TreeSp2Vec: a method for developing tree species embeddings using deep neural networks

Miguel A. González-Rodríguez (miguel.an.gon.ro@gmail.com)
Universidade de Santiago de Compostela https://orcid.org/0000-0002-0072-3124
Ulises Diéguez-Aranda
Universidade de Santiago de Compostela

Research Article

Keywords: species classification, national forest inventory, representation learning, multi–layer perceptron, artificial intelligence

Posted Date: March 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-957638/v3

TreeSp2Vec: a method for developing tree species embeddings using deep neural networks

González-Rodríguez, M.A.[a,*], Diéguez-Aranda, U.[b]

[a] *Guangzhou Institute of Geography, Guangdong Academy of Sciences, Guangzhou 510070, China*
[b] *Unidade de Xestión Ambiental e Forestal Sostible, Departamento de Enxeñaría Agroforestal, Universidade de Santiago de Compostela. Escola Politécnica Superior de Enxeñaría, R/ Benigno Ledo, Campus Terra, 27002 Lugo, Spain*

Abstract

In recent years, Representation Learning (RL), a subdiscipline of artificial intelligence, has proved a useful resource in many research fields for effectively mapping abstract categories into numeric scales. In this study, we explored a novel application of RL to forest ecology, labeled TreeSp2Vec, for developing tree species numeric representations using deep neural networks and data from a National Forest Inventory. Our approach consisted of a supervised species classification of individual trees using as input a set phytocentric and geocentric variables derived from forest inventory data and environmental cartography. Among the tested neural network architectures, a multi-layer perceptron with two hidden layers of 1024 units and an embedding layer of 16 units provided the best apparent and test performances (Matthew's Correlation Coefficient = 0.89). The developed latent representations ($\mathbf{W}$), or embeddings, were evaluated intrinsically by estimating their correlations with supplementary species descriptors that were not included in the training dataset. The evaluation analysis revealed some significant associations that proved the generality of the embedding model. Some latent dimensions (e.g., W_6 and W_{16}) were useful for differentiating species general features, such as conifers vs broad-leaved species, while other dimensions (e.g., W_2 and W_5) were related to forest ecosystem characteristics such as competition intensity (relative spacing index) and biodiversity (Simpson index). We concluded that the developed embeddings provided accurate and generalizable numeric representations of the considered tree species, which can be used as a ground for further cutting-edge forest modelling approaches

and open a new range of artificial intelligence applications in ecology.
Keywords: species classification, national forest inventory, representation learning, multi–layer perceptron, artificial intelligence

1. Introduction

Over the last decade, deep neural networks have revolutionized the areas of predictive and descriptive modelling (Sejnowski, 2018). Recent innovations in network architectures have paved the path for new and fruitful scientific disciplines (e.g., Computer Vision), frequently included in the broader scope of artificial intelligence. One of the most notable and recent milestones in the field of deep learning is the emergence of Representation Learning discipline (RL, Goodfellow et al., 2013; Huang et al., 2014). The latter focuses primarily on the mapping of qualitative objects into numeric scales, so abstract and hard to quantify differences between objects (i.e., "semantic" distances) can be expressed as algebraic distances. The resulting quantitative dimensions from this mapping procedure are commonly referred to as "embeddings" and the vector space defined by them is called "latent" space (Yu et al., 2013). From a practical modelling perspective, the most relevant implication of RL is the possibility of transforming or encoding categorical variables into ensembles of continuous variables. One of the first and most notorious applications of RL was the embedding of English words into a 100D latent space with the Word2Vec methodology (Mikolov et al., 2013), which is the backbone of current Natural Language Processing research. A diverse range of applications of RL have been developed in the recent years, such as sentence embedding (Tang et al., 2013), biomedical notes embedding (Wang et al., 2018), semantic-based image embedding (Irtaza et al., 2014) and embedding of states in dynamic systems (Lesort et al., 2018; Gelada et al., 2019).

Despite the recent successful uses of RL in several disciplines, applications in ecology and environmental sciences are still scarce or non-existent. The potential benefits of applying

*Corresponding author
Email addresses: miguelangel.gonzalez.rodriguez@rai.usc.es (González-Rodríguez, M.A.), ulises.dieguez@usc.es (Diéguez-Aranda, U.)

RL in ecology are varied and have to do with the ability of embedding models to transform usual ecological categories, such as species, populations, and communities, into quantitative variables. On the one hand, this transformation might allow to overcome frequent information bottlenecks in ecological research derived from the combination of categorical and numeric data. For instance, the projection of individuals belonging to a community into a latent space might allow to develop complex quantitative biodiversity indexes that integrate both categorical information (e.g., species and sociological classes) and quantitative variables (e.g., individual morphology and environmental conditions), which could be a crucial upgrade regarding the study of biodiversity. Another example could be to streamline the study of dissimilarity between populations, species and communities by turning it into a continuous problem where differences between classes could be represented by algebraic distances in the latent space. On the other hand, transforming ecological categories into quantitative variables might open a new range of high generality ecological modelling applications. For instance, the representation of different species in a latent space might allow to develop multi-species modelling approaches, able to describe ecological dynamics of many species simultaneously while keeping high predictive performance.

In the current study, we introduce a novel application of RL on forest ecology: an embedding model of tree species, labeled "TreeSp2Vec", based on deep neural networks. As a case study, we applied our modelling approach to the most relevant forest species in Spain using national forest inventory data and available environmental cartography.

2. Methods

2.1. Data sources and preprocessing

We used tree-level data of the 40 most frequent species in the Third Spanish National Forest Inventory, accounting to a total of 50K inventory plots and approximately 850K trees. The dataset encompassed a relatively wide range of forest types, including native secondary forests, afforestations with native species and commercial plantations of exotic species, existing in different mediterranean and eurosiberian ecoregions within the country.

For details about characteristics and methodology of the Spanish national forest inventory see the explanation by Alberdi et al. (2010). From this dataset, 44 "phytocentric" descriptors were computed, including: i) the diameter at breast height (cm) and total height (m) of each tree, ii) the arithmetic means of diameters and heights of all the trees in each plot, and iii) the count of the rest of trees in each plot for every considered species (i.e. 40 variables resulting from the sum of expanded one-hot encoded vectors). Additionally, we derived 26 quantitative "geocentric" descriptors from available cartography, including: i) physical and soil attributes obtained from the European Soil Data Centre (Ballabio et al., 2016, 2019), ii) physiography variables derived from a Digital Elevation Model of the National Geographical Institute of Spain (MDT200 <2009> CC-BY 4.0 ign.es), and iii) climate variables from the Wordlclim 2.1 dataset (Fick and Hijmans, 2017).

2.2. Embedding approach and model training

A common approach for learning representations is to transform the problem into a supervised learning task (Yang et al., 2015). Under this setup, the latent space is defined by a layer of intermediate transformations necessary for obtaining the final target. In the present study, the proposed encoding strategy was to perform a multi-label classification for identifying the species of every tree in the dataset as a function of the phytocentric and geocentric descriptors. Being u the number of descriptors and v the number of species considered (i.e., the input and output dimension, respectively), and w the dimension of the latent space, the embedding procedure consisted of an "encoder" map $F : \mathbb{R}^u \rightarrow \mathbb{R}^w$, that transforms an input sample into an embedded representation, and a "decoder" map $G : \mathbb{R}^w \rightarrow [0, 1]^v$, that translates the representation into a boolean target. Two types of embedding could be derived from this approach. On the one hand, the output of the encoder function F (hereafter, $\mathbf{E}$) is a sample-specific or *tree-level* embedding with shape $n \times w$, being n the number of samples in the dataset. On the other hand, the set of weight parameters in G with shape $v \times w$ (herafter, $\mathbf{W}$) represents a class-specific or *species-level* embedding.

Both transformations, F and G, were implemented using deep neural networks (Figure 1). The F map comprised a series of fuly connected layers including the embedding layer ($\mathbb{R}^w$) as

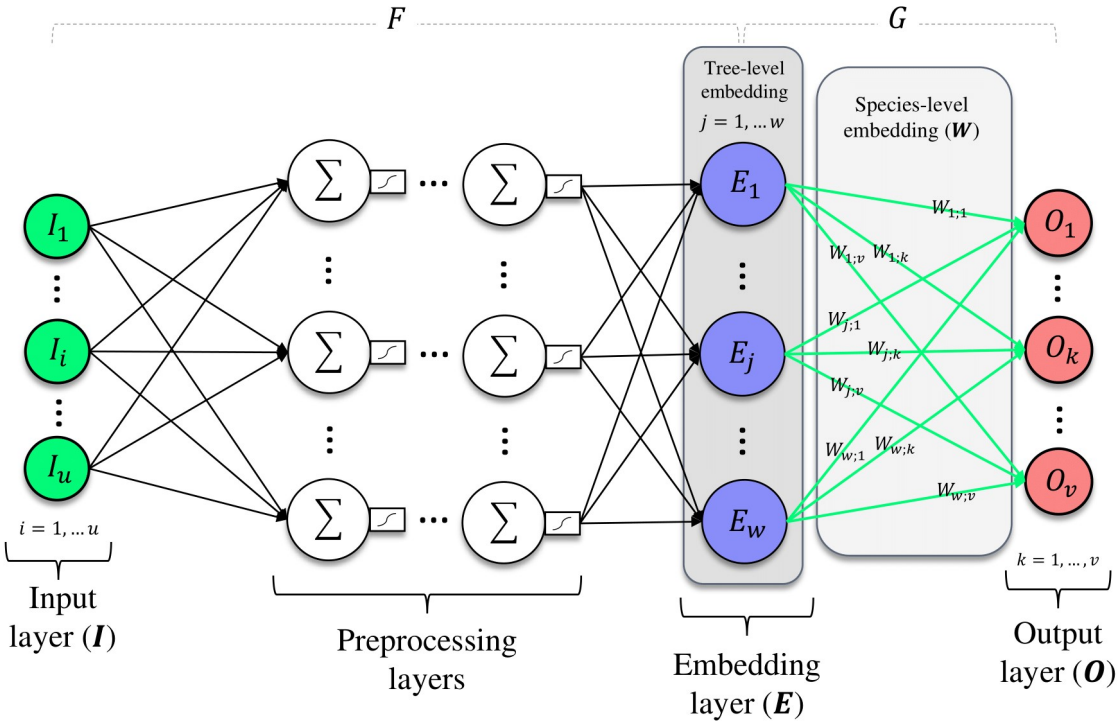

Figure 1: Diagram of the embedding approach using a deep neural network.

well as preprocessing layers with hyperbolic tangent activations for feature extraction. With regard to G, the embedding layer was fully connected to a v-dimensional output layer with softmax activation. We used the Keras API (Chollet et al., 2015) for Python 3 and GPU acceleration for optimizing different architectural variants with one to three preprocessing layers of 64-1024 units and an embedding layer with four to 32 units (scaling in powers of two). The different model variants were fitted using an Adam optimizer for minimizing the categorical cross-entropy loss over a maximum of 5000 epochs. Finally, the trained models were evaluated using a test subset of 20% of the initial dataset basing on the Matthew's Correlation Coefficient (MCC) implemented in the *scikit-learn* API (Pedregosa et al., 2011).

2.3. Evaluating the embeddings

To test the generality of the latent spaces developed, the ability of the model for inferring species and forest unseen characteristics was evaluated. We did so by checking potential associations between embeddings and supplementary species attributes that were not explicitly

included as inputs. Specifically, point-biserial and Spearman correlations with latent dimensions were estimated for three different groups of variables (hereafter, "test descriptors"):

- a set of qualitative descriptors encoded as binary variables representing whether a species is a conifer (conifer = 1, broad-leaved = 0), deciduous (deciduous = 1, evergreen = 0), native (native = 1, non-native = 0), invader (invader = 1, non-invader = 0), or of "commercial interest" (according to recent timber production statistics in Spain);
- the relative frequencies of occurrence of each species in the eight ecoregions of Spain, according to the TEOW regionalization (Olson et al., 2001); and
- the by-species mean values of a set of forest stand variables calculated for each inventory plot (e.g., basal area).

In addition, we also evaluated the embeddings subjectively through visualization. To further evaluate and illustrate the usability of the embeddings, an example of the method for developing multi-species predictive models is presented in Suplemental Material.

3. Results

3.1. Model building

The trained models achieved high classification performances, being the minimum test MCC score above 0.80. The more robust architecture in terms of classification performance (apparent MCC = 0.89 and test MCC = 0.87) had two hidden preprocessing layers of 1024 units and 16 embeddings (1024:1024:16). The models including a high number of parameters, such as the architectures with three preprocessing layers or 32 embedding units, yielded very similar yet slightly lower accuracy (see Table 1). Consequently, the model with two hidden layers of 1024 units and 16 embeddings (1.14×10^6 parameters), hereafter labeled "TreeSp2Vec", was selected as the best alternative for species representation learning. Some variations in performance across the set of species for this model were noticed (see the confusion matrix in Figure 2), being the minimum MCC = 0.66 for *Populus tremula* L. and the

maximum MCC = 0.98 for *Populus x canadensis* Moench. However, no consistent relationship was found between classification performance and species proportions in the dataset, since some very infrequent species, such as *Abies pinsapo* Boiss. or *Chamaecyparis lawsoniana* (A. Murray bis) Parl., were accurately classified (MCC = 0.96 and 0.94, respectively), while some relatively frequent species, such as *Quercus faginea* Lam., showed poorer performance (MCC = 0.78). The apparent MCC for every species is shown in Appendix A (Table A.1).

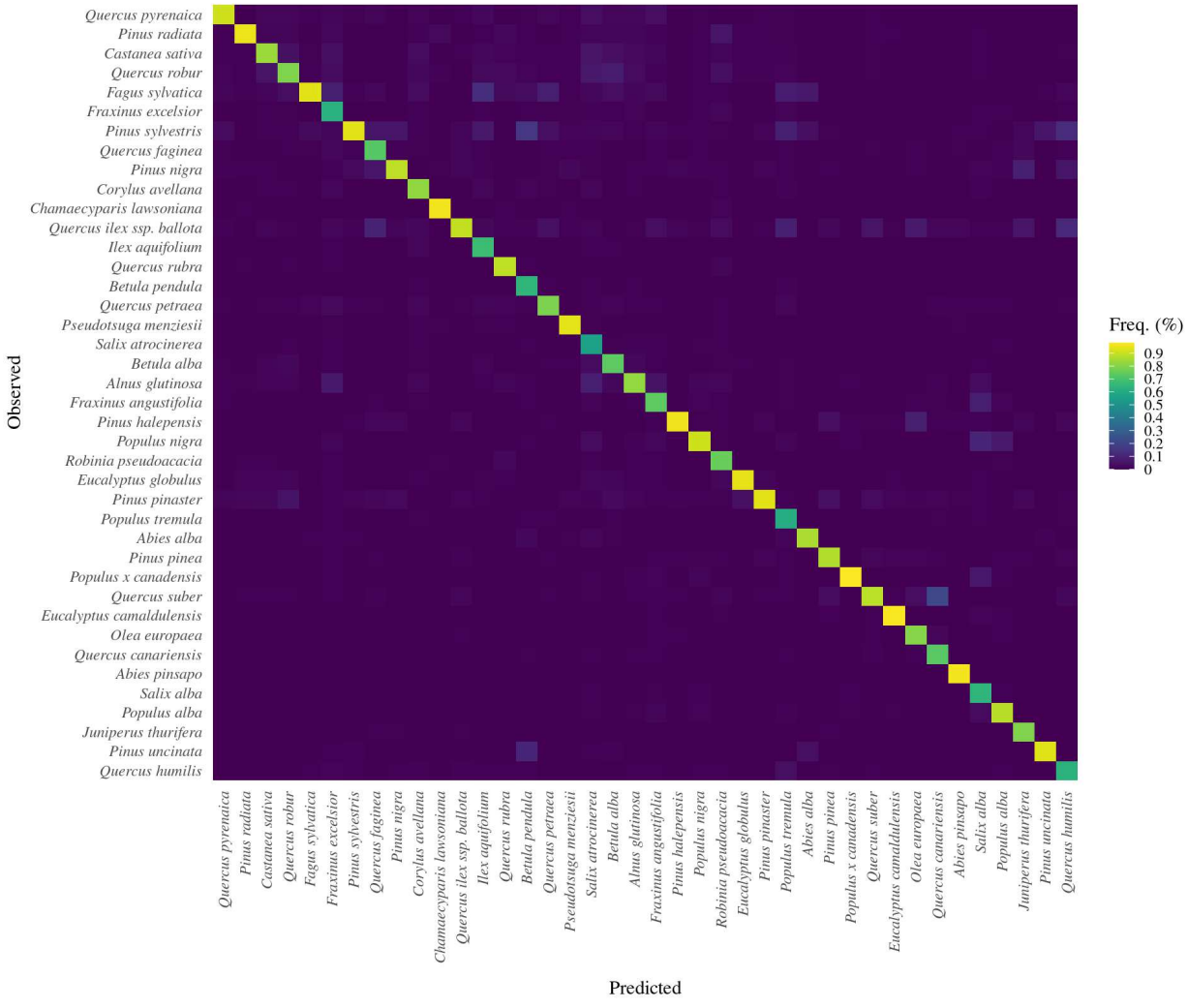

Figure 2: Confusion matrix (normalized frequencies) of the multi-label classification task using the TreeSp2Vec model for the 40 species considered.

Architecture	n° par.	MCC	MCC_{test}
1024:1024:16	1.14×10^6	0.892	0.871
1024:1024:32	1.16×10^6	0.889	0.870
1024:1024:8	1.13×10^6	0.890	0.869
1024:1024:1024:32	2.21×10^6	0.887	0.868
1024:1024:1024:16	2.19×10^6	0.885	0.866
512:512:32	3.17×10^5	0.887	0.866
512:512:512:32	5.79×10^5	0.885	0.865
1024:32	1.07×10^5	0.881	0.864
512:512:512:16	5.71×10^5	0.883	0.863
512:512:16	3.08×10^5	0.883	0.863

Table 1: Summary of MCC statistics and number of parameters in the ten best model architectures tested.

3.2. Embedding evaluation

The association analysis between the resulting species-level embeddings and the test descriptors (Figure 3) showed some noticeable correlations, for a confidence level of 90%, with the exceptions of W_1 and W_9. Concerning the species qualitative attributes, the latent dimensions W_6 and W_{16} showed intense correlations both with $Conifer$ and $Deciduous$ features, while the dimensions W_4 and W_5 were respectively correlated with $Native$ and $Invader$ attributes. W_2 and W_4 were strongly related to the commercial vs non-commercial dichotomy. Regarding the species presence in the TEOW ecoregions, correlations with latent dimensions were found for six out of the eight regions considered. Some of the strongest correlations were found between W_5 and the presence in the "Pyrenees conifer and mixed forests", between W_4 and W_{10} and the "Iberian sclerophyllous and semi-deciduous forests" and between W_{10} and the "Iberian conifer forests". The six stand variables were significantly associated with species embeddings. This was especially noticeable for the stand density, which was correlated with four latent dimensions. The two variables linked to α-diversity (species richness and the Simpson index), were similarly correlated with the dimensions W_2,

and W_6. W_{15} was strongly correlated with the dominant height (m) and the relative spacing (also named Hart-Becking index).

The visualization of latent dimensions also allowed to detect some patterns that made species easily distinguishable, at least, for some dimensions and species. A plot of tree-level embeddings of latent dimensions E_6 and E_{15} is shown in (Figure 4).

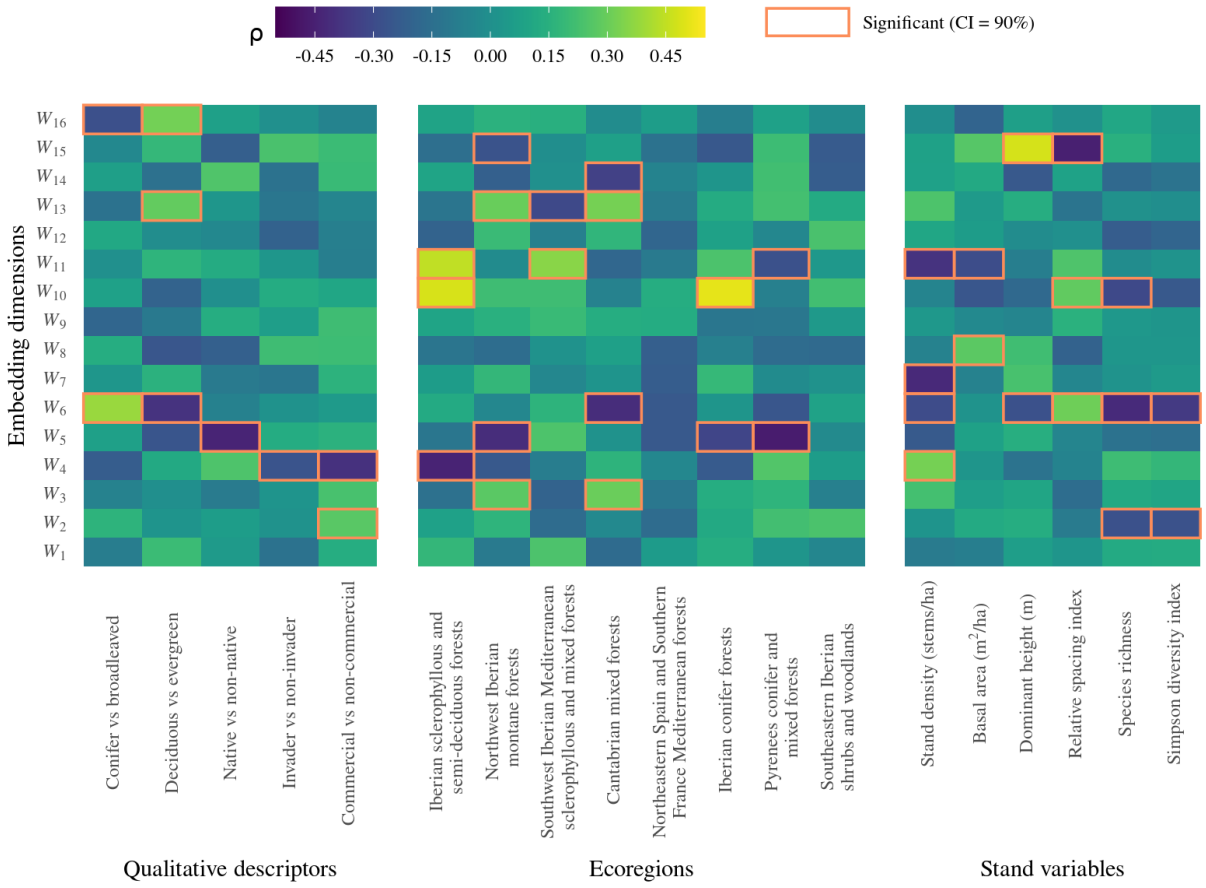

Figure 3: Heatmaps of Spearman and point-biserial correlations (ρ) between species-level embeddings and: 1) qualitative descriptors, 2) presence frequencies in the eight TEOW ecoregions, and 3) mean stand variables. Significant correlations for a Confidence Interval (CI) of 90% are highlighted.

4. Discussion

In the field of Representation Learning, the embeddings are considered to efficiently encapsulate the most discriminative portion of information hidden in the data for effectively representing the underlying factors that define abstract relationships between objects (Bengio et al., 2013). In the current study, we applied a Representation Learning approach for developing a latent space that allowed to quantify abstract characteristics of tree species present in Spanish forests. As input data, we considered a set of phyto- and geocentric descriptors that represented tree species attributes regarding three major axes: i) tree size (diameters and heights), ii) specific composition in the vicinity, and iii) habitat characteristics. The high classification performance yielded by the developed neural network models suggests that the proposed architectures were able to extract useful information retained in the inputs and efficiently translate it into a latent space. The latent dimensions resulting from the TreeSp2Vec model seemed to be semantically meaningful, as they showed significant correlations with unseen species features (see Figure 3). The geocentric information provided as input seemed to be effectively encoded by the model, as some latent dimensions proved to be good identifiers of the habitat variability encompassed by the considered ecoregions. In addition, the phytocentric information was effectively compressed by the encoding approach into synthetic representations related to stand variables and specific composition, allowing the latent dimensions to identify species and forest characteristics coherently. For instance, W_6 and W_{16} dimensions seemed to be useful for differentiating between conifers and broad-leaved species, as well as between deciduous and evergreens. W_2 proved to be an effective identifier of commercial plantations with low biodiversity (negative correlations with species richness and Simpson index). W_8 seemed to be an exclusive identifier of tree species basing on forest basal area. The strong relationship between stand density and some dimensions, such as W_4 and W_{11}, provided a useful criterion for identifying species associated with sparse forests, such as elm oaks and cork oaks, frequent in the Iberian sclerophyllous and semi-deciduous forests. W_3 was an exclusive identifier of the Northwest Iberian montane forests and the Cantabrian mixed forests, characterised by deciduous native species (W_{13})

and non-native plantations of evergreens (W_5 and W_6), which is consistent with the actual characteristics of forests existing in those regions.

In the current study, the embedding evaluation procedure was based on correlations with supplementary species descriptors, and could be classified as a "concept categorization" intrinsic evaluator, according to the classification proposed by Wang et al. (2019). This method differed notably from the most frequently used approaches in the family of intrinsic evaluators in Representation Learning studies (similarity and analogy evaluators), which compare statistical distances between pairs or groups of objects with subjective similarity rankings previously made by domain experts (Faruqui et al., 2016). In forestry, we lack of a dataset of human perceived similarity between species and, consequently, alternative methods for embedding interpretation are necessary. However, even though the evaluation method proposed in the present study proved to be effective for revealing the generality of the encoding approach, the coherency concerning species proximity in the latent space remains unclear. In this regard, visualization proved to be useful for assessing semantic coherency in some cases. For instance, in Figure 4 the dimensions E_6 and E_{15} allow to visually distinguish between species belonging to the same genus. In this Figure, species that never occur in the same inventory plot, such as the mountain pine and eucalypts, are represented far away in the subspace and do not overlap. In contrast, occasional neighbours, such as the Scots pines and cork oaks happening in forests in central and northern Spain or the elm oaks and the river red gums that can be present together in southwestern forests, overlap partially. Regardless of the evaluation approach, the coherency of the latent space could be enhanced by implementing regularization approaches, such as variational architectures, which have demonstrated very good properties in previous studies for generative-oriented encoding (Pu et al., 2016; Kingma and Welling, 2014). Derived from this, the development of a variational embedding model for tree species might be a key improvement line in future research.

Concerning model performance, the variations in MCC observed across the set of species might have implications regarding the uncertainty of the developed latent representations. It is reasonable to think that latent vectors corresponding to species with low MCC values will have more uncertainty regarding the meaningfulness of their semantic representations.

Though there is not an objectively definable threshold for accepting or rejecting the classification metrics yielded by the model, additional simulations carried out with a random naive classifier yielded very low MCC values (maximum observed MCC = 0.0007), which suggests that the species considered in this study were, overall, adequately classified. Even so, complementary methods for analysing the potential interactions between performance and semantic coherency might be necessary for clearing this matter in future research.

Considering the adequate results of model performance and semantic coherency, we believe that the tree species embeddings developed in the current study might be a valuable resource for other forest research areas. The species-level embeddings can be used as input for multi-species predictive modelling tasks by representing the different target species by their corresponding latent vector instead of a categorical variable. In this case, testing the predictive ability of an embedding-based multi-species model in comparison to single-species models can provide a criterion for semantic coherency evaluation from an extrinsic perspective (as shown in Supplemental Material). Apart from multi-species predictive modelling, the use of species-level embeddings could enhance previous methodologies for biological similarity estimation between forest types (Hao et al., 2019a), by providing a supervised framework for integrating non-linear relationships and interactions of a broad spectrum of tree and habitat descriptors. In this regard, two major improvement lines could be proposed: i) the development of new metrics for assessing the heterogeneity of the latent space in terms of biodiversity, similarly to traditional indexes (e.g., the Avalanche Index, Ganeshaiah et al., 1997), and, ii) the addition of new species descriptors for representing morphological diversity and physiological traits (Hao et al., 2019b). Moreover, the inclusion of spatial (e.g. spatial distribution indexes) and temporal (e.g. growth estimations derived from plot remeasurements) descriptors could be crucial improvements for the development of numeric representations of tree species considering forest structure and dynamics. These developments might provide relevant advantages for forest research in the future.

5. Conclusions

In this study, an approach for learning abstract representations of tree species basing on tree classification using deep neural networks was presented. The developed models yielded high classification accuracy, reaching the best performance (Matthew's Correlation Coefficient = 0.89) with a latent space of 16 dimensions. The evaluation of these embeddings revealed that the resulting model, labeled TreeSp2Vec, was able to capture and compress abstract characteristics of tree species in a semantically meaningful latent space, whose generality allowed to predict unseen forest characteristics. We conclude that the developed latent space of representations might be a useful resource for further research on forest multi-species modelling and opens a new range of artificial intelligence applications in ecology.

6. Acknowledgements

We thank the *Banco de Datos de la Naturaleza* (BDN) for the Third National Forest Inventory data provided. We also thank the IGN for the digital elevation model provided, the ESDAC for the soil attributes maps and the Wordlclim 2.1 project for the climate dataset (Fick and Hijmans, 2017).

7. Conflict of interest statement

The authors declare no conflict of interest.

8. Data availability

The input data used in this article is available at the *Banco de Datos de la Naturaleza* website (miteco.gob.es).

References

Alberdi, I., Condés, S., Martínez, J., Martínez, S., de Toda, S., Sánchez, G., Pérez, F., Villanueva, J., and Vallejo, R. (2010). Spanish national forest inventory. Martínez, S.; de Toda, S.; Sánchez, G.; Pérez, F.; Villanueva, J.A.; Vallejo, R.Alberdi, I.; Condés, S.; Martínez, J.; Martínez, S.; de Toda, S.;

Sánchez, G.; Pérez, F.; Villanueva, J.A.; Vallejo, R.¡br/¿Spanish national forest inventory. In National Forest Inventories. Pathways for Common Reporting; Tomppo, E.,¡br/¿Gschwantner, T., Lawrence, M., McRoberts, R.E., Eds.; Springer: Berlin, Germany, 2010; pp. 527–54.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., and Panagos, P. (2019). Mapping lucas topsoil chemical properties at european scale using gaussian process regression. *Geoderma*, 355:113912.

Ballabio, C., Panagos, P., and Monatanarella, L. (2016). Mapping topsoil physical properties at european scale using the lucas database. *Geoderma*, 261:110–123.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Chollet, F. et al. (2015). Keras.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*.

Ganeshaiah, K. N., Chandrashekara, K., and Kumar, A. R. V. (1997). Avalanche index: A new measure of biodiversity based on biological heterogeneity of the communities. *Current Science*, 73(2):128–133.

Gelada, C., Kumar, S., Buckman, J., Nachum, O., and Bellemare, M. G. (2019). DeepMDP: Learning continuous latent space models for representation learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2170–2179. PMLR.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests.

Hao, M., Corral-Rivas, J. J., González-Elizondo, M. S., Ganeshaiah, K. N., Nava-Miranda, M. G., Zhang, C., Zhao, X., and von Gadow, K. (2019a). Assessing biological dissimilarities between five forest communities. *Forest Ecosystems*, 6.

Hao, M., Ganeshaiah, K. N., Zhang, C., Zhao, X., and von Gadow, K. (2019b). Discriminating among forest communities based on taxonomic, phylogenetic and trait distances. *Forest Ecology and Management*, 440.

Huang, P., Huang, Y., Wang, W., and Wang, L. (2014). Deep embedding network for clustering. In *2014 22nd International Conference on Pattern Recognition*, pages 1532–1537.

Irtaza, A., Jaffar, M. A., Aleisa, E., and Choi, T.-S. (2014). Embedding neural networks for semantic association in content based image retrieval. *Multimedia Tools and Applications*, 72.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes.

Lesort, T., Díaz-Rodríguez, N., Goudou, J.-F., and Filliat, D. (2018). State representation learning for control: An overview. *Neural Networks*, 108:379–392.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., and Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51(11):933–938.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., and Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*, 10:290–300.

Sejnowski, T. (2018). *The Deep Learning Revolution*. The MIT Press. MIT Press.

Tang, D., Qin, B., Liu, T., and Li, Z. (2013). Learning sentence representation for emotion classification on microblogs. In Zhou, G., Li, J., Zhao, D., and Feng, Y., editors, *Natural Language Processing and Chinese Computing*, pages 212–223, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.-C. J. (2019). Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87.

Yang, S., Luo, P., Loy, C. C., Shum, K. W., and Tang, X. (2015). Deep representation learning with target

coding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Yu, W., Zeng, G., Luo, P., Zhuang, F., He, Q., and Shi, Z. (2013). Embedding with autoencoder regularization. In Blockeel, H., Kersting, K., Nijssen, S., and Železný, F., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 208–223, Berlin, Heidelberg. Springer Berlin Heidelberg.

Appendix A. Classification performance by species

Figure 4: Scatterplot of 6^{th} vs 15^{th} dimensions of tree-level embeddings for *Pinus uncinata* Ramond ex A.DC., *Pinus sylvestris* L., *Quercus ilex* L. *ssp. ballota*, *Quercus suber* L., *Eucalyptus globulus* Labill., and *Eucalyptus camaldulensis* Dehnh. Shaded ellipses correspond to confidence intervals = 99% of by-genus multivariate normal distributions.

Species	MCC
Chamaecyparis lawsoniana	0.938
Pinus sylvestris	0.903
Pinus uncinata	0.931
Pinus pinea	0.875
Pinus halepensis	0.945
Pinus nigra	0.875
Pinus pinaster	0.918
Pinus radiata	0.943
Abies alba	0.853
Abies pinsapo	0.955
Pseudotsuga menziesii	0.927
Juniperus thurifera	0.829
Quercus robur	0.794
Quercus petraea	0.819
Quercus pyrenaica	0.914
Quercus faginea	0.787
Quercus ilex ssp. ballota	0.878
Quercus suber	0.857
Quercus canariensis	0.755
Quercus rubra	0.879
Populus alba	0.892
Populus tremula	0.658
Alnus glutinosa	0.798
Fraxinus angustifolia	0.785
Populus nigra	0.913
Eucalyptus globulus	0.931
Eucalyptus camaldulensis	0.960
Ilex aquifolium	0.734
Olea europaea	0.814
Fagus sylvatica	0.926
Castanea sativa	0.841
Corylus avellana	0.786
Robinia pseudoacacia	0.750
Quercus pubescens	0.718
Fraxinus excelsior	0.662
Salix alba	0.738
Populus x canadensis	0.975
Betula alba	0.747
Salix atrocinerea	0.677
Betula pendula	0.718

Table A.1: Mathew's Correlation Coefficient of the classification task performed by the TreeSp2Vec model for each of the 40 species considered.

Supplemental Material: TreeSp2Vec: a method for developing tree species embeddings using deep neural networks

S-I. Example of embedding-based multi-species modelling

We present a simple illustrative example of how to use species-level embeddings for developing empirical multi-species models. The task is to develop a multi-species generalized height-diameter (*h-d*) equation for predicting the total height of individual trees as a function of their diameter at breast height.

S-I.1. Example methods

The proposed *h-d* model for this example is the Chapman-Richards equation (Richards, 1959):

$$h = 1.3 + a\Big(1 - \exp(-bd)\Big)^{c}, \tag{S1}$$

where h is the tree total height (m), d is the tree diameter at breast height (cm), and a, b and c are parameters. We fit the Chapman-Richards equation for the 30 most frequent species in the dataset by developing 1) one global generalized *h-d* model (i.e., with the same a, b and c values for all the species) and 2) 30 different single-species models. Then, the parameters (a, b and c) in each single-species model are expanded using the species-level embeddings. We do so by predicting a, b and c as a function of the previously developed 16 latent dimensions ($\mathbf{W}$) using a machine learning model. Specifically, we use multi-layer perceptrons (MLPs) with three hidden layers of 32 units and ReLU activations, calibrated through Monte Carlo validation. Finally, we apply the expanded a, b and c parameters for each species for re-predicting the total tree height using a multi-species Chapman-Richards model:

$$h = 1.3 + MLP_a(\mathbf{W})\Bigg(1 - \exp\Big(-MLP_b(\mathbf{W})d\Big)\Bigg)^{MLP_c(\mathbf{W})}, \tag{S2}$$

where MLP_x are the outputs of the multilayer perceptrons trained for predicing a, b and c as a function of $\mathbf{W}$.

S-I.2. Example results

As expected, considering the variety of species in the dataset, the global *h-d* model performs very poorly (see Table S1), yielding an R^2=0.49, being the unweighted mean over the set of species R^2=0.05. In comparison, the single-species models perform much better, although they also show strong variations in R^2 between different species (ranging from R^2=0.32 to 0.73, see Figure S1). The three parameters are succesfully predicted from the latent dimensions (see Figure S2), which accounts for the existence of meaningful associations with $\mathbf{W}$ (see Figure S3). Finally, the re-prediction of tree height using the multi-species Chapman-Richards model reveals, on average, only in a slight drop in performance with respect to the single-species models (see Figure S4). This result confirms the usefulness of the approach for using only one model for all the species while maintaining good performance.

Model	R^2	R^2_{min}	R^2_{mean}	R^2_{max}
Global	0.485	-0.502	0.0566	0.501
Single-species	0.736	0.322	0.504	0.7308
Multi-species	0.662	0.0929	0.455	0.7208

Table S1: Predictive performance of the developed *h-d* approaches.

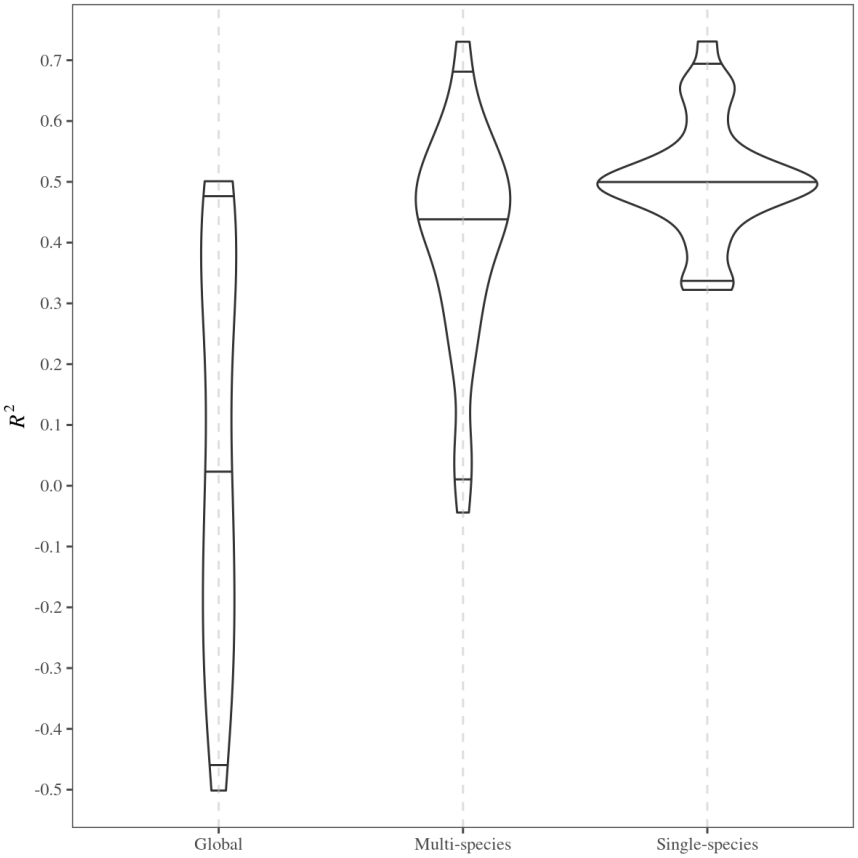

Figure S1: Violin plots representing the distribution of R^2 across species of the three *h-d* modeling approaches.

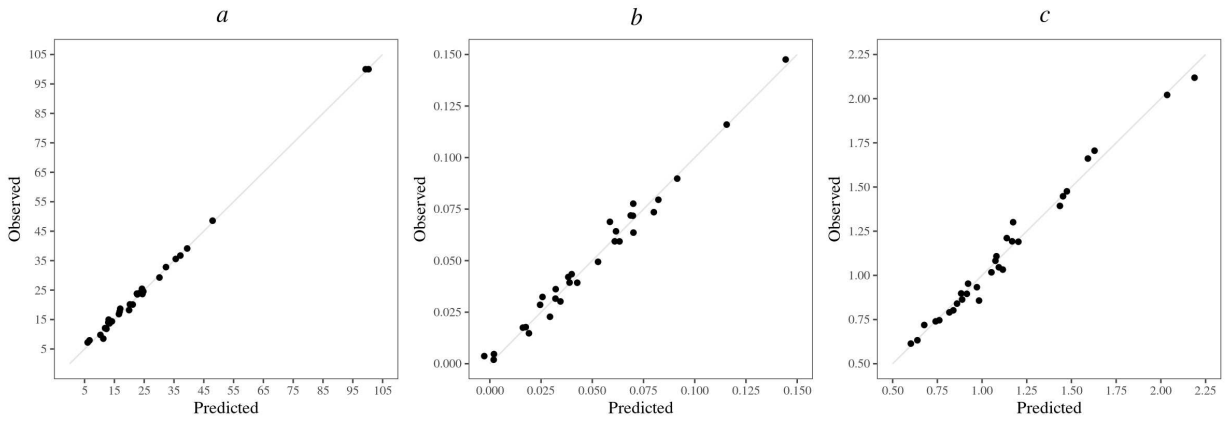

Figure S2: Observed vs predicted values of the parameters using the three MLP models.

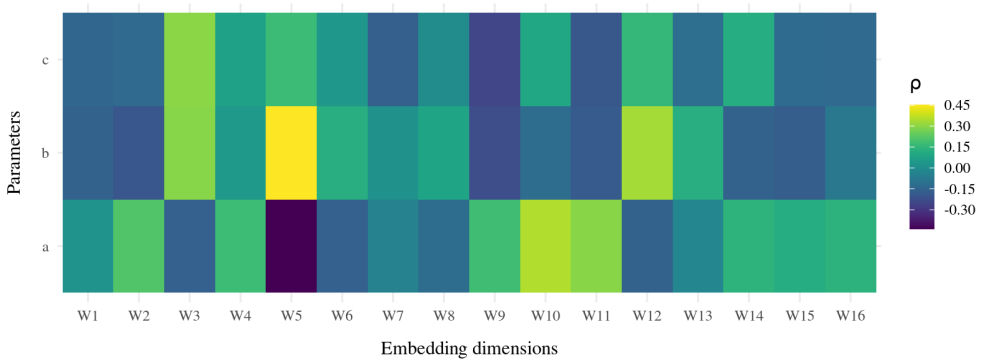

Figure S3: Heatmap of Spearman correlations (ρ) between species-level embeddings and the three parameters of each single-species h-d model.

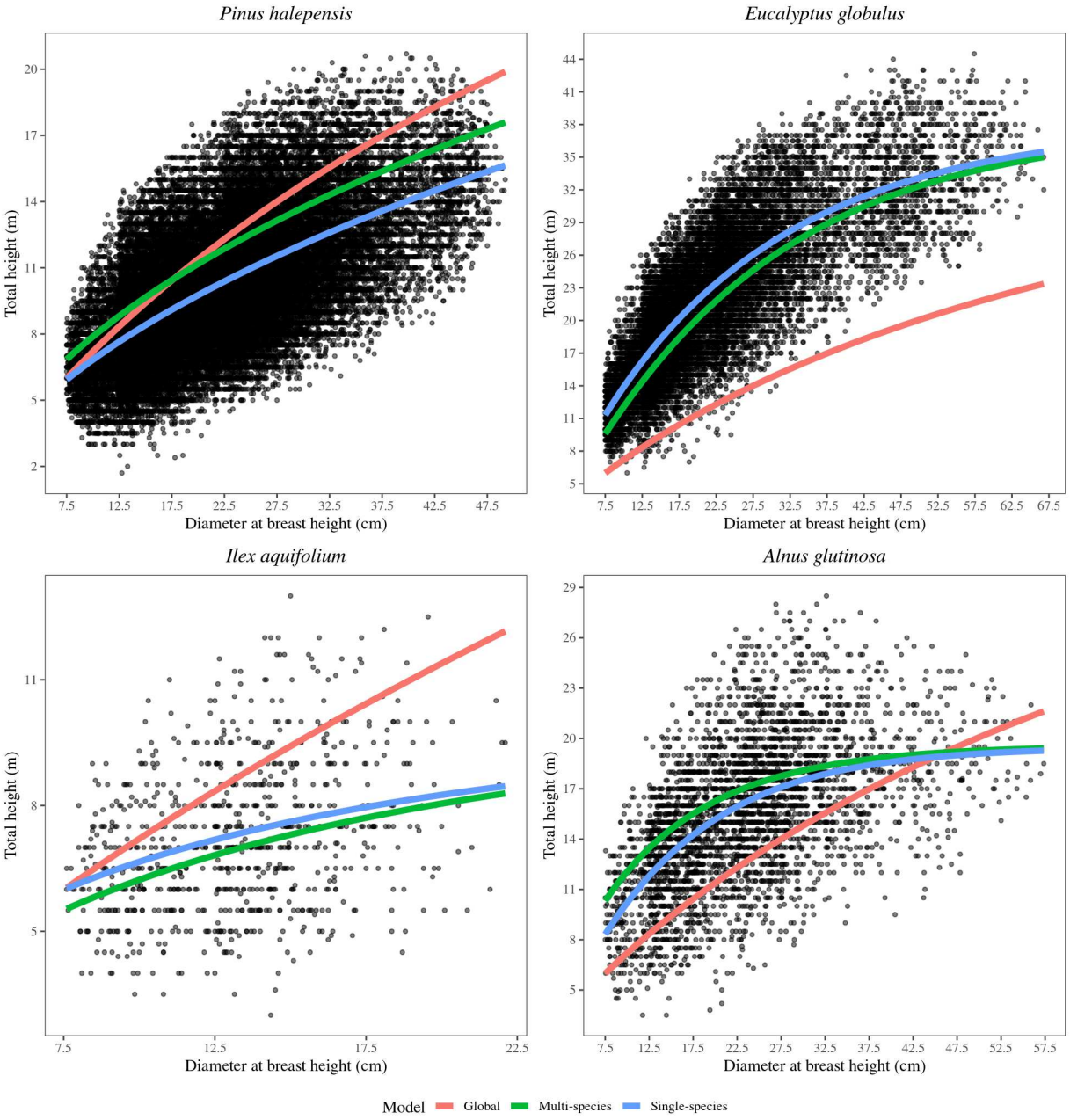

Figure S4: *h-d* scatter plots of four tree species with predicted trends for the three generalized *h-d* model approaches.