Predicting the spread of invasive marine species with open data and machine learning: Process and Challenges

Adrian Bumann, Chalmers University of Technology (Sweden), adrian.bumann@chalmers.se
Robin Teigland, Chalmers University of Technology (Sweden), robin.teigland@chalmers.se
Jannes Germishuys, Combine AB (Sweden), jurie.germishuys@combine.se
Benedikt Ziegler, Combine AB (Sweden), benedikt.ziegler@combine.se
Martin Mattson, Medins Havs och Vattenkonsulter AB (Sweden), martin.mattsson@medinsab.se
Eddie Olsson, RISE - Research Institutes of Sweden (Sweden), eddie.olsson@ri.se
Robert Rylander, RISE - Research Institutes of Sweden (Sweden), robert.rylander@ri.se
Markus Lindh, Swedish Meteorological and Hydrological Institute (Sweden), markus.lindh@smhi.se
Yixin Zhang, University of Gothenburg (Sweden), yixin.zhang@ait.gu.se
Torsten Linders, University of Gothenburg (Sweden), torsten.linders@gu.se

Background.

One of the world's most complex marine challenges is the spread of invasive species. Invasive species cause severe harm to marine ecosystems and the people who depend on them, with economic impact alone amounting to several billion dollars annually. Recent advances in data science and artificial intelligence (AI) along with the increasing availability of free marine and other data online are improving the possibility to tackle these challenges. This paper presents the efforts by Ocean Data Factory Sweden (ODF Sweden), a data-driven innovation consortium in Gothenburg, to apply machine learning (ML) to one use case – the prediction of the spread of the Killer Shrimp, or Dikerogammarus Villosus, into the Baltic Sea (Figure 1). We discuss our process to address this use case as well as some reflections on the process and its challenges, in particular when taking into consideration the FAIR (findable, accessible, interoperable and reusable) principles in data science.

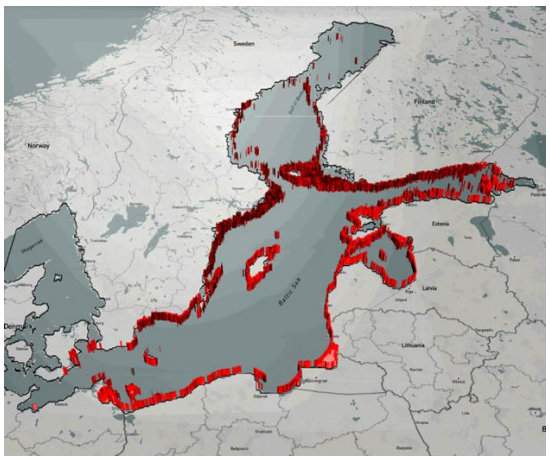

Figure 1: Visualizing the prediction of the spread of the Killer Shrimp into the Baltic Sea

Developing the Killer Shrimp Use Case.

Defining the questions for investigation. Killer Shrimp have already invaded Europe (e.g., Bollache et al. 2004) presumably through the ballast water of cargo ships as ocean expanses are too vast for the shrimp to traverse. The shrimp have been recorded in rivers in Western Europe, perhaps by travelling through inland waterways from the Black Sea, and more recently, it has been detected in the Baltic Sea. As this shrimp devastates the local ecosystems it invades, ODF Sweden and its partner, the Swedish Agency for Marine and Water Management (SwAM), decided to investigate whether ML could help predict the areas of the Baltic Sea suitable for the Killer Shrimp. In particular, we decided to explore these questions: 1) what are the factors that could lead to the spread of the shrimp into the

Baltic Sea region?, 2) how might various scenarios, such as changes in climate or shipping routes affect these factors?, and 3) how might this species' spread Baltic affect ecosystems and local industry?

Preparation of training and test datasets. We collected the following open datasets: 1) port locations in Europe (EMODNET), 2) ocean surface temperatures and salinity for Baltic Sea (SMHI) and North Sea regions (SeaDataNet), 3) presence data of D. Villosus from observations from 1928-2019 (GBIF), 4) marine data layers (Bio-Oracle), and 5) ocean temperature and salinity (CMEMS). The features of temperature, salinity, depth, substrate and wave activity were selected based on input of marine experts. Missing data were removed, and features were visualized. We noticed that the data were very skewed towards the absence class, i.e. there was extreme high-class imbalance. To address this, we used oversampling to increase the instances of the "presence" class by creating synthetic cases based on the original presence cases. We split the data 80/20 into a training set and a test set for evaluation.

***Building and refining the model*.** We used primarily tree-based models, a single decision tree and a Random Forest but also included Deep Neural Networks for more complex feature extraction. All models were trained with their standard configurations in scikit-learn and fast.ai Python libraries for easy replication. Models were scored on their ability to correctly predict the locations where the killer shrimp would be present. We were able to get a probability that a particular point belongs to our presence class. Throughout the entire process, we adapted our methods as new data became available and we learned more about the nature of the problem.

Visualizing our results. Since our features come in the form of rasters, using our trained models we were able to make predictions for each cell in the raster grid. The model output is then the probability of "presence" in that cell. We built a web application that helps visualize the probabilities from some of these models as well as the impact of future climate changes on these probabilities in the Baltic Sea. Specifically, we noticed the increased suitability of Åland and the Eastern coast of Sweden under future climate condition forecasts from SMHI (figure 1).

Discussion.

Following the FAIR principles, ODF Sweden worked with a principle of openness. Only open datasets were used; the project was documented on Jupyter notebooks on Kaggle; and Github was used to host our code repository. This resulted in several challenges. First, while the data are plentiful on open data platforms, they lie in multiple siloed systems without central access point or methodology. As a result, extracting and converting data took the bulk of the time. Second, when shifting to geospatial ocean data, we found that it was easy to be overconfident in our model predictions. When we simplified sample data points from a large area in the ocean and then split our datasets into training and test sets, the distribution of the training and test data were so similar that the test set effectively "leaked" into the training set. Third, each data provider has its own preferred coordinate reference systems (CRS). Since the Earth is spherical, each CRS represents a projection onto a flat 2D surface for visualization. Re-projecting between these systems is often necessary when performing comparisons and calculations. Python packages, such as GDAL and Rasterio, helped simplify this. Finally, another major challenge was interpreting inland data. Since we had no information available about inland water sources, we had to match these to the closest body of ocean water. This proved to be difficult and inaccurate, and we had to make assumptions such as "inland water is just as salty as sea water". This led to large biases in our initial results and led us to revisit this assumption and ultimately abandon this when we obtained additional presence data in the Baltic Sea. The model output also revealed the importance of appropriate data input in answering our questions. In our case, our results reveal that we need additional data to answer more insightful questions, such as future migration patterns to predict species abundance and data on other species in the area to understand the interaction effects on overall biodiversity in areas where the Killer Shrimp has been detected. Looking forward from a user perspective, we recommend that data providers improve and align documentation standards. We also hope that datasets will become more searchable and that new datasets will be promoted to boost research efforts to answer these important questions.

References

Bollache, L., Devin, S., Wattier, R., Chovet, M., Beisel, J. N., Moreteau, J. C., & Rigaud, T. (2004). Rapid range extension of the Ponto-Caspian amphipod Dikerogammarus villosus in France: potential consequences. *Archiv für Hydrobiologie*, *160*(1), 57-66.