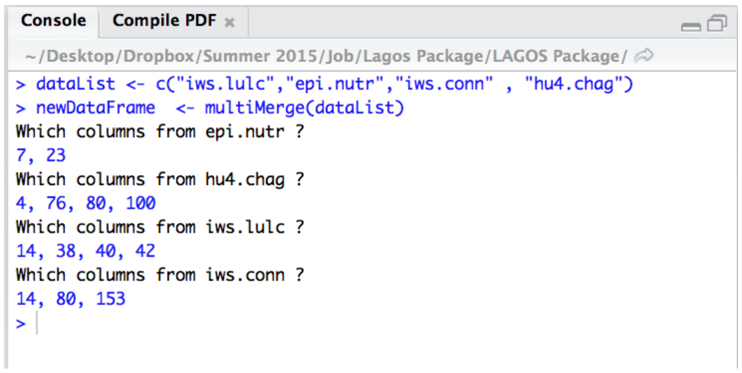
You can see "secchi" column number is 7, "sampleyer" is 12, and "lakes4ha_buffer500m _streamdensity_streams_sum_lengthm" is 4. Now we can use the merge function as follow,

```
>  newDataFrame <- merge( secchi[,c(7,12,2)],
+                 lakes4ha.buffer500m.conn[,c(4,12)],
+                 by.x="lagoslakeid",
+                 by.y = "lagoslakeid" )
```

In the above code, in addition to the original columns that we wanted, we also added the column "lagoslakeid", column number 2 in *secchi* and column number 12 in *lakes4ha.buffer500m.conn*. "lagoslakeid" is the common column between these two data frames and merge function needs a common column to do the merging between two data frame.

```
>  head(newDataFrame)
```

```
  lagoslakeid secchi sampleyear
1           1    1.2       2002
2           1    1.3       2002
3           1    1.4       2002
4           2    6.5       2006
5           2    5.8       2006
6           2    6.1       2006
  lakes4ha_buffer500m_streamdensity_streams_sum_lengthm
1                                             3271.3529
2                                             3271.3529
3                                             3271.3529
4                                              573.3263
5                                              573.3263
6                                              573.3263
```


Console
Compile PDF
~/Desktop/Dropbox/Summer 2015/Job/Lagos Package/LAGOS Package/

```
> dataList <- c("iws.lulc","epi.nutr","iws.conn" , "hu4.chag")
> newDataFrame  <- multiMerge(dataList)
Which columns from epi.nutr ?
7, 23
Which columns from hu4.chag ?
4, 76, 80, 100
Which columns from iws.lulc ?
14, 38, 40, 42
Which columns from iws.conn ?
14, 80, 153
> 
```

Figure 2: Using *multiMerge()* function, the function get a vector of names of data frames as input and then ask about the column numbers in the Console. The numbers must be comma seprated numbers

The common column should have a unique value for each row of the data frame we call these columns identifiers in our data frames, see Table 1 for the identifier column of each data frame in the package.

however,using *merge()* has two major difficulties. First, most of the data frames in our package do not have common identifiers, see Figure 1 for the connections of data frames. For example if you want to merge *lagos.source* data frame with *iws.conn*, you have to merge *lagos.source* with *lagos.program* then with *epi.nutr* then with *lagoslakes* and finally *iws.conn*. Second, merge function only accepts two data frames as input.
The second option is using *multiMerge()* function in the package. For instance, consider we want to merge some columns from *epi.nutr*, *hu4.chag*,*iws.lulc*, and*iws.conn*. First, we need to find the columns numbers we are interseted in from each data frame using *names()* funciton as we did for the above example. Then, we make a string vector that contains the name of data frames we wanted as it's elements.

```
> dataList <- c("iws.lulc","epi.nutr","iws.conn" , "hu4.chag")
```

Then call the*multiMerge()* function as follow

```
> newDataFrame  <- multiMerge(dataList)
```

When we call the multiMerge(), the function asks for the columns numbers we want from each data frame. We have to insert the columns numbers for each data frame in the **R** Console. The columns numbers must be seperated by ",", see Figure 2. There is no need to add any identifiers column or extra data frame in case of no connection between data frames.

name	**type**	**variables**	**observations**	**identifier**
county	geo	8	955	county_zoneid
county.chag	geo	147	955	county_zoneid
county.conn	geo	152	955	county_zoneid
county.lulc	geo	182	955	county_zoneid
edu	geo	10	91	edu_zoneid
edu.chag	geo	147	91	edu_zoneid
edu.conn	geo	153	91	edu_zoneid
edu.lulc	geo	166	91	edu_zoneid
hu4	geo	9	65	hu4_zoneid
hu4.chag	geo	147	65	hu4_zoneid
hu4.conn	geo	153	65	hu4_zoneid
hu4.lulc	geo	166	65	hu4_zoneid
hu8	geo	9	511	hu8_zoneid
hu8.chag	geo	147	511	hu8_zoneid
hu8.conn	geo	153	511	hu8_zoneid
hu8.lulc	geo	166	511	hu8_zoneid
hu12	geo	12	20257	hu12_zoneid
hu12.chag	geo	144	20257	hu12_zoneid
hu12.conn	geo	154	20257	hu12_zoneid
hu12.lulc	geo	163	20257	hu12_zoneid
iws	geo	12	51065	iws_zoneid
iws.conn	geo	156	51065	iws_zoneid
iws.lulc	geo	158	51065	iws_zoneid
state	geo	7	17	state_zoneid
state.chag	geo	147	17	state_zoneid
state.conn	geo	152	17	state_zoneid
state.lulc	geo	158	17	state_zoneid
lakes4ha.buffer100m	geo	3	51065	lagoslakeid
lakes4ha.buffer100m.lulc	geo	152	51065	lagoslakeid
lakes4ha.buffer500m	geo	3	51065	lagoslakeid
lakes4ha.buffer500m.conn	geo	12	51065	lagoslakeid
lakes4ha.buffer500m.lulc	geo	155	51065	lagoslakeid
lagoslakes	geo	108	141271	hub
epi.nutr	limno	87	164402	lagoslakeid, programname, eventida
lake.specific	limno	41	141471	hub
secchi	limno	13	656474	lagoslakeid, programname, eventidc
lagos.source	limno	3	28	sourceid
lagos.program	limno	13	40	programname, sourceid

Table 1: The list of data sets in the package

2.2 Data Diagnostic, Quality Control

2.3 Data Pre-processing

2.4 Descriptive Stats and Visualization

3 Data Analysis