

**Machine learning methods reveal processes affecting abundance at multiple scales. A commentary on ‘Global and regional drivers of abundance patterns in the hart’s tongue fern complex (*Aspleniaceae*)’**

**Joel H. Nitta**

Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8562, Japan

**Corresponding author details:** Joel E. Nitta, [joelnitta@gmail.com](mailto:joelnitta@gmail.com)

**Key words:** Abundance, *Asplenium*, conservation, fern, GBIF, machine learning.

We live in an era of unprecedented ecological disruption and human-driven extinction. Conservation policies seeking to mitigate such damage depend on documenting biodiversity while at the same time determining the processes generating it (Smith *et al.*, 1993). Due to the large amounts of data required and because of the complex and intertwined nature of ecology, this is a difficult endeavour and we can only rarely discern particular determinants of species distributions. In this issue of *Annals of Botany*, Heo *et al.* (2022) combine global sampling with recently developed machine learning methods to gain insight into the processes governing abundance in *Asplenium scolopendrium*, or hart’s tongue fern (HTF; Fig. 1). Abundance is a key parameter for understanding biodiversity patterns, as it governs the existence of local populations and ultimately species; a full appreciation of the processes regulating abundance is critical to conserving biodiversity in the Anthropocene (Brown *et al.*, 1995).

Hart’s tongue fern is a widely distributed species complex with major centres in Europe, North America and Asia. Among regions, HTF shows different patterns of abundance, with patchily distributed, smaller populations in North America, widely and continuously distributed populations in Europe and intermediate densities in Asia. A much smaller population centre in New Zealand likely stems from recent introduction and naturalization (Brownsey and Perrie, 2017). Hart’s tongue fern has been well collected and

documented, with thousands of occurrence records available in GBIF. Thus, HTF is an excellent system for understanding the processes affecting species abundance on global and local levels.

It is likely that processes at multiple scales interact to impact abundance (Brown, 1984). At broad (global) scales, climate change may shift species distributions along latitudinal gradients (Chen *et al.*, 2011). Also at larger scales, changes in land use influence species distributions (Daskalova *et al.*, 2020). At the same time, processes acting at regional to local scales, such as soil type distributions and local climates, further influence species abundance (Lafleur *et al.*, 2010; Williams and Newbold, 2020). The complex and interwoven nature of these processes has long been an impediment to understanding how they function in detail. However, Heo *et al.* (2022) were able to use machine learning, in combination with dense sampling, in an attempt to overcome these difficulties.

Mathematical models, and linear regression models in particular, have long been a key tool in ecology (Bolker *et al.*, 2009). Classic ecological approaches typically start by defining a reasonable model for a given phenomenon and then seek to fit the parameters of the model to the data. While such classic models are straightforward to interpret, they often have relatively low explanatory power. This stems from the extremely complex nature of ecology, which includes many interacting processes across multiple scales (Graham, 2003). Machine learning approaches, on the other hand, do not start from a particular model; rather, they attempt to learn the relationship between predictors and response variables via iterative algorithms. While these represent promising approaches for understanding ecological processes, they have not been widely applied in ecology, perhaps due to the dominant statistical paradigm that has focused on explanation rather than prediction (Elith *et al.*, 2008).

Heo *et al.* (2022) used boosted regression trees (BRTs), a type of machine learning that combines a large number of simple regression trees to obtain a robust prediction (Elith *et al.*, 2008). Their BRT model included variables at both global (e.g. magnitude of past climate change, anthropogenic impacts) and regional (topography, edaphic conditions, local climate) scales. The BRT model is well suited to analyses including interacting

variables at multiple scales as it is non-parametric and insensitive to outliers; furthermore, any potential interactions are accounted for during the learning process and do not need to be specified *a priori*. Since HTF was already known to show different distribution patterns between regions (Europe, North America and Asia), Heo *et al.* (2022) applied the BRT model separately to each region and compared the results.

On a global scale, Heo *et al.* (2022) found that HTF is mostly restricted to a single biome, temperate broadleaf and mixed forests. Furthermore, their niche modelling analysis showed that this species complex has likely shifted its range from lower to higher latitudes since the last glacial maximum (~21 ka). In addition to these common global processes, the BRT model revealed processes operating at finer scales that in some cases differed between regions.

One clear factor influencing abundance identified by Heo *et al.* (2022) was population isolation (measured as distance to the nearest population), which was found to be the most important variable across all four regions. However, population isolation was less important in North America, where populations tend to be restricted to scattered microhabitats, than in Europe or Asia, where populations tend to be more contiguous. It could be that distances between North American populations are greater than the typical dispersal ability of HTF. In contrast, populations in Europe and Asia may be subject to the ‘rescue effect’, whereby propagules from neighbouring populations are able to replenish a low-abundance population otherwise in danger of extirpation (Brown and Kodric-Brown, 1977).

Heo *et al.* (2022) found that the next most important variables after population isolation varied among regions. In Europe, precipitation seasonality was most important, whereas precipitation heterogeneity and past climate change were most important in Asia and North America, respectively. Again, these differences make sense in light of what is known about the geographic context of each area. In Europe, dense populations (hotspots) of HTF occur in the west where the climate is oceanic. In contrast, HTF hotspots in Asia are often on mountainous islands with steep climatic gradients, which is reflected in the importance of precipitation heterogeneity in the BRT



FIG. 1. *Asplenium scolopendrium* var. *americanum* growing in New York. Photo by James E. Watkins Jr, used with permission.

model for that area. The flora of North America has been strongly affected by glaciation, and many extant HTF populations there are relictual, persisting in concave karst landscapes with protected microclimates. This is consistent with both a relatively low contribution of population isolation and a greater contribution of past climate change observed in the BRT model for North America.

Heo *et al.* (2022) further interpreted the significance of their findings in the context of the conservation of this species complex. In particular, the results of their fine-scale case study of population change over time in North America provided evidence that conservation policies have an impact on survival of HTF. For paired populations in similar environments, both at higher latitudes, those outside of conservation zones tended to decrease in abundance over time, whereas those in protected areas increased. While many studies on the impact of climate change on biodiversity are overwhelmingly negative, this highlights the clear benefits to be gained from establishment of protected areas.

There are some caveats that must be recognized in this study. First, occurrence records in GBIF are frequently biased in their spatial and temporal sampling. For example, plant specimens are more likely to be collected along roads and during the spring (Daru *et al.*, 2018). Second, any model that includes spatial data may be subject to spatial autocorrelation, the tendency for data points that are closer together to be more similar. Spatial autocorrelation violates the assumption of independent sampling and may unduly influence models (Legendre, 1993). While Heo *et al.* (2022) partially accounted for this by rarefying the occurrence data, they did not check for spatial autocorrelation in the model residuals, so it is unclear if results may have been influenced by spatial autocorrelation. Future studies should

explicitly check for spatial autocorrelation and handle it appropriately if detected.

Questions of statistical best practices aside, the study of Heo *et al.* (2022) is a clear example of the insights that can be gained from the application of machine learning to large biodiversity datasets. Such investigations will surely become more insightful as datasets expand and analytical methods improve. Indeed, they are needed more urgently than ever as life on this planet is increasingly confronted with the effects of unchecked economic growth gained at the expense of our environment.

#### ACKNOWLEDGEMENTS

Eric Schuettelpelz provided helpful comments on an earlier draft of this manuscript.

#### LITERATURE CITED

- Bolker BM, Brooks ME, Clark CJ, *et al.* 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24**: 127–135. doi:10.1016/j.tree.2008.10.008.
- Brown JH. 1984. on the relationship between abundance and distribution of species. *American Naturalist* **124**: 255–279. doi:10.1086/284267.
- Brown JH, Kodric-Brown A. 1977. Turnover rates in insular biogeography: effect of immigration on extinction. *Ecology* **58**: 445–449. doi:10.2307/1935620.
- Brown JH, Mehlman DW, Stevens GC. 1995. Spatial variation in abundance. *Ecology* **76**: 2028–2043.
- Brownsey PJ, Perrie LR. 2017. Taxonomic notes on the New Zealand flora: lectotypes in the fern family Aspleniaceae. *New Zealand Journal of Botany* **55**: 249–257.
- Chen I-C, Hill JK, Ohlemüller R, Roy DB, Thomas CD. 2011. Rapid range shifts

of species associated with high levels of climate warming. *Science* **333**: 1024–1026. doi:10.1126/science.1206432.

- Daru BH, Park DS, Primack RB, *et al.* 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* **217**: 939–955.
- Daskalova GN, Myers-Smith IH, Bjorkman AD, *et al.* 2020. Landscape-scale forest loss as a catalyst of population and biodiversity change. *Science* **368**: 1341–1347. doi:10.1126/science.aba1289.
- Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* **77**: 802–813. doi:10.1111/j.1365-2656.2008.01390.x.
- Graham MH. 2003. Confronting multicollinearity in ecological multiple regression. *Ecology* **84**: 2809–2815. doi:10.1890/02-3114.
- Heo N, Leopold DJ, Lomolino MV, Yun S, Fernando DD. 2022. Global and regional drivers of abundance patterns in the hart's tongue fern complex (Aspleniaceae). *Annals of Botany*.
- Lafleur B, Paré D, Munson AD, Bergeron Y. 2010. Response of northeastern North American forests to climate change: will soil conditions constrain tree species migration? *Environmental Reviews* **18**: 279–289. doi:10.1139/a10-013.
- Legendre P. 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* **74**: 1659–1673. doi:10.2307/1939924.
- Smith TB, Bruford MW, Wayne RK. 1993. The preservation of process: the missing element of conservation programs. *Biodiversity Letters* **1**: 164–167. doi:10.2307/2999740.
- Williams JJ, Newbold T. 2020. Local climatic changes affect biodiversity responses to land use: a review. *Diversity and Distributions* **26**: 76–92.