PITcleanr\_lite: Workflow Documentation

Last modified: 2/6/2023

Modified by: Mark Roes

Overview

The following is a step-by-step guide to compile PIT-tag mark and observation data and process them into capture histories useful for analysis of fish or animal movement, growth, survival, etc. The workflow uses a draft version of the GitHub repository [PITcleanr\_lite](https://github.com/Mount-Hood-Environmental/PITcleanr_lite) which is intended to be a user-friendly, stand-alone group of R functions and scripts that leverages functionality and is a companion to the previously developed [PITcleanr](https://github.com/KevinSee/PITcleanr) R package. PITcleanr\_lite is designed to “compress” large, sometimes unwieldy PIT-tag observation datasets for a given list of PIT-tagged fish into observation records and capture histories that are more manageable to aid in fisheries analyses. PITcleanr\_lite accommodates observation data downloaded from the Columbia Basin PIT Tag Information System ([PTAGIS](https://www.ptagis.org/)), the Biomark, Inc. [BioLogic](https://data3.biomark.com/) web portal, raw “.log” files, and data from submersible PIT tag detection systems. PTAGIS is the centralized database for PIT-tagged fish in the Columbia River basin and houses observation data for many of the more permanent PIT-tag arrays located throughout the region. However, PTAGIS does not contain most observation data from either 1) observation sites outside of the Columbia River basin or 2) temporary or project-specific detection sites (e.g., litz cords). These observations can instead be found in BioLogic™ database and web portal, at least for arrays and infrastructure installed by Biomark, Inc.

In this document we describe the PITcleanr\_lite (2022) workflow, using an example from the Lemhi River watershed to compile a list of unique PIT-tagged fish that are marked and/or released at rotary screw traps (RST). Our focus is on juvenile Chinook salmon *Oncorhynchus tshawystscha* and steelhead *O. mykiss* released at two RSTs located just upstream from the confluence of Hayden Creek and the upper Lemhi River: the Upper Lemhi River Rotary Screw Trap (LEMTRP) and the Hayden Creek Rotary Screw Trap (HYDTRP). Each of these traps are intended to monitor the emigration timing, abundance, and survival from subpopulations spawning upstream; juveniles are then tracked through the lower Lemhi River (downstream from Hayden Creek to the confluence with the Salmon River) including through multiple stream habitat rehabilitation projects that are “wired” with litz cords to monitor use of e.g., newly created or restored side channels. Juveniles can also be interrogated or tagged and released at the Lower Lemhi River RST (LLRTP) located below many of the stream rehabilitation projects before leaving the Lemhi River to the Salmon River.

Required and Recommended Items

The following example requires:

1. A recent version of R software installed (<https://cran.rstudio.com/>)
2. A local copy of the PITcleanr\_lite GitHub repo (<https://github.com/Mount-Hood-Environmental/PITcleanr_lite>)
3. A PTAGIS account to create and execute queries (<https://www.ptagis.org/>)
4. Access to Biomark’s BioLogic™ web portal (<https://data3.biomark.com/>)

We additionally recommend the use of the following:

1. [RStudio](https://posit.co/download/rstudio-desktop/)

R and RStudio

For step-by-step instructions on downloading and installing both R and RStudio, please visit [Appendix A](https://rstudio-education.github.io/hopr/starting.html) of the Hands-On Programming with R website.

PITcleanr\_lite

Download the PITclear\_lite repository from GitHub ([link](https://github.com/Mount-Hood-Environmental/PITcleanr_lite)) onto your local machine by visiting the link, clicking on the green “Code” button, and “Download ZIP”. Download the ZIP folder and unzip it into a convenient and accessible location.

Tag or Mark Data

Any set of tagging or mark data compatible with PITcleanr\_lite can be transformed using the workflow. By default, importing data from a PTAGIS “complete tag history” query will form a master list of tags, such that tag observation data from other sources will only be included for tags listed within the PTAGIS data. Details about changing defaults are located in the “Script Configuration” section of this document.

In our example, we use two PTAGIS queries to construct a list of unique PIT-tagged juveniles tagged and/or released at the LEMTRP and HYDTRP screw traps, which we will track through the lower Lemhi River. This list of tags is then queried for their *Complete Tag History*. Alternatively, a user could provide their own list of PIT-tagged individuals in a .txt or .csv file format. To complete the PTAGIS queries, the user will need to do the following:

1. Navigate to <https://www.ptagis.org>
2. Login (or Register for a new account, if necessary)
3. Navigate to the “Advanced Reporting Home Page”
4. Click “Create Query Builder2 Report”

This will take you to a list of standard queries that are provided by PTAGIS.

### Tagging Detail Query

First, we will execute a “Tagging Detail” query in PTAGIS to identify all newly PIT-tagged Chinook salmon and steelhead that are released at RSTs within the Lemhi River Subbasin. For the purposes of this exercise, let’s focus on fish marked starting in 2020 to present. In the left panel of the Tagging Detail Query page, the user will see a series of indices that allows one to choose which attributes to include in their export, while also allowing filtering on select attributes.

1. Report type: Tagging Detail
2. **1** Select Attributes – add the following attributes to the default:
   1. Length
   2. Weight
   3. Note: any additional attributes may be added by the user and will be included in the export, but “Length” and “Weight” must be added for proper workflow operation.
3. Filters
   1. **4** Capture Method = Screw Trap
   2. **9** Mark Site Subbasin = 17060204:Lemhi
   3. **10** Mark Year = 2020 – Present
   4. **16** Species = Chinook, Steelhead
4. Save (bottom of screen)
   1. Give the query a descriptive name, such as “MarkDat\_Lemhi”
   2. **Important:** The query must be saved as “static”. Be sure to check the box as in the highlighted image below



* 1. You can organize your queries in folders, if desired.
  2. Click OK

1. Run Report

The above query provides us a list of all newly tagged fish released at Lemhi River RSTs. However, some juvenile Chinook salmon and steelhead are PIT-tagged using other capture methods (e.g., electrofishing) which can occur upstream of the RSTs. If the desire is to monitor all unique PIT-tagged juveniles released at the RSTs to 1) ensure representative marks and 2) bolster sample sizes, then an additional query is needed to identify juveniles that arrive at the RSTs with a PIT-tag. These fish are released with those from the *Tagging Detail Query* and can be identified using the following query:

Recapture Detail Query

1. Report type: Recapture Detail
2. **1** Select Attributes – add the following attributes to the default:
   1. Recap Length
   2. Recap Weight
   3. Note: Similar as above, any additional attributes may be added by the user to include in the export, but “Recap Length” and “Recap Weight” must be added for workflow operation.
3. Filters
   1. **10** Mark Site Subbasin = 17060204:Lemhi
   2. **19** Recap Capture Method = Screw Trap
   3. **27** Recap Year = 2020 – Present
   4. **28** Species = Chinook, Steelhead
4. Save (bottom of screen)
   1. Give the query a descriptive name, such as “RecapDat\_Lemhi”
   2. **Important:** The query must be saved as “static”. Be sure to check the box as in the highlighted image below



* 1. To can organize your queries in folders, if desired.
  2. Click OK

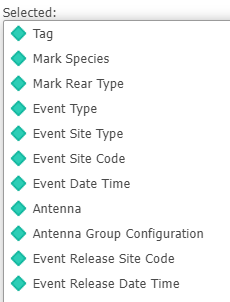
1. Run Report

The combined above *Tag Detail Query* plus *Recapture Detail Query* provides a list of all unique PIT-tagged Chinook salmon and steelhead released at RSTs in the Lemhi River, including those tagged at the trap and those that arrived at the trap with a PIT-tag. The user can then use a *Complete Tag History Query* in PTAGIS to query observations of those fish at interrogation sites (e.g., PIT-tag arrays) that are registered in PTAGIS. Note: This query only provides observations at site registered in PTAGIS, which are generally more permanent sites. To get observations at additional sites not registered in PTAGIS (e.g., at litz cords), we will need to download those separately from BioLogic™.

Complete Tag History Query

This query filters the entire PTAGIS database for tags that are recorded in the Tag Detail Query and Recapture Detail Query and returns the associated complete tag histories to be downloaded.

1. Report type: Complete Tag History
2. **1** Select Attributes – select the following attributes as listed below:



1. Filters
   1. **6** Event Site = HAYDNC, HYC, HYDTRP, LEMTRP, LRW, EVU, S2I, S2O, BHC, S3B, S3A, EVL, LLRTP, LLR
      1. This is a list of interrogation sites in the Lemhi River that are registered in PTAGIS as of the writing of this user guide.
   2. **31** Tag – Saved Report (UNION)

A picture containing table

Description automatically generated

1. Save (bottom of screen). Unlike the first two queries, this report can be saved as “prompted”.
2. Run Report
   1. **Important: Running this query for several mark years and event sites will likely cause PTAGIS to error out due to too many records. In this case, we recommend running the report multiple times, either splitting out the query by mark year or event sites. As an example, the HYC site has nearly 1 million records from mark years 2020-2022 alone. Running the report for all sites causes PTAGIS to crash, but if you run the report once with event site set to “HYC” only, then again for all other sites, the is resolved. Feel free to break up the queries by year or site to keep the number of records under the 1 million row threshold for PTAGIS operation.**
3. Export
   1. Icon: 
   2. Whole report
   3. CSV file format
   4. Remaining as default
   5. Export
4. Name the downloaded file using the following naming convention:
   1. “TagHist\_basin\_sitenames\_year.csv” e.g., “TagHist\_Lemhi\_HYC\_20-22.csv”
5. Save the downloaded .csv file within the “input” folder in PITcleanr\_lite and include the string “PTAGIS” in the file name (e.g., “PTAGIS\_lemhi\_20202022.csv”).

BioLogic™ Data

PTAGIS only provides the user with PIT-tag interrogations at sites registered with PTAGIS; however, many interrogations may also occur at sites not registered with PTAGIS e.g., at temporary, project-specific sites like the litz cords installed at stream rehabilitation projects in the Lemhi River to monitor the use of side channels and off-channel habitats, particularly by juveniles. In this case, much of those sites are installed and managed by Biomark, Inc. and so the data are available in the BioLogic™ database. Here, we provide an example of how to download observation data from BioLogic™ managed sites in the Lemhi River. **Note: BioLogic™ data is not required for PITcleanr\_lite to function.** The workflow can function with any assemblage of accepted data types.

After logging in to BioLogic™, the Site Module should show each Site Name that the user’s account has access to. We’ll need to download data from each site, individually. Note: some sites may have multiple or several readers, in which case, one download for the site will contain the observation data for each of its readers. In this example, we will download data for one site (0HR) with one reader and one site (0LL) with several readers. 0HR is an instream array located just upstream of the Henry’s Reach project in the Lemhi River designed to provide juvenile rearing habitat in multiple restored side channels. 0LL contains several readers, each a litz cord, laid across several side channels within the Henry’s Reach project.

1. Navigate to the BioLogic™ web portal (<https://data3.biomark.com/>).
2. Login
3. Click on Site Name: Henry’s Ranch Instream Array (0HR)
4. Navigate to the Download tab 
5. From: 01/01/2020
6. To: E.g., Present Day
7. Files to generate: TAG FILE
8. After the file generates, click on the filename to download.
9. Repeat steps 3-8 for the 0LL site.

Save BioLogic™ data downloads to the “input” folder, with the file name beginning with the site (e.g., 0HR, 0LL) and including the string “BIOLOGIC” (e.g., “0HR\_tagobs\_BIOLOGIC\_20220929”).

Submersible Data

Observation data obtained from submersible downloads is also compatible with PITcleanr\_lite. To incorporate these data, the input file must adhere to the following:

* Standard submersible download format (see example data in the “input/data\_examples/” folder)
* Located within the “input/” folder
* In .xlsx file format
* The file name must begin with the name of the reader/node, and include “SUB” in the file name. E.g., “nodename\_SUB\_mmddyyyy.xlsx” for a reader/node named “nodename”, or “SUB2\_mmddyyy.xlsx” for a submersible named “SUB2”.

Raw .log File Data

PITcleanr\_lite is also compatible with raw “.log” file downloads. These will be included in observation data when they adhere to the following:

* In standard .log file format (see example data in the “input/example\_data” folder)
* Located within the “input/” folder
* In textual .log format
* The file name must begin with the name of the reader/node. E.g., “nodename\_mmddyyy.log”

## Site Metadata

Site metadata information is contained in “config/site\_metadata.csv” file. This should include comprehensive metadata for all observation sites outside of the PTAGIS network.

Any additional sites or modifications to existing site metadata must be made within this file and follow the default column conventions. The following columns are required when adding or modifying an observation site:

* Reader Number: An integer corresponding with a unique Reader Name and Antenna Number for a given Site Code (e.g., the “reader” column from Biologic data)
* Reader Name: A unique identifier for the reader
* Site Code: Name of the site where the reader is located. E.g., all readers in the Henry’s Restoration reach are part of the “0LL” site.
* Antenna number
* Latitude in decimal degrees – required when generating maps
* Longitude in decimal degrees – required when generating maps
* RKM: The river kilometer location of the reader. This is required for the directionality component of the workflow. If two readers have the same RKM value, then movement between the nodes is considered “lateral” movement.
* **Optional: Restoration metadata. Required when estimating restoration use.**
  + Restoration Site (Yes/No): Indicates whether the observation site is located within or adjacent to a restoration site. A value of “Yes” indicates that a fish observed at this location is eligible to use a restoration area, not necessarily that they have entered one. In the Lemhi example, the reader “HRIA” is not directly within a restoration area, but is located at the mainstem directly above the restoration area, so a value of “Yes” is entered.
  + Restoration entrance or exit (Yes/No): Indicates whether an observation site marks the entrance or exit of a restoration use area. In the Lemhi example, some observation sites (e.g., HRSC1, HRSC2) are located on the edge of a restoration use area, so fish detections at these locations do not necessarily indicate direct use or entrance into the restoration area.
  + Restoration Interior (Yes/No): Used for locations within a restoration use site. If a fish is detected here, they are directly using a restoration area.

Tag Filtering

If the user desires to filter observations from some PIT-tags out of the results (e.g., test tags), those tag IDs can be entered into the “config/filter\_tags.csv” file. List all the tags to be filtered out in the “tag\_num” column.

Running the Scripts

1. In the PITcleanr\_lite parent folder, open the “PITcleanr\_lite” R project file using R or Rstudio.
2. Within R, open the “Workflow.r” script located within the “scripts” folder. This can be done by navigating to “File → Open File”, ctrl+o, or using the “Files” tab in the files and help pane.
3. If desired, configure arguments within the script to alter calculations and outputs (see *Script Configuration* below).
4. Highlight the entire script and click the “run” button.
   1. Packages listed under the “load packages” will automatically be downloaded from the CRAN repository if they have not been installed. These are required for the script to run.
5. This will output three .csv files with cleaned tag observation data, and a fourth .csv with the final paths for all detected fish. These are located in the “output” folder.
   1. “TagObs\_Compressed\_YY-MM-DD.csv” contains cleaned and compressed tag observation data. Each row describes the number and duration of observations for a single tag at one array.
   2. “TagObs\_Wide\_YY-MM-DD.csv” summarizes compressed tag observation data and pivots it wide. Each row is a unique tag and the associated number of detections at all observation sites.
   3. “TagObs\_Directionality\_YY-MM-DD.csv” contains tag records with associated movement direction, based on observation location RKMs.
   4. “TagObs\_FinalPaths\_YY-MM-DD.csv” provides the first and last observation for each unique PIT tag, along with an ordered list of all locations where the fish was detected. Restoration use information is also included when configured (see *Script Configuration* below).

Script Configuration

Script operation can be configured to change how data is summarized, as well as toggle some functionalities. Arguments that can be modified are listed below:

**readTagData()**

This function has two primary arguments that can be modified

* filter\_test\_tags: If there is a desire to filter out certain PIT tags, this argument should be set = TRUE, with a list of these tags in the filter\_tags spreadsheet located in the “config” folder. Else, set = FALSE
* filter\_to\_ptagis: When set =TRUE, input data from PTAGIS will act as the primary tag list, and observational data from outside of PTAGIS will only be included for tags that also exist within PTAGIS. When set =FALSE, tags are not filtered out.

**compressWrap()**

In this function that compresses and cleans PIT tag observation data, we can include any additional attributes or columns through the workflow as desired. For example, if fish lengths and weights from PTAGIS are desired to be paired with observations, then this argument will do so. First, the columns must be present within at least one of the data sources. Then, list these columns by name within the argument. For example, PTAGIS lengths and weights are named “event\_length\_mm” and “event\_weight\_g”, respectively. To include them through the workflow, include attributes\_list = c(“event\_length\_mm”, “event\_weight\_g”).

**addDirectionWrap()**

This function estimates the directionality component for compressed tag data and can generate a flowline map with observation locations if desired. Both of the arguments listed below must be configured to generate a map.

* generate\_map: If set =TRUE, then the function will attempt to create a map of all observation locations using NHD flowlines. An internet connection is required to download the appropriate flowlines.
* downstream\_site: This defines the furthest downstream site for the map, and is only used if generate\_map = TRUE. To use this argument, set downstream\_site = “site code” (e.g., “HYC”).

**finalPaths()**

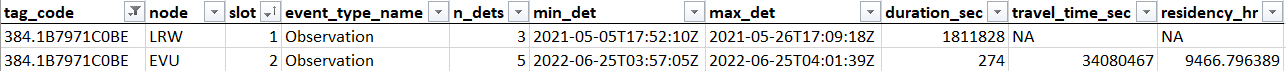
This function subsets observation data from the “addDirectionWrap()” function to the last observation of each fish, and can calculate restoration use when the correct information is supplied in the “site\_metadata” configuration file.

* calculate\_restoration\_use: Set “=TRUE” if restoration use time should be calculated, and site metadata is properly configured to do so. Else, set “=FALSE”.

Data Interpretation

“TagObs\_Compressed\_YY-MM-DD.csv”

This file contains all tag observation data. Each record is the number and duration of observations for a single tag at one node. For example, the first record below reads as: The first observation (slot = 1) for tag “384.1B7971C0BE” was at node LRW. There were 3 detections from 2022-05-05 at 17:52:10Z to 2022-05-26 at 17:09:18Z (which is 1,811,828 seconds). Then, the second node that the fish was detected at was EVU (slot = 2). The amount of time that elapsed between the final detection at LRW and first detection at EVU was 9,466 hours (or 34,080,647 seconds). As default, each detection of a unique tag at a new node is given a “slot”. In the example below, the fish was first detected at LRW, so this set of 3 detections at LRW is assigned slot #1. Next, the fish went to EVU for 5 detections, which is assigned slot #2. If the fish was detected at LRW **after** the 5 detections at EVU, then the next set of LRW detections would get a new row and have slot #3.



“TagObs\_Wide\_YY-MM-DD.csv”

This file takes the nodes and number of detections from the “compressed” file and pivots it wide. Here, each row describes the number of detections that the tag had at all nodes. In the example below, tag “3DD.003D57FB85” was detected 1 time at HYC, 2 times at LRW, 7 at SRSC1, 1 at HRSC6, and so on.



“TagObs\_Directionality\_YY-MM-DD.csv”

This file is generated by tracking observations of unique fish and movement between nodes. This is accomplished using the compressed observation data along with RKM metadata.

In the “direction” column there are 1 of 5 possible values, depending on the previous node where the fish was detected:

* **Start:** The first observation of the fish.
* **Downstream:** Between the previous observation and the current observation, the fish has moved downstream.
* **Upstream:** Between the previous observation and the current observation, the fish has moved upstream.
* **Lateral:** The previous and current fish detection was at two nodes that have the same RKM value.
* **No movement:** The previous fish detection was at the same location, but under a different “slot” number.
* **Unknown**: Fish movement could not be determined, perhaps RKM or other metadata was missing.

“TagObs\_FinalPaths\_YY-MM-DD.csv”

This file subsets data from the “directionality” file so that there is a maximum of one record per unique tag. This record describes the first and last times and locations where the fish was observed, and the “final\_path” column is an ordered list of all nodes where the fish has been detected. If configured, this file also includes information about restoration use, in the following columns:

**entered\_restoration\_site**

This column is a Yes/No value that is intended to provide information if a fish was observed to be eligible to use a restoration area. If a fish was ever observed at a restoration site as indicated by the “site\_metadata” file, then the value is “Yes”. The proportion of fish that were eligible to use a restoration area can be estimated by dividing the number of “Yes” values by the total number of unique tags.

**total\_restoration\_use\_hr**

This column is the sum total hours that a fish is assumed to have used a restoration area, calculated with the following logic:

* If a fish is repeatedly observed at a location that is listed as “Restoration interior” (see site\_metadata), then include the total time from the first to the last detection at the location.
* If a fish was observed at a “Restoration interior” location and then the next observation was either at another “Restoration interior” or “Restoration entrance or exit” location, then include the travel time between observations.
* If a fish was observed at a “Restoration entrance or exit” location and then immediately after at a “Restoration interior” location, then the travel time between locations is included.

This calculation method provides a conservative estimate for restoration use, given that some basic assumptions are met regarding fish movement and detection efficiency. An example calculation with these data is to calculate the proportion of eligible fish that used the restoration area can be calculated by dividing the number of rows with “total\_restoration\_use\_hr > 0” by “entered\_restoration\_site = Yes”.