**DATA BIOGRAPHY**

Your Name: Alana Santana, Rory Spurr

**1. RESOURCE OVERVIEW** (*what, why, who*)

1.1. Title of the dataset. *Descriptive title that usually includes data type, time period, location, and name of author, program, or institution.*

NOAA West Coast Region Permitting Data from 1994 to present, export from larger relational database.

1.2. Abstract. *One paragraph summary of the dataset in plain language. Include one sentence of broader context, followed by the dataset’s origin/purpose, and briefly expand on the elements of the title. A length of 200-250 words is a good target.*

The NOAA WCR permitting dataset is an amalgamation of permits requested through NOAA’s permits and application division. The portion of the dataset that we will be working with is an export of a larger relational database. This dataset acts more like a repository for permit requests and allows individuals from NOAA and the public to search through various research permits that are currently underway or have been issued in the past. Furthermore, the dataset allows for the permit and application division to see what types of permits are being issued/are ongoing and the various elements of those specific permits. For example, the database provides information on the permit type (4d, 10a1a Salmon, Tribal 4d, and BiOP DTA), whether the permit has been issued or not, where the research is occurring, what type of research will be ongoing, what type of ESA-listed species will the PI be researching, and whether take will occur during the period of this project. This is beneficial for NOAA’s permit and application division as it allows them to visualize active permits and mitigate the effects of take over time to assist with conservation/recovery efforts.

1.3. Purpose. *Brief description of why the data were collected, including the goals and intended outcomes (this may or may not include application to decision-making).*

The NOAA WCR permitting dataset was collected as a form of record-keeping of current and past permits. The data is automatically stored in this repository to be reaccessed by anyone searching for past or current permits.

1.4. Contacts. *Provide contact info for the people who managed the project, collected the data, generated the dataset, and/or managed the data. Contact information should include name, organization, role in the project, email and/or phone.*

Agency: NOAA

Division: Permits and Application Division

Office Phone: (301) 427-8401

1.5. Sponsors. *Who or what organization sponsored collection of the data (e.g., NOAA as a part of a mandated monitoring program)? Who funded collection of the data (if applicable)?*

NOAA is the federal agency overseeing the Protected Resource Division through the Department of Commerce. Under the Protected Resource Division, the Regional Permit Coordinator manages and allocates scientific research and take permits. All data regarding permits applications, issuance, withdrawal, or denial is recorded and stored within a database. NOAA is the sponsoring organization for the collection of this data. As a federally mandated program, the collection of data is federally funded.

1.6. Citation for the dataset. *Use the citation format below and include a link to the data source.*

Author, A.A. (YEAR). Name of data set. [Data set]. doi:XXXXX Available from: URL

NOAA (1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022), WCRpermitBiOp\_allregns\_all\_years\_\_7Jan2022. [Data set]. Available from:...

1.7. Keywords. *Include 3-5 keywords for the dataset. Think of these as search terms that someone might use to find the data.*

Permits, West Coast Region, Section 4(d), Section 10, NOAA.

**2. TEMPORAL AND SPATIAL EXTENTS** (*when, where*)

2.1. Temporal extent. *The entire time range (specific years) for observations included in the dataset.*

The dataset began collecting permitting data in April of 1994 and continues to collect and document permitting data to this day. However, the data made available for this project extends to January of 2022.

2.2. Temporal resolution. *The frequency at which data are collected or acquired. Be as specific as possible. Note whether measurements were taken at regular intervals or irregularly.*

Data was collected annually. For each permit inquiry, the date was logged for the request, issuance, withdrawal, and/or denial.

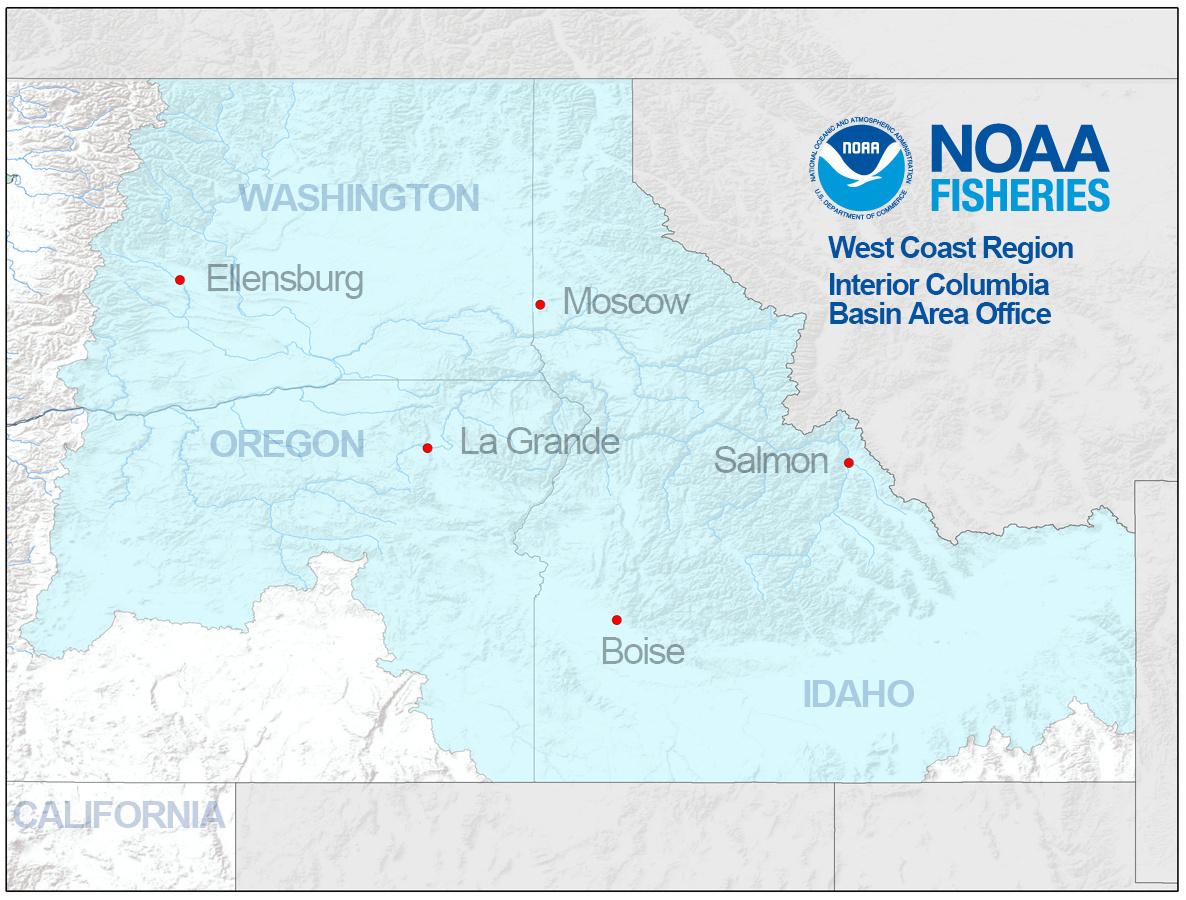
2.3. Spatial extent. *Boundaries of the data set. If possible, include both the (a) geographic description, and (b) coordinates describing north, south, east, and west boundaries of the area included in the data. You do not need to include granular geospatial data (e.g., survey tracks, buoy locations).*

The dataset collects data regarding the scientific research permits requested across the west coast region. This region expands from California, Oregon, Washington, and Idaho. Below is a map regarding the extent of WCR and where the data on permits was collected (Fig. 1 - 4).

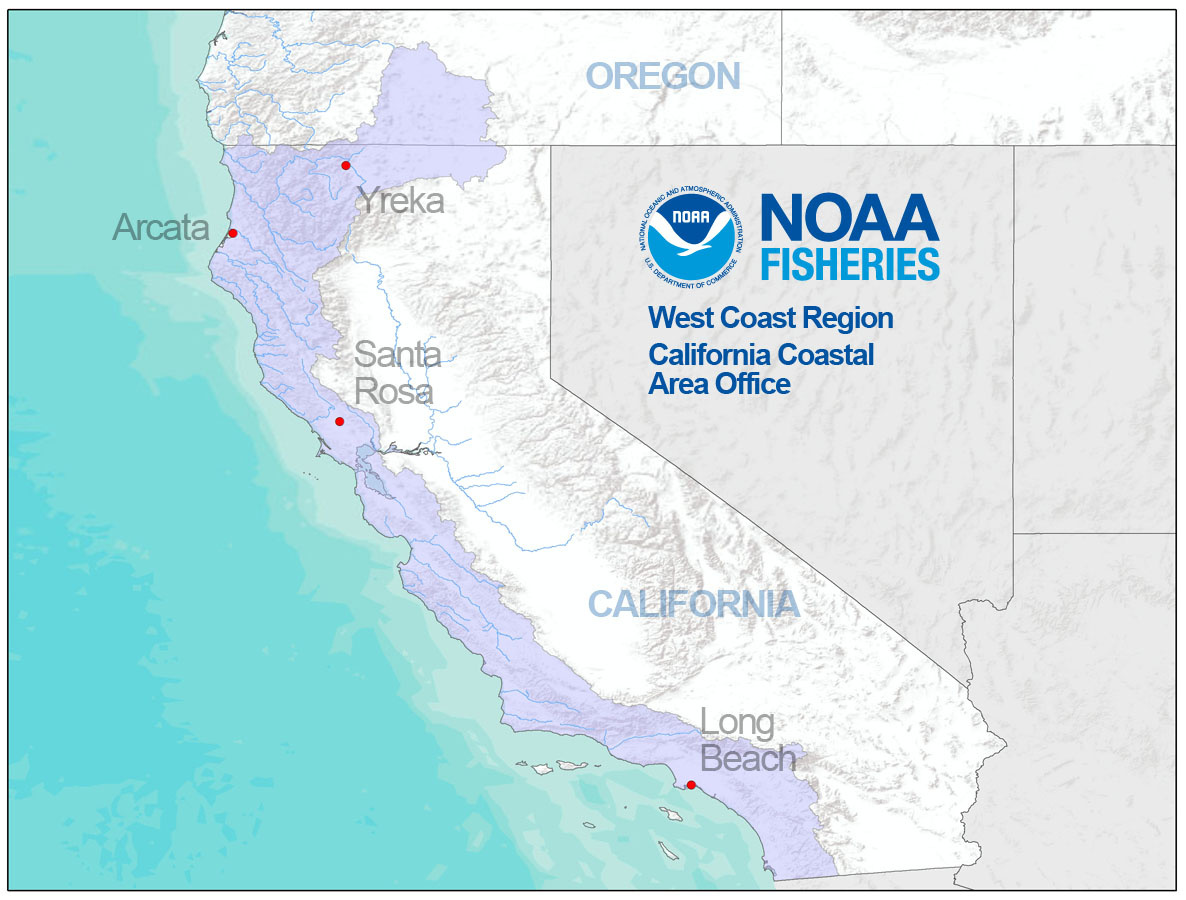
2.4. Spatial resolution. Specificity with which spatial data are recorded. For example, are locations of measurements recorded using GPS? Locality (place) names? Were measurements collected on a uniform grid and if so, at what spatial scale?



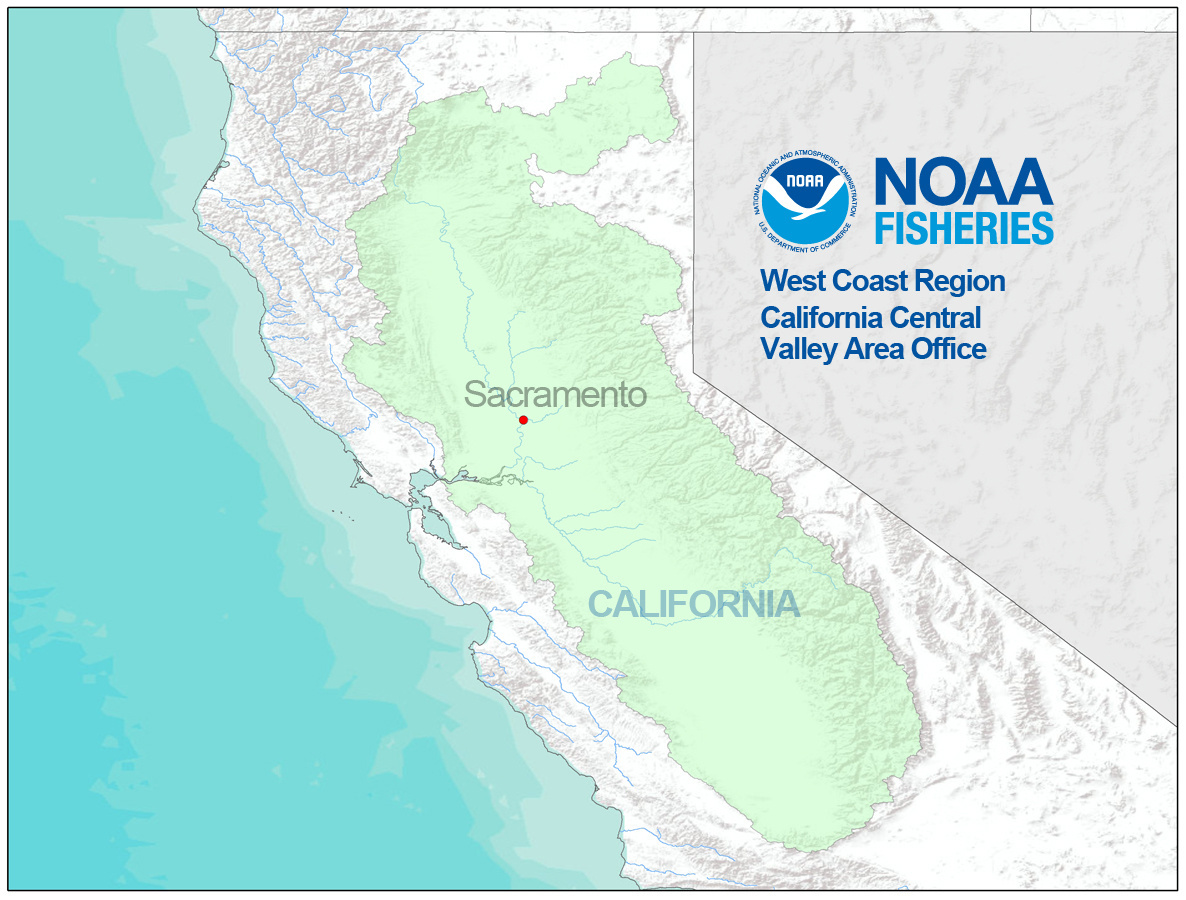
*Fig. 1 NOAA Fisheries Oregon-Washington Coastal Area.*



*Fig. 2 Washington, Oregon, Idaho Interior Columbia Basin Area.*



*Fig. 3 California Coastal Area.*



*Fig. 4 California Central Valley Area.*

**3. RESOURCE CONTENT** (*what*)

3.1. Digital context. *Names of data file(s), names of tables within data file(s), file format(s), and date the data were last modified. If you have multiple data files, describe any relationships among them (e.g., queried database tables saved as separate files and linked through an identifier?). For each data file and table within a data file, include a brief (1 sentence) narrative description of the contents.*

Data is downloaded as a “.csv” file from the APPS interface.

> Authorized permit data

> Reporting data

> Abundance data

3.2. Data components and data table attributes. *This section details the contents of each data table and/or data file and might be most effectively organized as a table (but it’s up to you). For each data file/data table, provide the names, definitions, and units of the attributes of any data in tabular format (e.g., column headers in a CSV file). Depending on the nature of the data, this could include: parameter name, measurement units, instrument type, precision, accuracy, taxonomic details, definitions of codes used, and any other important information for an analyst (e.g., quality review notes, missing values). Indicate whether data are raw values (not modified in any way after collection), processed values (corrected or calibrated), or derived values (an index or summarized value calculated based on other data).*

***WCRpermitBiOp\_allregns\_all\_years\_\_7Jan2022.csv***

|  |  |
| --- | --- |
| **Data Component** | **Attributes** |
| FileNumber | File number automatically assigned by the APPS system; used in correspondence about the application |
| ResultCode | Permit type or authority |
| AccountStatus | Indicates the current status of the Application |
| PermitStatus | Indicates the status of the permit that was applied for. (ex “Issued”, “In Review”) |
| DateIssued | Month, day, and year of the permit issuance. |
| DateExpired | Month, day, and year of the permit expiry. |
| AnnualTimeStart | Window when applicants begin to send in reports. |
| AnnualTimeEnd | Window when applicants cease to send in reports. |
| ProjectTitle | Name of the research project being conducted. |
| FirstName | First name of the principal investigator. |
| LastName | Last name of the principal investigator. |
| Organization | Name of the organization sponsoring the research. |
| HUCNumber | Displays the Hydrologic Unit Code (HUC) where that take is expected to occur. |
| WaterbodyName | Name of the major waterbody where the take is expected to occur |
| BasinName | Name of the water basin where the take is expected to occur |
| StreamName | Name or names of the streams involved where the take is expected to occur. Can be general (“Lower Columbia River and its tributaries”) or specific (“Battle Creek”). |
| LocationDescription | Specific description giving location of sampling site/trap etc. where take is expected to occur. Can be general if take is expected to occur all along the stream/waterbody. |
| CommonName | Common name of the species on which take is expected to occur. |
| Population | Specific population on which take is expected to occur.. Mainly defined by the river or geographic area. |
| Run | Specific run where take is expected to occur. Usually described as the time of year (Spring, Summer, Winter) when the salmon or other anadromous fish return to spawn. |
| Production | Indicates whether the fish are for hatchery production or wild. |
| LifeStage | Life stage of the species/individual the project expects to take. |
| Sex | Sex of the species/individual the project expects to take. |
| TakeAction | Description of the kind of take that is expected to occur (ex. “Broodstock collection”, “Capture”, “Handle”) |
| CaptureMethod | Gear used to capture species. |
| ExpTake | Number of individuals the project expects to take as a result of research. |
| IndMort | Number of incidental mortality as a result of research. |

***WCRPermitBiOp\_Pass report data 4d and S10\_22March22.csv***

|  |  |
| --- | --- |
| **Data Component** | **Attributes** |
| FileNumber | File number automatically assigned by the APPS system; used in correspondence about the application |
| ***ReportID*** | ***Five-digit unique code for each active project that reported take and mortality.*** |
| ResultCode | Permit type or authority |
| AnnualTimeStart | Window when applicants begin to send in reports. |
| AnnualTimeEnd | Window when applicants cease to send in reports. |
| ***DateReportPeriodEnd*** | ***Date in which the report on actual take and mortality is submitted/ends. Reported in YYYY - MM - DD format.*** |
| ***RegionID*** | ***State abbreviations for where permits are occurring. I.e. WA, CA, OR, ID.*** |
| FirstName | First name of the principal investigator. |
| LastName | Last name of the principal investigator. |
| Organization | Name of the organization sponsoring the research. |
| ProjectTitle | Name of the research project being conducted. |
| BasinName | Name of the water basin where the take is expected to occur |
| WaterbodyName | Name of the major waterbody where the take is expected to occur |
| StreamName | Name or names of the streams involved where the take is expected to occur. Can be general (“Lower Columbia River and its tributaries”) or specific (“Battle Creek”). |
| CommonName | Common name of the species on which take is expected to occur. |
| Population | Specific population on which take is expected to occur.. Mainly defined by the river or geographic area. |
| Production | Indicates whether the fish are for hatchery production or wild. |
| LifeStage | Life stage of the species/individual the project expects to take. |
| TakeAction | Description of the kind of take that is expected to occur (ex. “Broodstock collection”, “Capture”, “Handle”) |
| CaptureMethod | Gear used to capture species. |
| ExpTake | Predicted number of individuals the project expects to take as a result of research. |
| ***ActTake*** | ***Actual number of individuals a project takes that occurred as a result of research.*** |
| IndMort | Predicted number of incidental mortality as a result of research. |
| ***ActMort*** | ***Actual number of lethal take that occurred as a result of research.*** |

*Note - data components that are italicized and bolded are unique to WCRPermitBiOp\_Pass report data 4d and S10\_22March22.csv data file.*

**4. METHODS** (*how*)

4.1. Lineage statement. *Provide a summary of the methods used to collect the data. Ideally, this is a brief narrative description that includes citations to standard operating procedures, field manuals, or other references.*

Data is collected and recorded when applications are submitted through the Authorizations and Permits for Protected Species (APPS) online application system. When researchers, hatchery managers, or other professionals wish to complete research on an ESA-listed species or plan on encountering ESA-listed species, they submit their applications through the APPS online portal. Data is then automatically recorded on the type of species the research will be conducted on, the type of take occurring, and where the take will be occurring.

4.2. Process steps. *The general process steps that occurred between data collection and its current form (brief narrative description or bulleted list). Depending on the dataset, processing might include digitization, removing or identifying outliers via computer scripts, file processing, data summarization, or data transformations. This does not need to be exhaustive, but should include information that would be important for an analyst to be aware of when they are using the data for research. Include relevant citations.*

1. Created Reading and Filtering Computer Script
   1. Filtering out expired and non-issued permits.
      1. filter(PermitStatus == "Issued")
      2. filter(DateIssued >"2012-01-01")
      3. filter(DateExpired >= Sys.Date())
   2. Filtering for specific and relevant permit types - NMFS 10a1A Salmon, 4d, NMFS BiOp DTA, Tribal 4d.
      1. filter(ResultCode == c("NMFS 10a1A Salmon","4d", "NMFS BiOp DTA", "Tribal 4d"))
   3. Recoded all smolts, fries, larva to “juveniles” and subadults to “adults”.
      1. mutate(LifeStage = recode(LifeStage,

"Smolt" = "Juvenile",

"Fry" = "Juvenile",

"Larvae" = "Juvenile",

"Subadult" = "Adult"))

* 1. Recoded HUC numbers based on reclassification by government agencies
     1. mutate(HUCNumber = recode(HUCNumber,

`18020103` = 18020156,

`18020109` = 18020163,

`18020112` = 18020154,

`18020118` = 18020154,

`18040005` = 18040012,

`18060001` = 18060015,

`18060012` = 18060006))

* 1. Recoded and categorized species by their distinct population segments or evolutionary significant unit.
     1. mutate(Species = paste(Population, CommonName, sep = " "))
  2. Recoded and categorized species by production - Natural, Listed Hatchery, Unlisted Hatchery.
     1. mutate(Prod = recode(Production,

"Natural" = "Natural",

"Listed Hatchery" = "Listed Hatchery",

"Listed Hatchery, Clipped and Intact" = "Listed Hatchery",

"Listed Hatchery Intact Adipose" = "Listed Hatchery",

"Listed Hatchery Adipose Clip" = "Listed Hatchery",

"Unlisted Hatchery" = "Unlisted Hatchery"))

* 1. Filtered out non-invasive, non-lethal take actions (i.e. Observe/Harass and Sampling dead animals)
     1. filter(TakeAction == c("Capture/Handle/Release Animal",

"Capture/Mark, Tag, Sample Tissue/Release Live Animal",

"Intentional (Directed) Mortality",

"Unknown", # what is it

"Unintentional mortality",

"Collect, Sample, and Transport Live Animal",

"Captive animals (research, enhancement, public display)",

"Removal from wild (permanent)",

"Incidental take",

"Import/export/receive only",

"Recondition and release",

"Juvenile Releases", #should this be included?

"Broodstock collection",

"Rescue/Salvage" ))

1. Developing new fields within data files.
   1. Created a new field to display ESU or DPS under “Species”
      1. mutate(Species = paste(Population, CommonName, sep = " "))
   2. Created a new field to recode and condense production type under “Prod”
      1. mutate(Prod = recode(Production,

"Natural" = "Natural",

"Listed Hatchery" = "Listed Hatchery",

"Listed Hatchery, Clipped and Intact" = "Listed Hatchery",

"Listed Hatchery Intact Adipose" = "Listed Hatchery",

"Listed Hatchery Adipose Clip" = "Listed Hatchery",

"Unlisted Hatchery" = "Unlisted Hatchery"))

* + 1. filter(Prod != "Unlisted Hatchery")
  1. Created a new field to report total mortality that occurred.
     1. create\_totalmorts() - function included in *NMFSResPermits* package
  2. Created a new field to report total authorized take under “AuthTake.
     1. mutate(AuthTake = ExpTake + IndMort)
  3. Created a field (called “Location”) that describes the location name, based on hierarchical sort of the 4 location-based columns in the raw data. The hierarchy is WaterbodyName > BasinName > StreamName > LocationDescription, where the highest-level data that is not NA was used. For example, if a permit had an NA value for WaterbodyName, and a non-NA value for BasinName, the BasinName value would be used to describe that permits location in the new “Location” column.
  4. Created a field (called “SW\_FW”) that designates whether the species is being encountered in saltwater, freshwater or both. This was done using a keyword sorting algorithm.
     1. Saltwater keywords: "Sound", "Bay", "Ocean", "Strait", "Admiralty", "Whidbey", "Canal", "shelf", "Cape", "estuary", "Estuary”, “lagoon", "Lagoon", "Delta", "estuarine"
     2. Freshwater keywords: "Lake", "stream", "freshwaters", "Columbia"
     3. Then a function was created to sort the location names (created in “e”) based on the keywords found in the location name (e.g., a location name with freshwater keyword(s) would be assigned “FW” for freshwater. This function also checked for keywords found in both the saltwater and freshwater word banks and would assign both freshwater and saltwater to permits in these cases.
     4. It should be noted that this sorting function was not perfect, and due to this and ambiguous location names some of the permit’s freshwater or saltwater designations may not be entirely accurate.

4.3. Quality assurance and quality control. *Note anything the data creators did to ensure the completeness and accuracy of their dataset (e.g., instrument calibrations, automated procedures, manual/visual tests for outliers).*

Data has not been modified or changed, but filters pre-applied within permit application regarding species type and gear type. No true QC process other than the script written to collect the biological opinion has been QC’d prior to using it to access queries. Iterative QC and manual clean up.

4.4. Data completeness and constraints. *Were any data excluded from the dataset? If so, why? What are known cautions or problems, such as sampling bias? Are there ways the data should not be used (according to the data creators)?*

**Assumptions/Limitations**

1. *Data inputs* - As the database began collecting data in 1994, the permit application process has changed greatly over time. Therefore, there are areas that lack consistency and information. For example, early permit applications were not required to put in specific information regarding location, HUC number, act, and so on. As we are aware of this, we will either provide a quality check to fill in gaps based on provided information or will not include these values/permits into our project as they are not complete/missing necessary information. Additionally, in the StreamName section, NOAA allowed permit applicants to provide the name of the waterbody they planned to work with. As this was a fill in process rather than a drop down with provided names, information on the same waterbody can be unnecessarily duplicated. Therefore, we will need to provide a quality check there as well to ensure we are not duplicating waterbody names and provide more consistency across the nomenclature.
2. *Non-reported take -* Under the data file *WCRPermitBiOp\_Pass report data 4d and S10\_22March22.csv,* researchers are asked to report the actual take and mortality that occurred to ESA-listed species during the duration of their study. However, some organizations/projects can opt out or neglect to report on the actual take/mortality. As a result, these projects cannot contribute to our analyses due to lack of information and thus may be excluded from our overall final product. Thus, a limitation of this unreported data is we are missing total mortality data and therefore complicates further abundance data analysis.
3. *Changed HUC numbers -* Over time, HUC 8 codes have been rearranged and their boundaries redrawn. Many permits use old or outdated HUC 8 codes, causing issues when trying to map our permit data using the Watershed Boundary Dataset (USGS et al. 2022). Therefore, HUC 8 codes had to be updated to reflect any changes to their boundaries. Decisions were made using an unpublished document that summarizes HUC 8 code changes up until 2018 (Hanson et al. 2018). Some of these changes ran on assumptions using other fields (such as WaterbodyName or LocationDescription) and are detailed below:
   1. # 18020103 = 18020156 # very certain
   2. # 18020109 = 18020163 # very certain
   3. # 18020112 = 18020154 # very certain based on location descriptions
   4. # 18020118 = 18020154 # very certain based on location descriptions
   5. # 18040005 = 18040012 # very certain based on location descriptions
   6. # 18060001 = 18060015 # split between 18050006 as well, arbitrarily picked
   7. # 18060012 = 18060006 # chose this over Monterey Bay as population is South-Central Cal Coast

Citations:

U.S. Geological Survey (USGS), U.S. Department of Agriculture – Natural Resource Conservation Service (NRCS), U.S. Environmental Protection Agency (EPA) (2022). USGS National Watershed Boundary Dataset in FileGDB 10.1 format (published 20220526). Accessed May 15, 2022 at URL<https://prd-tnm.s3.amazonaws.com/index.html?prefix=StagedProducts/Hydrography/WBD/National/GDB/>

Hanson, K., Daw, S., Davenport, L., Jones, K., Niknami, L., & Buto, S. (2018). Criteria for Legacy Name and Code Changes. [Unpublished]