# Multi-output regression with structurally incomplete target labels : A case study of modelling global vegetation cover

þÿ B e i g a i t , R i t a

2022-12

þÿ B e i g a i t , R , R e a d , J & } l i o b a i t , I 2 0 2 2 , ' M u l t i - o u t p u t r e g r e s s i o n w
incomplete target labels : A case study of modelling global vegetation cover ' , Ecological
Informatics , vol. 72 , 101849 . https://doi.org/10.1016/j.ecoinf.2022.101849

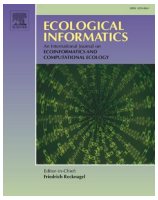
http://hdl.handle.net/10138/350570
https://doi.org/10.1016/j.ecoinf.2022.101849

Ecological Informatics 72 (2022) 101849

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



# Multi-output regression with structurally incomplete target labels: A case study of modelling global vegetation cover

Rita Beigaitė [a,*], Jesse Read [b], Indrė Žliobaitė [a,c]

[a] Department of Computer Science, University of Helsinki, Finland
[b] LIX, Ecole Polytechnique, Institut Polytechnique de Paris, France
[c] Finnish Museum of Natural History LUOMUS, University of Helsinki, Finland

ARTICLE INFO

*Keywords:*
Multi-output regression
Compositional data
Incomplete targets
Weakly supervised learning
Vegetation cover modelling

ABSTRACT

Weakly-supervised learning has recently emerged in the classification context where true labels are often scarce or unreliable. However, this learning setting has not yet been extensively analyzed for regression problems, which are typical in macroecology. We further define a novel computational setting of structurally noisy and incomplete target labels, which arises, for example, when the multi-output regression task defines a distribution such that outputs must sum up to unity. We propose an algorithmic approach to reduce noise in the target labels and improve predictions. We evaluate this setting with a case study in global vegetation modelling, which involves building a model to predict the distribution of vegetation cover from climatic conditions based on global remote sensing data. We compare the performance of the proposed approach to several incomplete target baselines. The results indicate that the error in the targets can be reduced by our proposed partial-imputation algorithm. We conclude that handling structural incompleteness in the target labels instead of using only complete observations for training helps to better capture global associations between vegetation and climate.

## 1. Introduction

Target variables are usually fully labeled in the classical supervised machine learning setting. In real-world predictive tasks, however, labels are often scarce and/or noisy. Various definitions and terms are used in the literature to describe variants of noise and scarceness of labels (Allison, 2001; Xie and Huang, 2018; Nikoloski et al., 2021; Sun et al., 2010; Gao et al., 2017; Alarcón and Destercke, 2021; Van Engelen and Hoos, 2020), and each setting requires tailored approaches for exploiting such target labels. In this study, we formulate a new computational setting for regression, where target labels are *structurally incomplete*. We computationally study this task via a case study in predictive modeling of global vegetation cover.

### 1.1. The vegetation modelling task

Our problem setting derives from a case study in which we aim to build a predictive model capturing large-scale associations between the global distribution of vegetation and prevailing climatic conditions. The goal of our task relates to manual climate classification schemes used in climate science (Kottek et al., 2006; Holdridge et al., 1967; Whittaker, 1962). These schemes are rule-based systems composed manually by experts. They do not provide sufficient resolution to predict local vegetation types worldwide, and even less so to extrapolate to new scenarios under predicted climate change.

Climate extremes and extreme weather events have changed over the last few decades (Ummenhofer and Meehl, 2017; Seneviratne et al., 2012). The frequency and intensity of daily temperature and precipitation extremes have been observed to increase due to human-induced climate change (Hulme et al., 1999; Mitchell et al., 2006; Seneviratne et al., 2012). The effects of climate extremes on vegetation distribution are highly uncertain. Typically, present, future, or past large-scale changes in vegetation cover are predicted using process-based models, such as Dynamic Global Vegetation Models (DGVMs) (Quillet et al., 2010). DGVMs simulate the underlying physiological processes, climatic and biotic interactions using differential equations (e.g., Snell et al., 2014). Those models are generally reliable in terms of process representation but challenging in setting their parameters. Experimentation with variants is computationally expensive and not yet sufficiently streamlined, at least quantitatively. Thus, there is high demand for simpler machine-learned models that could work with remote sensing data. This way, future vegetation shifts can potentially be projected

* Corresponding author.
*E-mail address:* rita.beigaite@helsinki.fi (R. Beigaitė).

https://doi.org/10.1016/j.ecoinf.2022.101849
Received 9 May 2022; Received in revised form 31 August 2022; Accepted 2 October 2022
Available online 8 October 2022

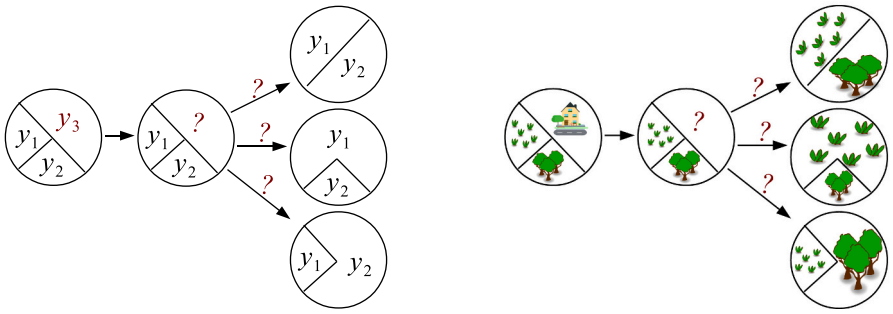



**Fig. 1.** A simplified example of structurally incomplete target labels ($y_j$ where $j = 1, 2$). Each pie chart represents the composition of the targets of one observation. After disregarding the human activity target label as noise, we want to reconstruct the original proportion of the remaining target labels of the composition.

more accurately following predicted climate changes drawn from dynamic earth-system models (Kelly and Goulden, 2008; Holsinger et al., 2019; Gomez-Ruiz and Lacher, 2019). Machine-learned models may also aid in conservation planning where vegetation classification plays an important role (Hoagland, 2000).

A related research question in ecology and biogeography is how to predict the *potential* natural vegetation (PNV). PNV is the expected state of mature vegetation, given a particular set of environmental constraints in the absence of human intervention (Chiarucci et al., 2010). At first, PNV models were constructed based only on expert knowledge, whereas nowadays, various statistical techniques and machine learning methods are more widely employed (Hemsing and Bryn, 2012). In Hengl et al. (2018), authors evaluate different machine learning methods, such as neural networks, random forests, gradient boosting, and k-nearest neighbours, for PNV mapping in a classification setting. The latter example describes global PNV mapping. However, most PNV studies focus on specific areas or regions (Raja et al., 2019; Vaca et al., 2011; Hemsing and Bryn, 2012).

### 1.2. Why the task is difficult

More than one type of vegetation is commonly present in any given area. In the case study accompanying our algorithmic analysis, we aim to predict the distribution of natural land cover types from climatic conditions. Those types can either be natural vegetation or lack vegetation in the form of snow and ice, as well as deserts. In the underlying data, each land point on Earth is represented as a set of fractions of various vegetation types. Our prediction task is thus a multi-output regression with compositional constraints (prediction outputs must sum up to one and be non-negative).

One way to address this task is to look at it from the compositional data analysis perspective. *Compositional data* is a term that refers to data where elements are non-negative and are in the form of proportions of some whole (Aitchison, 1982; Pawlowsky-Glahn and Buccianti, 2011). Such type of data is common in many fields: demography (Lloyd et al., 2012), economics (Ferrer-Rosell et al., 2015), chemistry and geology (Buchanan et al., 2012). The constraints in the compositional data setting are equivalent to *label distribution learning* (Geng, 2016) where, instead of labels, probabilities of the outputs are predicted in a classification task.

The main technical challenge in our prediction task is that many parts of the landscape are altered by excessive human activities and changed into urban areas or croplands. Nowadays, this can happen under almost any climatic conditions and, by and large, is unpredictable from climatic variables (Zanelli, 2021). Yet from the ecological perspective, the main task of interest is what would be a natural

**Table 1**
An example training instance for different weakly-labeled settings. Here $y_j$ indicates the targets of an instance. Red indicates an error in the training data (deviation from ground truth) – it is *not known* to the user where errors occur. Note that in the weak label setting, only 0s may be errors. Shading represents a constraint (in this example, that $\sum_{j=1}^{5} y_j = 1$ and $y_j \geq 0$). In our setting (structurally incomplete), the constraint may not be met, but errors are neither random deviations. Mixtures of these settings are possible, e.g., we may have structurally incomplete with missing labels, etc.

| Setting | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | Remark |
|---|---|---|---|---|---|---|
| Supervised learning | 0.1 | 0.3 | 0.2 | 0.4 | 0 | Ground truth labels available |
| Missing labels | 0.1 | ? | ? | 0.4 | 0 | Some values are missing, e.g., Allison (2001) |
| Unsupervised learning | ? | ? | ? | ? | ? | No labeled instances |
| Semi-supervised learning | ? | ? | ? | ? | ? | Some instances are unlabeled, some instances are labeled, e.g, Zhu (2005); Van Engelen and Hoos (2020), Kostopoulos et al. (2018); Levatić et al. (2018) |
| | 0.1 | 0.3 | 0.2 | 0.4 | 0 | |
| | ? | ? | ? | ? | ? | |
| Label-distribution learning | 0.1 | 0.2 | 0.3 | 0.4 | 0 | Constraint; e.g., Geng (2016); Aitchison (1982) |
| Structurally incomplete | 0.1 | 0.1 | 0 | 0.4 | 0 | Constraint, not met; our work |
| Partial/Noisy labels | 0.1 | 0.8 | -3 | 0.4 | 0 | Contains errors; e.g., Xie and Huang (2018) |
| Partially labeled data | 0.1 | ? | ? | 0.4 | 0 | Some targets are unlabeled, some targets are labeled, e.g, Nikoloski et al. (2021) |
| | ? | 0.3 | ? | ? | 0 | |
| | ? | 0.3 | 0.2 | 0.4 | ? | |
| Weak labels | 0.1 | 0 | 0 | 0.4 | 0 | Missing values set as 0; e.g., Sun et al. (2010) |
| Ambiguous labels | 0.1 | 0.3 | 0.2 | 0.4 | 0 | Multiple hypotheses per single instance; e.g., Gao et al. (2017); Alarcón and Destercke (2021) |
| | 0.1 | 0.4 | 0.2 | 0.4 | 1.1 | |
| | 0 | 0.2 | 0.1 | 0 | -0.2 | |

vegetation cover under given climatic conditions (Hengl et al., 2018). Thus, from the computational perspective, part of the information on what would be the natural distribution of vegetation, given current climatic conditions, is not available in the training data. Consider an example (Fig. 1) where an observation is composed of 50% of urban area, 25% grassland and 25% forest. When we discard human activity proportion considering it as noise, we are left with a 50% covered grid cell. Our targets are then incomplete as we do not know how those remaining 50% would have been distributed. For instance, they could have been equally distributed between forest and grassland (both types would occupy 50 % of land in total), or maybe the urban area was replacing only grassland or only forest (in this case, one type would occupy 75% and the other 25% of land in total) and so on.

We can make a simplifying assumption that human activity landscapes are only masking currently known vegetation types rather than making some unobserved vegetation extinct and consider these land-cover types as a *structural noise*. If we then regard land cover types associated with intensive human activity (e.g., croplands) as noise, the target labels of natural vegetation become incomplete under the compositional constraint. In other words, the natural vegetation fractions at many points on Earth do not sum up to one as they should.

### 1.3. Related statistical and machine learning settings

Statistically, the missing data problem is defined as a lack of information for some variables for some cases (Allison, 2001). The term *incomplete data* is sometimes used as a synonym for missing data. However, values of the vegetation cover fractions in our setting are only partially missing, making standard missing value imputation methods not directly applicable or relevant.

Incompleteness in our setting of *structurally incomplete* targets resembles weakly supervised learning problem (Zhou, 2018). Weak supervision refers to a problem setting where the data is labeled, but the labels are inexact, erroneous, or faulty. Multiple solutions exist for classification tasks (Yao et al., 2016; Sun et al., 2010; Dery et al., 2017; Xu et al., 2014) in weakly supervised learning (Zhou, 2018). However, this problem for the regression task, and particularly multi-output regression, to the best of our knowledge, does not yet have tailored solutions.

Various other terms have been used to describe weakly supervised learning tasks in the literature (Table 1). Weak labels (Sun et al., 2010) or partial labels (Xie and Huang, 2018) are mainly considered in the context of binary labels. They often include different types of noise coming from the labeling process and data sources and lacks constraints. Distribution learning (Gao et al., 2017) entails the constraint that the outputs must sum to unity, but this constraint is already met in the training data, unlike in our *structural incompleteness* setting.

Recently a manifold regularization technique (Berikov and Litvinenko, 2021) has been introduced for weakly supervised single-target regression problem where the learning sample includes labeled, unlabeled, and inaccurately labeled data. The authors use a normal distribution with different parameters for modelling the uncertain labeling. In Chung et al. (2022) Bayesian probabilistic model was proposed for weakly-supervised multi-output regression with partially labeled outputs. However, here the term *partially labeled* means the absence of a correct label that indicates its group membership, for instance, whether the observed body mass index belongs to the diabetic or non-diabetic patient group.

In our setting of *structurally incomplete* targets, in some observations, targets form a complete distribution (that sums up to one), whereas in other observations, targets are structurally incomplete. From this perspective, the task resembles semi-supervised learning where a vast amount of unlabeled data is used with a small number of labeled examples (Zhu, 2005; Van Engelen and Hoos, 2020; Kostopoulos et al., 2018). In Nikoloski et al. (2021), the authors have proposed using semi-supervised predictive clustering trees (Levatić et al., 2018) that can handle partially labeled examples for multi-target regression task of water quality assessment. In this setting, partially labeled examples are the ones that have some targets in the composition missing and some known. In contrast, even if the observations are incomplete in our setting of *structurally incomplete* targets, parts of correct information about each target of the observation are known. In addition, the constraint of compositional data (summation to unity) gives information on how much noise is present in each target. Therefore, our defined problem setting is unique from the machine learning perspective and only resembles certain aspects of various other learning settings. The task of natural vegetation prediction is well established in ecology. Our study presents a new computational perspective to this problem.

### 1.4. Previous work and the structure of the paper

An extended abstract version of an early version of this study was presented at the ICML Workshop *On the Art of Learning with Missing Values* (Artemiss) (Beigaite et al., 2020). We have also investigated added value of including extreme climatic variables in the modelling process when the task is reduced single-output classification problem (Beigaitė et al., 2022). While the task from the computational perspective was different, some of the data from that study are reused here. Apart from the difference in the computational setting and methods, the paper primarily focused on biological aspects rather than computational aspects of natural vegetation cover prediction.

In the remainder of the paper, we formally define the proposed computational setting of *structurally incomplete* targets, backing it up with a case study of modelling global vegetation cover distribution (Section 2.1). We propose an algorithmic solution for handling structural incompleteness in the targets (Section 2.2–2.3). We experimentally investigate how well such strategy handles the structural noise in the training labels and how it can help to improve the performance of predictive models (Section 3).

## 2. Formal definitions and proposed solutions

In this section, we define a novel problem setting of *structurally incomplete* targets. We propose baseline approaches along with a partial imputation algorithm for tackling this problem.

### 2.1. Formal definition of the computational task

Let the available training data be represented as matrices

$$\mathbf{X} \in \mathbb{R}^{n \times d} \quad \text{and} \quad \widetilde{\mathbf{Y}} \in \mathbb{R}^{n \times l}$$

where each $i$-th row represents an input observations $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,d}]$ and the corresponding structurally-incomplete outputs $\widetilde{\mathbf{y}}_i = \left[\widetilde{y}_{i,1}, \ldots, \widetilde{y}_{i,l}\right]$ (at least some of these rows are incomplete, in the sense that they do not sum to unity), respectively.

The goal is to recover/reconstruct a valid distribution for all rows by adding probability mass, if and as necessary, to the incomplete outputs such that they represent a valid probability distribution representative of the underlying unobserved ground truth.

In other words, to produce estimate $\widehat{\mathbf{y}} \in \mathbb{R}_+^l; \sum_{j=1}^{l} y_j = 1$, and $y_j \geqslant 0 \forall j$.

Examples of this task are illustrated diversely in Figs. 3 and 4.

### 2.2. Baseline strategies for learning from structurally incomplete targets

We define three possible baseline strategies dealing with the targets of our task:

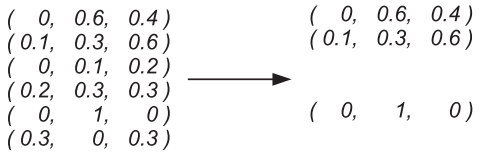



**Fig. 2.** A simplified numerical example of discarding incomplete targets.

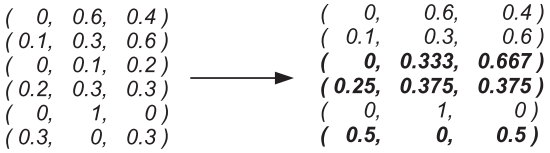



**Fig. 3.** A simplified numerical example of normalized targets.

1. Considering incompleteness of the targets as random noise and making no corrections, i.e., taking no action and using all observations as is.
2. Complete case approach borrowed from the missing data problem. That is, discarding all instances where targets do not follow summing up to unity constraint; i.e., conducting analysis only on complete observations $i$ where $\sum_j \widetilde{y}_{ij} = 1$ (a numerical example is provided in Fig. 2).
3. Normalizing targets of each incomplete observation to sum up to unity, i.e., setting $\mathbf{y}_i^{'} = \widetilde{\mathbf{y}}_i / \sum_j \widetilde{y}_{ij}$ (a numerical example is provided in Fig. 3).

### 2.3. Proposed partial imputation algorithm

We propose correcting each incomplete observation using our derived partial imputation Algorithm 1. The algorithm redistributes missing or unwanted information over the existing fractions of each label. I.e., the difference between 1 and the sum of target values in the incomplete observation is redistributed over the targets.

**Algorithm 1** Partial Imputation of Incomplete Targets

**Data:** Matrix $\widetilde{\mathbf{Y}} = [\tilde{y}_{ij}]$
**Result:** Imputed observations $\widehat{\mathbf{Y}}$
**foreach** *cluster* $t$ **do**
  Select all observations $\tilde{y}_{i,}$ which sum up to unity in the cluster $t$;
  For these observations, compute (by columns) the vector of means $m = \{m_1, m_2, \ldots, m_l\}$;
  **foreach** *incomplete observation* $\tilde{y}_{i,}$ *in the cluster* $t$ **do**
    Compute the difference $k = \{k_1, k_2, \ldots, k_l\} = m - \tilde{y}_{i,}$;
    Assign zero value to all $k_j < 0, j = 1, \ldots, l$;
    Compute proportion vector $p = \{p_1, p_2, \ldots, p_l\} = k * (1/\sum_{j=1}^{l} k_j)$;
    **for** $j = 1$ **to** $l$ **do**
      $\tilde{y}_{i,j} \leftarrow \tilde{y}_{i,j} + p_j * (1 - \sum_j \tilde{y}_j)$
    **end**
  **end**
**end**

Here, we make an assumption that observations with similar target compositions are forming clusters based on features observed in the dataset (cluster assumption). Then, for each distinct cluster, we can calculate averages of each fraction in complete observations with valid distributions. Instead of substituting incomplete observations with these averages, we suggest only filling in the missing parts of each target in the composition. The difference between the average of complete observations and the incomplete observation is used as a proportion guideline of how much of each fraction should be filled in (a numerical example is provided in Fig. 4).

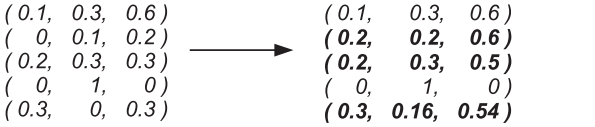



**Fig. 4.** A simplified numerical example of partially imputed targets. Incomplete targets belong to the same cluster where the average of complete targets is (0.2, 0.2, 0.6).

## 3. Experimental evaluation: predicting global vegetation cover

Climatic conditions are the strongest determinants of what vegetation types can exist where (Adams, 2009). Our task is to build a predictive model which would capture global associations between the current distribution of vegetation cover (Fig. 5) and and prevailing climatic conditions. While addressing this task, we must account for structured incompleteness in the natural vegetation.

In this experimental analysis, we firstly investigate how the noise/error in the data, which comes from the incompleteness of targets, can be reduced by the baselines strategies formulated in Section 2.2–2.3, and then we evaluate the performance of our proposed new algorithmic solution.

### 3.1. Dataset

We have a global dataset of $n = 52\,297$ land tiles each described by $d = 47$ climatic features – each associated with two different land cover classification schemes: a distribution over $l = 13$ natural vegetation (and its absence) types and a distribution over $l = 10$ types.

The climatic feature matrix $\mathbf{X}$ is assembled from: BIOCLIM (Fick and Hijmans, 2017), Climdex Climate Extreme Indices (CEI) (Sillmann et al., 2013) datasets and potential evaporation (PET) variable from cru_ts4.01 dataset (Harris et al., 2014) as in Beigaite et al. (2020). The BIOCLIM dataset includes climatic features such as mean annual temperature or mean annual rainfall. The CEI dataset includes various climatic extremes, such as consecutive days without rainfall.

We use ESA CCI LC land cover classification scheme (Poulter et al., 2015) and MODIS (Channan et al., 2014) land cover product (MCD12C1, year 2001) for creating two different natural vegetation cover distribution matrices $\widetilde{\mathbf{Y}}_1$ and $\widetilde{\mathbf{Y}}_2$, respectively. These matrices are created by discarding human activity (croplands, urban areas, croplands, and natural vegetation mosaic) and water columns which we consider as noise.

Additionally, we use elevation information extracted from the HYDRO1k geographic database (Verdin and Greenlee, 1998).

A more detailed description of the dataset can be found in Beigaitė et al. (2022).

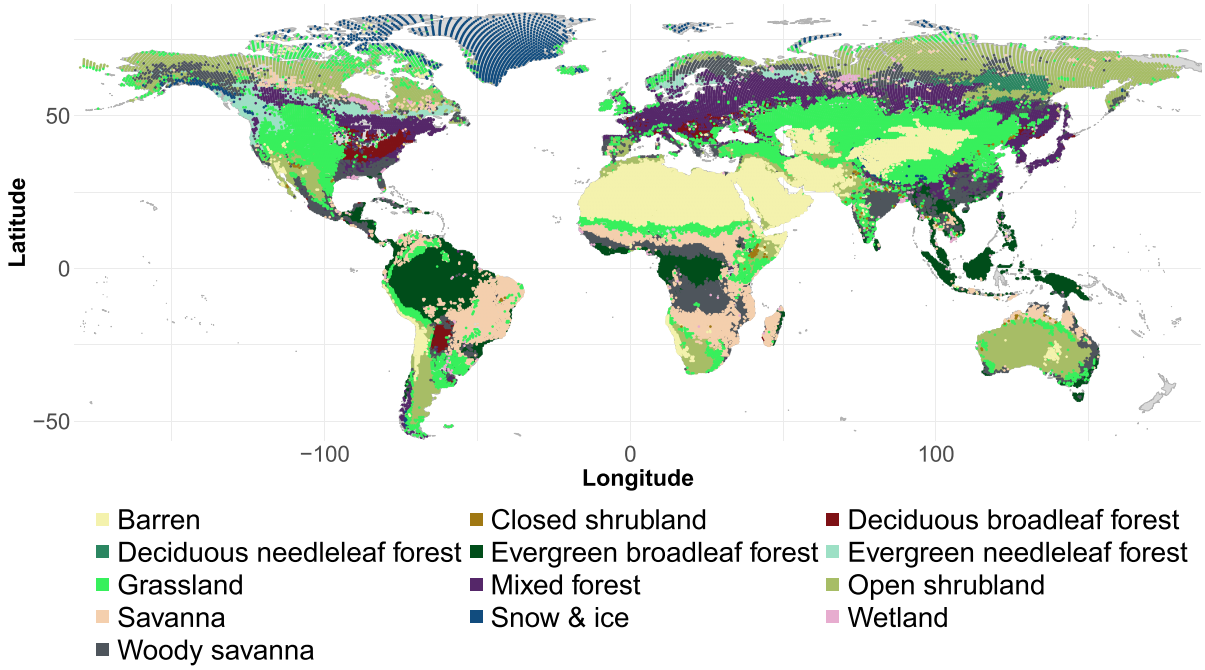



**Fig. 5.** Distribution of vegetation cover (indicated by dominant vegetation type).

### 3.2. Experimental protocol

We carry out two types of computational experiments: (1) evaluating how well incomplete target approaches work in reducing initial structural incompleteness in the data (i.e., direct evaluation), and (2) how the predictive model performance improves when using these approaches (indirect evaluation).

#### 3.2.1. Evaluation of data massaging strategies for handling structural incompleteness

Employing the ESA CCI LC land cover scheme, we evaluate how proposed approaches for handling structural incompleteness work when we artificially introduce incompleteness into the data. In this experiment, we take only complete observations of the matrix $\widetilde{\mathbf{Y}}_1$ and introduce incompleteness in the following three ways:

1. **Random down-scaling**. We reduce each target in the composition by a random percentage.
2. **Random reduction**. We reduce each target in the composition by a random fraction. In this case, some targets can be reduced to 0.
3. **Non-random (biased) reduction**. We reduce targets randomly, as in the random reduction. However, only a few selected targets are being reduced. This way, we imitate the biased incompleteness of only some target labels.

Then, we compare the mean absolute error (MAE) left in the data after we employ baseline approaches from Section 2.2 (taking no action and applying normalization) and partial imputation algorithm (Section 2.3) to this artificially incomplete data. MAE is averaged over 1000 repetitions for each experiment.

In the partial imputation algorithm, we use two types of clusters:

- Combinations of the same degree of latitude where observations are geographically similar (135 distinct latitude degrees with complete observations) and elevation range in the world. Elevation values are divided into 5 range categories: <500 m, [500, 1000], [1000, 2000], [2000, 3000], >3000 m. Total of 675 clusters.
- Climate clusters are derived using the K-Means algorithm on the feature matrix $\mathbf{X}$. For a fair comparison, we use not the optimal number of clusters but the same number (675) as in latitude and elevation clusters.

The true cluster average is used for the imputation of artificially incomplete targets.

#### 3.2.2. Evaluation of accuracy of predictive models

Using MODIS classification, we build predictive models and evaluate their performance when targets are handled with proposed structural incompleteness approaches. For solving the vegetation fraction prediction problem we employ three models internally handling multiple targets:

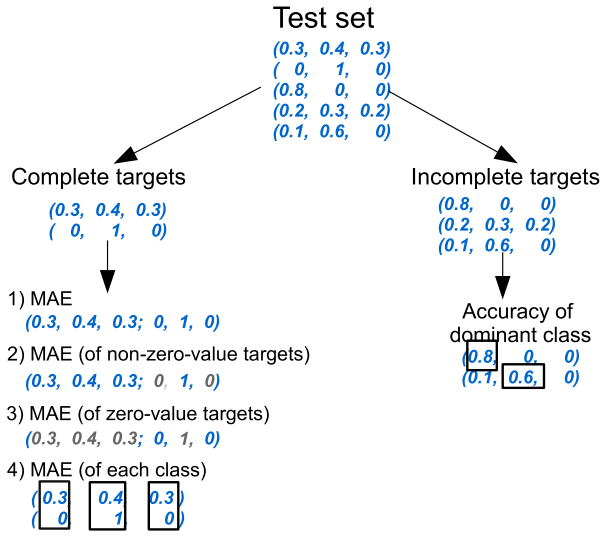



**Fig. 6.** A scheme of performance evaluation (with simplified numerical examples).

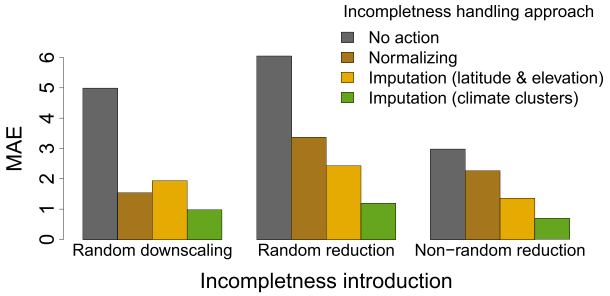



**Fig. 7.** A comparison of errors ($\times 10^2$) in the data when incompleteness handling approaches are applied after introducing structural incompleteness.

1. **A feed-forward neural network**. We chose a fixed architecture of a feed-forward neural network, of three layers. The first layer consists of 47 input neurons for each climatic feature. The hidden layer has 40 neurons with sigmoid activation functions. The third layer consists of 13 output neurons for each land cover type, with a softmax function, to satisfy the constraint of the outputs summing up to unity. The neural network is trained with *Adam* optimizer and the *mean absolute error* loss function using the *Keras* (Chollet, 2015) library. This loss function was chosen as it is more robust to outliers.
2. **Multivariate random forests**. The model is trained using the *Scikit-learn* (Pedregosa et al., 2011) library. We chose 50 trees in the forest, a squared error function for measuring the quality of a split and unlimited tree depth.
3. **K-nearest neighbours regression**. The model is trained using the *Scikit-learn* (Pedregosa et al., 2011) library. We select ten nearest neighbours and computing of Euclidean distance.

We randomly split the data into the training (70%) and testing (30%) subsets. Then, we train each model and examine how incomplete observations affect the accuracy of predictions. We carry out experiments for each incomplete data approach.

### 3.2.3. *Evaluation metrics when true target labels are not known*

We compare the performance of different approaches primarily via the Mean Absolute Error (MAE), which we have chosen since it is easily interpretable and naturally scales for predictions of fraction values between 0 and 1. With this metric, we evaluate the performance only on complete test set observations.

For the incomplete observations, we do not know the true underlying distribution of the target values; we only have noisy or incomplete observations. Thus, it would be misleading to assess the prediction accuracy on those observations. A further challenge is that the complete observations are not uniformly distributed worldwide. Therefore, to evaluate predictive performance on incomplete data, we assess how well the dominant vegetation cover type is predicted and measure the prediction accuracy similarly to a classification task. More specifically, we check whether the largest predicted fraction of each observation equals to the ground-true dominant vegetation cover type. The dominant vegetation cover type is considered to be the one that occupies the largest fraction in a grid cell. In such evaluation, we consider only those

**Table 2**
A comparison of prediction accuracy and errors ($\times 10^2$) with different strategies for handling structurally incomplete targets. Results of the neural network model. Bold font indicates where partial imputation based on climatic clusters leads to the same or better model performance compared to taking no action.

| Approaches → Measures of accuracy ↓ | Only complete targets | Normalized | Imputed (latitude and elevation) | Imputed (climate clusters) | No action |
|---|---|---|---|---|---|
| Accuracy (of incomplete targets) | 77 % | 88 % | 85% | 88% | 89% |
| MAE | 1.74 | 1.88 | 1.89 | **1.81** | 1.82 |
| MAE (non-zero-value targets) | 5.79 | 6.23 | 5.96 | **5.83** | 5.85 |
| MAE (zero-value targets) | 0.08 | 0.13 | 0.20 | 0.13 | 0.14 |
| MAE (Evergreen Needleleaf Forest) | 0.90 | 0.80 | 0.84 | 0.85 | 0.76 |
| MAE (Evergreen Broadleaf Forest) | 0.81 | 0.76 | 0.86 | **0.73** | 0.81 |
| MAE (Deciduous Needleleaf Forest) | 0.53 | 0.47 | 0.44 | **0.45** | 0.46 |
| MAE (Deciduous Broadleaf Forest) | 0.17 | 0.43 | 0.20 | **0.20** | 0.23 |
| MAE (Mixed Forest) | 1.17 | 1.07 | 1.14 | 1.11 | 1.09 |
| MAE (Closed Shrubland) | 0.30 | 0.31 | 0.31 | **0.31** | 0.31 |
| MAE (Open Shrubland) | 5.74 | 6.01 | 6.26 | 6.06 | 5.86 |
| MAE (Woody Savanna) | 2.60 | 2.98 | 2.86 | **2.70** | 2.76 |
| MAE (Savanna) | 1.65 | 2.11 | 2.08 | **1.98** | 2.03 |
| MAE (Grassland) | 4.57 | 5.01 | 4.91 | **4.71** | 4.87 |
| MAE (Permanent Wetlands) | 0.33 | 0.38 | 0.40 | **0.38** | 0.41 |
| MAE (Snow and ice) | 0.49 | 0.51 | 0.53 | **0.52** | 0.53 |
| MAE (Barren) | 3.31 | 3.54 | 3.68 | **3.47** | 3.55 |

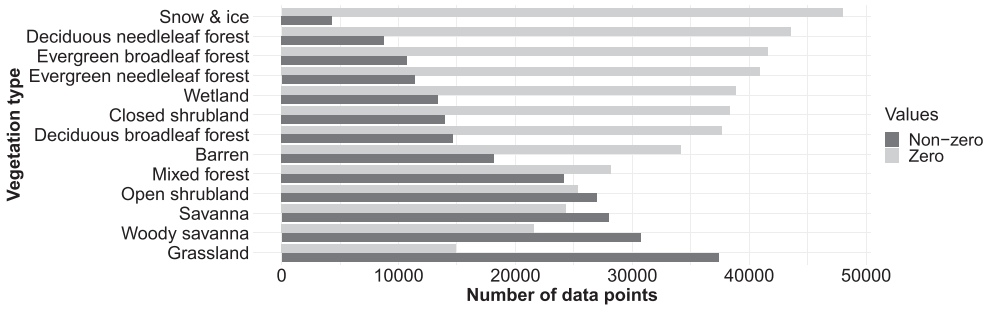



**Fig. 8.** Number of zero and non-zero values in distribution of each vegetation type.

**Table 3**
A comparison of prediction accuracy and errors ( $\times 10^2$) with different strategies for handling structurally incomplete targets. Results of the random forest model. Bold font indicates where partial imputation based on climatic clusters leads to the same or better model performance compared to taking no action.

| Approaches → Measures of accuracy ↓ | Only complete targets | Normalized | Imputed (latitude and elevation) | Imputed (climate clusters) | No action |
|---|---|---|---|---|---|
| Accuracy (of incomplete targets) | 77% | 92% | 90% | 91% | 92% |
| MAE | 1.68 | 1.72 | 1.77 | **1.69** | 1.71 |
| MAE (non-zero-value targets) | 5.26 | 5.31 | 5.37 | **5.21** | 5.43 |
| MAE (zero-value targets) | 0.23 | 0.26 | 0.33 | 0.25 | 0.22 |
| MAE (Evergreen Needleleaf Forest) | 0.89 | 0.78 | 0.79 | 0.77 | 0.76 |
| MAE (Evergreen Broadleaf Forest) | 0.84 | 0.99 | 1.16 | **0.93** | 1.21 |
| MAE (Deciduous Needleleaf Forest) | 0.41 | 0.43 | 0.44 | 0.42 | 0.41 |
| MAE (Deciduous Broadleaf Forest) | 0.22 | 0.23 | 0.22 | 0.21 | 0.20 |
| MAE (Mixed Forest) | 1.04 | 1.03 | 1.09 | **1.02** | 1.02 |
| MAE (Closed Shrubland) | 0.33 | 0.35 | 0.33 | 0.34 | 0.33 |
| MAE (Open Shrubland) | 5.37 | 5.39 | 5.56 | **5.35** | 5.36 |
| MAE (Woody Savanna) | 2.42 | 2.55 | 2.50 | 2.48 | 2.42 |
| MAE (Savanna) | 1.59 | 1.77 | 1.73 | **1.71** | 1.72 |
| MAE (Grassland) | 4.43 | 4.51 | 4.55 | **4.44** | 4.46 |
| MAE (Permanent Wetlands) | 0.38 | 0.48 | 0.46 | 0.45 | 0.44 |
| MAE (Snow and ice) | 0.53 | 0.46 | 0.50 | **0.49** | 0.51 |
| MAE (Barren) | 3.32 | 3.36 | 3.62 | **3.30** | 3.43 |

**Table 4**
A comparison of prediction accuracy and errors ( $\times 10^2$) with different strategies for handling structurally incomplete targets. Results of the k-nearest neighbours regression model. Bold font indicates where partial imputation based on climatic clusters leads to the same or better model performance compared to taking no action.

| Approaches → Measures of accuracy ↓ | Only complete targets | Normalized | Imputed (latitude and elevation) | Imputed (climate clusters) | No action |
|---|---|---|---|---|---|
| Accuracy (of incomplete targets) | 75% | 89% | 87% | **89%** | 89% |
| MAE | 1.76 | 1.81 | 1.87 | **1.78** | 1.83 |
| MAE (non-zero-value targets) | 5.64 | 5.73 | 5.79 | **5.63** | 5.90 |
| MAE (zero-value targets) | 0.19 | 0.23 | 0.31 | 0.23 | 0.20 |
| MAE (Evergreen Needleleaf Forest) | 0.88 | 0.80 | 0.83 | 0.81 | 0.80 |
| MAE (Evergreen Broadleaf Forest) | 0.91 | 1.15 | 1.25 | **1.06** | 1.36 |
| MAE (Deciduous Needleleaf Forest) | 0.45 | 0.47 | 0.49 | **0.47** | 0.47 |
| MAE (Deciduous Broadleaf Forest) | 0.22 | 0.18 | 0.18 | **0.17** | 0.17 |
| MAE (Mixed Forest) | 1.11 | 1.11 | 1.17 | **1.07** | 1.15 |
| MAE (Closed Shrubland) | 0.34 | 0.35 | 0.34 | **0.34** | 0.34 |
| MAE (Open Shrubland) | 5.69 | 5.72 | 5.90 | **5.69** | 5.79 |
| MAE (Woody Savanna) | 2.61 | 2.75 | 2.73 | 2.67 | 2.66 |
| MAE (Savanna) | 1.67 | 1.91 | 1.87 | 1.85 | 1.84 |
| MAE (Grassland) | 4.67 | 4.63 | 4.73 | **4.56** | 4.63 |
| MAE (Permanent Wetlands) | 0.38 | 0.50 | 0.48 | 0.48 | 0.47 |
| MAE (Snow and ice) | 0.52 | 0.52 | 0.55 | 0.55 | 0.54 |
| MAE (Barren) | 3.42 | 3.47 | 3.77 | **3.46** | 3.57 |

observations where the dominant type occupies more than 55% of a grid cell. In addition, for all complete observations, we individually measure prediction errors for each vegetation cover type and separately measure errors for non-zero and zero-valued targets. Such approach allows us to dissect the effects of predictive performance: (1) whether the types that can grow in a given environment are predicted correctly, and (2) to what extent the relative shares of those types in a given environment are predicted correctly.

The scheme or performance evaluation with simplified numerical examples is provided in Fig. 6.

## 4. Results and discussion

### *4.1. Performance of data massaging strategies for handling structural incompleteness*

Fig. 7 summarizes the results of structural target incompleteness handling approaches. As seen from the barplot, if we take no explicit action to address incompleteness in the targets, the MAE error is the lowest when target reduction happens non-randomly. In this scenario, fewer training instances are affected by incompleteness; thus, the overall error is smaller.

We can observe in the figure (Fig. 7) that both normalizing and partial imputation approaches significantly reduce the noise. In all cases, partial imputation based on climate clusters performs better than imputation based on latitude and elevation. This is due to climate clusters reflecting environmental conditions more precisely than the rough grouping by latitude and altitude.

Normalization performs slightly better than partial imputation based on latitudes and elevation when incompleteness is introduced by random down-scaling. However, in other scenarios, the performance of this approach appears to be much worse compared to the climate cluster imputation. This result can be attributed to the fraction of zero-valued

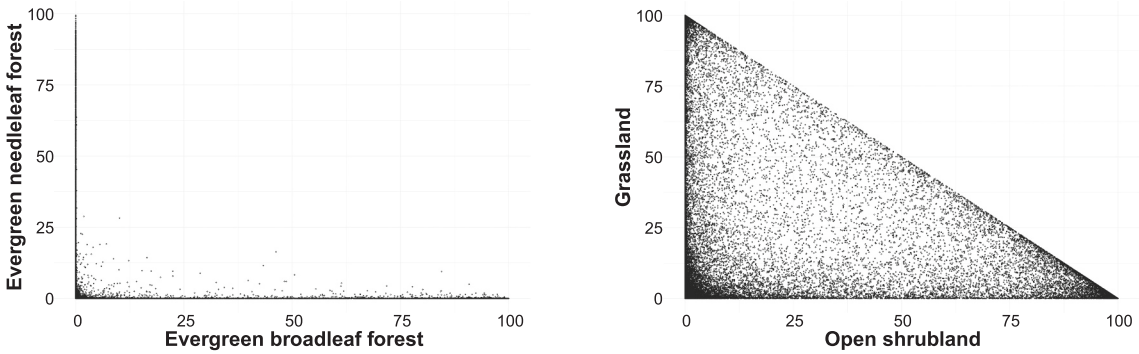



**Fig. 9.** Comparison of pairwise proportions of different vegetation cover types.

targets. If targets are reduced to zero, normalizing cannot recover these fractions and introduce more noise in the fractions as compared to the scenario where non-zero values would have remained present in the training data.

### 4.2. Performance of predictive models

The performance of compared predictive models is summarized in Tables 2–4. When using only complete targets for training, the MAE of complete observations of the test set is the lowest. However, experimental results on evaluating the prediction accuracy of dominant vegetation cover types show that using this approach, the prediction accuracy of incomplete observations is more than 10% less than using other approaches. This suggests that complete observations do not carry full information about the distribution of the natural vegetation cover worldwide. It confirms existing findings (Nikoloski et al., 2021) that better performance can be achieved if incompletely labeled data is used in model training instead of discarding it.

The highest accuracy using the neural network model and the random forest model is achieved when no action is taken while handling the targets. One of the explanations for this superiority could be that, with the lower error than in imputation and normalization cases, this model is better trained to predict the *open shurbland* type, which is the second largest vegetation type. Using k-nearest neighbours, the prediction error of the *open shurbland* type is lower compared to the error when no action is taken. In that case, the prediction accuracy of imputation based on the clusters approach and the no-action approach is the same.

The lowest MAE among other than the complete case approaches is achieved by using partial imputation when trained on the climate clusters. This can be observed for all predictive models. Even though the mean error is similar to the one when no action is taken, it can be seen that the error is decreased for nine out of 13 vegetation types using the neural network model, seven types using the random forest model, and eight types using the k-nearest neighbours model.

When analyzing zero and non-zero-value predictions, we notice that the error of zero-value targets increases in the case of latitude and elevation imputation. This suggests that we can be imputing vegetation types where they are not present.

If we analyze the errors of each vegetation type separately, it is clear that not all fractions of vegetation cover types can be predicted equally well. For instance, fractions of *grasslands*, *open shrublands* and *Barren or sparsely vegetated* are predicted with at least four times higher error than any other vegetation cover type. One of the possible reasons for this could be that these vegetation types can exist in very similar or, in some cases, the same climatic conditions. Fig. 9 represents a visual pairwise comparison of how proportions of different vegetation cover types vary in the same observations. *evergreen broadleaf forests* and *evergreen needleleaf forests* are easily separated as mostly only when one type occupies a fraction close to zero, another type exists in different sizes. Whereas *open shrublands* and *grasslands* coexist in various proportions. Thus, they are easily mixed up by the predictive model.

Overall, we see consistently low errors on the zero-value targets, indicating that the predictive models can capture and predict very well where each vegetation type can be present and where must be absent. For example, in Fig. 10 locations of predictions of *evergreen broadleaf forests* and *deciduous needleleaf forests* are visualised. These locations

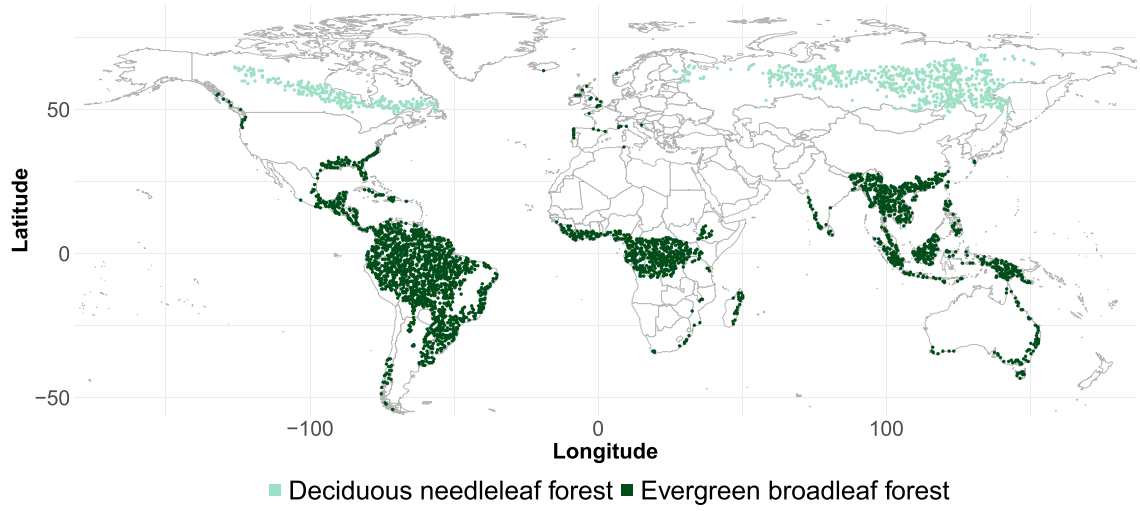



**Fig. 10.** Locations where *Deciduous needleleaf forest* and *Evergreen broadleaf forest* types are predicted to occupy fraction grater than zero.

match geographical areas where *evergreen broadleaf forests* and *deciduous needleleaf forests* can be observed.

While it is more difficult to predict the exact proportion of each vegetation type, our proposed algorithmic solution (partial imputation) brings the most noticeable improvement in predicting these non-zero values. In Fig. 8, we observe that many vegetation types have fractions equal to zero. The *grassland* has the most observations with non-zero values. We can observe that prediction error of this type decreases in all models when our porpoised imputation method is used.

### 4.3. Limitations of incomplete targets approaches

Discarding all incomplete observations leads to the loss of most data points in Europe, North America, and India. Thereby, results can be systematically biased as different parts of the world are not represented equally well in the training set.

If we make an assumption that humans occupy the land with each vegetation cover type equally, normalizing incomplete data could be a good option. However, agricultural competition leads to more intense use of fertile lands (Bičık et al., 2001), and it is more likely for some vegetation cover types to be converted to agricultural lands than others. Hence, by using this approach, we could inaccurately alter the distribution or not be able to recover vegetation-type fractions that are reduced to zero.

Partial imputation algorithm is based on the cluster assumption (Chapelle et al., 2002). In semi-supervised learning, the cluster assumption is considered to apply fairly well to the data which is likely to be clustered. Here we assume that the average of complete observations in a cluster is similar to a true cluster average, i.e., the average of all observations if they were complete. However, these two averages may differ. In that case, the imputation of vegetation fraction would be faulty.

Using latitudes and elevation as clusters is based on the assumption that similar vegetation cover can be found in similar geographical or climatic zones. These climatic zones can be broadly classified by latitudes as different latitudes on Earth receive different amounts of sunlight. However, the climate of the places, which are at the same latitude, can vary depending on altitude or continental position. Therefore, we also have to take into account the elevation of each observation.

### 4.4. Implications and future research directions

Our results highlight the importance of handling incomplete target labels when the noise in the data is not evenly distributed but structured. Other examples of such data could include demographic data, user modeling and recommendation systems data, soil composition data where we do not have the same quality information, or the same quality samples in all regions of the world. In our case study, some locations, e.g., India, appear much more affected by human activity. Failing to address this issue would lead to biased vegetation models for future climate scenarios.

The evaluation of the model accuracy on incomplete data is a challenging task. The compositional structure of the targets enabled us to use the classification accuracy of the dominant vegetation types as an alternative evaluation metric in the regression setting. However, there are still opportunities for further enhancements to both the evaluation of prediction accuracy on incomplete target labels and its handling. For example, as exact altitudinal zones in mountain regions around the world differ, the ranges suggested in this paper could be further tailored.

## 5. Conclusions

We defined a novel computational setting of weakly supervised multi-output regression and proposed an algorithmic approach for handling *structurally incomplete* learning targets. We evaluated how the proposed approach helps reduce the imprecision in the data and compared the performance of predictive models in our global vegetation modeling case study. Our experimental results show that with our proposed algorithmic approach of partial imputation, we can reach the lowest prediction error within different climatic contexts (clusters).

The proposed partial imputation algorithm allows us to preserve the existing information about the vegetation types in training data. At the same time, the algorithm potentially enhances this information by redistributing the fraction of human activity in urban areas and croplands across the training set.

When incomplete data is included in the training set, different parts of the world are more equally represented, and the model avoids overfitting to complete observations only. Imputing incomplete data helps to achieve more accurate predictions of vegetation compositions. This allows us to model global associations between natural vegetation cover and climate potentially more accurately, which is pivotal for understanding the future vegetation response to climate change.

In addition to macroecology and conservation, this computational task setting can potentially apply to other domains, including, for example, soil composition, demographics data or user modeling and recommendation systems under the constraints of eliminating undesired or irrelevant information from predictions when such information is present in the historical data.

## Funding

Research leading to these results was supported by the Academy of Finland (Grant No. 314803).

## Availability of data and material

The data set derived from MODIS, CLIMDEX and BIOCLIM data will be available upon publication at https://github.com/ritabei/Vegetation-cover.

## Code availability

The code of the experiments will be available upon publication at https://github.com/ritabei/Vegetation-cover.

## Authors' contributions

RB, JR and IŽ planned the study. RB designed the computational approaches and carried out the experiments. RB, JR and IŽ analyzed the results. RB prepared the first draft. RB, JR and IŽ finalized the manuscript.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data set derived from MODIS, CLIMDEX and BIOCLIM data will be available at https://github.com/ritabei/Vegetation-cover.

## Acknowledgments

We thank Hui Tang for initial pre-processing of the data.

## References

Adams, J., 2009. Vegetation-climate interaction: how plants make the global environment. Springer Science & Business Media.

Aitchison, J., 1982. The statistical analysis of compositional data. J. Roy. Stat. Soc.: Ser. B (Methodol.) 44 (2), 139–160.

Alarcón, Y.C.C., Destercke, S., 2021. Multi-label Chaining with Imprecise Probabilities. In: European Conference on Symbolic and Quantitative Approaches with Uncertainty. Springer, pp. 413–426.
Allison, P.D., 2001. Missing data, vol. 136. Sage publications.
Beigaite, R., Read, J., Zliobaite, I., 2020. Multi-output prediction of global vegetation distribution with incomplete data. ICML Workshop on the Art of Learning with Missing Values (Artemiss).
Beigaitė, R., Tang, H., Bryn, A., Skarpaas, O., Stordal, F., Bjerke, J.W., Žliobaitė, I., 2022. Identifying climate thresholds for dominant natural vegetation types at the global scale using machine learning: Average climate versus extremes. Glob. Change Biol.
Berikov, V., Litvinenko, A., 2021. Weakly Supervised Regression Using Manifold Regularization and Low-Rank Matrix Representation. In: International Conference on. Mathematical Optimization Theory and Operations Research 447–461.
Bičık, I., Jeleček, L., Štěpánek, V., 2001. Land-use changes and their social driving forces in czechia in the 19th and 20th centuries. Land Use Policy 18 (1), 65–73.
Buchanan, S., Triantafilis, J., Odeh, I., Subansinghe, R., 2012. Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data. Geophysics 77 (4), WB201–WB211.
Channan, S., Collins, K., Emanuel, W., 2014. Global mosaics of the standard modis land cover type data. University of Maryland and the Pacific Northwest National Laboratory, College Park, Maryland, USA.
Chapelle, O., Weston, J., Schölkopf, B., 2002. Cluster kernels for semi-supervised learning. Adv. Neural Inf. Process. Syst. 15.
Chiarucci, A., Araújo, M.B., Decocq, G., Beierkuhnlein, C., Fernández-Palacios, J.M., 2010. The concept of potential natural vegetation: an epitaph? J. Veg. Sci. 21 (6), 1172–1178.
Chollet, F.. Keras. https://keras.io.
Chung, S., Kontar, R., Wu, Z., 2022. Weakly supervised multi-output regression via correlated gaussian processes. INFORMS Journal on Data Science.
Dery, L.M., Nachman, B., Rubbo, F., Schwartzman, A., 2017. Weakly supervised classification in high energy physics. J. High Energy Phys. 2017 (5), 145.
Ferrer-Rosell, B., Coenders, G., Martínez-Garcia, E., 2015. Determinants in tourist expenditure composition—the role of airline types. Tour. Econ. 21 (1), 9–32.
Fick, S.E., Hijmans, R.J., 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315.
Gao, B.B., Xing, C., Xie, C.W., Wu, J., Geng, X., 2017. Deep label distribution learning with label ambiguity. IEEE Trans. Image Process. 26 (6), 2825–2838.
Geng, X., 2016. Label distribution learning. IEEE Trans. Knowl. Data Eng. 28 (7), 1734–1748.
Gomez-Ruiz, E.P., Lacher Jr., T.E., 2019. Climate change, range shifts, and the disruption of a pollinator-plant complex. Sci. Rep. 9, 14,048.
Harris, I., Jones, P.D., Osborn, T.J., Lister, D.H., 2014. Updated high-resolution grids of monthly climatic observations–the cru ts3. 10 dataset. Int. J. Climatol. 34 (3), 623–642.
Hemsing, L.Ø., Bryn, A., 2012. Three methods for modelling potential natural vegetation (pnv) compared: A methodological case study from south-central norway. Nor. Geogr. Tidsskr.-Nor. J. Geograph. 66 (1), 11–29.
Hengl, T., Walsh, M.G., Sanderman, J., Wheeler, I., Harrison, S.P., Prentice, I.C., 2018. Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. PeerJ 6, e5457.
Hoagland, B., 2000. The vegetation of oklahoma: a classification for landscape mapping and conservation planning. Southwest. Nat. 385–420.
Holdridge, L.R., et al., 1967. Life zone ecology. Life zone ecology. (rev. ed.).
Holsinger, L., Parks, S.A., Parisien, M.A., Miller, C., Batllori, E., Moritz, M.A., 2019. Climate change likely to reshape vegetation in north america's largest protected areas. In: Conservation Science and Practice, p. e50.
Hulme, M., Barrow, E.M., Arnell, N.W., Harrison, P.A., Johns, T.C., Downing, T.E., 1999. Relative impacts of human-induced climate change and natural climate variability. Nature 397 (6721), 688–691.
Kelly, A.E., Goulden, M.L., 2008. Rapid shifts in plant distribution with recent climate change. PNAS 105 (33), 11823–11826.
Kostopoulos, G., Karlos, S., Kotsiantis, S., Ragos, O., 2018. Semi-supervised regression: a recent review. J. Intell. Fuzz. Syst. 35 (2), 1483–1500.
Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the köppen-geiger climate classification updated. Meteorol. Z. 15 (3), 259–263.
Levatić, J., Kocev, D., Ceci, M., Džeroski, S., 2018. Semi-supervised trees for multi-target regression. Inf. Sci. 450, 109–127.
Lloyd, C.D., Pawlowsky-Glahn, V., Egozcue, J.J., 2012. Compositional data analysis in population studies. Ann. Assoc. Am. Geogr. 102 (6), 1251–1266.
Mitchell, J.F., Lowe, J., Wood, R.A., Vellinga, M., 2006. Extreme events due to human-induced climate change. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 364 (1845), 2117–2133.
Nikoloski, S., Kocev, D., Levatić, J., Wall, D.P., Džeroski, S., 2021. Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: a case study of water quality assessment in ireland. Ecol. Inform. 61, 101,161.
Pawlowsky-Glahn, V., Buccianti, A., 2011. Compositional data analysis. Wiley Online Library, ISBN 978-0-470-71135-4.
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., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Poulter, B., MacBean, N., Hartley, A., Khlystova, I., Arino, O., Betts, R., Bontemps, S., Boettcher, M., Brockmann, C., Defourny, P., et al., 2015. Plant functional type classification for earth system models: results from the european space agency's land cover climate change initiative. Geosci. Model Dev. 8 (7), 2315–2328. https://doi.org/10.5194/gmd-8-2315-2015.
Quillet, A., Peng, C., Garneau, M., 2010. Toward dynamic global vegetation models for simulating vegetation–climate interactions and feedbacks: recent developments, limitations, and future challenges. Environ. Rev. 18 (NA), 333–353.
Raja, N.B., Aydin, O., Çiçek, İ., Türkoğlu, N., 2019. A reconstruction of turkey's potential natural vegetation using climate indicators. J. Forest. Res. 30 (6), 2199–2211.
Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., et al., 2012. Changes in climate extremes and their impacts on the natural physical environment. doi:10.1017/CBO9781139177245.006.
Sillmann, J., Kharin, V., Zhang, X., Zwiers, F., Bronaugh, D., 2013. Climate extremes indices in the cmip5 multimodel ensemble: Part 1. Model evaluation in the present climate. J. Geophys. Res.: Atmos. 118 (4), 1716–1733.
Snell, R.S., Huth, A., Nabel, J.E., Bocedi, G., Travis, J.M., Gravel, D., Bugmann, H., Gutiérrez, A.G., Hickler, T., Higgins, S.I., et al., 2014. Using dynamic vegetation models to simulate plant range shifts. Ecography 37 (12), 1184–1197.
Sun, Y.Y., Zhang, Y., Zhou, Z.H., 2010. Multi-label learning with weak label. In: Proceedings of the AAAI conference on artificial intelligence, pp. 593–598.
Ummenhofer, C.C., Meehl, G.A., 2017. Extreme weather and climate events with ecological relevance: a review. Philos. Trans. R. Soc. B: Biol. Sci. 372 (1723), 20160,135.
Vaca, R.A., Golicher, D.J., Cayuela, L., 2011. Using climatically based random forests to downscale coarse-grained potential natural vegetation maps in tropical mexico. Appl. Veg. Sci. 14 (3), 388–401.
Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440.
Verdin, K., Greenlee, S., 1998. Hydro1k documentation, US Geological survey.
Whittaker, R.H., 1962. Classification of natural communities. Bot. Rev. 28 (1), 1–239.
Xie, M.K., Huang, S.J., 2018. Partial multi-label learning. In: Thirty-Second AAAI Conference on Artificial Intelligence.
Xu, Y., Zhu, J.Y., Eric, I., Chang, C., Lai, M., Tu, Z., 2014. Weakly supervised histopathology cancer image segmentation and classification. Med. Image Anal. 18 (3), 591–604.
Yao, X., Han, J., Cheng, G., Qian, X., Guo, L., 2016. Semantic annotation of high-resolution satellite images via weakly supervised learning. IEEE Trans. Geosci. Remote Sens. 54 (62), 3660–3671.
Zanelli, D., 2021. Predicting human activities patterns based on climate and related data. Master's thesis. University of Padua.
Zhou, Z.H., 2018. A brief introduction to weakly supervised learning. Natl. Sci. Rev. 5 (1), 44–53.
Zhu, X.J., 2005. Semi-supervised learning literature survey. University of Wisconsin-Madison Department of Computer Sciences.