PITcleanr Vignette

Kevin E. See

2017-11-22

Contents

Introduction

The PITcleanr package was developed to help query the necessary data to fit a DABOM model (**D**am **A**dult **B**ranch **O**ccupancy **M**odel) in order to estimate adult escapement to various tributaries above a tagging location. A key assumption of a DABOM model is that fish travel along a single upstream route. Therefore, the model will fail if presented with detections by a single fish in multiple tributaries above a branching node. PITcleanr is designed to help clean the PIT tag detection data to identify non-linear upstream pathways and help the user determine which observations to keep.

System requirements

PITcleanr relies on the following R packages which can be downloaded via CRAN or by using the function `install.packages()`:

- `dplyr`, `lubridate`, `httr`, `purrr` and `stringr`: can all be installed by installing the `tidyverse` package

In addition, many of the various queries in PITcleanr require connection to the internet.

```
library(tidyverse)
library(httr)
library(stringr)
library(lubridate)
library(PITcleanr)
```

Data sources

PITcleanr starts with data from an adult trap database, to determine which tags are considered part of the valid tag sample, and what the trap date is for each tag. For the Lower Granite dam, the data can be retreived from the "QCI Downloads" folder on the Idaho Fish and Game's IFWIS website. The trapping data is stored in an Microsoft Access database that is updated weekly and named "LGTrappingExportJodyW.accdb". Permission must be obtained from Idaho Fish and Game before viewing, accessing and retreiving data from the site. This database contains two tables, which can be exported as .csv files. Once this is done, PITcleanr has a function, `filterLGRtrapDB` to determine which tags are valid, and when they were caught in the adult trap.

Other than the adult trap database, PITcleanr relies on queries from the Columbia Basin Research Data Access in Real Time (DART) and the Columbia Basin PIT Tag Information System (PTAGIS).

Workflow

For the purposes of this vignette, we will focus on data to be fed into the Lower Granite version of DABOM, using data from spring/summer Chinook that crossed Lower Granite in the spring and summer of 2015.

```
spp = c('Chinook', 'Steelhead')[1]
yr = 2015
```

Configuration Table

One of the first steps is to compile a configuration file containing information about all of the detection sites that may be present. PITcleanr will create such a configuration file for all sites in the Columbia River basin, using the `buildConfig()` function.

```
org_config = buildConfig()
```

```
## [1] "Querying INT sites' metadata"
## [1] "Querying INT sites' configuration information"
## [1] "Querying MRR sites' metadata"
```

A user can then make custom changes depending on the model set-up. For example, in the Lower Granite version of DABOM, all antennas in VC1 and VC2 are combined into a single node.

```
my_config = org_config %>%
  mutate(Node = ifelse(SiteID %in% c('VC2', 'VC1', 'LTR', 'MTR', 'UTR'),
                       SiteID,
                       Node),
         Node = ifelse(SiteID %in% c('CROTRP',
                                     'CRT',
                                     'REDTRP',
                                     'REDR',
                                     'RRT'),
                       'above_SC2',
                       Node),
         Node = ifelse(SiteID == 'AFC',
                       ifelse(grepl('MAINSTEM', AntennaGroup),
                              'AFCBO',
                              'AFCAO'),
                       Node),
         Node = ifelse(SiteID %in% c('TUCH', 'TFH'),
```

```
                        'TUCH_TFH',
                        Node),
         Node = ifelse(SiteID == 'MCCA',
                        'STR',
                        Node),
         Node = ifelse(SiteID == 'CARMEC',
                        'CRCA0',
                        Node),
         Node = ifelse(SiteID == 'BIG2C',
                        'TAYA0',
                        Node),
         Node = ifelse(SiteID == 'WIMPYC',
                        'WPCA0',
                        Node),
         Node = str_replace(Node, '^BTC', 'BTL'),
         Node = ifelse(SiteID %in% c('YANKFK', 'CEY'),
                        'YFKA0',
                        Node),
         Node = ifelse(SiteID == 'SAWT',
                        'STL',
                        Node),
         Node = ifelse(SiteID == 'LOOH',
                        'LOOKGC',
                        Node),
         Node = ifelse(SiteID == 'RPDTRP',
                        'RAPH',
                        Node),
         Node = ifelse(SiteID == 'CHARLC',
                        'CCAB0',
                        Node),
         Node = ifelse(Node == 'KEN',
                        'KENB0',
                        Node),
         Node = ifelse(Node == 'HYC',
                        'HYCB0',
                        Node),
         Node = ifelse(Node == 'LLR',
                        'LLRB0',
                        Node),
         Node = ifelse(Node == 'LRW',
                        'LRWB0',
                        Node),
         Node = ifelse(SiteID == '18M',
                        paste0('X', Node),
                        Node)) %>%
  distinct()

# add HBC if it isn't in there yet
if(!'HBC' %in% unique(my_config$SiteID)) {
  my_config = my_config %>%
    bind_rows(tibble(SiteID = 'HBC',
                     Node = 'HBC',
                     SiteType = 'INT',
```

```
                    RKM = '522.303.416.049.002',
                    RKMTotal = 1292))
}

# a sample of this configuration file
head(my_config)
```

```
## # A tibble: 6 x 19
##   SiteID ConfigID AntennaID  Node ValidNode  StartDate EndDate Comment
##    <chr>    <dbl>     <chr> <chr>     <lgl>     <dttm>  <dttm>   <lgl>
## 1    158      100        D1 158A0        NA 2011-11-29      NA      NA
## 2    158      100        D2 158A0        NA 2011-11-29      NA      NA
## 3    158      100        D3 158B0        NA 2011-11-29      NA      NA
## 4    158      100        D4 158A0        NA 2011-11-29      NA      NA
## 5    158      100        D5 158A0        NA 2011-11-29      NA      NA
## 6    158      100        D6 158B0        NA 2011-11-29      NA      NA
## # ... with 11 more variables: SiteType <chr>, SiteName <chr>,
## #   ModelMainBranch <lgl>, AntennaGroup <chr>, ArrayOrder <lgl>,
## #   SiteDescription <chr>, SiteTypeName <chr>, RKM <chr>, RKMTotal <dbl>,
## #   Latitude <dbl>, Longitude <dbl>
```

Node Network

The next step is to define how the various sites that are to be used with DABOM are related to each other. PITcleanr contains a function to describe that network for the Lower Granite version of DABOM, `writeLGRNodeNetwork()`. This serves two purposes: it defines which sites are to be used in the DABOM model, and it describes how they are related to each other along the stream network.

```
site_df = writeLGRNodeNetwork()

# remove some sites that have been combined with others (see the modifications to the configuration fil
site_df = site_df %>%
  filter(!SiteID %in% c('MCCA',
                        'WIMPYC',
                        'YANKFK', 'CEY',
                        'SAWT',
                        'LOOH',
                        'CARMEC',
                        'BIG2C',
                        'RPDTRP'))

# a sample of what this looks like
head(site_df)
```

```
## # A tibble: 6 x 12
##   SiteID                          path      Step1    Step2 Step3 Step4
##   <fctr>                         <chr>      <chr>    <chr> <chr> <chr>
## 1    LAP         Clearwater.Lapwai.LAP Clearwater   Lapwai   LAP
## 2    MIS     Clearwater.Lapwai.LAP.MIS Clearwater   Lapwai   LAP   MIS
## 3    SWT     Clearwater.Lapwai.LAP.SWT Clearwater   Lapwai   LAP   SWT
## 4    WEB Clearwater.Lapwai.LAP.SWT.WEB Clearwater   Lapwai   LAP   SWT
## 5    JUL       Clearwater.Potlatch.JUL Clearwater Potlatch   JUL
## 6    KHS   Clearwater.Potlatch.JUL.KHS Clearwater Potlatch   JUL   KHS
## # ... with 6 more variables: Step5 <chr>, Step6 <chr>, Step7 <chr>,
```

```
## #   Step8 <chr>, Step9 <chr>, Step10 <chr>
```

Parent-Child Table

Next, we build a parent-child dataframe describing which child nodes are directly upstream of each parent node in our system. The `createParentChildDf` function contains several arguements:

- `sites_df`: This is the dataframe we built using the `writeLGRNodeNetwork()` function.
- `config`: This is the configuration file we built using the `buildConfig()` function.
- `startSite`: This defines which site all paths will start from. This is typically the dam where the adult trap is located.
- `startDate`: The first date (in YYYYMMDD format) when fish may start being caught in the adult trap. Configurations that ended prior to this date are excluded.

```
parent_child = createParentChildDf(site_df,
                                   my_config,
                                   startSite = 'GRA',
                                   startDate = ifelse(spp == 'Chinook',
                                                      paste0(yr, '0301'),
                                                      paste0(yr-1, '0701')))

# add one site in Kenney Creek, if it's been dropped
if(sum(grepl('KENA0', parent_child$ChildNode)) == 0) {
  lineNum = which(parent_child$ChildNode == 'KENB0')

  parent_child = parent_child %>%
    slice(1:lineNum) %>%
    bind_rows(parent_child %>%
                slice(lineNum) %>%
                mutate(ParentNode = 'KENB0',
                       ChildNode = 'KENA0')) %>%
    bind_rows(parent_child %>%
                slice((lineNum + 1):n()))
}
```

Valid Tags

The user must define the path to the adult trap database, which contains the data PITcleanr uses to determine which tags are valid to be fed into DABOM. PITcleanr contains an example of the trap database (`trap_chnk2015`), which only contains spring/summer Chinook in 2015, so if `trap_path` is set to `NULL`, this will be used. If the user has access to the trap database, set `trap_path` equal to the file path, including the file name and extension, of the actual table (named tblLGDMasterCombineExportJodyW.csv in my version of the trap database). Currently it must be in CSV format.

From the trap database, PITcleanr searches for tags that are part of the valid tag list for a particular species and year, and returns a dataframe with their tag code and the date of trapping. Currently, within the Lower Granite trap database, PITcleanr filters for tags of the appropriate species and spawn year, and then filters for returning adults (`LGDLifeStage == 'RF'`), adipose-intact fish (`LGDMarkAD == 'AI'`), with non-missing PIT tag numbers (`!is.na(LGDNumPIT)`) and that are marked as valid (`LGDValid == 1`). Besides filtering out these valid tags, the function `filterLGRtrapDB` provides an arguement to save those tag IDs as a text file, to make it easier to upload to PTAGIS.

```
data(trap_chnk2015)
head(trap_chnk2015)
```

```
## # A tibble: 6 x 52
##   MasterID CollectionDate SpawnYear LGDTrapID BioSamplesID
##      <int>          <dttm>     <chr>     <int>        <chr>
## 1   260635     2015-04-23    SY2015        NA         <NA>
## 2   260640     2015-04-23    SY2015        NA         <NA>
## 3   260636     2015-04-23    SY2015        NA         <NA>
## 4   260639     2015-04-23    SY2015        NA         <NA>
## 5   260671     2015-04-23    SY2015        NA         <NA>
## 6   260672     2015-04-23    SY2015        NA         <NA>
## # ... with 47 more variables: CollectionLocation <chr>, LGDFLmm <int>,
## #   LGDWeight <chr>, SRR <chr>, LGDSpecies <int>, PtagisSpecies <int>,
## #   GenSpecies <chr>, LGDRun <int>, PtagisRun <int>, GenRun <chr>,
## #   LGDRear <chr>, PtagisRear <chr>, GenRear <chr>, LGDLifeStage <chr>,
## #   LGDSex <chr>, GenSex <chr>, GenStock <chr>, GenStockProb <dbl>,
## #   GenParentHatchery <chr>, GenBY <int>, GenPBT_ByHat <chr>,
## #   GenPBT_RGroup <chr>, GenComments <chr>, LGDFishComments <chr>,
## #   BioScaleFinalAge <chr>, PtagisEventSites <chr>,
## #   PtagisLastEventSite <chr>, PtagisLastEventDate <dttm>,
## #   PtagisEventLastSpawnSite <chr>, RepeatSpawner <chr>,
## #   BiosamplesValid <lgl>, LGDValid <int>, LGDInjuryiesAll <chr>,
## #   LGDMarksAll <chr>, LGDMarkAD <chr>, LGDTagsAll <chr>, LGDNumPIT <chr>,
## #   LGDNumRT <chr>, LGDNumJaw <chr>, LGDOpComments <chr>,
## #   PTAgisSxCGRAObs <chr>, OpTrapSampleRate <dbl>, OpWaterTemp <chr>,
## #   CreatedDate <dttm>, CreatedUser <chr>, Editdate <dttm>, EditUser <chr>
```

```
valid_df = filterLGRtrapDB(trap_path = NULL,
                           species = spp,
                           spawnYear = yr,
                           saveValidTagList = T,
                           validTagFileNm = 'ValidTags.txt')
```

Complete Tag History

The next step is to compile the complete tag history of all valid tags in our sample. This can be done in two ways.

Through PTAGIS

After logging into PTAGIS, select the "Data" tab, "Advanced Reporting" and then the "Launch" button. Now, select the "New Query Builder2 Report" icon and the "Complete Tag History" option. The **DABOM** functions require the following attributes to be selected; *Tag*, *Event Date Time*, *Event Release Date Time*, *Event Site Code*, *Antenna*, and *Antenna Group Configuration*. Other attributes can be selected and they can appear in any order. After selecting attributes, upload *ValidTags.txt* in the 'Tag Code - List or Text File' section and then run the report and save the output. Next read the outputted observation file into R or use the example observation file, "chnk15_obs".

```
data(chnk2015_obs)
observations = chnk2015_obs
```

Through DART

DART has built an API query allowing users to query the complete tag history of any particular tag. Be warned however, on one test machine, this query could return the records of about 10 tags per minute, so for several thousand tags, it will take awhile.

```
observations = valid_tag_df %>%
  select(TagID) %>%
  as.matrix() %>%
  as.character() %>%
  as.list() %>%
  map_df(.f = queryCapHist,
         configuration = my_config)
```

Process Raw Observations

`PITcleanr` contains a single function, `processCapHist_LGD` that processes the raw observations and returns several key pieces of information. First, it constructs a table of all the pathways a fish might take that are "valid" in DABOM, meaning they reflect continual upstream movement along a single branch of the stream network (using an internal function, `getValidPaths`). Next, it uses the trap database to determine the trap date of each tag. The next step is to assign each observation to an appropriate node from the DABOM model, as defined in the configuration file, using an internal function, `assignNodes()`. It requires the complete tag histories, the dataframe of valid tags and their trap date, the site configuration dataframe, and the parent-child dataframe to filter out observations from sites not contained in the DABOM version. If `truncate` is set to `TRUE` (the default), the `assignNodes()` function will filter out observations that occurred prior to the trap date, and remove consecutive observations at the same node to simplify the file.

Finally, the processing function determines whether each observation should be considered valid for `DABOM`. It does this by determining the final node a fish was observed at, and querying the valid path dataframe for all the downstream sites that may be encountered on the way to that node. It also proposes an extended path of the first set of nodes upstream of the final node, and acknowledges that observations along that extended path may also be valid (e.g. a fish swims a particular route upstream, spawns and then swims partway downstream, past a node or two. The most upstream node should be considered a valid observation, and the likely spawning location.).

It returns a dataframe with a row for each observed node, showing the tag code, the site code, the model node, the minimum observed date-time, and two columns called "AutoProcStatus" and "UserProcStatus". There are additional columns as well, called:

- `NodeOrder`: how many nodes would a tag cross, including the one identified in that row, to reach the current node? This is regardless of whether a tag was observed at those nodes.
- `Direction`: based on previously observed node, is the tag moving upstream or downstream? If this is `NA`, the current node is not along a valid path containing the previous node.
- `ValidPath`: Taken together, do all the observed nodes for a particular tag fall along a single valid path?
- `ModelObs`: Marked `TRUE` for all observed nodes, in sequence by observation, that initially constitute a valid path. Does not assume that the final observation is along the true valid path.
- `SiteDescription`: The site description of the site, as provided by PTAGIS.

"AutoProcStatus" is PITcleanr's best attempt to determine which observations should be kept (and which should be deleted) before bringing the data into DABOM. "UserProcStatus"" is there for the user to change things. Any tag that has at least one flagged observation ("AutoProcStatus" == `FALSE`) will have all of the observations labeled as blanks in the "UserProcStatus" column. This allows the user to filter initially on the rows with `UserProcStatus == ''`, and see all the observations for each flagged tag. The user can then determine which observations should be included as "valid" by marking the UserProcStatus as `TRUE`, or which ones should be deleted by marking the UserProcStatus as `FALSE`, guided by the suggestions in

the AutoProcStatus column, information contained in the other columns, and by the user's knowledge of fish behavior and the system in question. The *UserComment* field is meant to record reasons why certain observations are deleted.

We imagine the workflow to involve initially filtering on all the blank UserProcStatus rows. Then, for each tag, determine whether each observation should be kept (mark *UserProcStatus* == `TRUE`) or discarded (mark *UserProcStatus* == 'FALSE') based on suggestions from the *AutoProcStatus*, *ValidPaths* and *ModelObs* fields. For any observation destined to be discarded, a reason should be provided in the *UserComment* for re-producibility. When no blank *UserProcStatus* rows remain, save the file. It is now ready for importation into DABOM.

```
proc_list = processCapHist_LGD(species = spp,
                               spawnYear = yr,
                               configuration = my_config,
                               parent_child,
                               trap_path = NULL,
                               filter_by_PBT = T,
                               observations = observations,
                               truncate = T,
                               save_file = T,
                               file_name = 'ProcessedCH.xlsx')
```

```
## [1] "2017-11-22 08:59:55  1 of 3025 tags. Current tag: 384.3B2395D0A4"
## [1] "2017-11-22 09:00:17  500 of 3025 tags. Current tag: 384.3B23AD2E1B"
## [1] "2017-11-22 09:00:39  1000 of 3025 tags. Current tag: 3DD.00773A0BCE"
## [1] "2017-11-22 09:01:02  1500 of 3025 tags. Current tag: 3DD.00773AA565"
## [1] "2017-11-22 09:01:25  2000 of 3025 tags. Current tag: 3DD.00773AD944"
## [1] "2017-11-22 09:01:47  2500 of 3025 tags. Current tag: 3DD.00773B083C"
## [1] "2017-11-22 09:02:12  3000 of 3025 tags. Current tag: 3DD.00773BB453"
```

Summarise final spawning location and biological information

`PITcleanr` provides a function to determine a fish's final spawning location, and combine it with lots of biological data from the trap database at Lower Granite, `summariseTagData`. It relies on a processed capture history file returned by the function `proc_list`. Internally, this relies on another function, `estimateSpawnLoc`, which determines the final spawning location of each tag, by finding the furthest upstream node a tag was observed at within it's valid observations.

```
tag_summ = summariseTagData(proc_list)

head(tag_summ)
```

```
## # A tibble: 6 x 30
##            TagID             LastObs LastObsSite LastObsNode MasterID
##            <chr>              <dttm>       <chr>       <chr>    <int>
## 1 384.3B2395D0A4 2015-05-19 10:00:00      LGRLDR         GRA   269174
## 2 384.3B2396E59B 2016-12-02 08:01:00        KOOS        KOOS   262153
## 3 384.3B239716AE 2015-06-05 10:00:00      LGRLDR         GRA   271660
## 4 384.3B23978F9F 2015-04-28 10:00:00      LGRLDR         GRA   261134
## 5 384.3B2397AD2E 2015-05-22 10:00:00      LGRLDR         GRA   270044
## 6 384.3B2397B7A8 2015-05-06 10:00:00      LGRLDR         GRA   263830
## # ... with 25 more variables: CollectionDate <dttm>, SpawnYear <chr>,
## #   BioSamplesID <chr>, LGDFLmm <int>, SRR <chr>, GenRear <chr>,
## #   LGDLifeStage <chr>, GenSex <chr>, GenStock <chr>, GenStockProb <dbl>,
```

```
## #   GenParentHatchery <chr>, GenBY <int>, GenPBT_ByHat <chr>,
## #   GenPBT_RGroup <chr>, BioScaleFinalAge <chr>, PtagisEventSites <chr>,
## #   PtagisLastEventSite <chr>, PtagisLastEventDate <dttm>,
## #   PtagisEventLastSpawnSite <chr>, RepeatSpawner <chr>,
## #   BiosamplesValid <lgl>, LGDValid <int>, LGDInjuryiesAll <chr>,
## #   LGDMarksAll <chr>, LGDMarkAD <chr>
```