Moving past mean similarity in homogenization research

David J. Harris

Abstract

Studies of beta diversity often use the average Jaccard similarity among pairs of sites to quantify a landscape's biotic homogeneity. Over more than a decade, community ecologists have suggested a range of ecological hypotheses and frameworks to explain the variation in species' effects on this average similarity. However, many of these results can be explained simply by counting the proportion of sites occupied by each species and adding a small correction for non-independent assortment. Specifically, average Jaccard similarity can be decomposed exactly into an occupancy-based component that explains most of the variance and a covariance-based component that explains the rest. To the extent that homogenization research can be summarized by these two terms, it contains little information about ecological processes. Future ecological studies on beta diversity and biotic homogenization should therefore focus on 1) explaining changes in occupancy, 2) explaining changes in the covariance term, and 3) explaining changes in local similarity at sub-landscape scales. Finally, the decomposition presented here shows that mean Jaccard similarity may have problems as a measure of homogenization, and ecologists should consider alternative metrics.

Introduction

As human influences spread some species globally and drive others extinct, community ecologists have become increasingly concerned about the loss of local biotic distinctiveness, or *biotic homogenization* (McKinney and Lockwood 1999, Rooney et al. 2007). In general, homogenization occurs when landscapes become dominated by cosmopolitan invaders and lose their local endemics, although ecologists have observed a great deal of context-dependence (McKinney 2004, Olden and Rooney 2006, Rooney et al. 2007, Rosenblad and Sax 2016). While several frameworks have been proposed for understanding these effects at the level of individual site pairs, they are cumbersome to use in practice. For example, Olden and Poff (2003)'s framework involves classifying biotic turnover events into 14 categories, while Rosenblad and Sax (2016)'s framework involves six. In either framework, a researcher may need to keep track of thousands of such events across thousands of pairs of sites. Moreover, these pairs of sites will not be statistically independent of one another, as each site will belong to many pairs. Worse, several types of turnover event are compatible with either homogenization or differentiation, depending on the prior state of the community (Rooney et al. 2007, Rosenblad and Sax 2016). This complexity makes it difficult to test a framework's ability to explain the observed context dependence or its ability to make useful predictions

about future biotic changes.

Homogenization is typically quantified as an increase in the average Jaccard similarity among pairs of sites on a landscape ($\bar{J}$, defined more precisely in Equation 1 below). While many papers provide verbal arguments about the behavior of this metric, Harris et al. (2011) showed that it is not sensitive to the site-level patterns emphasized by popular frameworks. As discussed below, nearly all of the variance in mean Jaccard similarity is determined by species' occupancy rates, and the remaining variance is determined by a simple covariance term; there is no room for any additional ecological information in the metric. By collapsing thousands of site-level comparisons into two easy-to-calculate quantities, this decomposition can provide researchers a more intuitive and less cumbersome way to study biotic homogenization at the landscape level than previous frameworks. As the behavior of landscape-level mean similarity has been fully characterized, homogenization research can now focus on other factors (discussed at the end of this paper).

Average Jaccard similarity

The Jaccard similarity between site i and j is defined as the proportion of species that are shared between them. More precisely, it is the number of species that are shared between the two sites (S_{ij}) divided by the number of species that occur in at least one (T_{ij}). The landscape-level mean Jaccard similarity value that homogenization research has largely focused on is therefore given by

$$\bar{J} = \frac{1}{\binom{n}{2}} \sum_{i \neq j} \left(\frac{S_{ij}}{T_{ij}} \right), \tag{1}$$

where n is the number of sites on the landscape (e.g. islands in an archipelago) and $\binom{n}{2}$ is the number of distinct site pairs, equal to $n(n-1)/2$. Most existing frameworks focus on tracking biotic turnover's effects on individual S_{ij} and T_{ij} values (e.g. Olden and Poff 2003, Rosenblad and Sax 2016). However, there is good reason to believe that this level of detail is not necessary for understanding $\bar{J}$ (Harris et al. 2011). To evaluate this claim, I examined 47 large data sets from the USDA PLANTS database [one for each of the contiguous US states except Maryland; USDA NRCS (2010); Appendix 1]. Independently shuffling the list of sites occupied by every species did not appreciably affect $\bar{J}$ (mean absolute deviation of 0.006 on a scale from 0 to 1; Appendix 1).

The central role of occupancy

If similarity and homogenization do not depend strongly on which sites are occupied by which species, then what does matter? In Harris et al. (2011), two colleagues and I showed (empirically and with an appeal to the law of large numbers) that average similarity depends primarily on the proportion of sites occupied by each species. Specifically, we defined an approximation to mean Jaccard similarity, J^*, given by the average value of S_{ij} divided by the average value of T_{ij}. Substituting in formulas for these two averages, we derived

$$J^* = \sum_k \binom{p_k n}{2} \Big/ \sum_k \left[\binom{n}{2} - \binom{(1-p_k)n}{2} \right], \tag{2}$$

where p_k is the proportion of sites occupied by species k.

Using this approximation, we

Despite the omission of any information at the level of individual sites or site pairs, we showed that the approximation explained 99.8% of the variance in $\bar{J}$ across the 47 USDA PLANTS data sets. It also explained an average of 98.8% of the variance in species-level effects on average similarity (Figure 1). The *blender* package (Harris 2014) for R (R Core Team 2015), introduced alongside this paper allows users to easily perform these calculations. Similar results were presented around the same time by Chase et al. (2011) and by Vergara et al. (2011).

The above equation can be simplified further using Harris et al. (2011)'s notion of "effective occupancy," denoted p^*. Effective occupancy collapses a vector of species-level occupancy values down to a single number that describes the whole community, and can be calculated as

$$p^* = \frac{J^*(2n-1)+1}{(J^*+1)n}. \tag{3}$$

By interpolating between the species' p_k values, p^* acts as a "center of gravity" for average similarity: species whose p_k values exceed p^* pull its value up, while species with smaller p_k values pull p^* down. As a result, the addition of exotic species will generally cause net differentiation until they occupy at least as many sites as their native counterparts; from then on, their spread will make the landscape more homogenous than it would have been if they hadn't invaded at all (Figure 1B). Extirpations' effects on $\bar{J}$ can be predicted in the same fashion. The existence of such a critical point has been clear to homogenization

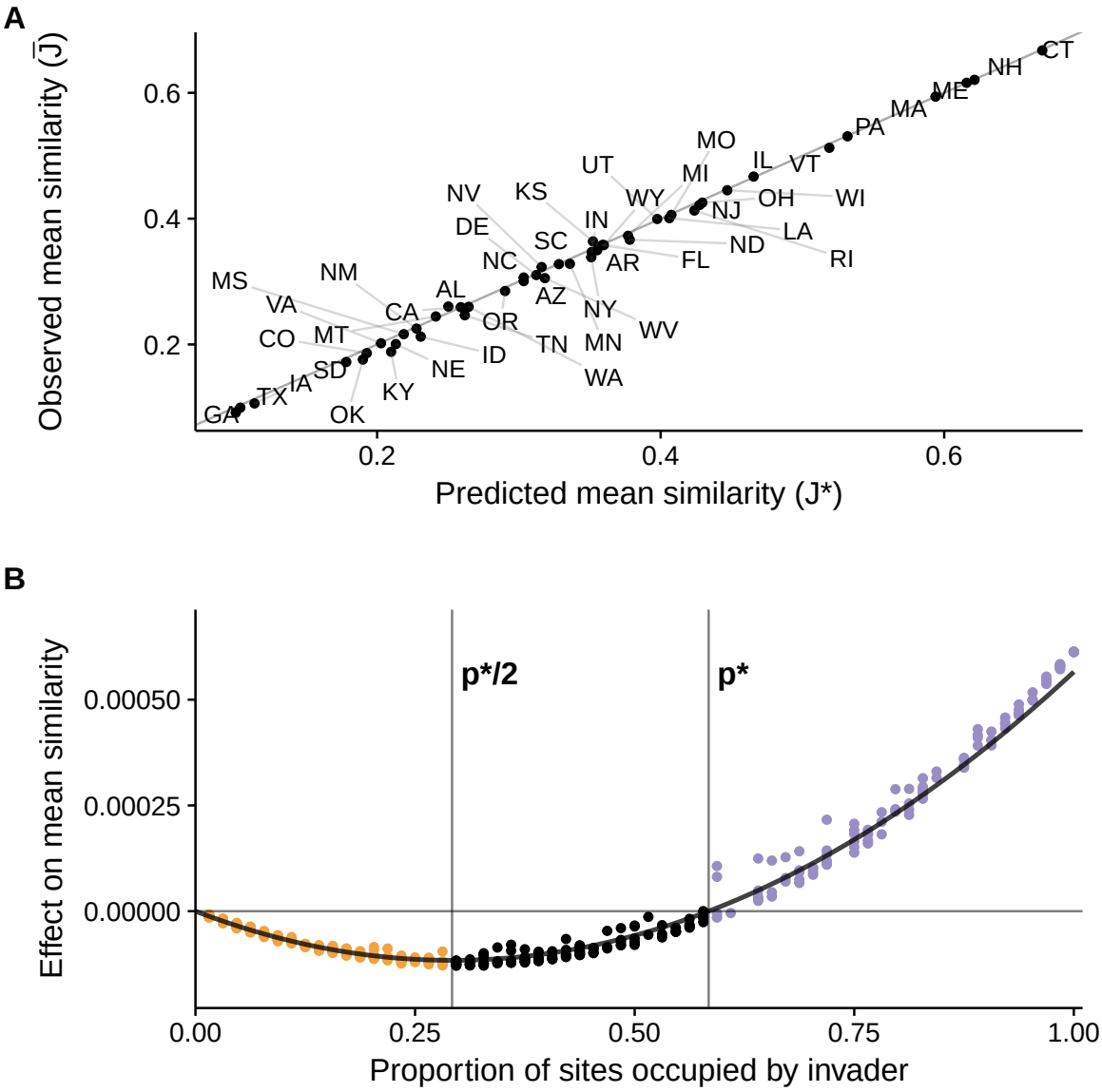

Figure 1: **A.** Across data sets from 47 US states, mean similarity is almost entirely determined by species' occupancy rates (as calculated with the J^* approximation in Equation 2). **B.** The proportion of sites occupied by different exotic species in Louisiana explains almost all of the variance in their effects on mean similarity (Harris et al. 2011). Louisiana was chosen because it had the median R^2 value of the 47 available landscapes. As a species invades and spreads across the landscape, it passes through three phases, indicated by colors. The boundaries between these phases are given by $p^*/2$ and p^*, described in Equation 3. *Phase I*: the invader is rare, and increasing its occupancy magnifies its differentiating influence. *Phase II*: spreading the invader makes it more similar to the background of native species, so its net effect approaches zero as the invader spreads. *Phase III*: the invader is more common than the background of native species, and incresaes mean similarity as it spreads.

researchers for more than a decade (McKinney 2004, Rooney et al. 2007, Rosenblad and Sax 2016), but its location has not been discussed outside of Harris et al. (2011) (nor has the location of an equally important point at $p^*/2$, shown in Figure 1).

As discussed in Harris et al. (2011), a large portion of the homogenization literature can be explained by this single value Figure 1. For example, the observed tendency for recent invaders to reduce mean similarity while more established ones increase it can be explained by these species' typical positions along the path in Figure 1B. Likewise for the differences between local range expansions versus novel introductions and for the differences among taxa with different dispersal capabilities. To the extent that occupancy generally explains more than 95% of the variance in $\bar{J}$, the residuals will be too small for other mechanisms (such as explanations based on the biological properties of the taxon being studied or the spatial scale of the observations) to play much of a role. At most, these other factors will explain the residuals of results like (Figure 1B); however these residuals can already be explained more simply, as discussed in the next section.

The role of covariance

One point that Harris et al. (2011) made but did not emphasize is that the residuals from the J^* approximation are given by an identity from Welsh et al. (1988):

$$\bar{J} = J^* - \frac{\mathrm{cov}(T_{ij}, S_{ij}/T_{ij})}{\mathrm{mean}(T_{ij})}, \tag{4}$$

where cov refers to the population covariance (rather than the more familiar sample covariance). Equation 4 shows how $\bar{J}$ can be exactly decomposed into an occupancy component and a covariance component.[1] In other words, any observed effect on $\bar{J}$ that does not act through occupancy must act through the covariance term.

In 2011, we largely disregarded the covariance component of this decomposition because it was usually small (and approaches zero as the size of the data set increases under many circumstances), but it has important consequences for the way we think about homogenization. In general, if one holds p^* constant, landscapes with high variance across species' occupancy rates will have lower mean similarity, while sites with distinct site classes will have higher mean similarity. Figure 2 shows three hypothetical landscapes with different covariances to demonstrate the latter effect. Counter-intuitively, the mean similarity metric

[1]Note that our 2011 treatment of this decomposition switched J^* and $\bar{J}$ in Appendix 2's Equation B1 and some of the subsequent discussion in that Appendix is backwards.

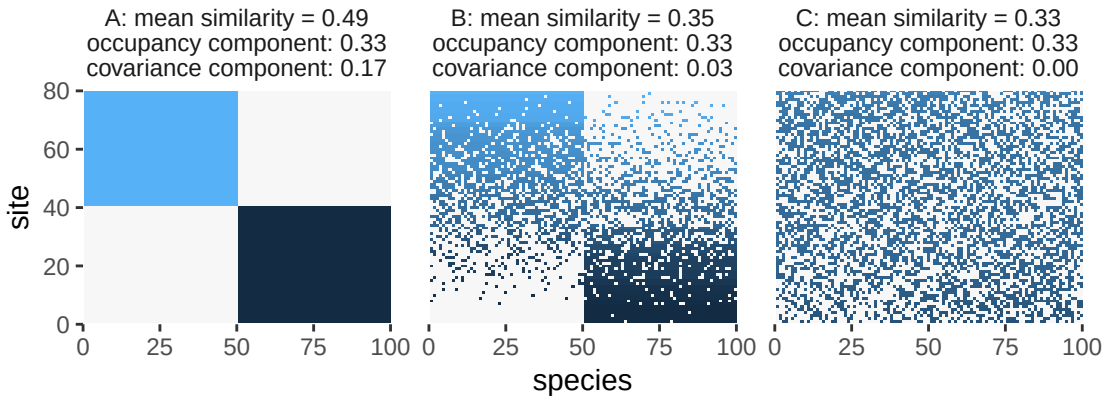

Figure 2: The effect of covariance on $\bar{J}$ (Equation 4), holding each species' occurrence rate constant at 50% and the occupancy component constant at 0.33. In **A**, there are only two kinds of communities: light blue communities that contain species 1-50 and dark blue communities that contain species 51-100. Here, the covariance between T_{ij} and S_{ij}/T_{ij} is strongly negative and covariance contributes substantially to $\bar{J}$. In **B**, the two community types are not as distinct, and the covariance effect is much smaller. In **C**, there are no distinct community types and the covariance is not distinguishable from zero. Counter-intuitively, randomly smearing species across the landscape causes anti-homogenization, in terms of mean similarity.

implies that landscapes with distinct types of communities are less homogenized than ones with randomly-arranged species. This apparent conflict deserves further scrutiny, and may indicate problems with the continued use of $\bar{J}$ to measure homogenization.

Avenues for future research

If nearly all of the variance studied in homogenization research can be explained by occupancy, and the rest can be explained by a simple covariance term, where does this leave the field? At the end of Harris et al. (2011), we listed three paths forward, each of which remains promising five years later.

1. To the extent that occupancy explains most of the variance, homogenization researchers should focus on explaining and predicting changes in species' occupancy rates. These rates already important in other areas of community ecology, and predictive methods from those fields (Pearse et al. 2013, Harris 2015) could help make better forecasts about homogenization than simpler extrapolation approaches.

2. Researchers should increase their focus on the covariance effects that cause deviations from J^*. Alternatively, researchers could focus on the observed deviations from

permutation-based null distributions (Chase 2007, Chase et al. 2011). Predictive models that account for correlations among species' occurrence probabilities (Warton et al. 2015) may also be useful for characterizing the covraiance term.

3. If $\bar{J}$ can be calculated exactly from occupancy rates and a covariance term, then a great deal of local information must be lost during the averaging process. To use that information effectively, homogenization researchers will need to think about similarity at scales below the landscape level (as they already do in analyses of spatial turnover). As the focus of homogenization research shifts toward sub-landscape scales, a deeper understanding of biotic turnover on individual pairs of sites (as provided by the frameworks of Olden and Poff (2003) and Rosenblad and Sax (2016)) will become increasingly valuable.

Finally, the results in this paper suggest one additional area that needs further exploration. In light of the potential discrepancy between $\bar{J}$'s behavior and the intuitive idea that randomly shuffling species should be a homogenizing process (Figure 2), community ecologists should consider the use of other metrics for landscape-level homogenization. In particular, ecologists should investigate whether other similarity metrics (especially abundance-based metrics) can be decomposed in a similar way, and whether metrics without this problem can be used instead.

Acknowledgements

This work was funded by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4563 to E. P. White.

References

Chase, J. M. 2007. Drought mediates the importance of stochastic community assembly. Proceedings of the National Academy of Sciences 104:17430–17434.

Chase, J. M., N. J. Kraft, K. G. Smith, M. Vellend, and B. D. Inouye. 2011. Using null models to disentangle variation in community dissimilarity from variation in a-diversity. Ecosphere 2:art24.

Harris, D. J. 2014. Blender: Analyze biotic homogenization of landscapes.

Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model.

Methods in Ecology and Evolution.

Harris, D. J., K. G. Smith, and P. J. Hanly. 2011. Occupancy is nine-tenths of the law: Occupancy rates determine the homogenizing and differentiating effects of exotic species. The American naturalist 177:535.

McKinney, M. L. 2004. Do exotics homogenize or differentiate communities? Roles of sampling and exotic species richness. Biological Invasions 6:495–504.

McKinney, M. L., and J. L. Lockwood. 1999. Biotic homogenization: A few winners replacing many losers in the next mass extinction. Trends in Ecology & Evolution 14:450–453.

Olden, J. D., and N. L. Poff. 2003. Toward a mechanistic understanding and prediction of biotic homogenization. The American Naturalist 162:442–460.

Olden, J. D., and T. P. Rooney. 2006. On defining and quantifying biotic homogenization. Global Ecology and Biogeography 15:113–120.

Pearse, I. S., D. J. Harris, R. Karban, and A. Sih. 2013. Predicting novel herbivore–plant interactions. Oikos 122:1554–1564.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rooney, T. P., J. D. Olden, M. K. Leach, and D. A. Rogers. 2007. Biotic homogenization and conservation prioritization. Biological Conservation 134:447–450.

Rosenblad, K. C., and D. F. Sax. 2016. A new framework for investigating biotic homogenization and exploring future trajectories: Oceanic island plant and bird assemblages as a case study. Ecography:n/a–n/a.

USDA NRCS. 2010. The PLANTS database, accessed 2010-11-23 from http://plants.usda.gov and retrieved from the blender package, (version 0.1.2; harris 2014). national plant data team, greensboro, NC 27401-4901 USA.

Vergara, P. M., J. Pizarro, and S. A. Castro. 2011. An island biogeography approach for understanding changes in compositional similarity at present scenario of biotic homogenization. Ecological Modelling 222:1964–1971.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. Hui. 2015. So many variables: Joint modeling in community ecology. Trends in Ecology & Evolution 30:766–779.

Welsh, A. H., A. T. Peterson, and S. A. Altmann. 1988. The fallacy of averages. The American Naturalist 132:277–288.