f i s h R m a n

v 1.1.0

Handbook

Index

1. Introduction *(from fishRman: A Shiny R Dashboard improving Global Fishing Watch data availability)*

One of the burdens of fisheries scientists, researchers, and enthusiasts, is the scarcity or lack of consistent, extensive data on the subject. When such data do exist, they are often only available:

- To universities or other research institutions;
- Through bureaucratic ordeals;
- For a fee.

This issue has been tackled by Global Fishing Watch, an independent, international non-profit organization promoting ocean sustainability through greater transparency, visualizing, tracking and sharing data about global fishing activity for free.

While the datasets are indeed publicly available, they are also rather large and quite difficult to manage, since they require proficiency in coding. In fact, at present, the most notable reading material instructing on the use of the datasets targets an audience who is proficient in the languages R, Python, JavaScript, or SQL to download, filter, summarise, and visualise the data.

Overcoming these two major barriers, fishRman sets itself as a completely web-based, one-stop solution that provides an intuitive user interface for querying, downloading, and analyzing Global Fishing Watch data on fishing effort.

This document aims to explain how to use the software to solve real-world analysis problems.

1.1 Global Fishing Watch Datasets of AIS-based Fishing Effort and Vessel Presence (from https://globalfishingwatch.org/data-download/datasets/public-fishing-effort)

This dataset contains the Global Fishing Watch AIS-based fishing effort and vessel presence datasets. Data is based on fishing detections of >114,000 unique AIS devices on fishing vessels, of which ~70,000 are active each year. Fishing vessels are identified via a neural network classifier, vessel registry databases, and manual review by GFW and regional experts. Data are binned into grid cells 0.01 (or 0.1) degrees on a side and measured in units of hours. The time is calculated by assigning an amount of time to each AIS detection (which is the time to the previous position), and then summing all positions in each grid cell.

Data are available in two formats:

1. **Fishing effort by flag state and gear type at 100th degree resolution**
2. **Fishing effort by MMSI at 10th degree resolution**

Versions

The fishing effort data is updated periodically to include more recent data and as improvements are made to our models and input data. Data from different major/minor versions should not be combined, as each version uses a different set of inputs.

Current version

- **2.0**: Fishing effort data for 2012-2020 using our latest AIS algorithms, neural network models, and vessel registry database.

For additional information about these results, see the associated journal article: D.A. Kroodsma, J. Mayorga, T. Hochberg, N.A. Miller, K. Boerder, F. Ferretti, A. Wilson, B. Bergman, T.D. White, B.A. Block, P. Woods, B. Sullivan, C. Costello, and B. Worm. "Tracking the global footprint of fisheries." Science 361.6378 (2018).

1.2 Global Fishing Watch Data in fishRman

The mission of **fishRman** is to make analyzing GFW data easy, and it does so via a user-friendly interface. This means that users are met by a simplified terminology in the front-end, which hides a less friendly one in the back-end. Although it is advised to make yourself acquainted with the data first, via the already mentioned sources, it is worth explaining some nuances in the present work.

Table 1: Front-end to back-end correlation, with explanation of the data in the column. The fields superscripted with '100' are only available for the GFW table gridded at 100th degree of resolution. The fields superscripted with '10' are only available for the GFW table gridded at 10th degree of resolution.

Front-end	Back-end	Explanation
Date	date	The date (yyyy-mm-dd) when the data got recorded.
Latitude	cell_ll_lat	The latitude at which the data got recorded, binned (rounded) to the decimal place of reference.
Longitude	cell_ll_lon	The longitude at which the data got recorded, binned (rounded) to the decimal place of reference.
Vessel hours	hours	The amount of hours spent in activities different from the act of fishing, in that particular grid cell, on that particular date.
Fishing hours	fishing_hours	The amount of hours spent fishing, in that particular grid cell, on that particular date.
Flag^{100}	flag	The flag state of the boat streaming the data.
Geartype^{100}	geartype	The main geartype (fishing gear) of the boat streaming the data.
MMSI present^{100}	mmsi_present	The number of different MMSI numbers, in that particular grid cell, on that particular date. The MMSI number is a vessel's unique ID.
MMSI^{10}	mmsi	The list of MMSI numbers recorded, in that particular grid cell, on that particular date.
AIS data at 10th degree	See explanation	The full Google BigQuery address of this table is global-fishing-watch.gfw_public_data.fishing_effort_v2
AIS data at 100th degree	See explanation	The full Google BigQuery address of this table is global-fishing-watch.gfw_public_data.fishing_effort_byvessel_v2

2. Query

Querying data is a simple, straightforward process. Just access the application at this URL: https://shyentist.shinyapps.io/fish-r-man/.

Alternatively, experienced users with demanding needs can download the code from the GitHub repository and 'comment out' the lines of code marked accordingly. This allows them to access their own Google account, and ignore the limits set by the application in terms of maximum loading time, maximum size of queried and loaded data. After Google Authentication, the application is launched.

On the left side of the landing page there is a sidebar. Starting from the top, and referring to the picture below, it features a select list (1), with which users can choose the table to query. There are two tables: AIS data gridded at 0.01 degrees, by gear type and flag state, and at 0.1 degrees, by MMSI (a vessel's unique ID).

Following, a checkbox group (2), which updates in accordance with the selected table, allows users to choose the fields by which they would like to query the table. Ticking a box enables the input box of the related field (3), to be filled by the user.

Note here that, for range inputs such as dates, latitude, longitude, vessel hours, fishing hours, and MMSI, the minimum is included in the range, while the maximum is not. For instance, querying by date range '2012-01-01 to 2013-01-01' will return data for the entire 2012 year, and will not include the first day of 2013.

As the fields are filled by the user, the SQL Query is constructed and presented to the user (4).

The ‘billing project’ field (5) is to be left empty from all users who have not edited the code in order to authenticate with their own Google account. Those who have will instead have to write the name of their project in Google BigQuery to which they wish to bill the data usage.

Once all choices have been made and fields have been filled, users can click the filter button (6). This initiates the query and locks the application until one of two paths is completed:

1. Unsuccessful: Data exceeds 1 million rows, the query is aborted, and users are invited to be more specific in their search.
2. Successful: Data is ready for (7) download as a csv (Comma Separated Value) or gpkg (GeoPackage) file, and the (8) output table has been filled.

The time spent to filter, download, and render data is directly proportional to the amount of queried data, so it is advised to be specific when choosing ranges, especially date, latitude, and longitude.

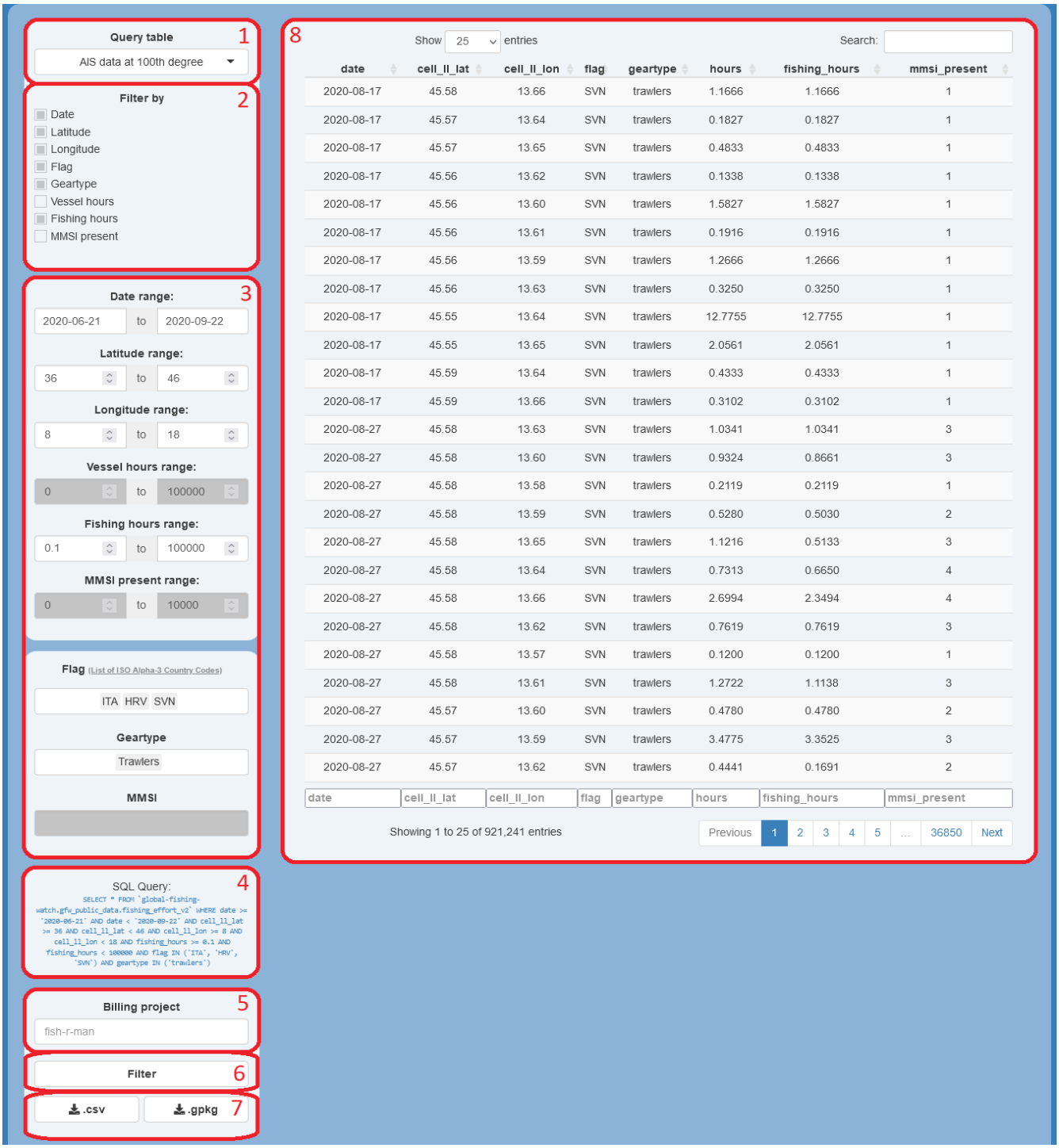

Figure 1: 'Query' tab after a query is run, with inputs and outputs marked and numbered in red.

3. Analysis

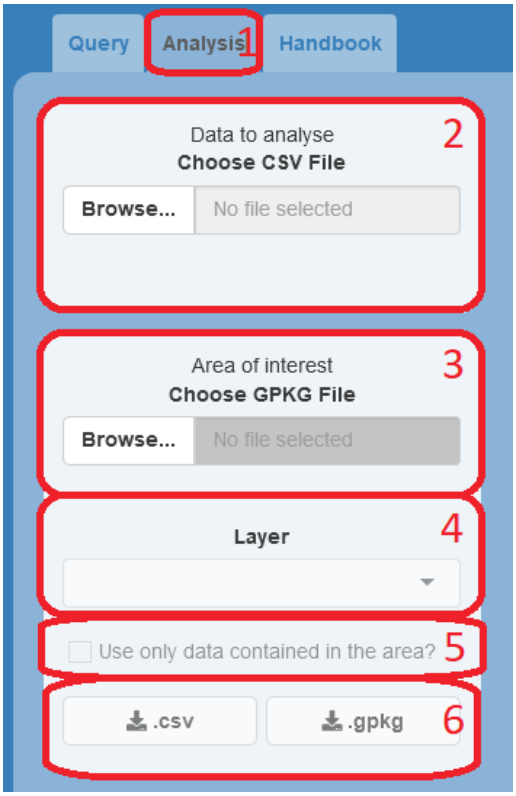

Figure 2: 'Analysis' tab, 'Data to analyse' and 'Area of interest' sidebars, with inputs and outputs marked and numbered in red.

Users can analyse their data by clicking on the 'Analysis' tab (1) at the top-left corner of the screen. Here, users can upload csv files (2) downloaded as described in paragraph '2. Query'. Maximum file size is 150 MB. The uploaded dataframe is treated the same as freshly queried data, so both kinds of data will be referred to as 'queried data'.

When queried data are present, users can upload the gpkg file (3) of an area where they wish to focus their analyses. The area can have any shape, so users are not limited to the squares set by latitude and longitude, but the chosen layer (4) must be a POLYGON or MULTIPOLYGON. This means that uploading the perimeter of the area of interest, which would be a LINE or MULTILINE object, will not work. Upon checking the 'clip' box (5), the software will subset the data to only retain the points falling within the area of interest.

If no point is found, a message is shown to the users, and the box is automatically unchecked. If at least one point is found, no message is returned, and users will be able to perform the analyses described in the next paragraph on 'clipped data', as we will call the clipped subset of queried data. Clipped data are also made available for download in csv or gpkg format (6).

Note here that, if the 'clip' box is checked, clipped data will always be analysed instead of the entire queried data, with the same methods described for queried data later on in this work.

3.1 Descriptive

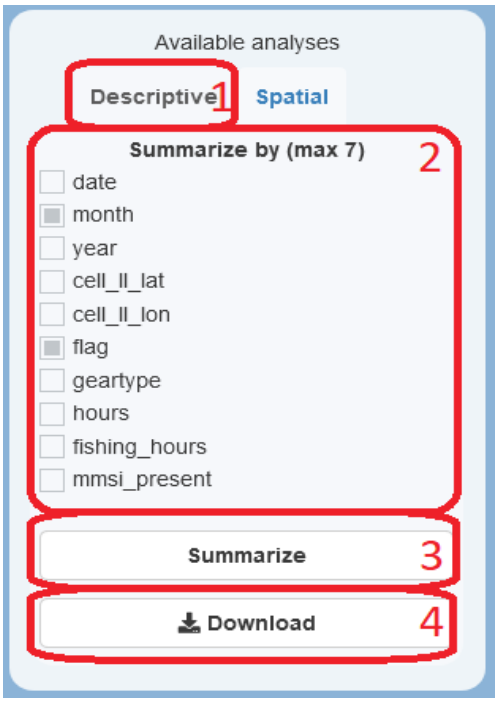

Figure 3: 'Descriptive' tab, with inputs and outputs marked and numbered in red.

The 'Available analyses' sidebar in the 'Analysis' tab hosts the 'Descriptive' tab (1), which allows users to choose any field (2), even multiple fields at once, by which to summarise the queried data (3). The resulting summary is rendered into a table showing measures of central tendency (where the distribution is centered i.e. mean, median, and mode) and measures of variability (the spread of the distribution i.e. minimum and maximum variables). It is then possible to (4) Download the summary as a csv file.

Note here that the measures are calculated per each combination of the factors present in the chosen fields.

For instance, one could summarise by flag, obtaining a summary per each flag (i.e. 'ITA', 'FRA'), or by geartype, obtaining a summary per each geartype (i.e. 'trawlers', 'longliners'). If one were to summarise by both fields, the result would be a summary per each combination flag-geartype:

1. ITA – trawlers,
2. ITA – longliners,
3. FRA – trawlers,
4. FRA – longliners.

If no data matches a particular combination, no summary is returned for that combination.

3.2 Spatial

Opening the 'Spatial' tab (1) of the ‘Available analyses’ sidebar, if queried data is present in the work environment, the 'Visualize' button (2) can be clicked, showing a plot of the data (3), and a menu on the right side (4). Within this menu, users can choose additional global layers to plot (5), the field to plot (6), the percentage for the top n-th percentile to include (7), the latitude and longitude to zoom in/out (8), and the resolution to use (9). These recalculations are performed when the 'Re-Visualize' button (10) is clicked. The plot can be downloaded as a png (Portable Network Graphics) file (11).

Note here that the default, and minimum, value for the resolution depends on the table that was originally queried, and defines the length of the sides of each cell of the spatial grid, in degrees of latitude and longitude. Data spreading over larger areas need a higher value for this field for the plot to be clear. Increasing this value leads to a new aggregation of data to match the new resolution. **Also note that** the ‘top n-th percentile’ is exactly what the name suggests: ‘the top X%’. If users select 90% for ‘Total fishing hours’, then the entries are sorted in decreasing order for that column, and summed one by one until 90% of the sum of all entries is reached (or as soon as it is surpassed).

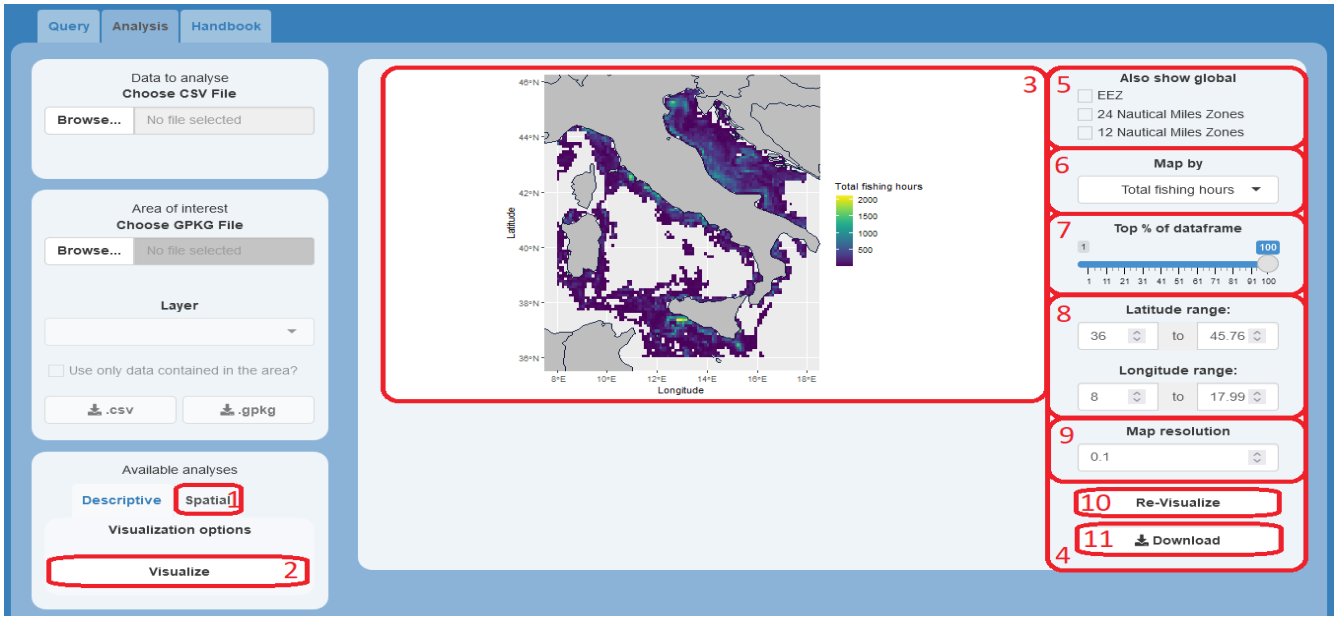

Figure 4: 'Spatial' tab, with inputs and outputs marked and numbered in red.

4. Examples

In the second half of 2020, multiple news outlets reported on the Chinese fleet exerting an intense fishing pressure barely off the borders of the Marine Protected Area of the Galápagos islands. The Guardian, in particular, reported some interesting numbers from a work by Oceana. As an example, the reader will be instructed on how to research this topic and fact-check this piece of news.

Note here that Oceana used advanced techniques to account for fishing activity exerted by vessels with AIS transmitters turned off, while **fishRman** works solely on 'visible' activity. The results will thus slightly differ, while also being equally correct in their context and with their premises.

This guide will fact-check 3 Points of the report by Oceana:

1. Chinese vessels account for 99% of the fishing effort;
2. The main target of the Chinese fleet was squid;
3. Spatial distribution of the fleet (visual comparison of the maps produced).

First of all, research the coordinates of the area to investigate. Data from a broader area than the islands alone should be queried in the 'Query' tab, adjusting the query through a process of trial and error until the reader is satisfied with the spatial extent of the analysis. The table to query is the one with AIS data gridded at 0.01 degrees, since it retains information about the flag state of the vessels.

As indicative values, this work suggests a latitude between -9 and 5, and a longitude between -100 and -80. The time period is stated in the sources as from 2020-07-13 to 2020-08-14. This suggested query is more formally written as this SQL Query:

```
SELECT * FROM `global-fishing-watch.gfw_public_data.fishing_effort_v2` WHERE date >= '2020-07-13' AND date < '2020-08-14' AND cell_ll_lat >= -9 AND cell_ll_lat < 5 AND cell_ll_lon >= -100 AND cell_ll_lon < -80
```

The reader is encouraged to find the interval of coordinates that best fits the analysis, bearing in mind that their results will diverge from the ones below if they do so.

When **fishRman** is done fishing the data, it is time to leave the 'Query' tab in favor of the 'Analysis' tab. Here, the focus is on the 'Descriptive' tab, in the 'Available analyses' section of the sidebar. Since the query encompasses multiple months, a way to calculate the total fishing hours exerted is to aggregate by 'year' by checking the relative checkbox and hitting the 'Summarize' button. The analysis returns a 'Total fishing' value of 79520.49 hours.

Focusing now on the Chinese fleet alone, the reader is invited to uncheck the abovementioned 'year' checkbox, checking the 'flag' checkbox instead (leaving the 'year' box checked returns the same result, it is only unchecked to focus the attention of the reader onto the new topic). This will calculate the total fishing hours exerted by each nation represented in the queried data. The analysis returns a 'Total fishing' value for CHN (Chinese) vessels of 70658.1493 hours.

In order to have the fishing effort of the Chinese fleet as a percentage of the total, one just needs to use the simple equation:

(Chinese fishing hours / Total fishing hours) * 100 = (70658.1493 / 79520.49) * 100 = 88.855274%

That amount is in line with the idea of one fleet dominating the fishing arena, and is accurate and correct with the right premises, even though it is still far from the 99% reported by Oceana.

Readers with minimal knowledge of GIS software, of which QGIS is the most renowned Open-Source representative, might take this analysis a step further. In particular, they might download from marineregions the shapefile for the Exclusive Economic Zone (EEZ) of the Galápagos islands, the borders of which correspond to the Marine Protected Area currently being analysed, create a buffer area outside the borders, and export the layer as a GeoPackage. This helps ignoring the fishing effort exerted *within* the EEZ or *too far* from it and focusing on the one exerted *immediately out* of it. For this guide, a buffer of 2 degrees of latitude-longitude was created. The reader is encouraged to find the buffer distance (width) that best fits the analysis, bearing in mind that their results will diverge from the ones below if they do so.

To use this buffer area, the reader must upload the GeoPackage file for the buffer as the 'Area of interest', choose the correct layer, and check the 'Use only data contained in the area?' checkbox. Now, they must repeat the 'Descriptive' analyses described above, one for 'year' and one for 'flag', and use the same equation for calculating the percentage. This will bear the following results:

Table 2: Fishing hours by all fleets combined and by the Chinese fleet alone in comparison.

Chinese fishing hours	Total fishing hours	Chinese/Total percentage
70615.9944	73101.13	96.600414%

As expected, focusing the analyses on the area immediately outside the MPA returns results that are comparable to those published by Oceana.

Point 1 has successfully been checked.

The next point focuses on the Chinese fleet, thus needing the reader to query the data again with the same parameters, this time adding a ‘flag’ filter for Chinese (CHN) vessels.

Note here that a researcher must be consistent in their analysis, or explain why they chose not to. The reader must thus ignore, or repeat, the optional buffer area step accordingly.

The reader must then move to the ‘Descriptive’ tab, in the ‘Available analyses’ section of the sidebar. Here, they must check the ‘flag’ checkbox, and click the ‘Summarize’ button. The resulting tables have been merged and the numbers rounded to the first decimal for clarity.

Table 3: Fishing hours and vessel hours for each geartype employed by the Chinese fleet. The numbers are reported before and after filtering by the buffer area of interest, and the difference between the two.

	Without buffer		With buffer		Difference	
Geartype	**Total fishing**	**Total hours**	**Total fishing**	**Total hours**	**Total fishing**	**Total hours**
Squid jigger	65189.2	196004.7	65167.0	195648.2	22.2	356.5
Fishing	5466.4	16006.1	5449.0	15966.1	17.4	40.0
Drifting longlines	2.6	124.6	0.0	13.2	2.6	111.4
All geartypes	70658.2	212135.4	70616.0	211627.5	42.2	507.9
Squid jiggers %	**92.3%**	**92.4%**	**92.3%**	**92.4%**	**0.0%**	**0.0%**

It is clear there is little difference between using or not the buffer area in terms of sheer results, since they all match, and confirm what was expected to be found in Point 2: Squid jiggers, whose main target species is squid, exerted the majority of the fishing effort for the Chinese fleet in the area.

For the last point, the reader must then move to the ‘Spatial’ tab, in the ‘Available analyses’ section of the sidebar. Here, they must click ‘Visualize’. This guide uses the following parameters:

- Also show global: EEZ
- Map by: Total fishing hours
- Top % of dataframe: 99%
- -98 < Longitude < -83
- -7 < Latitude < 5
- Map resolution: 0.05

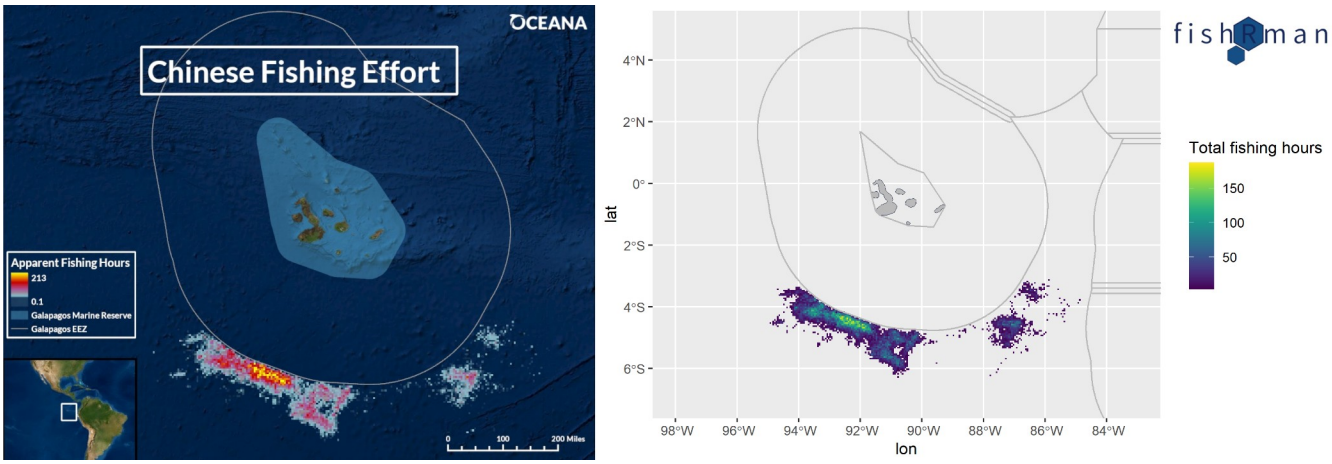

Figure 5: Visualization by Oceana (left) and fishRman (right) in comparison.

The spatial distributions of the fleet match in detail, with the same shapes, hotspots, and empty areas being reported.

The fact-check of Point 3 ends this example.

The reader is now capable of performing analyses and delivering visualizations that are comparable for accuracy, precision, and truth to facts, to those present in the report by Oceana.

5. References

[1] Buonomo, P., [2021]. *fishRman: A Shiny R Dashboard improving Global Fishing Watch data availability*. Journal of Open Source Software, 6(66), 3467, https://doi.org/10.21105/joss.03467

[2] Global Fishing Watch [2021]. https://globalfishingwatch.org/

[3] Buonomo P. [2021]. https://github.com/Shyentist/fish-r-man

[4] Collins D. [2020]. *Chinese fishing armada plundered waters around Galápagos, data shows*. https://www.theguardian.com/environment/2020/sep/17/chinese-fishing-armada-plundered-waters-around-galapagos-data-shows

[5] Valentine M. [2020]. *Oceana Finds 300 Chinese Vessels Pillaging the Galapagos for Squid*. https://doi.org/10.5281/ZENODO.4118526

[6] QGIS Development Team [2021]. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

[7] Flanders Marine Institute [2021]: MarineRegions.org. www.marineregions.org.

[8] Kroodsma D.A., Mayorga J., Hochberg T., Miller N.A., Boerder K., Ferretti F., Wilson A., Bergman B., White T.D., Block B.A., Woods P., Sullivan B., Costello C., and Worm B. [2018]. "*Tracking the global footprint of fisheries.*" Science 361.6378.