

Supplementary materials

# Spatial and temporal patterns of grassland species diversity and their driving factors in the Three Rivers Headwater Region of China from 2000 to 2021

Mingxin Yang <sup>1,2,4</sup>, Ang Chen <sup>2</sup>, Wenqiang Cao <sup>1</sup>, Shouxin Wang <sup>1</sup>, Mingyuan Xu <sup>1</sup>, Qiang Gu <sup>1</sup>, Yanhe Wang <sup>1,3</sup>, Xiuchun Yang <sup>2,\*</sup>

The following Supporting Information is available for this manuscript:

## 1. Introduction to the four variable selection methods

The GA is a computational model that simulates the biological evolution process of natural selection and genetics of Darwin's theory of biological evolution, which is a method of searching for the optimal solution by simulating the natural evolution process, and is characterized by the ability of stochastic global optimization search [1]. REF is a feature selection method that selects the optimal subset of features by iteratively training the model and eliminating features of unimportant variables [2]. STEP is the construction of a model by incrementally adding or deleting features to find the optimal subset of features, and the main goal of the method is to minimize the model's error or evaluation metrics while avoiding overfitting [3]. LASSO obtains a more refined model by constructing a penalty function that makes it compress some coefficients while setting some coefficients to zero, thus retaining the advantage of subset shrinkage as a form of biased estimation for dealing with data with complex covariance [4].

## 2. Introduction to four machine learning methods

The XGboost is to integrate many tree models together to form a very strong classifier, which can effectively avoid over-fitting by introducing the number of subtrees and the value of subtree leaf nodes, etc. in the loss function, which fully takes into account the regularization problem [5]. RF is a non-parametric machine learning algorithm that uses multiple decision trees to train samples and integrate predictions, it averages the predicted values of multiple decision trees and integrates them when dealing with regression prediction problems, compared to decision tree algorithms, random forests are more resistant to interference and have a stronger ability to generalize the model [6]. KNN finds out the K training samples in the training set closest to it based on the distance metric, and the average of the real-valued output labels of these K samples is used as the prediction result [7]. SVM deals with regression problems by minimizing the prediction error and maximizing the classification interval [2].

## 3. Figures and tables necessary to supplement the manuscript

Table S1. Descriptive statistics of species diversity in ground survey sample plots from 2005-2021.

Year	Number of sample plots	Min	Max	Average	SD	CV
2005	81	4.1	15.3	8.64	3.02	34.92
2006	143	5.0	21.9	10.16	3.28	32.23
2009	151	4.3	19.3	10.85	3.41	31.43
2010	176	5.0	24.1	12.03	3.68	30.62

2011	180	4.3	22.1	11.11	3.49	31.43
2013	193	3.4	23.5	10.61	3.58	33.72
2014	161	3.7	24.0	10.34	3.31	31.99
2015	196	3.3	18.3	10.62	2.97	28.00
2016	194	3.3	24.0	10.87	3.46	31.85
2017	190	3.4	19.6	9.86	3.31	33.54
2018	174	3.7	17.7	9.10	3.15	34.64
2021	297	3.0	18.0	9.07	3.32	36.59
2005 ~ 2021	2136	3.0	24.1	10.27	3.33	32.58

Table S2. Calculation formula of vegetation index.

Vegetation index	Calculation formula	References
Normalized Difference Vegetation Index (NDVI)	$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$	[8]
Green Normalized Vegetation Index (GNDVI)	$\text{GNDVI} = \frac{\text{NIR} - \text{Green}}{\text{NIR} + \text{Green}}$	[9]
Kernel Normalized Difference Vegetation Index (KNDVI)	$\text{KNDVI} = \tanh\left(\left(\text{NDVI}\right)^2\right)$	[10]
Enhanced Vegetation Index (EVI)	$\text{EVI} = \frac{2.5(\text{NIR} - \text{Red})}{\text{NIR} + 6.0\text{Red} - 7.5\text{Blue} + 1}$	[11]
Ratio Vegetation Index (RVI)	$\text{RVI} = \frac{\text{NIR}}{\text{Red}}$	[11]
Soil-adjusted Vegetation Index (SAVI)	$\text{SAVI} = \frac{(1 + L)(\text{NIR} - \text{Red})}{\text{NIR} + \text{Red} + L}$	[12]

Note: NIR for near-infrared bands; Red for red bands; Blue for blue bands; Green for green bands; L for the soil adjustment factor, L = 0.5.

Table S3. Grading of trends of species diversity changes.

Slope	Z value	Trends of change
$\geq 0.0005$	$\geq 1.96$	significant increase
$\geq 0.0005$	-1.96 ~ 1.96	slight increase
-0.0005 ~ 0.0005	-1.96 ~ 1.96	stable
$< -0.0005$	-1.96 ~ 1.96	slight decrease
$< -0.0005$	$< -1.96$	Significant decrease

Figure S1. Optimal model training set and test set fitting.

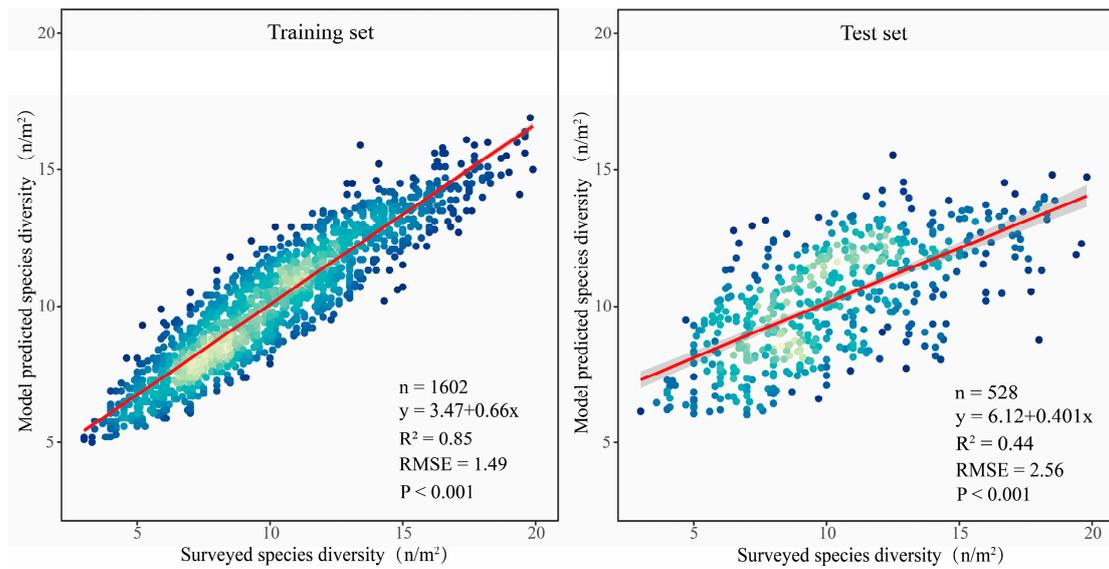
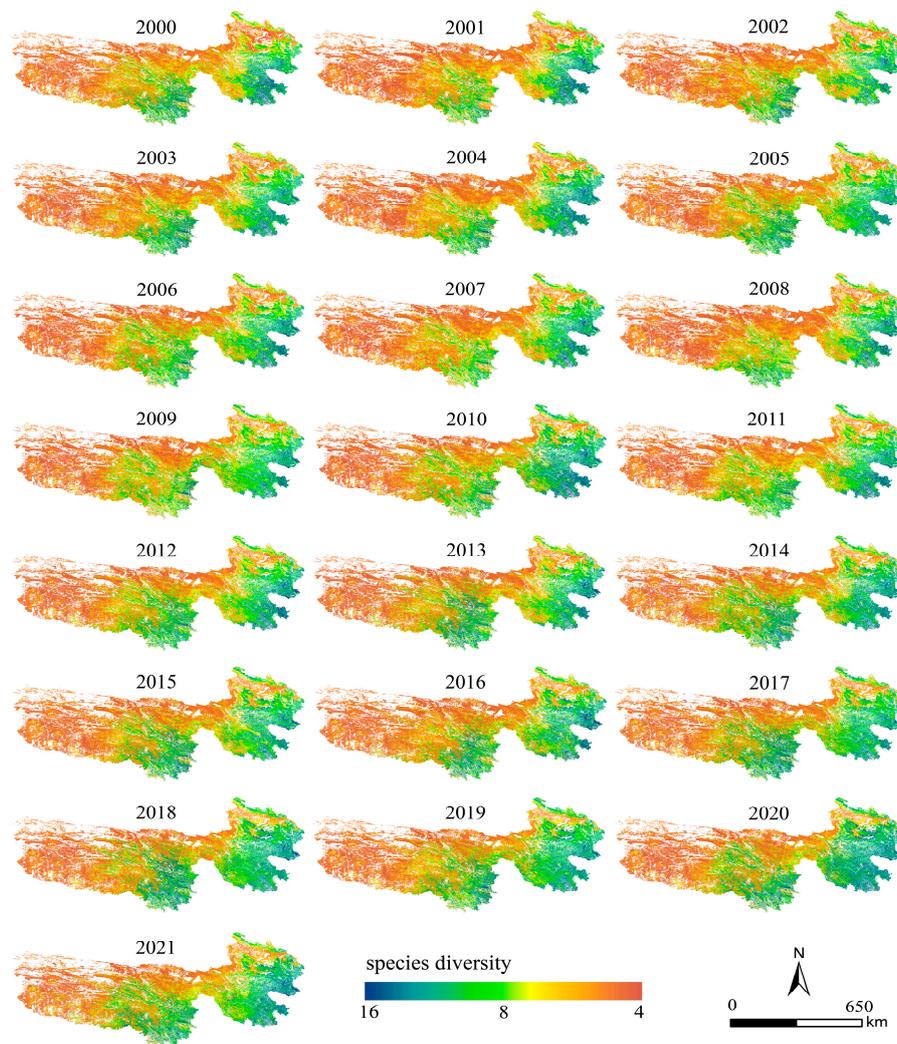


Figure S2. Spatial distribution of grassland species diversity in the Three River Headwaters Region during 2000 ~ 2021.



## References:

1. Andersen, C.M.; Bro, R. Variable selection in regression—a tutorial. *J. Chemometr.* **2010**, *24*, 728-37. DOI: 10.1002/cem.1360.
2. Duan, K.B.; Rajapakse, J.C.; Wang, H.; Azuaje, F. Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data. *IEEE T. Nanobiosci.* **2005**, *4*, 228-34. DOI: 10.1109/TNB.2005.853657.
3. Smith, G. Step away from stepwise. *Journal of Big Data.* **2018**, *5*, 32. DOI: 10.1186/s40537-018-0143-6.
4. Tibshirani, R. Regression shrinkage via the lasso. *J R Statist Soc.* **1996**, *58*, 267-88.
5. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937-67. DOI: 10.1007/s10462-020-09896-5.
6. Yang, M.; Chen, A.; Zhang, M.; Gu, Q.; Wang, Y.; Guo, J.; Yang, D.; Zhao, Y.; Huang, Q.; Ma, L.; *et al.*. Relationship between plant species diversity and aboveground biomass in alpine grasslands on the Qinghai – Tibet Plateau: Spatial patterns and the factors driving them. *Frontiers in Ecology and Evolution.* **2023**, *11*. DOI: 10.3389/fevo.2023.1138884.
7. Deng, Z.; Zhu, X.; Cheng, D.; Zong, M.; Zhang, S. Efficient kNN classification algorithm for big data. *Neurocomputing.* **2016**, *195*, 143-8. DOI: 10.1016/j.neucom.2015.08.112.
8. Tucker, C.J. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sens. Environ.* **1979**, *2*, 127-50. DOI: 10.1016/0034-4257(79)90013-0.
9. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289-98. DOI: 10.1016/S0034-4257(96)00072-7.
10. Camps-Valls, G.; Campos-Taberner, M.; Moreno-Martínez, Á.; Walther, S.; Duveiller, G.; Cescatti, A.; Mahecha, M.D.; Muñoz-Marí, J.; García-Haro, F.J.; Guanter, L.; *et al.*. A unified vegetation index for quantifying the terrestrial biosphere. *Science advances.* **2021**, *7*, c7447. DOI: 10.1126/sciadv.abc7447.
11. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195-213. DOI: 10.1016/S0034-4257(02)00096-2.
12. Huete, A.R. A Soil-Adjusted Vegetation Index (SAVI). *Remote Sens. Environ.* **1988**, *3*, 295-309. DOI: 10.1016/0034-4257(88)90106-X.