WELSH GOVERNMENT FISHERIES DATA WORKFLOW

To begin working with the Welsh fisheries data, it must be retrieved from the various sources, whether that be T3, IFISH or GeoFISH. The scripts provided in folder ‘0\_data\_access’ will do this, with a unique script being available for each data source. The title of each script indicates which source of data it should be used for.

0\_data\_access/0\_GEOFISH\_EFLALO\_and\_TACSAT\_WO10m.sql

These scripts are to be run using SQL, with the outputs being saved as:

eflalo\_ft.csv

eflalo\_le.csv

eflalo\_spe.csv

tacsat.csv

When the above data has been acquired from the chosen source, it will need to be loaded into R, where the analysis will take place. Again, there are source-specific scripts in the ‘0\_data\_access’ folder to do this. The word source in the title of the document should match the source the data was acquired from.

0\_data\_access/1\_Load\_SOURCE\_TACSAT\_EFLALO\_in\_R\_environment.R

This necessary and and format the CSVs, readying them for analysis by assigning column headers where necessary, and offering a brief insight into what the data looks like, for QC/QA purposes. Certain columns such as those containing date/time information are specifically formatted into year:month:day, hour:minute:second. This is to ensure consistency throughout the workflow.

Additionally, new columns and data are created here, such as assigning the data source and month/year columns, which are retrieved from other, already existing columns in the datasets. These will be used later in the analysis.

Following the data acquisition and loading into the R environment, it is useful to take a look into the data. The various ‘0\_data\_exploration’ documents do this, by exploring the data in a visual manner, with plots. Again, choosing the right script based on the source of the data is important due to subtle differences in the column names etc.

1\_data\_preprocessing/0\_data\_exploration.R

This script produces a number of plots which help to highlight outliers in the data, such as unexpectedly long number of days for a trip. Finally, an output for both eflalo and tacsat data are produced, which will be used in the subsequent script.

The next script is eflalo and tacsat preprocessing, where necessary checks are made on the data that will eliminate most of the anomalies or incorrect pieces of information.