Overview

This repository contains all code necessary to reproduce the analysis found in the paper “Global expansion of marine protected areas and the redistribution of fishing effort”. Most data necessary for reproducing the analysis are also available directly in this repository, including our fishing effort outcome variable and all model feature data except those relating to bunker fuel prices (more information on this below).

Paper Title: Global expansion of marine protected areas and the redistribution of fishing effort

Journal: Proceedings of the National Academy of Sciences of the United States of America (PNAS)

Authors: Gavin McDonald, Jennifer Bone, Christopher Costello, Gabriel Englander, Jennifer Raynor

Corresponding author: Gavin McDonald (gmcdonald@bren.ucsb.edu (mailto:gmcdonald@bren.ucsb.edu))

Reproducibility

Package management

To manage package dependencies, we use the `renv` package. When you first clone this repo onto your machine, run `renv::init()` and then select Option 1 (“Restore the project from the lockfile”) to ensure you have all correct package versions installed in the project. Please see the renv website (https://rstudio.github.io/renv/articles/renv.html) for more information.

Data processing and analysis pipeline

To ensure reproducibility in the data processing and analysis pipeline, we use the `targets` package. Targets is a Make-like pipeline tool. Using `targets` means that anytime an upstream change is made to the data or models, all downstream components of the data processing and analysis will be re-run automatically when the `targets::tar_make()` command is run. It also means that once components of the analysis have already been run and are up-to-date, they will not need to be re-run. All objects are cached in a `_targets` directory. Please see the targets website (https://github.com/ropensci/targets) for more information.

For targets to function correctly, the user only needs to make one change to update the cache directory. Run the command `tar_config_set(store = glue::glue("{targets_data_directory}/_targets"))` , where `targets_data_directory` is a local directory available to your machine. This local directory is where you will store the interim `targets` files. When choosing a directory, note that the final generated `targets` files will be very large (roughly 30GB, which includes trained random forest models, a large suite of hypothetical MPA network data, full simulation results, etc). Once this has been done, and once you have obtained the necessary fuel price data from Bunker Index (more details below), you can simply run `targets::tar_make()` to reproduce the entire analysis.

The `targets` pipeline is defined in the file `_targets.R` . In order to see what the targets pipeline looks like, you can run `targets::tar_manifest()` or `targets::tar_visnetwork()` , which also shows which targets are current or out-of-date. In overview, the pipeline:

1. Uses raw data to generate model features and the fishing efort outcome variable for the training data

2. Trains models using the training dataset

3. Tests the performance of the models using a variety of out-of-sample tests and model specification robustness checks
4. Generates the business-as-usual and hypothetical MPA network data for the simulations
5. Runs simulations using the business-as-usual and hypothetical MPA network scenarios and trained models
6. Finally, generates all figures, tables, and statistics included in the paper.

Data availability

We include most of the raw datasets directly in this repository and process most of these directly in the `targets` pipeline (raw datasets are found in `data/raw` , and processing functions are found in `r/_functions_data_wrangling.R`). These datasets are used to build most of the machine learning model features including: all MPA-related features derived (from the 2020 version of MPA Atlas), El Niño Southern Oscillation and Pacific Decadal Oscillation features (from NOAA), distance to seamounts (from Yesson et al. 2020), mesopelagic zone (from Sutton et al. 2017), all EEZ-related features (from Marine Regions), World Bank region (from the R `countrycode` package), and Global Fishing Index governance capacity (from Spijkers et al. 2023). We also include the raw data necessary to build the hypothetical MPA networks for the simulations (including the spatial files representing the networks from Visalli et al. 2020, Sala et al. 2021, and the Ecologically or Biologically Significant Marine Area networks from Dunn et al. 2014). Raw Global Fishing Watch data (originally described by Kroodsma et al. 2018) were obtained directly from Global Fishing Watch through Google BigQuery (which is performed in `03_data_wrangling_gfw.Rmd`) and processed versions of the datasets are made available in `data/model_features` (including our fishing effort main outcome variable, lagged fishing effort, AIS reception quality for type A and B transponders, latitude, longitude, distance to shore, distance to port, and bathymetry). Several model features require a multitude of very large downloads of publicly available raw data (NOAA's ERDDAP Sea Surface Temperature, Aqua MODIS Chlorophyll data, and the REMSSS wind data). This processing is done in `01_data_wrangling_erddap.Rmd` and `04_data_wrangling_wind.Rmd` , and the processed versions of the data are made available in `data/model_features` . Therefore, future researchers will have access to these processed versions of these data for reproducing our analysis, and will also have access to the exact code we used to process all data. More information on our data sources for each model feature are available in Table 1 of the paper.

As a important note: the only model feature dataset we are not able to include in this repository, either in raw or processed form, is the bunker fuel price data from Bunker Index. The bunker fuel price index data used in our analysis are subject to restricted use and are not available for public redistribution. Information on obtaining these data directly from Bunker Index can be found at https://bunkerindex.com/ (https://bunkerindex.com/) . Once you have obtained and locally downloaded these data, you can save them locally as `data/raw/bix_world_ifo_380_index.csv` and the `targets` pipeline will process them. Always be mindful to follow Bunker Index's terms and conditions, which are subject to change at any time (https://bunkerindex.com/terms (https://bunkerindex.com/terms)). Ensure that the dataset you are using is the global BIX World IFO 380 index, covers the time period from 2016 through 2021, and that the CSV has the columns `Date` and `Price` (where `Price` is in units of USD/MT) so that the function `wrangle_fuel_price_data` can properly aggregate the daily price data to annual means and standard deviations of price.

Repository Structure

The repository uses the following basic structure:

```
mpa-fishing-effort-redistribution
  |__ data
      |__ model_features
      |__ raw
  |__ figures
        |__ figure_data
  |__ r
  |__ renv
  |__ tables
```

`data/raw` contains all raw data files that are used directly in the `targets` pipeline, other than bunker fuel price data (see note above). `data/model_features` contains all pre-processed model feature data (which are generated in `01_data_wrangling_erddap.Rmd`, `02_data_wrangling_spatial_measures.Rmd`, and `03_data_wrangling_gfw.Rmd`).

All R code can be found in the `r` directory. The files in this directory are described as follows:

- `_functions_data_wrangling_erddap.R`: This contains functions for wrangling the spatial ERDDAP environmental data from NOAA (these eunfctions are used in `01_data_wrangling_erddap.Rmd`)
- `_functions_modeling.R`: This contains all functions necessary for training the machine learning models, testing their performance, and generating predictions from them. This also contains all functions for generating the hypothetical MPA networks for the simulations.
- `01_data_wrangling_erddap.Rmd`: Code for wrangling spatial ERDAPP spatiotemporal environmental data from NOAA. This wrangles SST and SST anomaly data (saved as `data/model_features/errdap_sst.csv`); and Chlorophyll-a (saved as `data/model_features/errdap_chl.csv`). This code also generates the global 1x1 degree grid used for the entire analysis, (saved as `data/model_feautres/global_grid.csv`).
- `02_data_wrangling_spatial_measures.Rmd`: This script processes various spatiotemporal and spatial data including bathymetry, distance from shore, and distance to port (saved as `data/model_features/gfw_static_spatial_measures.csv`), as well as ocean classification (saved as `data/model_features/oceans.csv`). This script also produces the no-take MPA shapefile that is used in the analysis (`data/raw/current_2020_no_take_mpa_geopackage.gpkg`). This shapefile is a version of MPA Atlas that has been filtered to `no_take == "All"`, `status %in% c("Designated","Established")`, and which has updated no-take zone implementation dates that were compiled during this project for the precise date after which no fishing was allowed.
- `03_data_wrangling_gfw.Rmd`: This script downloads and aggregates all fishing effort data from Global Fishing Watch (saved as `data/model_features/gfw_fishing_by_pixel_year.csv`). Fishing hours by 1x1 degree pixel by year is the main outcome variable of the analysis. It also generates features related to AIS reception quality (saved as `data/model_features/gfw_reception_quality`.).
- `04_data_wrangling_wind.Rmd`: Generates spatiotemporal wind data from CCMP (saved as `data/model_features/remss_wind.csv`).
- `05_analysis_generate_paper_figures_tables.Rmd`: Using the model performance data, variable importance data, and outputs of the simulations, this script generates all figures, tables, and summary statistics necessary for the paper. Figures are saved in the `figures` directory and tables are saved in the `tables` directory. The exact processed data that go into all figures, both in the main text and in the SI, are saved in the `figures/figure_data` directory.

The `figures` directory contains PDFs of all figures for the paper, which are generated in the `targets` pipeline. Main text figures have the prefix `main-text-`, and supplementary information figures have the prefix `si-`. All processed data necessary to reproduce all figures, both for the main text figures and from the supplementary information figures, are in the `figures/figure_data` directory. The `tables` directory contain the final model performance table for the paper, which is also generated in the `targets` pipeline. The `renv` directory and `renv.lock` lockfile include package version information for reproducing the analysis. These files are automatically used when you run `renv::init()` and then select Option 1 (“Restore the project from the lockfile”).