Mapping Post-Climate Change Biogeographical Regions with Deep Latent Variable Models

Christopher Krapu
GeoAI Group
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37830
krapucl@ornl.gov

Abstract

Forecasting future spatial patterns in biodiversity due to shifts in climate is challenging due to nonlinear interactions between species as recorded in their presence/absence data. This work proposes using variational autoencoders with side information to identify low-dimensional structure in species' joint co-occurrence patterns, leveraging this simplified representation to provide multivariate predictions of their habitat extent under future climate scenarios. We pursue a latent space clustering approach to map biogeographical regions of frequently co-occurring species and apply this methodology to a dataset from northern Belgium, generating predictive maps illustrating how these regions may expand or contract with changing temperature under a future climate scenario.

1 Introduction

Projecting trends in global biodiversity amidst anthropogenic climate change is a challenging problem due to uncertainties in climate forecasts, nonlinear relations between species' fitness and temperature, and climate-mediated network effects between species [Van der Putten et al., 2010, Garcia et al., 2018]. The disruption of established patterns of species-species interactions threatens to sever critical links required by ecosystems to function properly. For example, changes to the timing and seasonality of ecosystem events such as predation and pollination are likely to be disrupted by climate change [Bartomeus et al., 2011] with concomitant ecosystem-wide effects. Historically, these second-order effects have been challenging to accommodate in empirical species distribution models and thereby lessened the value of existing ecological forecasts which were often dominated by univariate models [Sinclair et al., 2010, Elith et al., 2011]. Unfortunately, past research has shown that omission of interactions between species can compromise the validity of species' abundance forecasts under future climate change [Mod et al., 2015]. Over the past several decades, collection of data tabulating the co-occurrence of multiple species has driven the development of empirical *joint* species distribution models. A shared commonality amongst several previous research efforts in statistical ecology [Pollock et al., 2014, Taylor-Rodríguez et al., 2017, Tikhonov et al., 2017, Krapu and Borsuk, 2020] is that these models account for species co-occurrence in a linear fashion, and thereby potentially limit the degree to which more complicated covariance structures may be analyzed.

Our interest in this work is to explore the usage of deep generative models for the purpose of joint species distribution modeling. In particular, we seek to analyze the suitability of using a variational autoencoder (VAE) for identifying meaningful low-dimensional structure from plant species presence/absence data with relatively large ($K \approx 2500$) dimension. We use this learned representation as an embedding space for application of clustering algorithms to identify *biogeographical* regions with strong species assemblage covariance patterns. Locations assigned to the same clusters in the embedding space also share strong correlation patterns in geographic space and appear to cleanly

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

represent biologically meaningful ecoregions shared requirements across species for survival and propagation. Furthermore, we correlate estimates of per-location latent variables with exogenous variables including annual rainfall and temperature, allowing us to provide forecasts for the expansion and contraction of these regions under increased temperature as supplied by the predictions from the MIROC-ESM model within the Coupled Model Intercomparison Project, version 6 (CMIP6).

We adopt the standard framework for describing variational autoencoders as reviewed in Kingma and Welling [2019] and clarify that we are interested in constructing a probability model $p(\mathbf{x}, \mathbf{z})$ over vector-valued random variables X, Z which represent, respectively, the K-dimensional binary vector $\mathbf{x}$ of species presence or absence and the corresponding latent vector $Z \sim \mathcal{N}(0, I_L)$ which is a real-valued latent representation of $\mathbf{x}$ in a low dimensional space. As is standard when working with binary data, we use a Bernoulli *decoder* model p_θ such that $p_\theta(\mathbf{x}|\mathbf{z}) = \prod_k^K \hat{p}_k^{x_k}$ and $(\hat{p}_1, ..., \hat{p}_K)^T = f_\theta(\mathbf{z})$ with $f_\theta : \mathbb{R}^L \to [0,1]^K$ implemented as a neural network with weight vector θ. Rather than jointly conduct inference for the parameters θ and the latent variables Z directly, amortized inference (with regard to $\mathbf{z}$) is employed via the *encoder* model q_ϕ to approximate the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ as $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), v_\phi(\mathbf{x}))$ with neural networks $\mu_\phi : [0,1]^K \to \mathbb{R}^L$ and $v_\phi : [0,1]^K \to \mathbb{R}^L_{>0}$ providing the approximate posterior mean and posterior variance of $\mathbf{z}|\mathbf{x}$, albeit at the extra computational expense of estimating the encoder parameters ϕ.

2 Experiments

2.1 Species presence data

The observational analyzed in this study is predominantly drawn from the Florabank1 biodiversity database [Van Landuyt et al., 2012] which records species presence and absence as catalogued from historical records of plant observations in Flanders, Belgium (Figure 1A) compiled since 1800. Most of the records in the database are derived from a systematic, grid-based survey of wild plant species. The version of the dataset used in our analyses contains data for $K = 2,448$ distinct species recorded across $N = 12,647$ spatial cells with area of one square kilometer each. The data exhibit a substantial negative imbalance, with presences reported for only 5% of the approximately 30 million potential presences. 95% of grid cells contained fewer than 288 species observed, while 95% of species were present in fewer than 4,418 grid cells.

2.2 Generative model

To analyze the Florabank1 data, we partitioned our data into subsets of 70% training, 15% validation, and 15% test. We implemented a variational autoencoder with symmetric encoder and decoder models using fully-connected layers, skip connections, batch normalization, and a sigmoid activation function for the final decoder layer. The training objective function for this model is the sum of a Gaussian KL-divergence term and class-weighted binary cross-entropy to address class imbalance. It can be written in terms of a single observation $\mathbf{x}$ and predicted presence probability vector $\hat{\mathbf{p}} = f_\theta(\mathbf{z})$ as

$$L(\mathbf{x}, \hat{\mathbf{p}}) = \beta \cdot \mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(\mu_\phi(\mathbf{x}), v_\phi(\mathbf{x}))||\mathcal{N}(0, I_L)\right) - \sum_k^K x_k \log \hat{p}_k + \lambda(1 - x_k)\log(1 - \hat{p_k}) \quad (1)$$

with weighting parameter λ to account for true/false imbalance in the dataset and parameter β controlling trade-off between Gaussian regularization and reconstruction error after Higgins et al. [2017]. To identify a suitable model structure, we performed a grid search over 1440 combinations of hyperparameters and architectures (Appendix A1), assessing model performance with Tjur's R^2 [Tjur, 2009] as applied to VAE reconstruction applied to the held-out test subset. Each model was trained until decreasing validation subset R^2 was observed for multiple consecutive epochs. We found that the reconstruction fidelity was maximized at $R^2 = 0.83$ with a latent dimension of 32, 2 layers with 256 hidden units each, $\lambda = 10$, and $\beta = 1$. We made use of this model for all following analyses in this study.

2.3 Region mapping via latent space clustering

As the goal of this work is to generate biogeographical maps of species' distributions, we used the encoder model from the previously described training procedure to encode all N observations

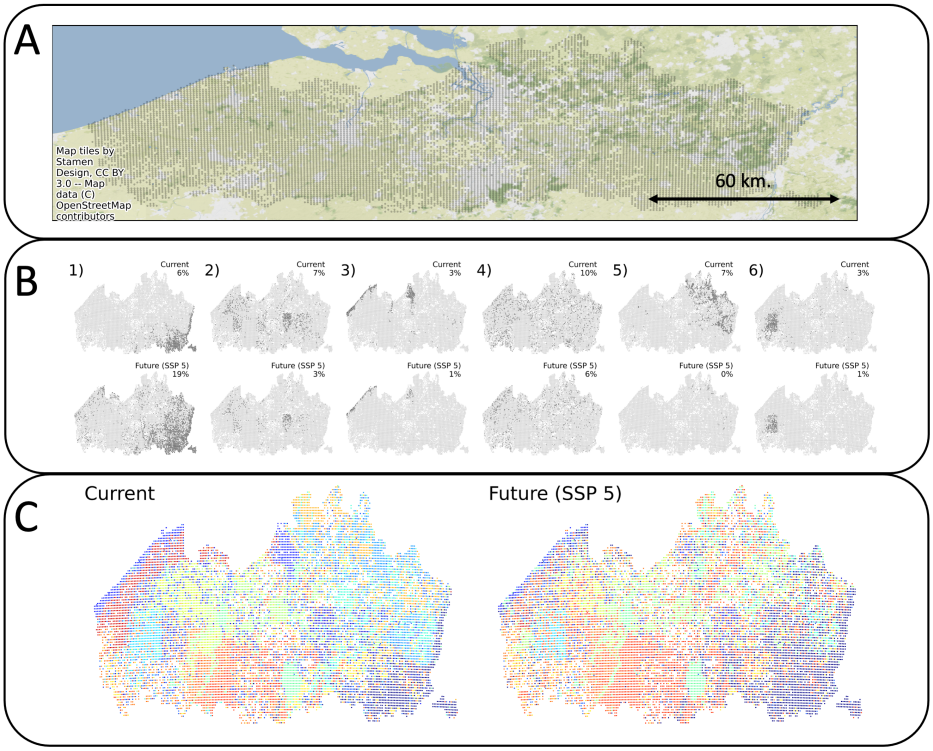

Figure 1: Biogeographical regions obtained with embeddings from the VAE latent space. Each color in the subplots above corresponds to a distinct cluster identified from the latent or embedded representation of the data. Note that colors are not aligned across subplots.

into the L-dimensional latent space such that $\mathbf{u_n} = \mu_\phi(\mathbf{z_n})$ by applying K-means clustering as implemented in `scikit-learn` [Pedregosa et al., 2011] with $M = 15$ clusters to the $N \times L$ matrix $U = [\mathbf{u_1}^T, ..., \mathbf{u_n}^T]$. This choice of M clusters was chosen subjectively; currently, no systematic comparison has yet been performed to identify an optimal number of clusters; future iterations of this work could investigate nonparametric clustering techniques making use of the Dirichlet process mixture model [Ferguson, 1973] or the silhouette score [Rousseeuw, 1987] to identify an appropriate value. Maps indicating the cluster assignment for each point in geographical space are provided in Figure 1C; for comparison, similar mappings were generated using UMAP [McInnes et al., 2018] and latent Dirichlet allocation [Blei et al., 2003] to yield embedded representations. These results are provided and described in Appendix A2.

To relate the latent space clustering to actual collections of co-occurring species, we calculated, for each cluster centroid in the latent space, the resulting species presence probability vector implied by the decoder model. The top occurring species for each cluster are shown in Table 1 for three clusters. There appears to be at least a modest degree of easily understandable co-occurrence structure in these clusters; for example, the third cluster/region rates multiple plants as lowest scoring which typically only appear in a sandy coastal environment such as *Armeria maritima* and *Suaeda maritima*.

2.4 Future projections under climate change

To predict how anticipated alterations to long-term averages in temperature might affect the spatial distribution of biographical regions, we used linear regression with ridge regularization to obtain a decomposition of $U = YB + W$ where Y denotes a $N \times P$ matrix of P side information variables or covariates observed at N locations, B is a $P \times L$ matrix of regression coefficients linking coordinates in the embedding space to the covariates, and W contains residual terms representing variation in the latent embeddings which cannot be captured with this linear model. The covariates we used include mean annual temperature, rainfall, elevation, slope, density of human settlement, and indicators

Table 1: High- and low-probability species for three biogeographic region clusters

Region 1		Region 2		Region 3	
High probability	**Low probability**	**High probability**	**Low probability**	**High probability**	**Low probability**
Crataegus macrocarpa	Juncus tenageia	Botrychium matricariifolium	Mentha villosa	Equisetum sylvaticum	Atriplex laciniata
Mentha rotundifolia	Carex reichenbachii	Ranunculus acris	Euphorbia prostrata	Myosotis nemorosa	Diphasiastrum tristachyum
Scabiosa columbaria	Andromeda polifolia	Rumex acetosella	Prunus persica	Geranium versicolor	Malva pusilla
Eranthis hyemalis	Myriophyllum heterophyllum	Cynosurus cristatus	Ranunculus tripartitus	Carex strigosa	Ammophila arenaria
Amaranthus graecizans	Littorella uniflora	Cardamine pratensis	Campanula poscharskyana	Polystichum aculeatum	Armeria maritima
Anthriscus cerefolium	Rhynchospora fusca	Anthoxanthum odoratum	Digitalis lutea	Dactylis polygama	Euphorbia paralias
Rubus lindleianus	Leucojum aestivum	Vicia lutea	Veronica spicata	Helleborus viridis	Vicia faba
Rubus vestitus	Wahlenbergia hederacea	Hieracium umbellatum	Leucojum aestivum	Gagea spathacea	Suaeda maritima
Rubus rudis	Euonymus fortunei	Campanula rapunculus	Cosmos bipinnatus	Anemone ranunculoides	Wahlenbergia hederacea
Cydonia oblonga	Rhododendron luteum	Centaurea jacea	Corylus ma	Luzula forsteri	Thlaspi caerulescens

for the presence of surface water. Further details are provided in Appendix A3; we note that the fraction of variance in the latent embeddings explained by the covariates is approximately 8%. We then constructed a future covariate matrix Y_f by replacing current mean annual temperature values for the study domain with forecasts of the same quantity for the period 2021-2040 under the SSP5 scenario as simulated via the MIROC-ESM model [Watanabe et al., 2011]. This scenario projects increases of roughly 1.0° to 2.0° C. across the spatial extent of the Florabank1 database. We then calculated the future $U_f = Y_f B + W$ and then classified the new values in U_f with the pre-existing K-means clustering algorithm, thereby assigning the new embedding points to clusters identified with the current-day data. Maps indicating the extent of each biogeographical region for current-day settings as well as the future scenario are shown in Figure 1C.

2.5 Limitations

As a report on our intermediate findings, we note several shortcomings to be addressed in future iterations of this work. Our choice of the standard unit Gaussian prior distribution for the VAE latent space is not well-suited to clustering as it does not favor clear demarcations between groupings of points. We would like to use a more suitable prior distribution after Dilokthanakul et al. [2017] which would also allow for clustering within the VAE framework. Additionally, a small portion of the variation in latent space embeddings is captured by the environmental predictor variables, suggesting that we should explore using a higher-capacity model for this secondary prediction task and/or obtain additional covariates.

3 Discussion and closing remarks

As indicated in Figure 1B, the effects of changing temperature are varied for the different biogeographical regions. The pre- and post-change spatial patterns of region label retain their rough overall relative position in the spatial domain, but exhibit substantial expansion or contraction. For example, we see that Region 5 shrinks from having a substantial presence over north-central Flanders to only a small remaining region on its northern border. Conversely, Region 1 grows sub substantially from a foothold in the southeast corner of the region to encompass much of the eastern part of Flanders.

As our predictions for the spatial extent of different biogeographical regions is conducted primarily through adjust covariates in the second-stage linear model, the quality of the generated forecasts is highly dependent upon the the representative properties of the predictor variables with regard to the true underlying biological processes. It is virtually certain that, in at least some of these predictions, variations in temperature translate into shifts in regional extent primarily because temperature covaries strongly with unobserved quantities such as hillshade prevalence or soil type which are the actual drivers of biodiversity are unlikely to change as dramatically with climate change. Further effort in this line of analysis must be directed towards including a comprehensive set of explanatory variables for this purpose.

In light of the challenges posed by this type of analysis, we are greatly encouraged by finding spatially cohesive and clearly meaningful biogeographical regions automatically from data using an unsupervised approach, and we anticipate that further research in this direction will aid researchers in identifying species co-occurrence patterns that extend beyond simple linear, pairwise interactions. We anticipate that analyses similar to those presented in this work will become important for analyzing large biodiversity datasets which are now becoming more common in ecology. We also opine that

such ecological data is fertile ground for methodological advances in machine learning due to its geospatial, high-dimensional and semistructured nature.

Acknowledgments and Disclosure of Funding

We acknowledge that this manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doepublic-access-plan). We would like to thank Philipe Ambrosio Diaz for helpful comments during the development of this work. This research was supported by Nvidia via a research GPU grant.

References

Ignasi Bartomeus, John S. Ascher, David Wagner, Bryan N. Danforth, Sheila Colla, Sarah Kornbluth, and Rachael Winfree. Climate-associated phenological advances in bee pollinators and bee-pollinated plants. *Proceedings of the National Academy of Sciences*, 108(51):20645–20649, December 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1115559108. URL `https://www.pnas.org/content/108/51/20645`. Publisher: National Academy of Sciences Section: Biological Sciences.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. URL `http://dl.acm.org/citation.cfm?id=944937`.

George Buttner, Jan Feranec, Gabriel Jaffrain, László Mari, Gergely Maucha, and Tomas Soukup. The Corine Land Cover 2000 Project. In *EARSeL eProceedings*, volume 3, page 16, 2004.

Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders. *arXiv:1611.02648 [cs, stat]*, January 2017. URL `http://arxiv.org/abs/1611.02648`. arXiv: 1611.02648.

Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. A statistical explanation of MaxEnt for ecologists: Statistical explanation of MaxEnt. *Diversity and Distributions*, 17(1):43–57, January 2011. ISSN 13669516. doi: 10.1111/j.1472-4642.2010.00725.x. URL `http://doi.wiley.com/10.1111/j.1472-4642.2010.00725.x`.

European Environment Agency. High Resolution Layer: Water & Wetness Probability Index (WWPI) 2015, 2018. URL `https://www.eea.europa.eu/ds_resolveuid/b60e5329e9d142fdb131eacb61464ca9`.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, March 1973. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176342360. URL `https://projecteuclid.org/journals/annals-of-statistics/volume-1/issue-2/A-Bayesian-Analysis-of-Some-Nonparametric-Problems/10.1214/aos/1176342360.full`. Publisher: Institute of Mathematical Statistics.

Francisca C. Garcia, Elvire Bestion, Ruth Warfield, and Gabriel Yvon-Durocher. Changes in temperature alter the relationship between biodiversity and ecosystem functioning. *Proceedings of the National Academy of Sciences*, 115(43):10989–10994, October 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1805518115. URL `https://www.pnas.org/content/115/43/10989`. Publisher: National Academy of Sciences Section: Biological Sciences.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. page 13, 2017.

Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000056. URL http://arxiv.org/abs/1906.02691. arXiv: 1906.02691.

Christopher Krapu and Mark Borsuk. A spatial community regression approach to exploratory analysis of ecological data. *Methods in Ecology and Evolution*, 11(5):608–620, 2020. ISSN 2041-210X. doi: 10.1111/2041-210X.13371. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13371. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13371.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, September 2018. ISSN 2475-9066. doi: 10.21105/joss.00861. URL https://joss.theoj.org/papers/10.21105/joss.00861.

Heidi K. Mod, Peter C. le Roux, Antoine Guisan, and Miska Luoto. Biotic interactions boost spatial models of species richness. *Ecography*, 38(9):913–921, September 2015. ISSN 09067590. doi: 10.1111/ecog.01129. URL http://doi.wiley.com/10.1111/ecog.01129.

NASA JPL. NASA Shuttle Radar Topography Mission Global 3 arc second, 2013. URL https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL3.003.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Martino Pesaresi and Sergio Freire. GHS settlement grid, following the REGIO model 2014 in application to GHSL Landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015), 2016. URL http://data.europa.eu/89h/jrc-ghsl-ghs_smod_pop_globe_r2016a.

Laura J. Pollock, Reid Tingley, William K. Morris, Nick Golding, Robert B. O'Hara, Kirsten M. Parris, Peter A. Vesk, and Michael A. McCarthy. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406, May 2014. ISSN 2041210X. doi: 10.1111/2041-210X.12180. URL http://doi.wiley.com/10.1111/2041-210X.12180.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Steve Sinclair, Matthew White, and Graeme Newell. How Useful Are Species Distribution Models for Managing Biodiversity under Future Climates? *Ecology and Society*, 15(1), February 2010. ISSN 1708-3087. doi: 10.5751/ES-03089-150108. URL https://www.ecologyandsociety.org/vol15/iss1/art8/. Publisher: The Resilience Alliance.

Daniel Taylor-Rodríguez, Kimberly Kaufeld, Erin M. Schliep, James S. Clark, and Alan E. Gelfand. Joint Species Distribution Modeling: Dimension Reduction Using Dirichlet Processes. *Bayesian Analysis*, 12(4):939–967, December 2017. ISSN 1936-0975, 1931-6690. doi: 10.1214/16-BA1031. URL https://projecteuclid.org/journals/bayesian-analysis/volume-12/issue-4/Joint-Species-Distribution-Modeling-Dimension-Reduction-Using-Dirichlet-Processes/10.1214/16-BA1031.full. Publisher: International Society for Bayesian Analysis.

Gleb Tikhonov, Nerea Abrego, David Dunson, and Otso Ovaskainen. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, 8(4):443–452, April 2017. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.12723. URL https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12723.

Tue Tjur. Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *The American Statistician*, 63(4):366–372, November 2009. ISSN 0003-1305, 1537-2731. doi: 10.1198/tast.2009.08210. URL http://www.tandfonline.com/doi/abs/10.1198/tast.2009.08210.

Wim H. Van der Putten, Mirka Macel, and Marcel E. Visser. Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1549): 2025–2034, July 2010. doi: 10.1098/rstb.2010.0037. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2010.0037. Publisher: Royal Society.

Wouter Van Landuyt, Leo Vanhecke, and Dimitri Brosens. Florabank1: a grid-based database on vascular plant distribution in the northern part of Belgium (Flanders and the Brussels Capital region). *PhytoKeys*, 12(0):59–67, May 2012. ISSN 1314-2003, 1314-2011. doi: 10.3897/phytokeys.12.2849. URL http://www.pensoft.net/journals/phytokeys/article/2849/abstract/florabank1-a-grid-based-database-on-vascular-plant-distribution-in-the-northern-part-of-belgiu

Jeffrey Verbeurgt, Michel Van Camp, Cornelis Stal, Thierry Camelbeeck, Lars De Sloover, Hans Poppe, Pierre-Yves Declercq, Pierre Voet, Denis Constales, Peter Troch, Philippe De Maeyer, and Alain De Wulf. The gravity database for Belgium. *Geoscience Data Journal*, 6(2):116–125, 2019. ISSN 2049-6060. doi: 10.1002/gdj3.74. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/gdj3.74. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/gdj3.74.

S. Watanabe, T. Hajima, K. Sudo, T. Nagashima, T. Takemura, H. Okajima, T. Nozawa, H. Kawase, M. Abe, T. Yokohata, T. Ise, H. Sato, E. Kato, K. Takata, S. Emori, and M. Kawamiya. MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, 4(4):845–872, October 2011. ISSN 1991-959X. doi: 10.5194/gmd-4-845-2011. URL https://gmd.copernicus.org/articles/4/845/2011/. Publisher: Copernicus GmbH.

Appendix

A1. Hyperparameter search

In creating our variational autoencoder, we performed an exhaustive search over the following settings: latent dimension - $\{8, 12, 16, 24, 32\}$, number of hidden units - $\{128, 256, 512, 1024\}$, β - $\{1.0, 2.0, 5.0\}$, λ - $\{1.0, 2.0, 5.0, 10.0\}$, number of layers - $\{1, 2, 3, 4, 6, 8\}$. The parameter settings maximizing validation Tjur's R^2 were 32, 1.0, 10.0, and 2 respectively. This model had 1,282,512 parameters and was trained to completion in 55 epochs requiring approximately 50 seconds of training on an Nvidia Titan X GPU with Adam ($\eta = 0.01$). These models were implemented in TensorFlow 2.4.1.

A2. Comparative embedding results

As a visual comparison with the latent space embeddings obtained with the variational autoencoder described in the main body of the text, we also applied K-means clustering to low-dimensional representations of the original dataset obtained using UMAP and latent Dirichlet allocation (LDA). These results are presented in Figure 2. We used the implementation of UMAP from the `umap-learn` Python package (version 0.5.1) and the online variational Bayes implementation of LDA posterior inference from `scikit-learn`. In each case, 15 clusters were used. Both UMAP and LDA, we used an embedding space with 32 dimensions or topics respectively, analogous to the latent dimensions of the VAE.

A3. Linear model covariates

We employed the covariates for our second-stage linear model from multiple sources of data regarding elevation, surface water, human settlement, and geology. We obtained gravimetric data from the Royal Observatory of Belgium's website, tabulating measurements of Earth's gravitational constant

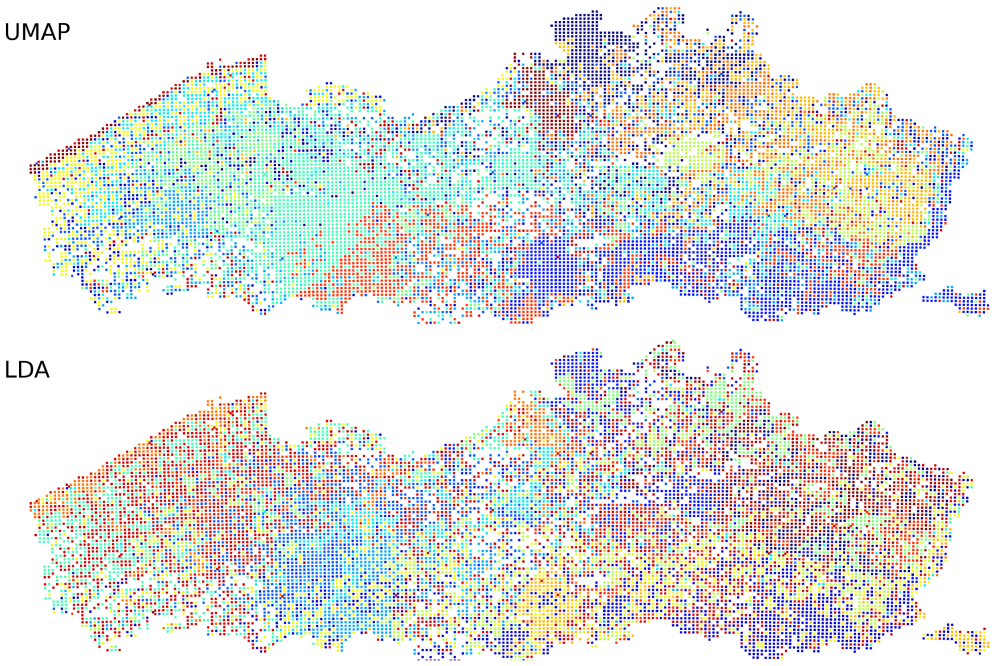

Figure 2: Biogeographical regions obtained with clusters calculated with embeddings from UMAP and LDA. Each color in the subplots above corresponds to a distinct cluster identified from the latent or embedded representation of the data. Note that colors are not aligned across subplots.

g at approximately 58,000 locations in Belgium [Verbeurgt et al., 2019]. To assign these values to our grid cell-indexed locations, we used nearest-neighbor interpolation. This data was used to provide a continuous measure of subsurface geology over the study region. We also incorporated information about the average elevation and slope for each grid cell as obtained from the Shuttle Radar Topography Mission [NASA JPL, 2013]. We incorporated two sources of information regarding human settlement; the first is the Global Human Settlement Layer [Pesaresi and Freire, 2016] which categorizes every square kilometer into an ordinal scale of 1, 2, and 3 for rural, urban, and dense urban center respectively. We used this ordinal scale without modification. To augment this data, we also used the Corine land use/land cover layer for surface imperviousness. [Buttner et al., 2004]. As availability of moisture and surface water plays a major role in determining ecological niches available to plant species, we used the wetness probability index from the Copernicus Land Monitoring Service [European Environment Agency, 2018] which assigns each location a score of 0–100 corresponding to the pixel's estimated wetness. We discretized this dataset into values indicating that the original value falls within one of three intervals: 10–30, 30–100 and 100. With the last category corresponding to permanent water. For all variables used, we computed the zonal mean with regard to the individual rasters. Each Florabank1 grid cell was used as a zone and the averaging was conducted over each raster cell which the grid cell touched.