The Journal of Open Source Software

stemflow: A Python Package for Adaptive Spatio-Temporal Exploratory Model

Yangkang Chen[1,2], **Zhongru Gu**[1,3], **and Xiangjiang Zhan**[1,2,3,4]

1 Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China **2** University of Chinese Academy of Sciences, Beijing, China **3** Cardiff University-Institute of Zoology Joint Laboratory for Biocomplexity Research, Chinese Academy of Sciences, Beijing, China **4** Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Renata Diaz

Reviewers:

- @jedalong
- @earth-chris

Submitted: 21 September 2023
Published: unpublished

Summary

Stemflow is a user-friendly Python package for Adaptive Spatio-Temporal Exploratory Model (AdaSTEM (Fink et al., 2013)) that follows the style of scikit-learn BaseEstimator class (Pedregosa et al., 2011). It provides one-line model creation, fitting, prediction, and evaluation. It implements spatio-temporal train-test-split and cross-validation functions. After model training, feature importance could be evaluated with spatio-temporal dynamics. Stemflow also provides functions for visualizing ensembles structured in model training and generating GIF file for predicted results to animate the spatio-temporal movement of animal population.

Statement of need

Spatio-temporal big data is an emerging but valuable resource for ecological studies as human enter the era of big data (Farley et al., 2018). Data from broad-scale survey like citizen science projects is increasingly important in modern ecological research (Dickinson et al., 2010). Intensity of survey activities grows rapidly as more people are involved in citizen science in recent years, resulted in exponential accumulation of observational data (Di Cecco et al., 2021; Sullivan et al., 2014). However, daily species observation records uploaded by non-professionals in citizen science program are known to have larger bias than professionally structured research, both in terms of data veracity and spatio-temporal balance of the datasets, which necessitates elaborate modeling methods to mine its merits (Dickinson et al., 2010; Farley et al., 2018).

Some species distribution modeling (SDM) approaches were brought forward to adjust for bias in citizen science and model on the unobserved components (Bird et al., 2014). Still, many failed to account for the autocorrelation of space and time (F. Dormann et al., 2007), which is especially crucial in modeling inherently spatio-temporal biological events with variations at different scales (Chave, 2013; Levin, 1992), such as seasonal migration. Adaptive Spatio-Temporal Exploratory Model (AdaSTEM) is a semi-parameterized machine learning model that leverages the spatio-temporal adjacency information of sample points to model occurrence or abundance of species (Fink et al., 2013). A QuadTree algorithm (Samet, 1984) is implemented to split data into smaller spatio-temporal grids (called stixels) conditional on the data abundance, with more abundant data allowing stixels to be divided into finer resolution (up to a maximum). Stixels with data size less than a certain threshold will not be modeled; instead, these stixels will be labeled as unpredictable. This procedure controls the degree of model extrapolation and reduces overfitting. A base model is trained for each stixel, that is, targets are only modeled on their adjacent information in space and time. Splitting-training is carried out several times to generate multiple ensembles. Finally, prediction results were aggregated across these ensembles.

AdaSTEM shows the capacity of supporting large scale spatio-temporal ecological data modeling in many studies (Fink et al., 2020; Fuentes et al., 2023; La Sorte et al., 2022), espetially for modeling animal abundance at different scales (Fink et al., 2013). One well-known application of AdaSTEM is the weekly abundance map of eBird Status and Trend product (Fink et al., 2022), which was widely used as data sources of abundance data of bird populations (Bird et al., 2014; Jarzyna & Stagge, 2023; Lin et al., 2022). The application of AdaSTEM could be extended to other fields with similar data structure and spatio-temporal dependence, for example, epidemiology. Despite the foreseeable significant role of spatio-temporal big data in the coming decades of scientific research, the development of tools has not necessarily kept pace.

Stemflow is positioned as a user-friendly Python package to meet the need of general application of modeling spatio-temporal large datasets. Scikit-learn style object-oriented modeling pipeline enables concise model construction with compact parameterization at the user end, while the rest of the modeling procedures are carried out under the hood. Once the fitting method is called, the model class recursively splits the input training data into smaller spatio-temporal stixels using QuadTree algorithm. For each of the stixels, a base model is trained only using data falls into that stixel. Stixels are then aggregated and constitute an ensemble. In the prediction phase, stemflow queries stixels for the input data according to their spatial and temporal index, followed by corresponding base model prediction. Finally, prediction results are aggregated across ensembles to generate robust estimations (see Fink et al. (2013) and stemflow documentation for details).

For survey projects that include abundance information like eBird (Sullivan et al., 2014), the targeted modeling values are often zero-inflated, owning to the fact of low observation probability in many species. Zero-inflation could lead to poor regression model performance (Campbell, 2021). In stemflow, we implement hurdle model classes that embed two sequential models: a classifier to classify the absence and presence states, followed by a regressor to model the abundance for prediction samples classified as presence. Hurdle model classes can be conjunctively used with AdaSTEM model classes in two ways: Use hurdle model as the base model for AdaSTEMRegressor (as in Johnston et al. (2015)), or use AdaSTEMClassifier and AdaSTEMRegressor as the classifier and regressor in hurdle model. We demonstrate the comparison of these two architectures in stemflow documentation.

One advantage of applying stemflow in scikit-learn style is that there is a variety of "base models" to choose from scikit-learn or scikit-learn-style repertoire. The choices vary from linear models to boosting and bagging tree-based models. Maxent model (C. B. Anderson, 2023) is also supported to play the role of "base model", which largely expands the potential application for presence-only modeling (see documentation).

While there exists mounting open source packages for species distribution modeling (mostly in R, (Norberg et al., 2019); and one in Python (C. B. Anderson, 2023)), most of them solely leverage environmental variables and do not support integration of spatio-temporal information during model construction (but see C. B. Anderson (2023); S. C. Anderson et al. (2022); Dobson et al. (2023)). This disadvantage is usually noted along with the overconfidence of the model extrapolation capacity both for Maxent-based and ensemble-based models (A. Lee-Yaw et al., 2022). To our knowledge, stemflow is the first package designed to solve the spatio-temporal dependency conjugated with bias in data abundance distribution in species distribution modeling. With the rapid accumulation of data and development of machine learning techniques, stemflow will show potentials in spatio-temporal modeling, and could be applied to other fields (e.g., epidemiology and weather prediction) in future.

Acknowledgements

This project was based upon work supported by National Natural Science Foundation of China grants 32125005 (to X.Z.), 32270455 (to Z.G.), CAS Project for Young Scientists in

Basic Research (YSBR-097 to X.Z. and Z.G.). We thank Dr. Daniel Fink at Cornell Lab of Ornithology for the R scripts of STEM model and QuadTree which inspired our implementation in Python.

References

A. Lee-Yaw, J., L. McCune, J., Pironon, S., & N. Sheth, S. (2022). Species distribution models rarely predict the biology of real populations. *Ecography*, *2022*(6), e05877. https://doi.org/10.1111/ecog.05877

Anderson, C. B. (2023). Elapid: Species distribution modeling tools for Python. *Journal of Open Source Software*, *8*(84), 4930. https://doi.org/10.21105/joss.04930

Anderson, S. C., Ward, E. J., English, P. A., & Barnett, L. A. K. (2022). *sdmTMB: An R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields* [Preprint]. Ecology. https://doi.org/10.1101/2022.03.24.485545

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods in Ecology and Evolution*, *12*(4), 665–680. https://doi.org/10.1111/2041-210X.13559

Chave, J. (2013). The problem of pattern and scale in ecology: What have we learned in 20 years? *Ecology Letters*, *16*(s1), 4–16. https://doi.org/10.1111/ele.12048

Di Cecco, G. J., Barve, V., Belitz, M. W., Stucky, B. J., Guralnick, R. P., & Hurlbert, A. H. (2021). Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. *BioScience*, *71*(11), 1179–1188. https://doi.org/10.1093/biosci/biab093

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dobson, R., Challinor, A. J., Cheke, R. A., Jennings, S., Willis, S. G., & Dallimer, M. (2023). dynamicSDM: An r package for species geographical distribution and abundance modelling at high spatiotemporal resolution. *Methods in Ecology and Evolution*, *14*(5), 1190–1199. https://doi.org/10.1111/2041-210X.14101

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., & Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, *30*(3), e02056. https://doi.org/10.1002/eap.2056

Fink, D., Auer, T., Johnston, A., Strimas-Mackey, M., Ligocki, S., Robinson, O., Hochachka, W., L, Jaromczyk, Rodewald, A., Wood, C., Davies, I., & Spencer., A. (2022). *eBird status* *and trends, data version: 2021; released: 2022*. Cornell Lab of Ornithology, Ithaca, New York. https://doi.org/10.2173/ebirdst.2021

Fink, D., Damoulas, T., & Dave, J. (2013). Adaptive Spatio-Temporal Exploratory Models: Hemisphere-wide species distributions from massively crowdsourced eBird data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *27*(1), 1284–1290. https://doi.org/10.1609/aaai.v27i1.8484

Fuentes, M., Van Doren, B. M., Fink, D., & Sheldon, D. (2023). BirdFlow: Learning seasonal bird movements from eBird data. *Methods in Ecology and Evolution*, *14*(3), 923–938. https://doi.org/10.1111/2041-210X.14052

Jarzyna, M. A., & Stagge, J. H. (2023). Decoupled spatiotemporal patterns of avian taxonomic and functional diversity. *Current Biology*, *33*(6), 1153–1161.e4. https://doi.org/10.1016/j.cub.2023.01.066

Johnston, A., Fink, D., Reynolds, M. D., Hochachka, W. M., Sullivan, B. L., Bruns, N. E., Hallstein, E., Merrifield, M. S., Matsumoto, S., & Kelling, S. (2015). Abundance models improve spatial and temporal prioritization of conservation resources. *Ecological Applications*, *25*(7), 1749–1756. https://doi.org/10.1890/14-1826.1

La Sorte, F. A., Horton, K. G., Johnston, A., Fink, D., & Auer, T. (2022). Seasonal associations with light pollution trends for nocturnally migrating bird populations. *Ecosphere*, *13*(3), e3994. https://doi.org/10.1002/ecs2.3994

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, *73*(6), 1943–1967. https://doi.org/10.2307/1941447

Lin, H.-Y., Binley, A. D., Schuster, R., Rodewald, A. D., Buxton, R., & Bennett, J. R. (2022). Using community science data to help identify threatened species occurrences outside of known ranges. *Biological Conservation*, *268*, 109523. https://doi.org/10.1016/j.biocon.2022.109523

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., … Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*(3), e01370. https://doi.org/10.1002/ecm.1370

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, Édouard. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Samet, H. (1984). The Quadtree and Related Hierarchical Data Structures. *ACM Computing Surveys*, *16*(2), 187–260. https://doi.org/10.1145/356924.356930

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003