Developing synthetic datasets for reef modelling

Rose Crocker, Barbara Robson, Takuya Iwanaga and Ken Anthony

Australian Institute of Marine Science, Cape Cleveland, Queensland, Australia
Email: r.crocker@aims.gov.au

Abstract: Synthetic data mimics the statistical properties of real-world datasets while removing reference to sensitive or confidential information in the original dataset (Quintana, 2020). Use of synthetic data began as a means to allow the discussion of confidential census data and has gained popularity in pharmaceutical and artificial intelligence research due to demand for retaining data privacy and need for large training data sets for machine learning models. Synthetic data is also useful for general model testing and development, with many methods available for generating data from machine learning models (Raghunathan, 2020).

Synthetic data is not widely used in conservation and environmental management, but its benefits are being increasingly recognized and used. Reviews of data sharing in ecology have stressed the need for better access to important data sets to fully capitalize on the wealth of data being generated (Reichmann et. al. 2011). Synthetic data offers a means of addressing essential macroecological and biodiversity questions where collection and/or integration of additional real-world data is prohibitively expensive, existing data cannot be made publicly available, or data collection has faced policy and funding challenges (Poisot et. al. 2015).

In the context of reef modelling, synthetic data can be used to support model analyses that can be published without referring to specific sites or reefs. This is desirable in the context of decision support for restoration of the Great Barrier Reef, as the Reef has many stakeholders and release of early modelling results for intervention scenarios at specific sites would be premature until a clear policy framework has been defined. Real reef data and the results of analysing real reef data may be sensitive to Traditional Owners and other stakeholders who may wish to keep the analysis of investment reef areas private. For these reasons, it is desirable to create synthetic datasets which realistically represent the ecological and environmental conditions of important reefs, but do not reference actual GBR reef sites.

We showcase a synthetic data pipeline developed for the reef decision support model ADRIA (Adaptive Dynamic Reef Intervention Algorithm), using methods from the Python package Synthetic Data Vault (Patki, N., Wedge, R. & Veeramachaneni, K. 2016) and others. The synthetic data models are developed to replicate the statistics of important case study reefs for publication, model testing and method validation without revealing sensitive site information when publishing model results. This pipeline includes developing models for tabular (reef compositional data), temporal (wave and heat stress data) and spatial data (larval connectivity). Conditional sampling methods which maintain spatial relationships across datasets are used to develop synthetic reef data packages which mimic the spatial and temporal properties of the original dataset. The utility of the synthetic data is demonstrated on sample reef datasets, and methods used for anonymizing the data are discussed. Finally, the results are discussed in the context of formalizing synthetic data for reef modelling.

REFERENCES

Chen, R.J. et al. 2021. Synthetic data in machine learning for medicine and healthcare. Nat. Biomed. Eng., 5, 493–497.

Patki, N., Wedge, R. & Veeramachaneni, K. The Synthetic Data Vault. IEEE DSAA 2016.

Poisot, T. et. al. 2015. Synthetic datasets and community tools for the rapid testing of ecological hypotheses. 39:4, 402-408.

Quintana, D.S. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. 2020. Elife. 9:e53275.

Raghunathan, T., 2021. Synthetic Data. Annu. Rev. Stat. Appl., 8:1, 129-140.

Reichman O.J. et al., 2011. Challenges and Opportunities of Open Data in Ecology. Science, 331, 703-705.

Keywords: *Synthetic data, ecological modelling, stakeholders, reef modelling, restoration*