Date of acceptance Grade

Instructor

Report

Han Lin

Helsinki

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Laitos — Institution — Department Department of Computer Science

Tekijä — Författare — Author
Han Lin

Työn nimi — Arbetets titel — Title

Report

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages 51 pages

Tiivistelmä — Referat — Abstract

In order to understand relationships between organisms and environment and reconstruct the environment in the past, where occurrence of animal species is known from fossils and climate is unknown, we build predictive models using machine learning algorithms. Our target is terrestrial net primary productivity (NPP) which represents fixed energy stored in vegetation as response value for prediction since it is one of the main climate determinants and previous research has shown that NPP can be robustly predicted from dental traits of plant eating mammals [LPE^{+}12]. Since data are not uniformly distributed, models built on enormous training data may generate low prediction accuracy. We propose predictive models that are built from modern day species occurrence data selected in different ways for different geographic regions such that the training data are more similar to testing data.

Considering data are not independently distributed over geographic space and fairness of comparison between our proposed models and global models, we also propose vertical spatial cross validation(VSCV) for evaluating performance of predictive models and we test spatial leave one out cross validation(SLOO).

For local models, we propose three types: baseline models, hierarchical clustering based models(HCM) and advanced hierarchical clustering based models(AHCM). Hierarchical clustering are utilised for identifying similarity of data. In addition, we test ordinary least squares regression, decision tree, random forest, rotation forest and gradient boosting regressor for both global models and local models. Root mean squared error(RMSE) and mean absolute error(MAE) are utilised for measuring performance of regression models. Moreover, we discuss fairness of comparison between local models vs (VSCV) and local models vs SLOO. Finally, a case study is conducted by applying the best model on fossil data and reconstructing environment at Turkana Basin in the past.

Experimental results illustrate that AHCM has the best performance. In addition, number of species for a data point can influence performance of local models and the effect is different in different area in Africa. Last but not the least, We demonstrate applicability of our models with a case study of fossil data from the Turkana Basin in Africa over the last 7 millions of years. Trend of NPP over time for fossil is that NPP firstly decreases slowly and it reaches the lowest value when it is around 2 to 3 Ma. Then, NPP starts increase and tend to be stable. NPP in time period that is larger than 4 Ma and smaller than 7 Ma is higher than NPP in present day in Turkana Basin.

Avainsanat — Nyckelord — Keywords
machine learning

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

It is known from evolutionary theory that organisms interact with and are influenced by physical environment [Dar09]. Relationships between organisms and environment can be described quantitatively by using mathematical models utilizing physical characteristics of organisms as features [Ž16]. Climate and other characteristics of environment can be predicted from occurrence of organisms at present-day where climate and occurrence of animal species is known. Those models are applied to the past where occurrence of animal species is known from fossil and climate is unknown. This study is aimed at building accurate predictive models that would help to analyze and understand the relationships and reconstruct the climate in the past over geological times. Understanding the past helps to understand evolutionary process over the ongoing climate change [BHG+17].

In machine learning context, we typically assume that training data are independently and identically distributed(i.i.d) [H+06]. However, the real species occurrences data are not uniformly distributed over geographic space and distribution is changing over time [ŽPEF17]. We test an idea that we can build predictive models with less training data which are selected in different ways for different geographic regions such that the training data are more similar to testing data instead of building global models that use all the available data. In our study, two types of models are tested: global models and local models. Models built on all available data are global models and models made on a part of data that are available are local models. In accordance with probably approximately correct learning framework [Val84], the generalization error decreases when the number of training data increases, which means a model performs the best when there are infinity training data. But data utilised are not identically and independently distributed(i.i.d), generalization error may increase while the number of training data increase [H+06]. Then how to find good local models is encountered as a research question. So this study propose solutions to this problem.

Our problem setting is that given occurrences of animals and their physical traits, predictive models can be built for inferring productivity of the environment. But those models can not be applied to fossil data since species are different in the past. Instead, we make models on average traits of animal communities. Traits can be measured at present and in the past. Thus, we can apply such models to the past.

Furthermore, three types of local models are illuminated and one of them is a baseline model. In local models settings, data consists of two groups and they are the testing data and rest data that would be selected as training data. Considered the rest data as a set, a group of training data is a subset of the rest data and the criteria for selecting data from the rest data as training data is similarity compared to test data. All models are evaluated by root mean squared error(RMSE) and mean absolute error(MAE) of test data. Furthermore, prediction result of global models, local models, VSCV and SLOO are compared and discussed based on performance of models on test data. In addition, fairness of comparison of VSCV vs local models and SLOO vs local models are discussed. Finally, the best model are applied to predict climate on fossil data in Turkana Basin as a case study.

Furthermore, evaluation of such models is challenging since species occurrence data are not uniformly distributed over geographic space. Cross validation(CV) has been widely used for evaluating performance of predictive models and overcoming overfitting assuming input data are i.i.d [JWHT14]. However, species occurrences data are non-independently distributed over geographical space and they are spatial autocorrelated(SAC) so regular cross validation can mislead to overfitting since data are closely related with each other when they are close in geographical space [PPNH17] [LRPB13]. In other words, data points from nearby ended up in the training and testing pool, it would be almost as if a copy of some training data points is added to the testing data. Thus we modify CV and propose vertical spatial cross validation(VSCV) for assessing performance of regression models and we test spatial leave one out cross validation(SLOO) [LRPM+14]. The idea of designing VSCV is similar to SLOO. A group of data that are adjacent to test data are discarded since those data are likely to be correlated to test data. In addition, those data are discarded also in the purpose of considering fairness of comparison with local models.

The structure of this paper is as follows: Section 2 describes steps of building global models and local models. Section 3 begins with illustration of data and experiment setup. Then, prediction results of global models and local models are presented and analyzed. Section 4 is a case study for fossil data and local models and global models strategies are applied on fossil data for predicting characteristics of environment in

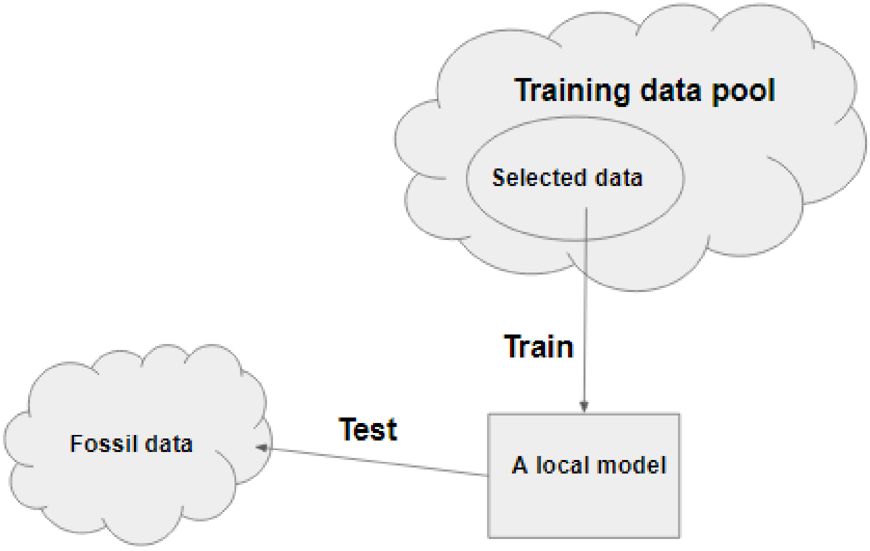

Figure 1: local models

ancient time. Section 5 is an overview of related work. Finally, the last section is conclusion and discussion.

2 Proposed models

In this section, we propose 3 types of local models. Local models are models that are built on data selected from training data pool as shown in figure 1.

2.1 Predictive modeling setting

Our units of analysis instances are areas of land, such as a national park as a grid cell. Input features describe characteristics of animals occurring in those areas. The target for prediction is climate of that area, measured as productivity, rainfall and temperature variable. Assumed that we have data of some part of the modern world where animal occurrences, their characteristics as well as climate variables of those areas are known, our goal is to build predictive models that could be applied to fossil data from regions that are not the same part of the world. We will refer to testing data as "fossil data". Besides, latitude and longitude of a data point describes location of that data point.

2.2 Local Models

In machine learning settings, input data consists of two parts: testing data and potential training data. We propose local models built on data that are selected from those potential training data. The selection criteria is based on similarity compared to fossil data, which means only data that matches fossil data closely are selected. Since data distribution spatially is not uniform. We expect that predictive accuracy would potentially be improved by selecting less training data which match fossil data more closely. Thus in this section, we propose three types of local models.

2.2.1 Baseline Models

To a reasonable approximation, regions with same latitude can be expected to have similar climate and environment, we expect to train the data that are located in the same level of latitude as fossil data. Firstly, two horizontal latitude boundaries can be obtained for fossil data. The top latitude boundary is the largest latitude for fossil data and the bottom latitude boundary is the smallest latitude for the fossil data. Secondly, training data have the same two boundaries as fossil data. Thirdly, a baseline model can be built on the training data by using a regressor. More precisely, a baseline model is made on a part of training data which are located in a region within two boundaries of fossil data. However, in some situations that the number of fossil data can be much larger than the number of selected training data obtained in the second step, this baseline models is not adequate since the number of the training data is too small. Thus the second baseline models are created for improving this baseline models.

The second type of baseline models are also based on the approximation that regions with the same latitude value in both the southern hemisphere and the northern hemisphere have similar environment. This kind of baseline models are similar to the first baseline models. The first step is the same in the second baseline models. In the second step, training data consists of two groups. one group has the same two latitude boundaries as the testing data. The other group has two boundaries with latitude that are symmetric value of boundaries of the testing data where equator is a symmetry axis. In the last step, a type 2 baseline model is built on training data obtained in the second step. We expect that by adding more training data in the second step, the accuracy of this type of baseline model can be improved.

Figure 2 describes the process of building baseline models. Circles and triangles represent input data points. Circles are fossil data and triangles are training data pool. Horizontal lines represent latitudes and vertical lines represent longitude. For type 1 baseline models, boundaries for selecting training data are those two red thick horizontal lines. Models built on triangles that are in red area between two red lines are type 1 baseline model. For type 2 baseline models, training data selected are triangles that are in blue area and red area.

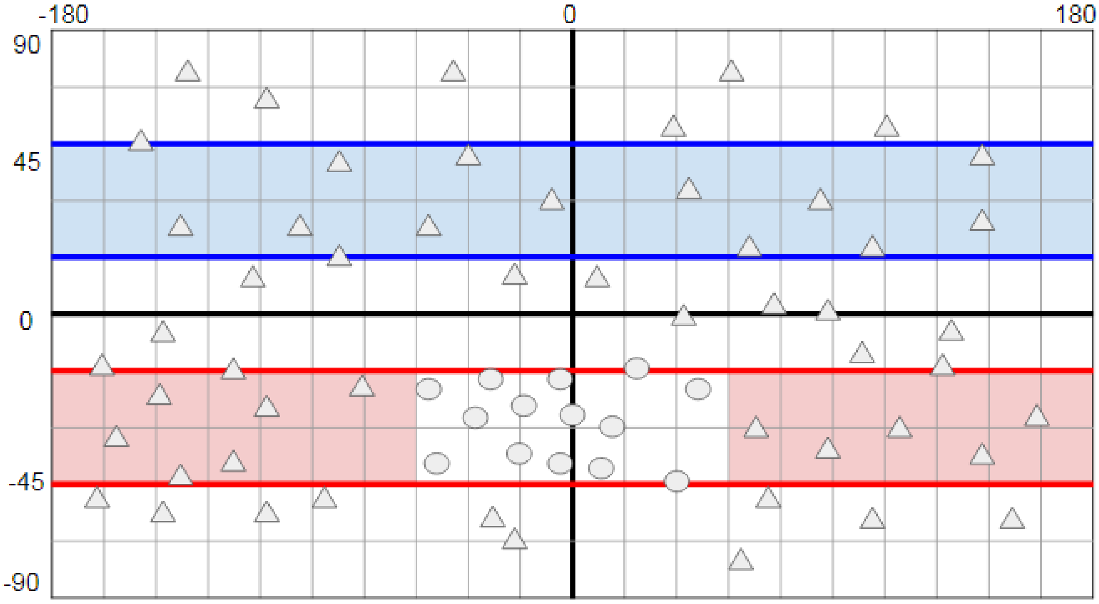

Figure 2: baseline models

2.2.2 Hierarchical clustering based models

In section 2.2.1, training data were manually selected data that are located in regions where climate and environment is estimated to be similar as regions where fossil data are located. Moreover, selected training data are estimated to be similar to fossil data by us. Actually, similarity between two data points can be measured by euclidean distance. Thus distance based clustering method can be utilised for finding groups of data that are similar to fossil data. Hierarchical clustering can describe how clusters are hierarchically related to each other. Thus a sequence of cluster labels illustrating a rank of similarity to fossil data can be obtained. So we use hierarchical clustering to automatically select data that match fossil data closely.

Building a hierarchical clustering based model(HCM) consists of five steps. Firstly, clustering input data, including both data in training data pool and fossil data, based on selected features to several clusters, for example k clusters. The value of k is smaller than the total number of data. Secondly, a collection of unique cluster labels for test data can be obtained, for example $S = \{x_1, x_2, ..., x_n\}, n \leq k$. Thirdly, started from the first element of set S, the cluster x_1 of fossil data are chosen as the testing data in the first loop. According to the result in the first step, a sequence of cluster labels based on similarity compared to the x_1 cluster can be obtained, for instance $R = \{y_1, y_2, ..., y_k\}$. Fourthly, the first m clusters in training data pool are selected for building a predictive model. If we mark training data as a set T, $T = \{cluster_{y_1}, cluster_{y_2}, ..., cluster_{y_m}\}$ and $m < k$. In the fifth step, repeated the third step to the fourth step, cluster in fossil data is changed from x_2 to x_n. Therefore, hierarchical clustering based models are built for all fossil data. Moreover, value of m can be selected by using cross validation.

Figure 3 gives a simple example of process of building a hierarchical based model. All kinds of shapes in the image are data points. Different shape also represent a cluster. For example, round shapes represent cluster 1 and square shapes are cluster 2. Triangular shapes are cluster 3 and cluster 4 are represented by diamond shapes. So it means that both fossil data and data in training data pool are clustered to 4 clusters in this example. Besides, fossil data only contains one cluster which is cluster 1. Assuming that the sequence of cluster labels R is $\{1, 2, 3, 4\}$ for cluster 1 and m parameter is chosen as 2, we select cluster 1 and cluster 2 as training data as shown in image and a hierarchical clustering model can be built on those training data. Algorithm 1 also shows process of building hierarchical clustering models.

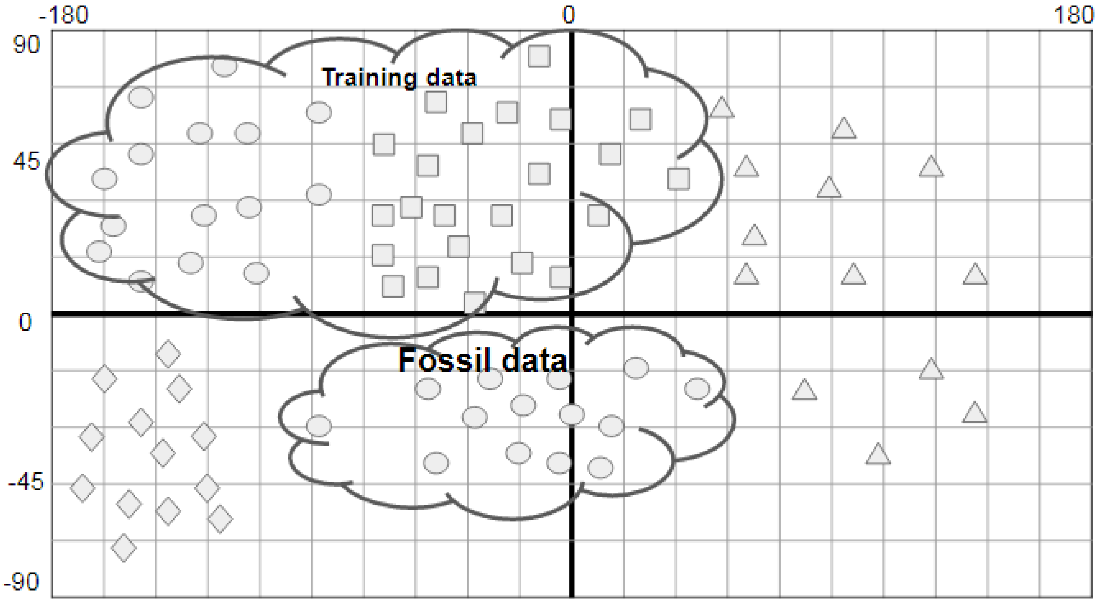

Figure 3: Hierarchical clustering based models

```
Algorithm 1: Hierarchical clustering based models
input  : Data: Fossil ⋃ TrainDataPool, m
output: Hierarchical based models
Fossil [labels], TrainDataPool [labels] ← hierarchicalCluster(Data);
S ← Unique(Fossil [labels]);
for $i$ ← 1 to length(S) do
    SubTestData ← Fossil [labels==S [i]];
    obtain a set R that is a sequence of cluster labels for cluster S [i];
    TrainData ← ObtainTrainData(TrainDataPool, $m$, R);
    Model ← Regressor(TrainData);
end
```

2.2.3 Advanced hierarchical clustering based models

Advanced hierarchical clustering based models(AHCM) are improved version of HCM. After the first step in building HCM, it is possible that the amount of data points in a cluster in fossil data can be large and the amount of selected training data is relatively small. Thus, we expect that partitioning some large clusters in fossil data into several small parts and building models for each small part separately has potential to improve accuracy of prediction. Thus, we propose advanced hierarchical clustering based models. They are based on hierarchical clustering based models.

In the first step, clustering input data into k clusters and Select clusters of fossil data with number of data that is larger than N. We mark those selected clusters of fossil data as $S = \{clusterx_1, ..., clusterx_i\}$, where $x_i \leq k$. In addition, for the rest clusters of fossil data, HCM are utilised for making predictive models. In the second step, Started from cluster x_1, it is clustered into j clusters by using hierarchical clustering. In next step, started from a cluster of cluster x_1, they are concatenated to training data pool as new input data. In the fourth step, the process of hierarchical clustering based models are repeated. In this step, data that are original from the fossil data in the new input data are still a group of testing data for making predictions. Likewise, data that are original from training data pool are still a group of data that are used for building models. In last step, the second step to previous step are repeated until all clusters in fossil data have tested. Algorithm 2 shows the process from the second step to the fourth step.

Algorithm 2: Advanced hierarchical clustering based models

```
input  : Data: fossil $\bigcup$ TrainDataPool, cluster $x_i$ selected
output: Advanced hierarchical clustering models
SubTestData ← fossil [labels== $x_i$];
SubTestData [newlabels] ← hierarchicalCluster(SubTestData);
uniqueLabels ← Unique(SubTestData [newlabels]);
for label in uniqueLabels do
    TestData ← SubTestData [newlabels==label];
    run Algorithm 1(TestData,TrainDataPool);
end
```

3 Proposed model evaluation procedures

In this section, we propose vertical spatial cross validation. Since species occurrence data are not uniformly distributed over geographic space, regular cross validation can mislead to overfitting.

3.1 Vertical spatial cross validation

Figure 4 and Figure 5 give examples of data distribution over geographical space. Those vertical lines represent longitudes and horizontal lines represent latitudes. Those triangles are data points. There are three steps for vertical spatial cross validation. Firstly, input data are partitioned vertically into k equal sized test folds as shown in Figure 5 and it gives an example of partitioning data into 5 folds, thus width of each fold in the image is different since data are not uniformly distributed in the geographic space. Secondly, for the a test fold as shown in Figure 4, two blue thick solid lines are boundaries for the test fold. Thus, the whole data were partitioned into three parts: the test fold, data that are on the left of the left boundary, data that are on the right of the right boundary. For data that are on the left side, those data whose geographical distances to the left boundary are smaller than ξ are discarded; For data that are on the right side, those data whose distances to the right boundary are smaller than ξ are dropped as well. Data excluding the test fold data and data that are discarded are utilised as training data. This process is described in Figure 4. Data points that are located in the red area are discarded and there are cross signs on those data points. Thus, in the second step, started from the first fold whose left boundary has the smallest value, models can be built on their corresponding training data and prediction can be made for the first fold. Thirdly, we repeat the second steps until all k folds are tested. Algorithm 3 illustrates the whole process.

Algorithm 3: Vertical spatial cross validation

```
input  : Data, k
output: Error of a model

[fold 1, fold 2, ..., fold k] ← PartitionData(Data, k) ;
for TestFold in [fold 1, fold 2, ..., fold k] do
    leftBoundary, rightBoundary ← GetBoundries(TestFold) ;
    TrainDataL ← GetTrainingDataL(Data, TestFold, leftBoundary, $\xi$) ;
    TrainDataR ← GetTrainingDataR(Data, TestFold, rightBoundary, $\xi$) ;
    Model ← Regressor(TrainDataL, TrainDataR) ;
    prediction ← fit(Model, TestFold) ;
    Error ← getError(prediction, TestFold) ;
end
```

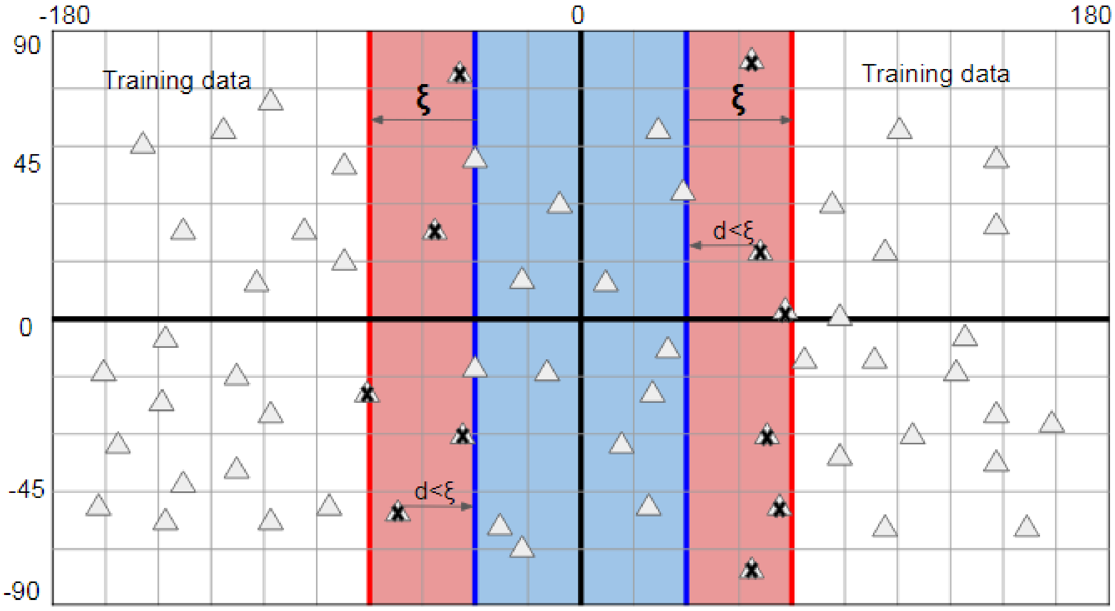

Figure 4: Vertical spacatical cross validation

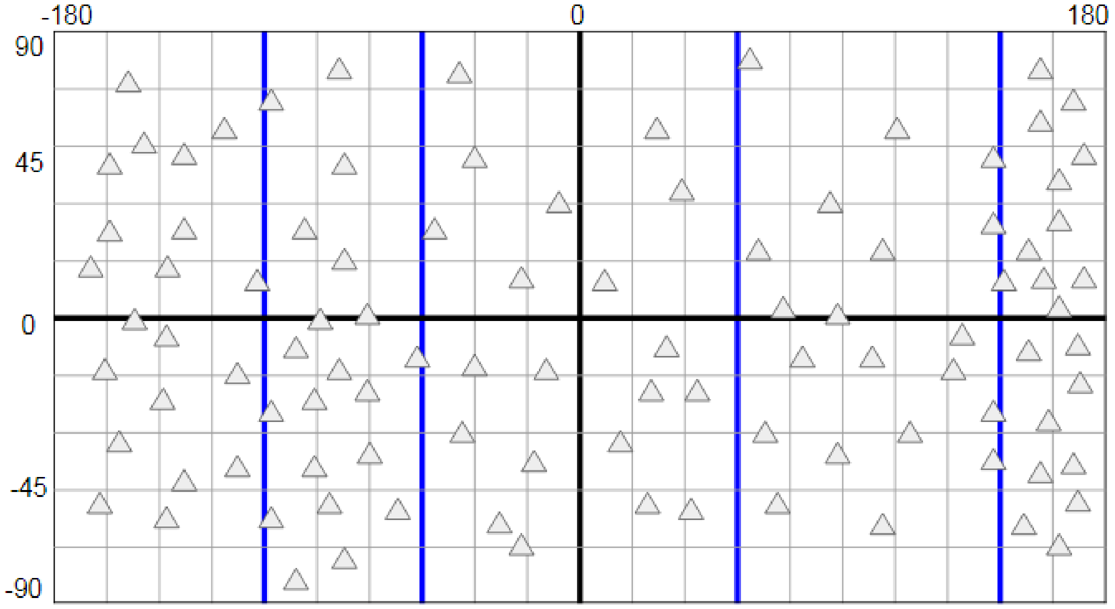

Figure 5: k folds

4 Experimental procedures

In this section, data, experiment set up and parameter settings illustrate in section 3.1 and 3.2. Section 3.3 and 3.4 illuminates the result for proposed models.

4.1 Data

In this study, three datasets are utilised for building models. One of those datasets shows dental traits(taxa × traits). It describes quantitative characteristics of animal teeth. The others reveals climate for each site in the world(sites × bioclimate) and occurrences of taxa for each site(sites × taxa). Table 1 lists all dental traits and possible value for each type in the dental traits dataset. It is the functional dental trait scoring scheme of [Ž16]. In dental traits dataset, it provides values of all dental traits for each taxon.

In the rest two datasets, a site represents a square grid of $50 * 50$ kilometers in the world map. In the sites × bioclimate dataset, there are 19 bioclimate variables describing the climate for each site. In those variables, we use two variables: annual mean temperature(AMT) in Celsius$*10$ and annual precipitation(AP) in millimeters. As illuminated in [LPE+12], NPP(net primary productivity) is the most relative variable to dental traits of plant eating mammals since NPP measures the fixed energy stored in vegetation. NPP is calculated in the following steps: (1)$NPPt = 3000/(1+\exp(1.315-0.119\times AMT))$, (2)$NPPp = 3000\times(1-\exp(-0.000664\times AP))$, (3)$NPP = \min(NPPt, NPPp)$ where NPP is grams carbon in $m^{-2}year^{-1}$dry matter. This climate dataset is from the WorldClim dataset http://www.worldclim.org/. In the dataset of sites × taxa, if a taxon occurrences in a site, it is marked 1 otherwise it is 0. Thus this dataset shows taxa that appears in each site. This dataset is from the list of International Union for Conservation of Nature https://www.iucn.org/. In those two datasets, sites in Australia are excluded since dental traits of the majority herbivore in Australia are different compared dental traits of herbivore in the rest of the world.

In addition, fossil data are data points located in Turkana Basin in Kenya. Each data point represent a site. Fossil data also have 8 features and there are 138 data points. Furthermore, there is also a feature showing time period of the fossil.

4.2 Experimental setup

In the first step, sites that the number of species occurred are smaller than 3 are discarded, taking it account that information of dental traits distribution in those

traits	value
hypsodonty(HYP)	$\in \{1,2,3\}$
longitudinal lophs count(LOP)	$\in \{0,1,2\}$
horizodonty(HOD)	$\in \{1,2,3\}$
acute lophs(AL)	$\in \{0,1\}$
obtuse lophs(OL)	$\in \{0,1\}$
structural fortifications of cusps(SF)	$\in \{0,1\}$
occlusal topography(OT)	$\in \{0,1\}$
coronal cementum(CM)	$\in \{0,1\}$

Table 1: dental traits [GTFŽ17]

sites are not enough for buiding high accuracy of predictive models because of limited number of species [GTFŽ17]. In the next step, the dental traits dataset and occurrences of taxa dataset are aggregated to be the input dataset shown distribution of dental traits. In this input dataset, each data point is the mean value of a dental trait over all species occurred in a site. This process is shown in Figure 6. In this figure, all x value in sites × taxa are either 0 or 1. For any k and p, assuming $1 \leq k \leq j$ and $1 \leq p \leq m$, $z_{kp} = \frac{\sum_{n=1}^{i} x_{kn} * y_{np}}{\sum_{n=1}^{i} x_{kn}}$. However, there are several missing data for a few dental traits in the taxa × traits dataset. Those missing data are skipped in the process of aggregation. In the third step, NPP value on each site are calculated based on the formula illustrated in the section 3.1. Finally, both the input data and data shown NPP for each site are ready for building models and there are 28886 number of data. The input data reveal the distribution of dental traits and NPP reveals the environment in the present day for the whole world. Thus models describe the relationship between them.

In addition, five different regressors are utilised to build models in this project. They are ordinary least squares regression, decision tree(CART), random forest, rotation forest, gradient boosting regressor. Except for the first linear regression model, the rest models are tree based and the last three modes are ensemble models. Tree based regression models can perform well when the relationship between dental traits and NPP is not linear.

Furthermore, we choose Africa as the test continent for both global models and local models. This is because recent Africa environment is relatively least affected by human activities and this way is expected to be similar to fossil data. The amount of Africa data is 8235. In addition, we measure performance of models by calculating root mean squared error and root mean absolute error on testing data.

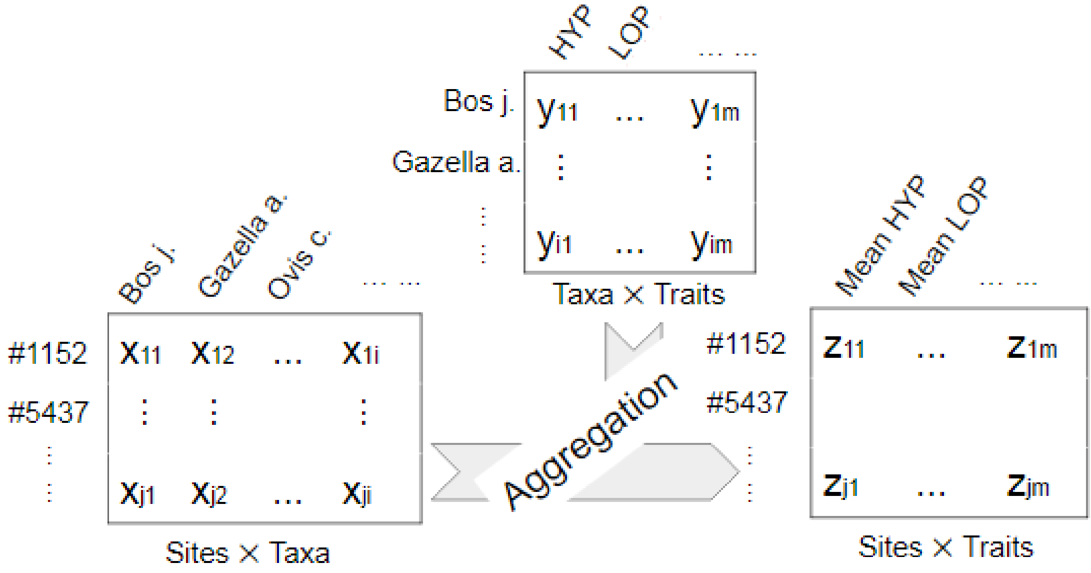

Figure 6: Data aggregation [GTFŽ17]

4.2.1 Global models

For global models, all data that are not in Africa are training data. We test five different regressors separately. Table 2 shows parameters of five regressors. For decision tree and random forest, models are built with parameters changed from 1 to 30. For regressors with more than one parameter, models are built with one parameter changed from 1 to 30 and the rest fixed. In addition, learning rate are tested with value between 0 and 1 with interval value 0.01. Number of subsets of rotation forest can not be larger than 8. Since the total number of features utilised are 8. Thus this parameter are tested from 1 to 8.

4.2.2 Local models: baseline models

So Africa data are test data and training data are selected data that are not located in Africa when building local models. Africa data are not included in training data

regressors	parameters
decision tree	number of depths
random forest	number of estimators
gradient boosting regressor	number of estimators, learning rate and number of depth
Rotation Forest	number of subsets and number of trees

Table 2: parameters for models

since Africa data act as fossil data for validating models in real case. Africa data are partitioned into ten horizontal layers and the height of each layer is not larger than 5 degrees(almost 555 kilometers) for both two types of baseline models. Parameters settings for rotation forest is the same as global models' and parameters settings for other regressors are tuned until RMSE and MAE reach the smallest value or are stable in some small value. In addition, data with latitude that is larger than 12.74 in the northern hemisphere are not included as test data since there are a few data points and their distribution in the map is dispersive.

4.2.3 Local models: Hierarchical clustering based models

The whole data are partitioned into ten clusters using hierarchical clustering. Figure 7 shows distribution of different clusters in the world map and Figure 8 is the dendrogram revealing similarity of different clusters and ten leaves nodes are the first cluster to the tenth cluster from the left corner leaf to the right corner leaf. A rank of clusters labels as mentioned in section 2.3 is generated from this dendrogram. For example, if a cluster labeled 6 in Africa are selected as a group of test data, a rank of cluster labels for selecting training data(without Africa data) is $\{6, 5, 9, 10, 7, 8, 3, 4, 1, 2\}$. The most similar data to test data which are labeled 6 is the 6th cluster in the training data. Furthermore, clusters in red have higher similarity than clusters in green since the 6th cluster is red and clusters in red can be merged in a bigger cluster as shown in image 8. In addition, the 6th cluster and 5th cluster can be merged as a cluster so the 5th cluster in training data is in the second position in the set of the rank of cluster labels. Moreover, cluster 3 and 4 have higher rank than cluster 1 and 2 since the distance of cluster 3 and 4 is smaller than cluster 1 and 2. Finally, in this study, Ward's linkage is chosen since it provides more reasonable clustering.

Each cluster in Africa are tested respectively and those regressors: ordinary least squares regression, decision tree(CART), random forest, rotation forest and gradient boosting regressor are tested for each cluster. A regressor that has the minimum RMSE and MAE is chosen for a cluster in Africa. The criteria for selecting a group of training data from data without Africa data for a cluster of test data is based on the minimum RMSE and MAE.

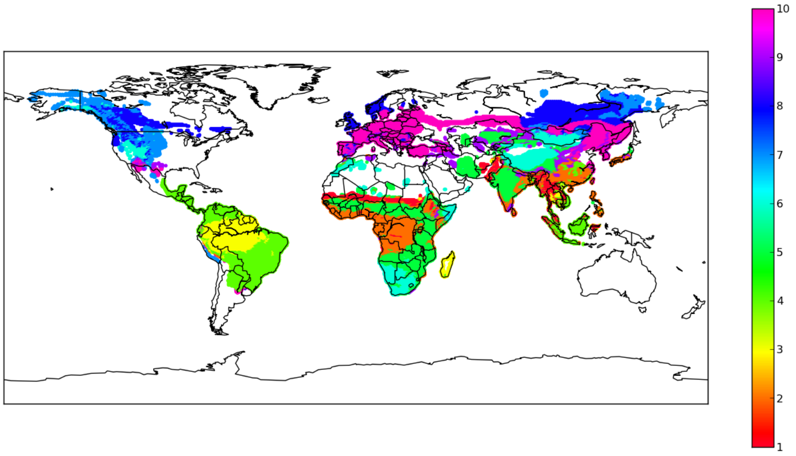

Figure 7: clusters on the world map

4.2.4 Local models: Modified hierarchical clustering based models

In order to improve performance of hierarchical clustering based models, an optimization strategies are utilised in making predictions. Since he number of data in a cluster in test data can be large and the region which those data located in can also be large. Thus data in a large cluster are partitioned into some small groups and models are tested on those groups seperately.

This optimization strategie consists of several steps. Firstly, a cluster in test data is selected. Secondly, that cluster is partitioned in a horizontal way into some layers and the height of layers are almost the same. This step is the same as the first step in building the first baseline models in section 2.2. Thirdly, started from the first layer data of the cluster, they are test data. As mentioned in section 2.3, a set R

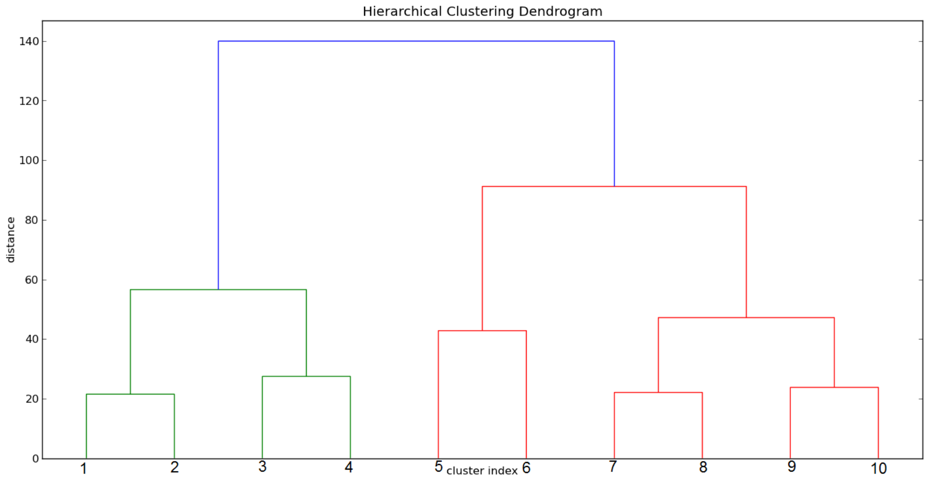

Figure 8: dendrogram

which is a rank of cluster labels based on similarity for a cluster can be obtained. Thus, a set $R = \{y_1, y_2, ..., y_k\}$ for this cluster can be obtained. The fourth step is the same as the step four in section 2.3. Started from the y_1 cluster in training data, a following cluster of the cluster in previous round is appended in next round until all training data are included in building a model. In step five, the best model of the first layer can be obtained according to accuracy. In the last step, the third step to the fifth step are repeated until all the best models of all layers are obtained. Thus part of that cluster that performs the best can be discovered. Algorithm 4 reveals the whole process. If there are more than one clusters that need to be analyzed, this algorithm can be ran for each cluster respectively.

Algorithm 4: Local Models: Modified hierarchical clustering based models

```
input  : Data: TestData + TrainData, $m$ layers, cluster $x_i$ selected
output: An array of error for all layers
SubTestData ← TestData [labels== $x_i$];
[layer$_1$, layer$_2$, ...layer$_m$] ← SubTestData;
obtain a set R that is a rank of cluster labels for cluster $x_i$;
for layer in [layer$_1$, layer$_2$, ...layer$_m$] do
    for $j$ ← 1 to length(R) do
        TempTrain ← TrainData [labels==R [j]];
        TotalTrain ← Combine(TotalTrain,TempTrain);
        Prediction ← Regressor(TrainData, layer);
        error ← GetError(Prediction, layer [realValue]);
        ArrayError ← append(error, ArrayError);
    end
    minError ← Min(ArrayError);
    AllErrors ← append(minError, AllErrors);
end
```

The cluster 1 in Africa are partitioned into three horizontal layers and the height of each layer is smaller than 333 kilometers and data of cluster 1 with latitude that is smaller than 9.6 are discarded since data with latitude that is below 9.6 are distributed dispersedly and the number of those data are not large. Each layer are tested as the way in hierarchical clustering based models. Training data are selected also from the data without the Africa data. In the process of selecting training data, a rank of cluster labels of test data is also generated as the order to select data as training data. Those five regressors are tested for each layer for the purpose of obtaining the best model with minimum RMSE and MAE.

The cluster 2 in Africa are partitioned into 6 layers and height of each layer is almost 555 kilometers. Data of the cluster 2 with latitude that is above 14.99 or below -14.54 are not included as test data with the same reason mentioned above. The cluster 5 in Africa are partitioned into ten layers as the way for the cluster 2 and data of the cluster 5 with latitude that is above 12.74 are also discarded. Finally, the cluster 6 are partitioned into four layers. Each layer in those clusters is tested as the same way for the cluster 1.

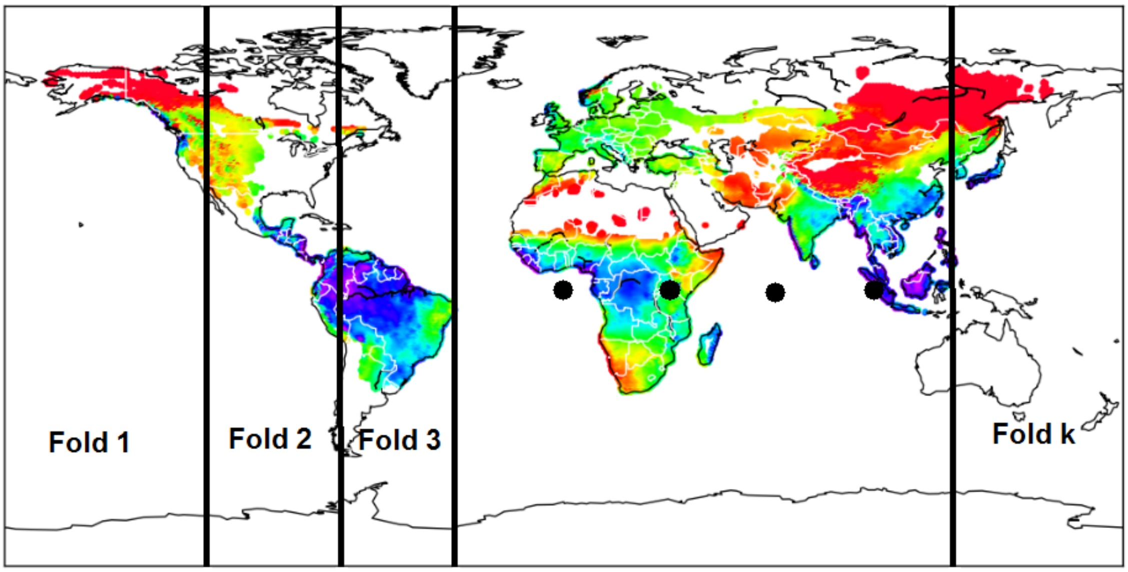

Figure 9: k test folds

4.2.5 Local models: Advanced Hierarchical clustering based models

Like in type 1 advanced hierarchical clustering based models, cluster 1, 2, 5 and 6 in Africa are partitioned into several small groups of data. Cluster 1 are clustered to 3 small clusters. Cluster 2 are partitioned into 6 clusters. Cluster 5 are clustered to 10 clusters. Cluster 6 are partitioned into 4 clusters. Each obtained small cluster are tested respectively and a new set of the rank of cluster labels for a small cluster is generated by combining that cluster with the rest of data without Africa and clustering them again. In this model, there is no test data that are discarded. Hierarchical clustering with ward's linkage method are utilised for clustering data.

4.2.6 Evaluation procedures of models

In this section, strategies of validating both global models and local models are illustrated and evaluation protocols are also described.

4.2.6.1 Vertical spatial cross validation and SLOO

In machine learning, cross validation is used commonly as a method to improve performances of models. Thus, we train and test models as the way in cross validation. For using vertical spatial cross validation, three steps are conducted in our experiment. Firstly, input data were partitioned vertically into 11 equal sized test folds as shown in Figure 9, thus width of each fold in the world map is different since data were not equally distributed in the map. Thus there are 2626 number of data for each test fold. Secondly, for the xth test fold as shown in Figure 4, two blue thick solid lines are boundaries for the xth test fold. Thus, the whole data were partitioned into three parts: the test fold, data that is on the left of the left boundary, data that is on the right of the right boundary. For data that is on the left side, those data whose distances to the left boundary are smaller than 300 kilometers are discarded; For data that is on the right side, those data whose distances to the right boundary are smaller than 300 kilometers are dropped as well. Finally, data excluding the test fold data and data that are discarded are utilised as training data. Data that are located in the grey area are discarded. Thirdly, models can be built from training data for each test folds and prediction are made for each test folds. This process is marked as spatial 11 folds cross validation.

For using spatial leave one out cross validation, in each training and testing loop, one data poin acts as testing data and some of the rest data are training data. The process contains several steps. In the first step, a data is selected as a test data. In the second step, all the data that is located in the point where is at least 500 kilometers away from the test data are training data. This process is marked as spatial leave one out cross validation. In addition, normal 11 fold cross validation and leave one out cross validation are also tested for comparison with global models solutions that we proposed.

Considering the running time of rotation forest is large, it is not tested in normal leave one out cross validation and spatial leave one out cross validation and the rest four models are tested.

In our experiment root mean squared error(RMSE), mean absolute error(MAE) and r square are utilised for measuring performance of models. RMSE gives relatively high weights to large errors and large errors are undesirable in our experiment so it is used. MAE is utilised to measure the accuracy of prediction. Both RMSE and MAE are negatively-oriented scores. Furthermore, in some situations r square is not utilised for comparison performances of models since the variance of different groups of test data can be different and this can influence the value of r square.

In our experiment, decision tree(CART), random forest, gradient boosting regressor are libraries in sklearn in python. OLS regression is from statsmodels package. Rotation forest is tested from source codes. Parameters for each model is shown in Table 3 and parameters settings of four models for k fold cross validation are also the same as Table 3.

More importantly, since the number of clusters selected in the training data are chosen by RMSE and MAE on the whole Africa data which are testing, thus this may cause overfit. Finally, in the last step of our experiments, we divide data on African continent into 3 parts vertically and each part has the same amount of data points. We evaluate our models as steps in vertical spatial cross validation. Namely, in the first step, we choose a part as testing data and we discard a small amount of data whose distance to boundaries of testing data are smaller than 300 kilometers. Then, we select number of clusters in training data based on data that are not discarded and are not testing data. Those data are called validation data. So we test three rounds until all data in Africa are acted as testing data once.

models	leave one out cross validation(normal and spatial)
decision tree	20 layers
random forest	normal: 10 estimators; spatial: 25 estimators
gradient boosting regressor	7 estimators,learning rate is 1.2 and maximum depth is 10
Rotation Forest	k is 2 and number of trees is 25

Table 3: parameters for models

5 Result Analysis

In this section prediction results of all models are described and analyzed. In addition, results of two types of evaluation strategies are discussed.

5.1 global models

Figure 10 to Figure 12 illuminate trends of training error and testing error with the change of a parameter of a regressor. In those images, green lines and light blue lines represent training error, and yellow lines and dark blue lines represent testing error. Here, training error are RMSE and MAE calculated by predictions made by models on data points that are not located in Africa data. Testing error are for Africa data. For four regressors with parameters, with increasing of parameter value, training error is decreasing until it reaches the smallest value and the error tends to be stable. The smallest training error can be around 360 for RMSE. Namely, it means that a global model has a smallest training error when a parameter of a regressor is large. As shown in figures, training error reaches the smallest when the parameter reaches the largest value 30. However, the trend of testing error with the increasing of a parameter is not the same as training error. The testing error can reach the smallest error when a parameter is a small value. As shown in Figure 10, RMSE of testing error reaches the smallest value which is 649 when number of depth is 13. However, when the number of depth is 13, training error is not the smallest value. In addition, if a regressor needs more than one parameter, different parameter can have different influence in building a good model. For example, the best RMSE is 512 when the number of depths is 7 and other two parameters are default value for gradient boosting regressor as shown in Figure 11. However, in Figure 12, the smallest RMSE is 525. Thus, the parameter of number of depth can have more effect on building a good model.

Table 4 shows error of five global models with the best parameter setting. Gradient boosting regressor has the best prediction result. However, the minimum RMSE is 512, so we need to find a better strategy to build better models with high accuracy. Thus we propose local models.

5.2 local models

In this section, results of all local models are discussed.

models	RMSE	MAE
OLS	623	516
decision tree	649	501
random forest	523	421
gradient boosting regressor	512	416
Rotation Forest	529	433

Table 4: Minimum error for testing data

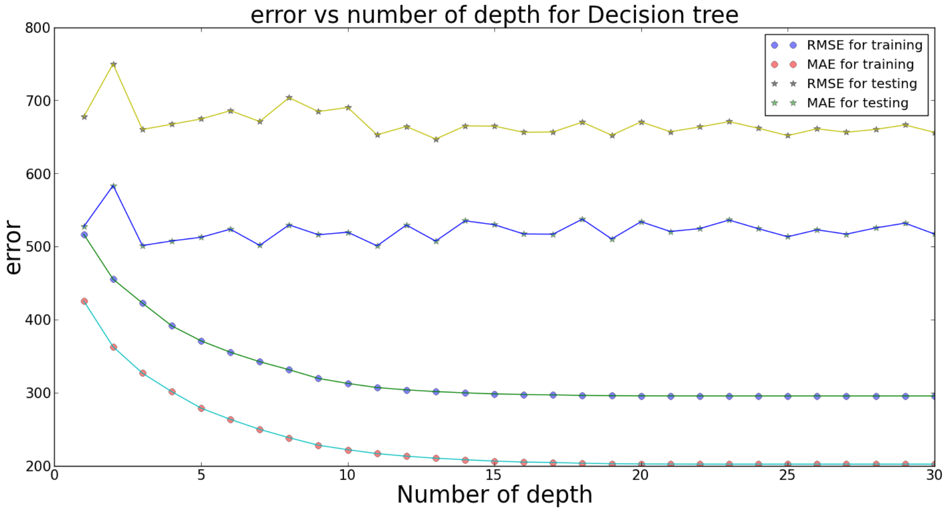

Figure 10: This image shows the change of error with the increasing of number of depths of decision tree

5.2.1 Baseline models

Table 5 illustrates performance of models over all layers so r^2, RMSE and MAE are calculated over the whole test data. Compared results in type 1 baseline model with results in type 2 model, performance of type 2 model improved a lot. One possible reason is that the number of training data for each layer increased. In addition, random forest and gradient boosting regressor have the best performance. Figure 15 and Figure 16 show the number of test data and training data in each layer and error for each layer of gradient boosting regressors. In Figure 15 and Figure 16, heights of red bars represent the number of test data and green bars stand for number of training data. In addition, those lines represent error value for layers. Figure 14 and Figure 13 show RMSE and MAE in a clearer way. Further-

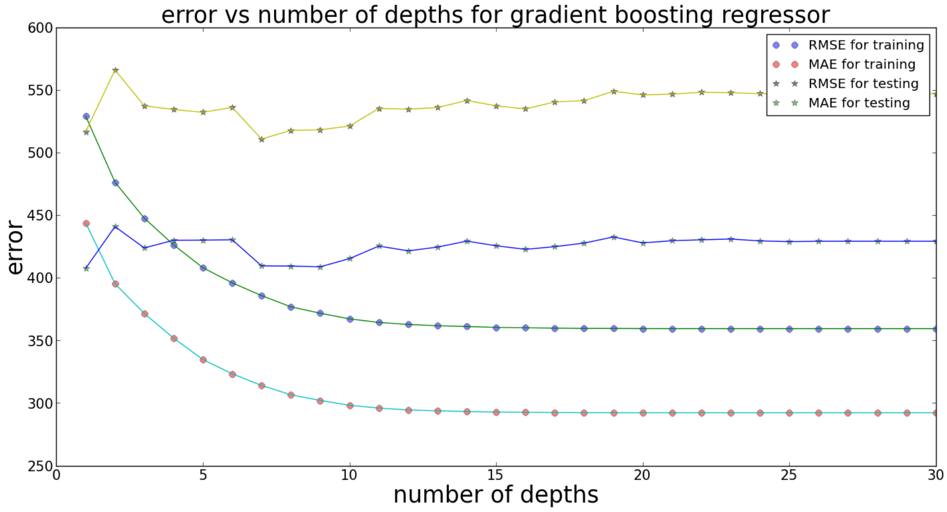

Figure 11: This image shows the change of error with the increasing of number of depths of gradient boosting regressor

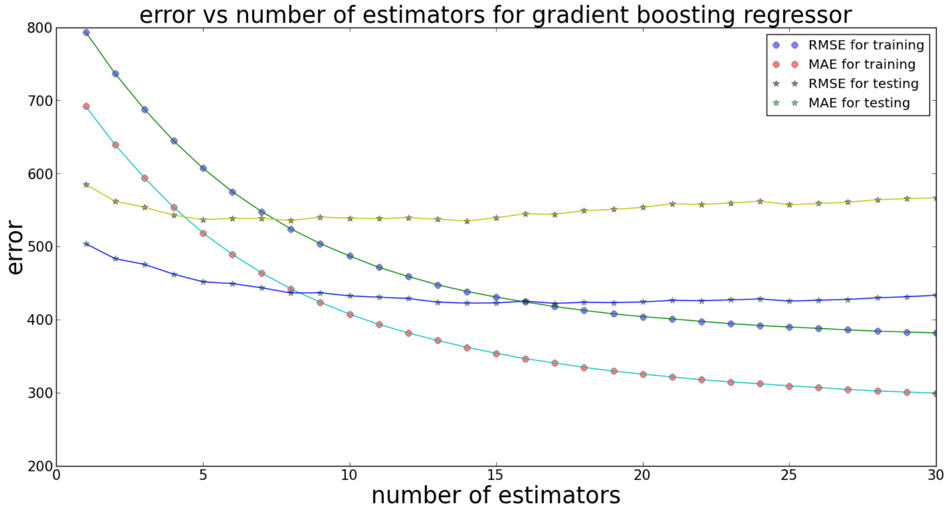

Figure 12: This image shows the change of error with the increasing of number of estimators of gradient boosting regressor

more, performance of models in layers increase with the number of training data increasing except for the model of the sixth layer. Although the number of training data in the type 2 baseline model in the sixth layer were almost two times larger than that in the type 1 baseline model, error value in the sixth layer of type 2 baseline model are increased. More importantly, the model in the sixth layer of type 1

Type 1 baseline model

	r^2	RMSE	MAE
OLS regression	$-1.79 * 10^{24}$	$7.40 * 10^{14}$	$2.62 * 10^{14}$
decision tree	-1.20	820	676
rotation forest	-4.27	1269	1157
gradient Boosting regressor	-0.710	723	567
random forest	-0.727	726	559

Type 2 baseline model

	r^2	RMSE	MAE
OLS regression	-3.43	1162	870
decision tree	-0.341	640	503
rotation forest	-3.01	1106	994
gradient Boosting regressor	0.0410	541	422
random forest	0.00291	552	418

Table 5: result of baseline models

baseline model have the best performance compared to models in rest layers. Thus, some data appended to training data in type 2 baseline model are not related to test data as a result of higher error value in type 2 baseline model. Therefore, to some extent, increasing number of training data can enhance performance of models but if part of training data are not that similar to test data, this can also impair performance of models. However, performance of the best model of baseline models is not that good and more advanced local models are needed for better performance.

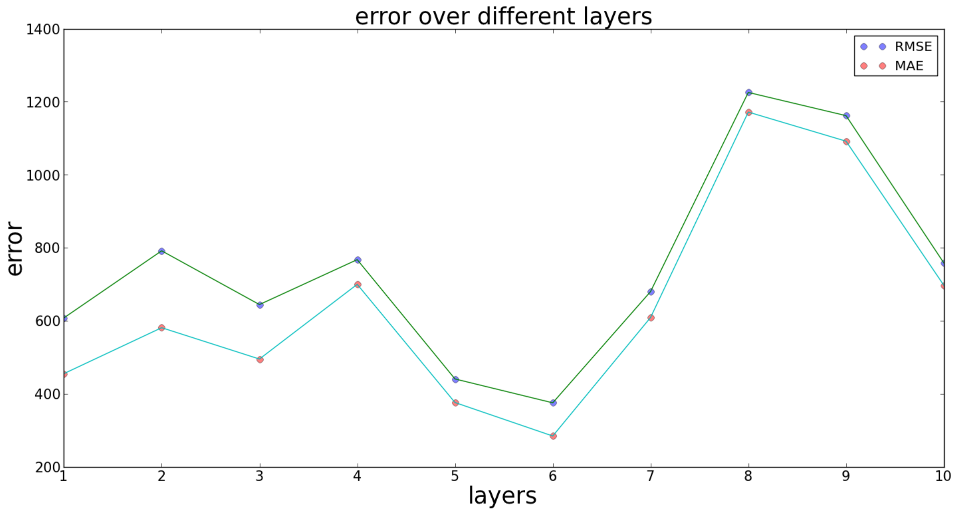

Figure 13: type 1 baseline model

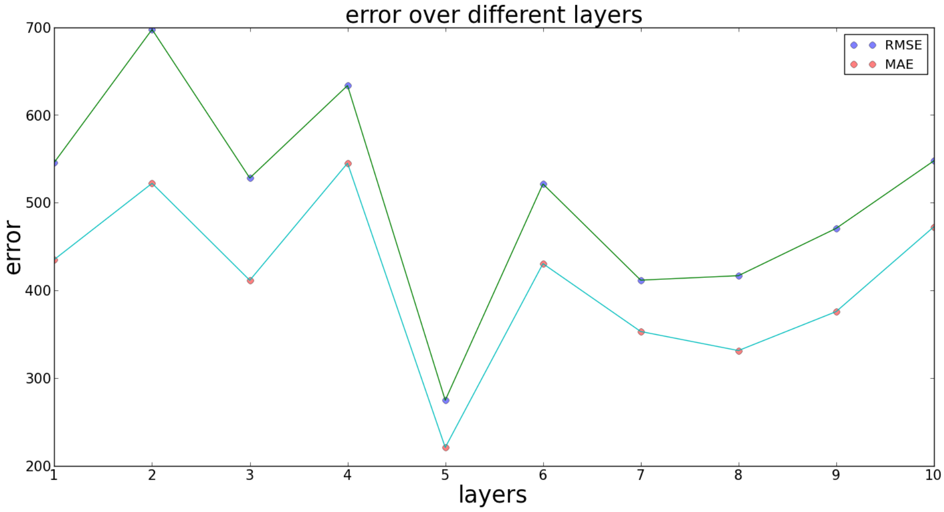

Figure 14: type 2 baseline model

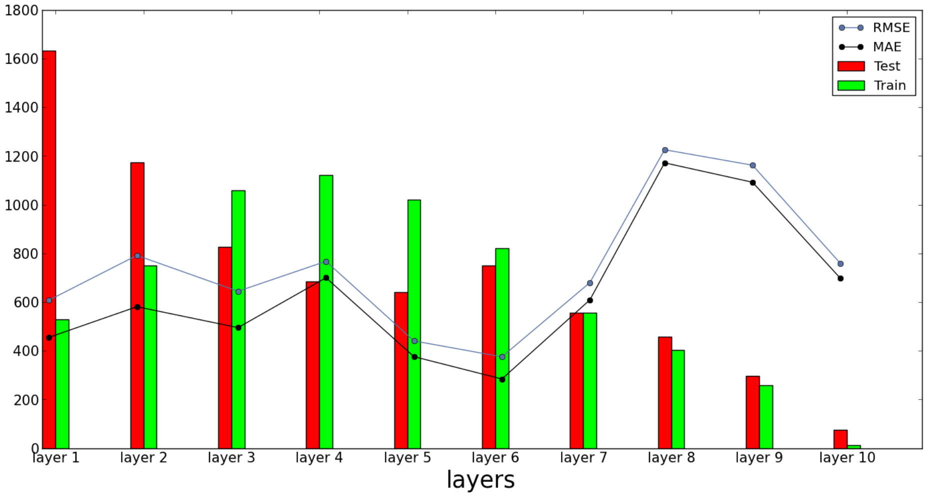

Figure 15: type 1 baseline model

5.2.2 Hierarchical clustering based models

Africa data are labelled as 8 different clusters and they are cluster 1 to 7 and cluster 9. Figure 19 shows number of all clusters. A height of a red bar in Figure 19 represents the total number of a cluster and a cluster of Africa data is a group of test data. The cluster labelled 5 is the largest cluster. It consists of 3668 number of data and it covers savannas area. The cluster labelled 6 covers desert area and

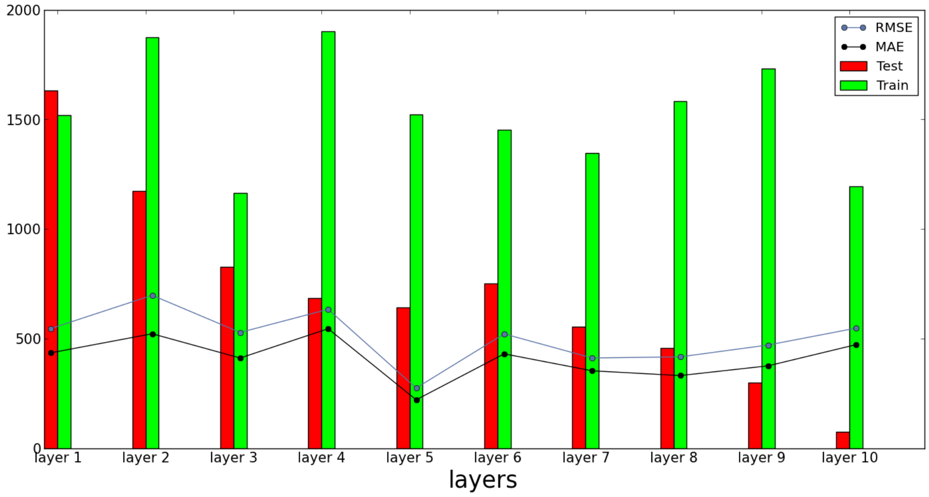

Figure 16: type 2 baseline model

cluster 2 covers forest. In addition, Figure 17 and Figure 18 provides an example showing the process of selecting training data. In Figure 17, those green bars represent the change of training data and red bars represent test data which are cluster 6 in Africa. Figure 18 reveals the change of error value with the increase of selected training data. The horizontal axis shows the training data increase from 1 cluster to 10 clusters. Both RMSE and MAE decrease while the training data increase but when there are more than 6 clusters, error increase acutely. As mentioned in section 2.3.2, those 6 clusters consists of a sequence of clusters: $\{6, 5, 9, 10, 7, 8\}$ and those 6 clusters can be merged in a bigger cluster as shown in Figure 8. Thus, the group of clusters: $\{3, 4, 1, 2\}$ diminish performance of model. In addition, the model has the best performance when training data consist of only two clusters. Thus, compared to global models, local models cost less running time since even all clusters of training data are selected, the amount of training data in local models is smaller than global models' because the whole Africa data are not included in the training data.

Figure 19 shows the number of test data for each cluster in Africa and the amount of selected training data. Two lines in the figure represents the change of error. Figure 20 shows error clearly. Both the cluster 4 and cluster 7 have relatively low error value since the number of data in those two clusters are very small. Cluster 3 has the worst performance and those data are mainly located in Madagascar. The performance of model for both the forest area and desert area are almost same and it is better than the performance of models for savannas area. Moreover, as shown in Figure 19, the amount of training data is smaller than the amount test data for the cluster 2 and the performance of models for the cluster 2 is good. This reveals that it is reasonable to find a group of data that are most related to test data when data are not i.i.d.

Figure 21 to Figure 24 illuminate the change of error with the number of species. In Figure 21, error reaches the lowest point when sites that has 10 species. However, error increase when the number of species is over 10. This trend shows that if the number of species of data for desert area is around 10, the model is possible to have good performance. For forest area, as shown in Figure 22, error decrease with increasing of number of species. Thus, if data with high number of species in forest area, the prediction performance on those data can be good. Figure 23 shows a different trend over savannas area. Data with small number of species(< 5) or with large number of species in savannas can be predicted with small error. Figure 24 illuminates that limited number of species on sites can cause high error in Madagascar. Furthermore, Figure 25 shows that the number of species in Madagascar are limited. Thus this can be the reason that the performance of models in Madagascar is small.

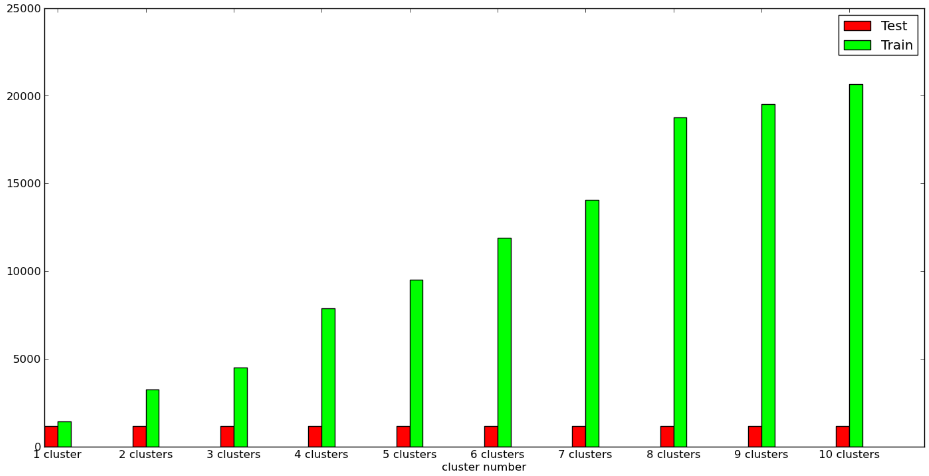

Figure 17: number of data

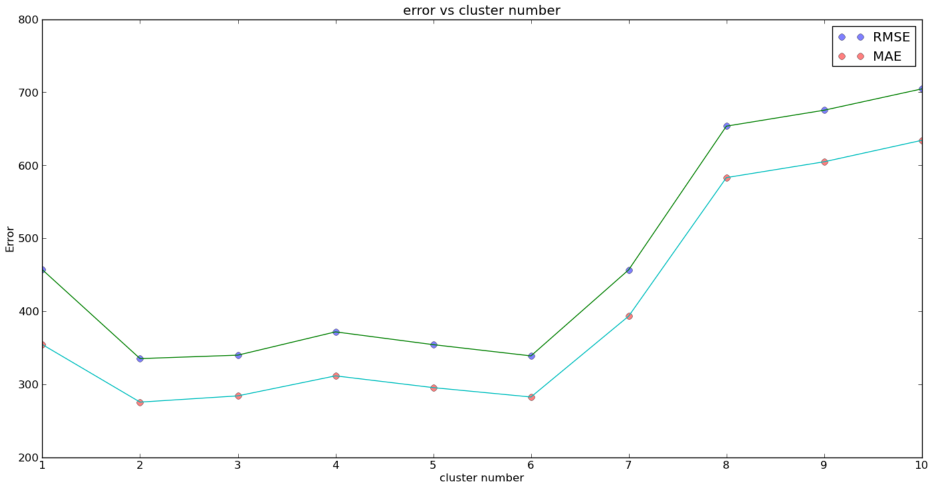

Figure 18: error for the sixth cluster

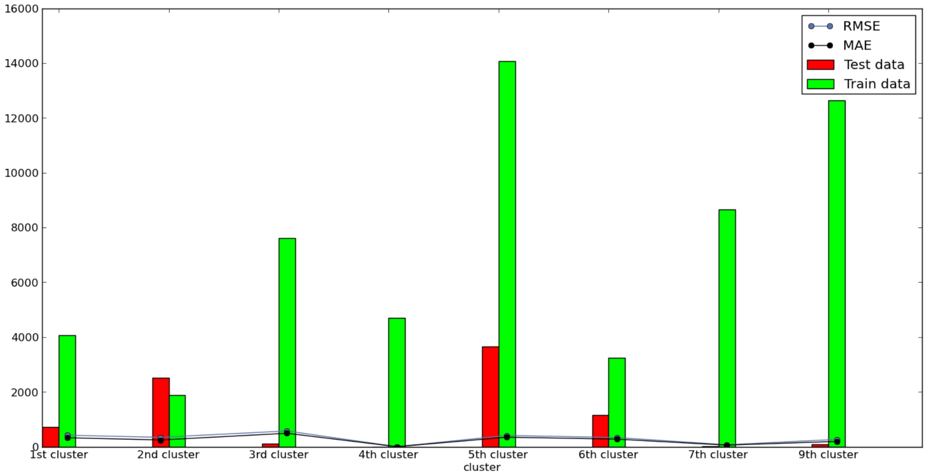

Figure 19: number of data

5.2.3 Modified Hierarchical clustering based models

As shown in last section, the number of data in a cluster in Africa can be large. For example, cluster 5 contains 3668 number of data and the number of similar data in other continents is relatively small. So if the number of training data is too small, it can also influence performance of models. Thus the main purpose of both two types of advanced hierarchical clustering based model is to improve performance of

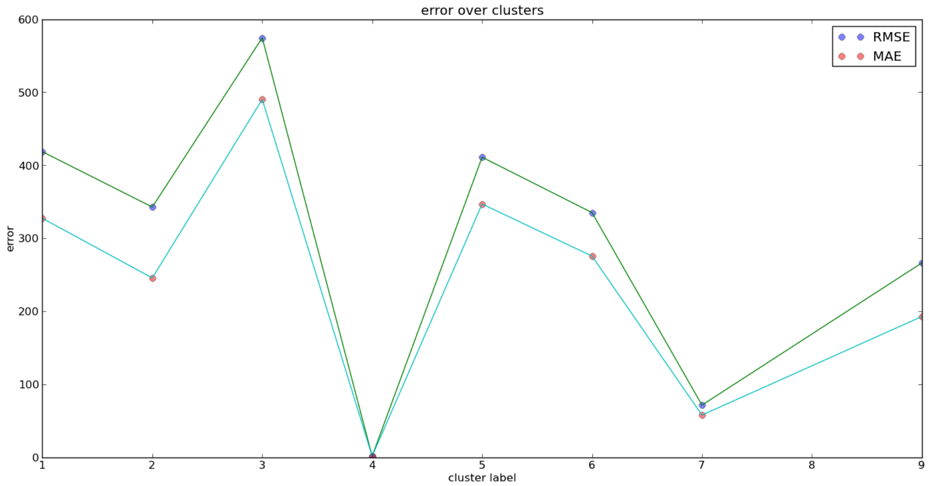

Figure 20: error over clusters

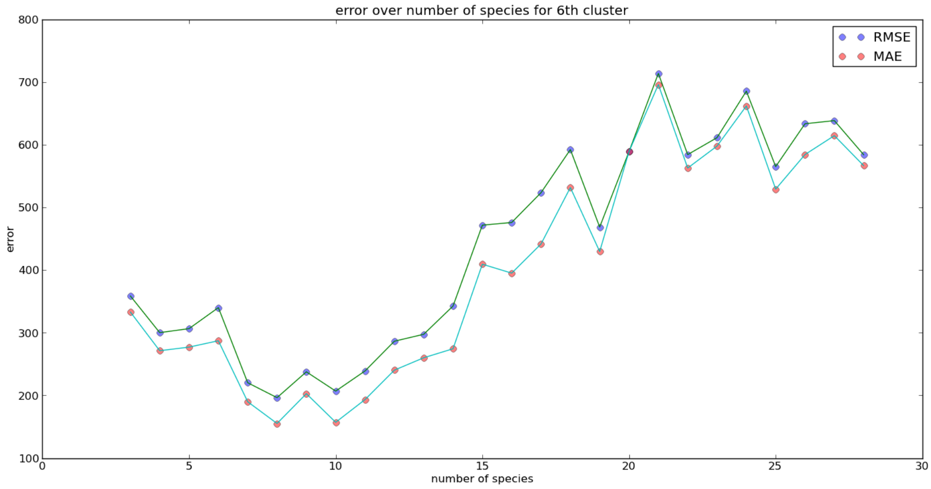

Figure 21: error over species for 6th cluster

models by partitioning a large group of test data into several small groups of data and testing them respectively. Thus different regressors can be selected for those small groups of data and this can improve performance of models. For example, if cluster 5 in Africa are partitioned into two groups of data, one of them has small error value when using OLS regressor but the other one has good prediction result with decision tree. Thus in these two types of advanced models, OLS regressor can be chosen for the former one and decision tree can be selected for the latter one.

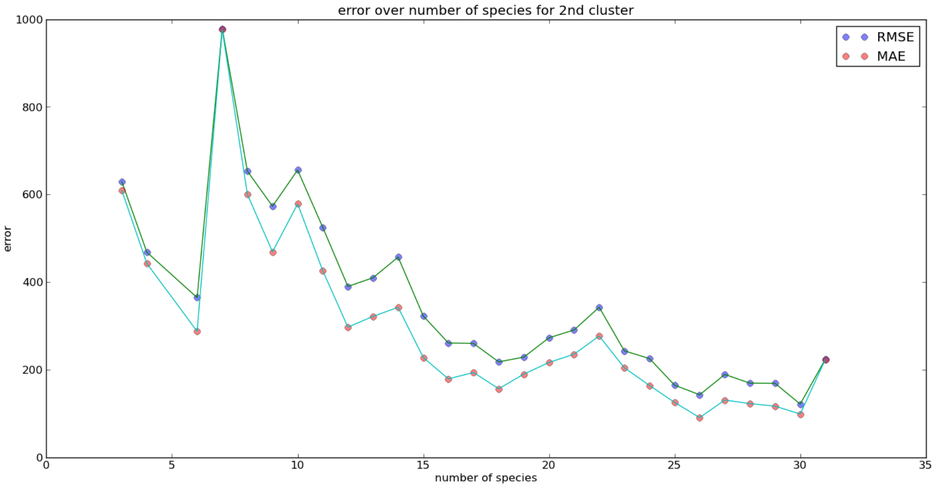

Figure 22: error over species for 2nd cluster

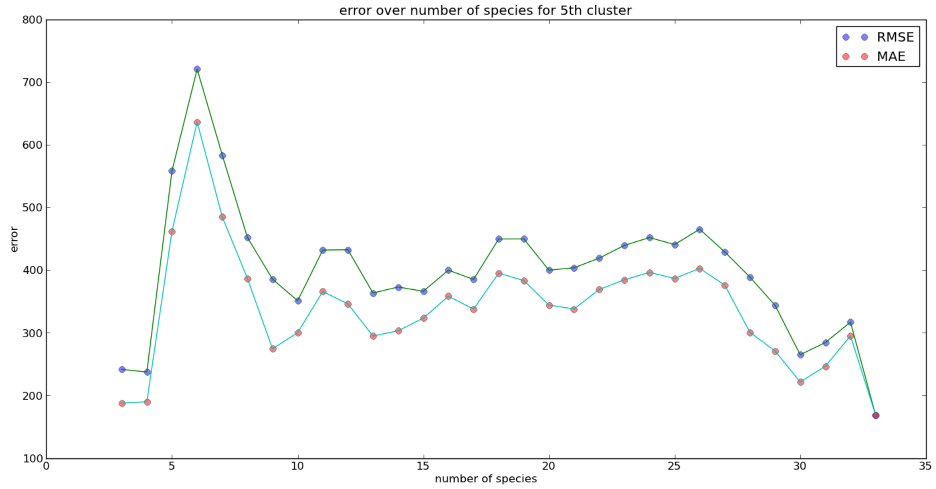

Figure 23: error over species for 5th cluster

However, if the cluster are not partitioned, only one regressor can be selected for all the data in the cluster. Therefore, cluster 1, 2, 5 and 8 are partitioned into different horizontal layers since they all consists of relatively large amount of data. In addition, the rest clusters in Africa are tested in the same way as in last section.

Figure 26 to Figure 29 reveals performance of models on different layers for those 4 clusters. Here a layer 1 means a top layer in the map for a cluster and a layer 2

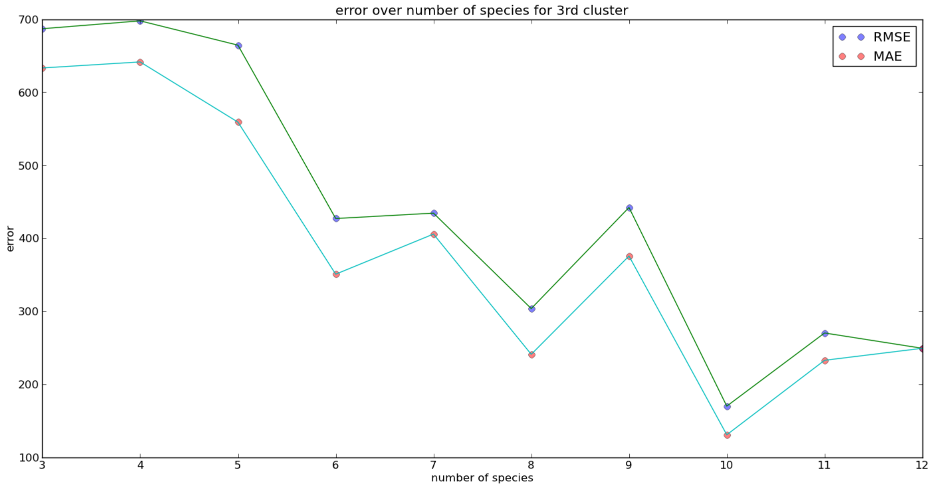

Figure 24: error over species for 3rd cluster

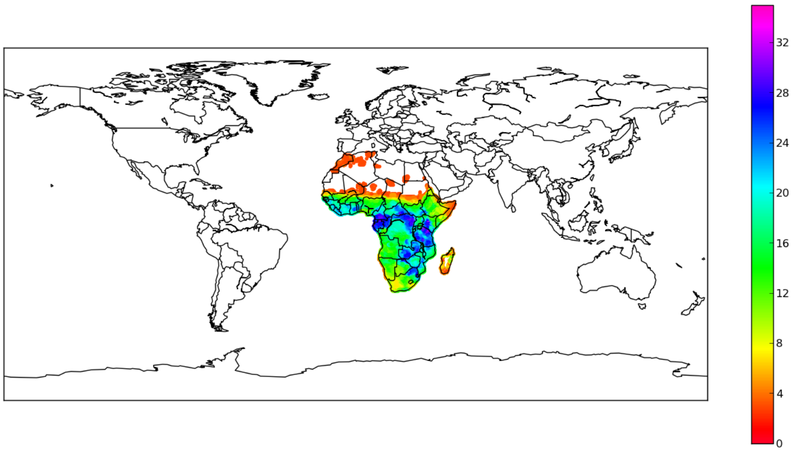

Figure 25: distribution of number of species in Africa

represents the layer that is the second top layer in the map for a cluster. Figure 30 reveals RMSE and MAE for different clusters. Compared error value with the result in section 3.4.2, performance of models for cluster 2, 5 and 6 are improved. However error value for the first cluster is higher than that in section 3.4.2. As shown in Figure 26, the error value of the first layer is near 700. Figure 32 illustrates relationship between error and number of species for this layer. Thus, in this layer, error value decreases while number of species increase. According to Figure 25 and Figure 5, the majority of data in the first layer of the cluster one have only 3 or 4 number of species and this can be one of the reason that influence performance of models.

Moreover, Figure 31 to Figure 33 present almost the same trend between error and number of species as in section 3.4.2.

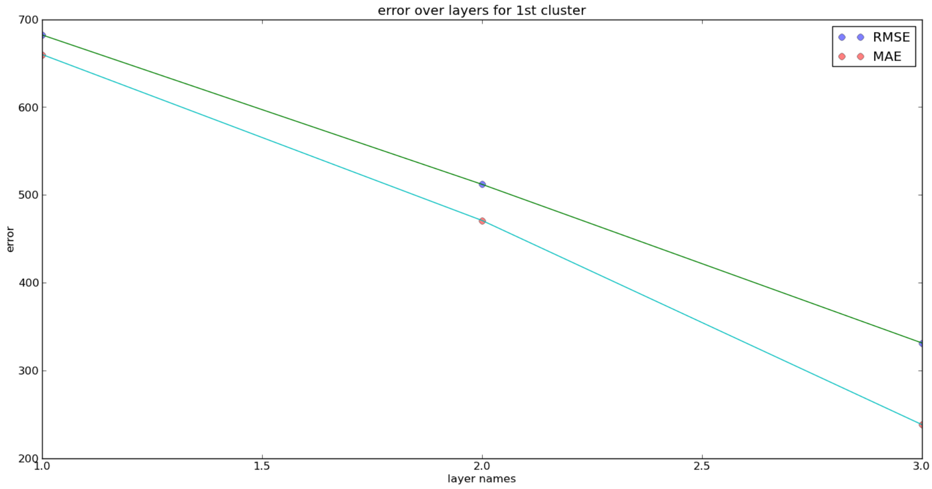

Figure 26: error over layers for 1st cluster

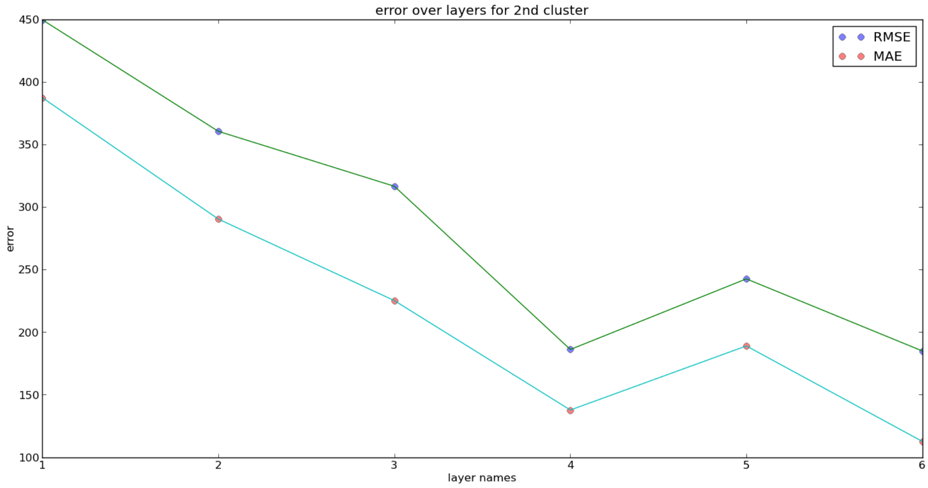

Figure 27: error over layers for 2nd cluster

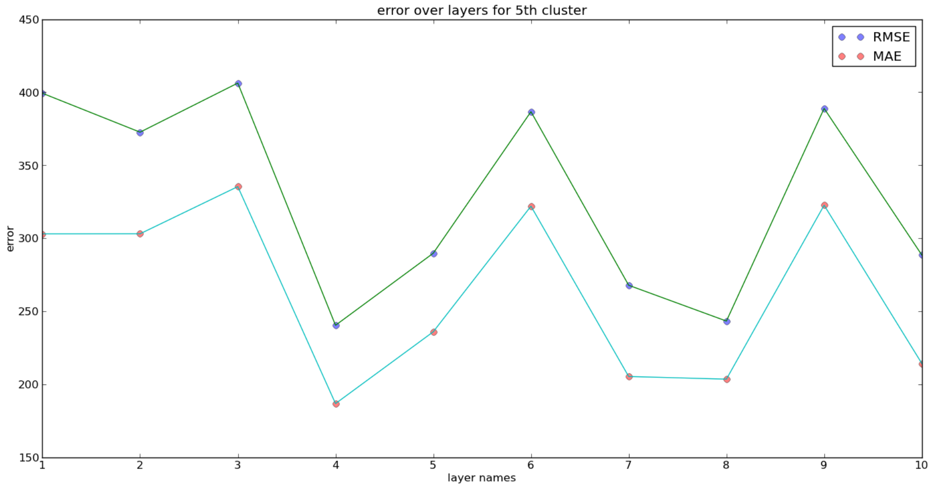

Figure 28: error over layers for 5th cluster

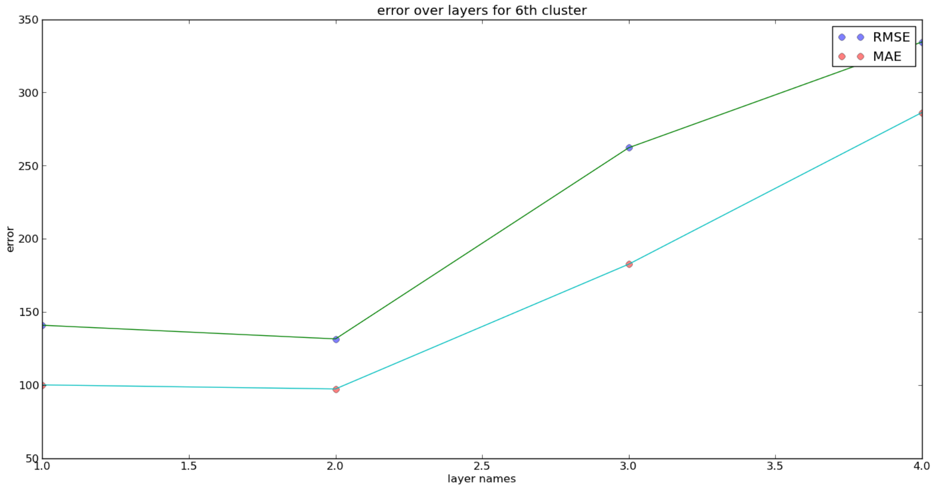

Figure 29: error over layers for 6th cluster

5.2.4 Advanced Hierarchical clustering based models

In last section, those four clusters with large number of data are partitioned manually into different small groups. In this section, those data are partitioned by hierarchical clustering. By using clustering, data in a new small group are similar since they are in the same cluster. Thus, in the next step, similar data obtained for training data are more accurate. In the process of clustering data that consists of a

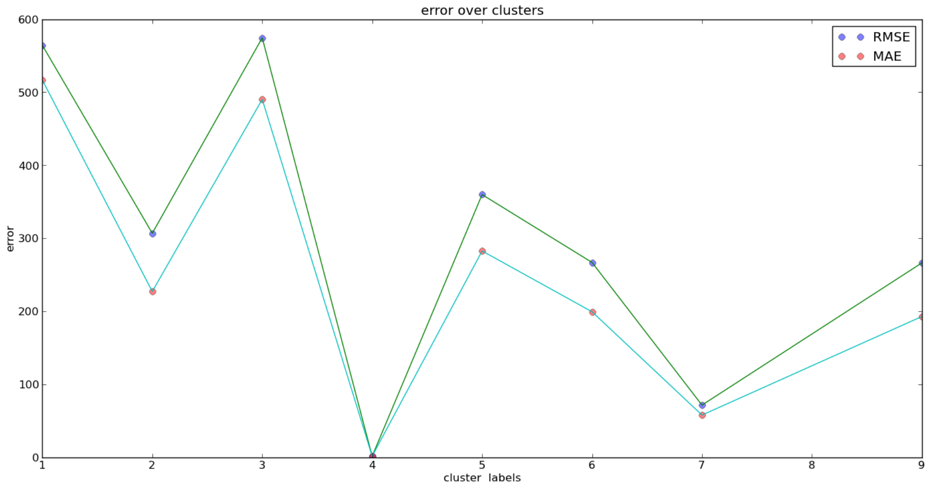

Figure 30: error over clusters

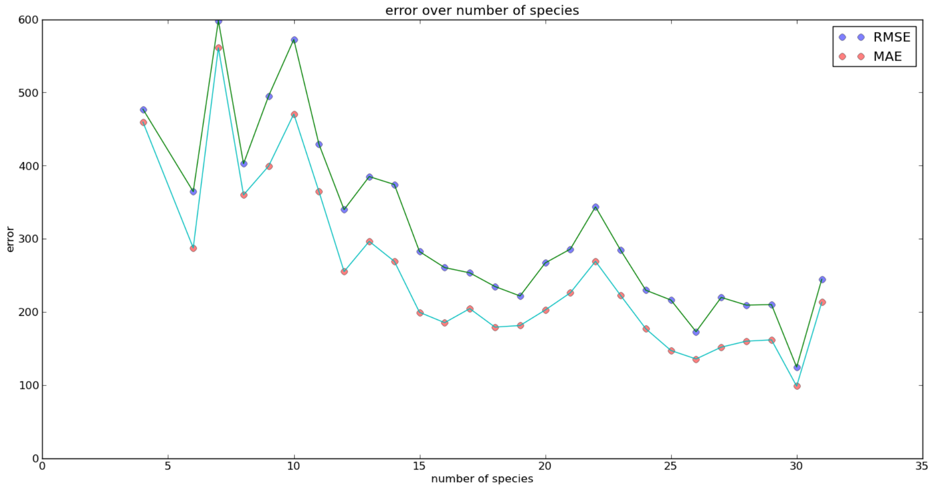

Figure 31: error over species for 2nd cluster

new small cluster and data without Africa data, more similar data can be obtained. Thus this can improve performance of models.

Figure 35 to Figure 37 illustrate the same trend for the change of error with number of species as in section 3.4.2 and 3.4.3. For savannas area, performance of models is worse when number of species is around 5 and larger than 3.

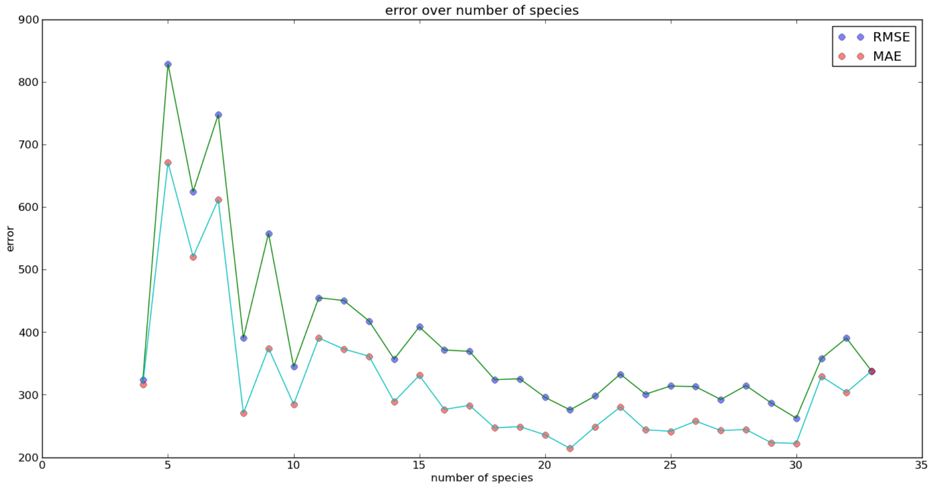

Figure 32: error over species for 5th cluster

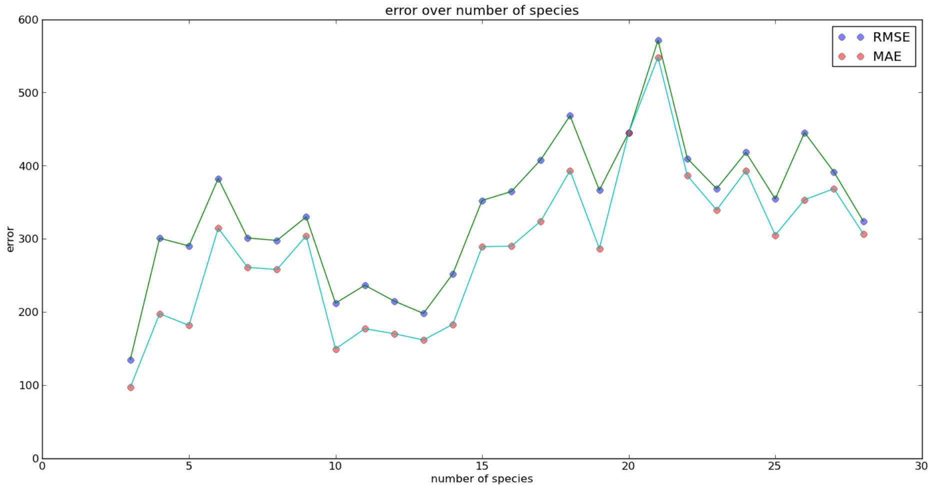

Figure 33: error over species for 6th cluster

Figure 38 illuminates error over original 8 clusters. RMSE and MAE are calculated over 8 clusters generated in 3.4.2. Performance of models improved for those cluster 1, 2, 5 and 6, compared with the result in 3.4.3. But RMSE value for cluster 1 is still larger than the result in 2.4.2, the MAE is improved. This reveals actually the prediction of model is near the true value but there are a few large prediction error

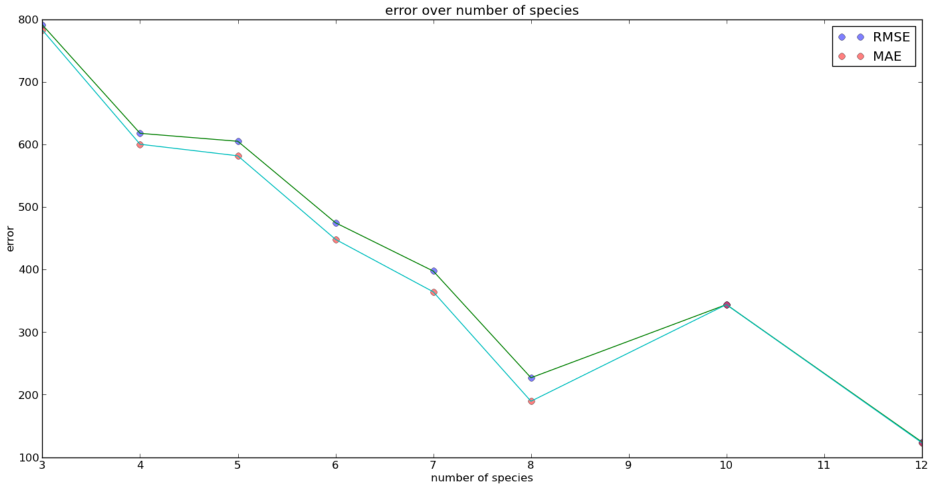

Figure 34: error over species for the first layer for the cluster 1

for some data. Figure 39 shows error over three small clusters and the cluster 1 in the original cluster 1 generate those large error. Figure 41 and Figure 42 shows locations of those data. In real situation shown in Figure 41, those blue data points reveal that those area is forest area but in prediction as shown in Figure 41, those data points are predicted like area with less plants. Thus, those data creates large error so that RMSE is not improved. Moreover, Figure 40 illustrates the change of error over number of species. This reveals that data with number of species that is larger than 5 have large error. As shown in Figure 23, those blue data points have number of species larger than 5. This can be also the reason that those data were not well predicted.

Finally, RMSE and MAE are calculated over all the data in Africa and Table 6 shows the result. Advanced hierarchical clustering based models(type 2) has the best performance over other two types of local models.

5.3 Evaluation procedures

Because of data is not i.i.d, r^2, RMSE and MAE are calculated over all data. Table 7 and Table 8 shows results of all models in four different situations. Random forest is shown to be the best model since both RMSE and MAE value are the smallest and

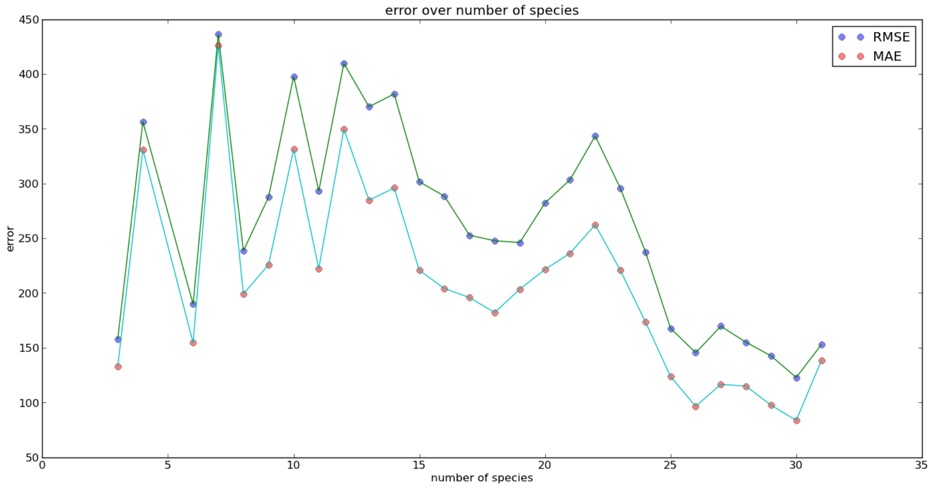

Figure 35: error over species for 2nd cluster

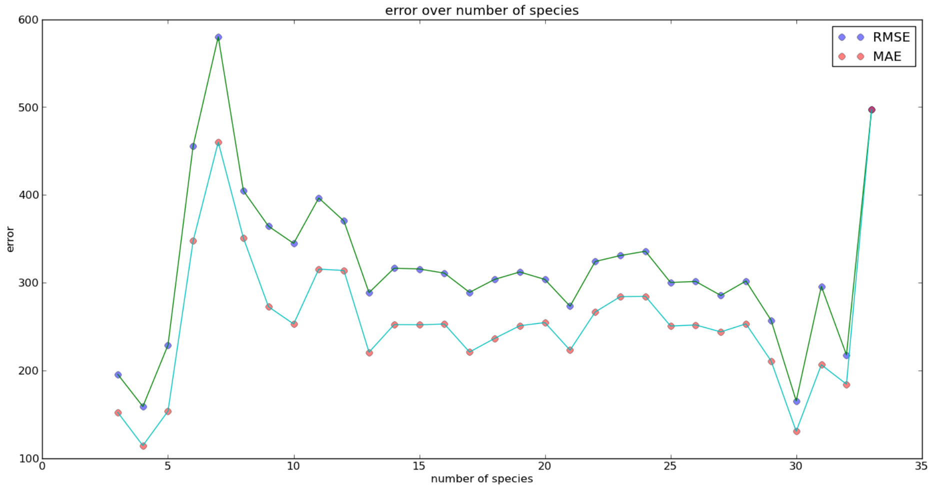

Figure 36: error over species for 5th cluster

r^2 is the largest. Compared OLS regression with tree based regression models, tree based regression models perform better since even the largest RMSE and MAE of tree based regression models is smaller than RMSE and MAE of OLS regression in each table. Moreover, compared the general performances of all models in normal 11 fold cross validation with performances of models in spatial 11 fold cross validation, models have more accurate prediction in normal 11 fold cross validation. This same trend also lie in comparison between normal leave one out cross validation and spa-

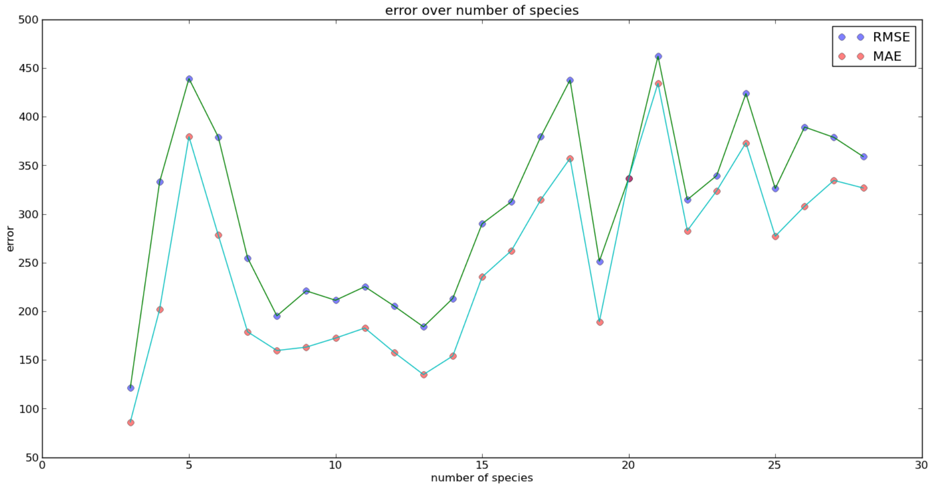

Figure 37: error over species for 6th cluster

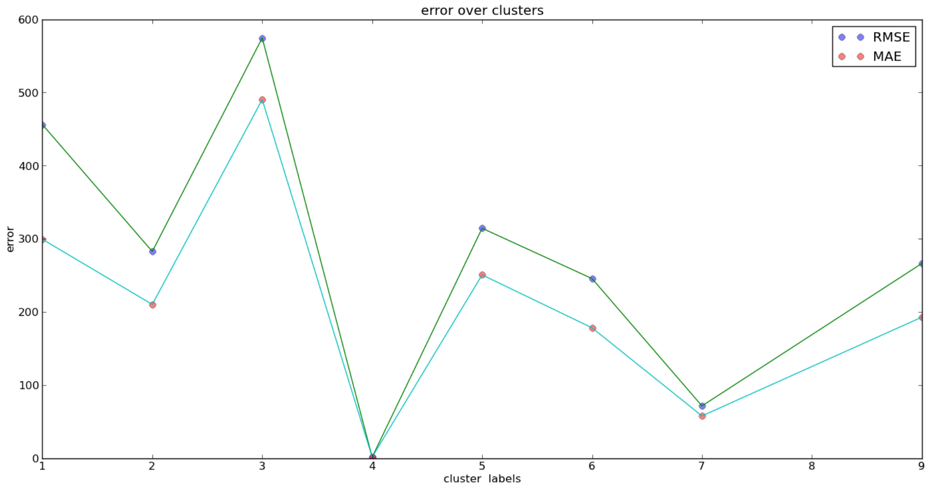

Figure 38: error over clusters

tial leave one out cross validation. This trend appears because autocorrelated data of each group of test data are pruned in spatial 11 fold cross validation and spatial leave one out cross validation. Furthermore, when compared a model in normal k fold cross validation with the same model in leave one out cross validation, the performances of that model is almost the same. However, when compared a model in spatial 11 fold cross validation with the same model in spatial leave one out cross validation, the performance of the model improves a lot in spatial leave one out cross

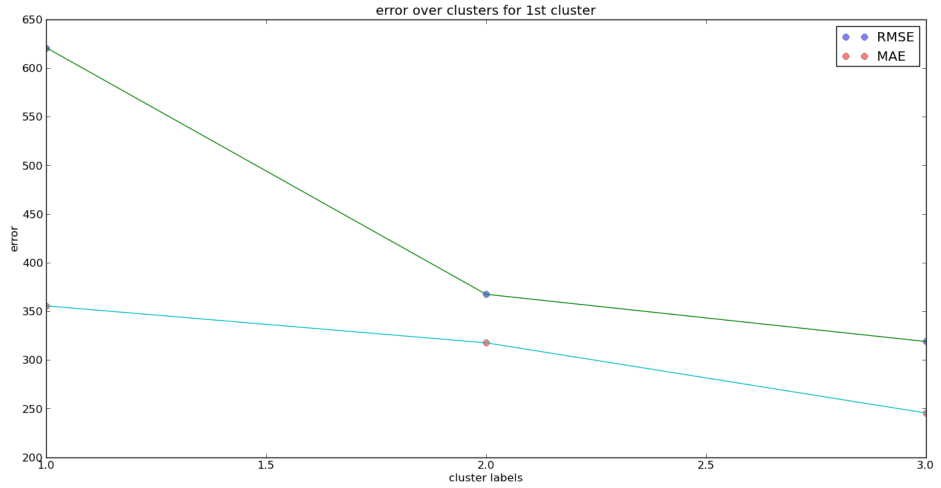

Figure 39: error for 1st cluster

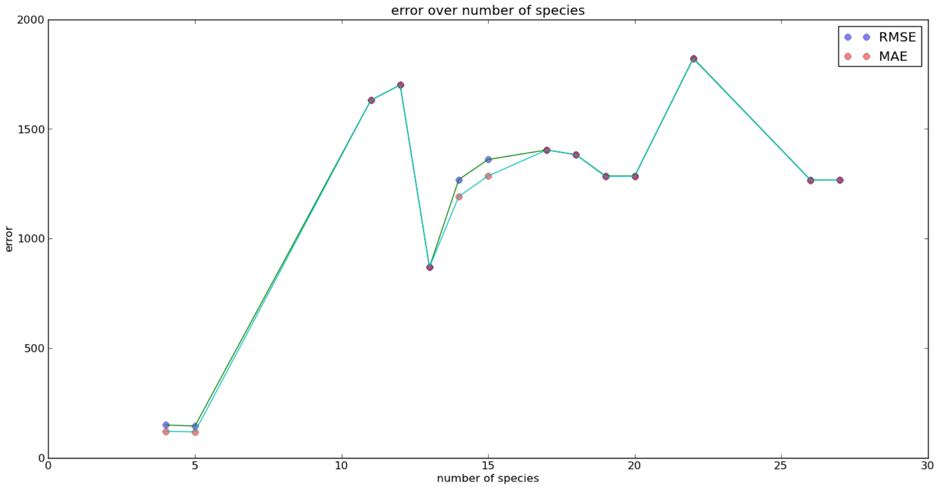

Figure 40: error over species

validation. One of the reason can be that less data were discarded in spatial leave one out cross validation compared to number of data discarded in spatial 11 fold cross validation. However the running time of spatial leave one out cross validation is much more larger since the same number of models as the number of data are built in spatial leave one out cross validation. Rotation forest is not tested in leave one out cross validation since it took a long time. Therefore, pruning autocorrelated data reduces performances of models indeed and if the number of data is super large,

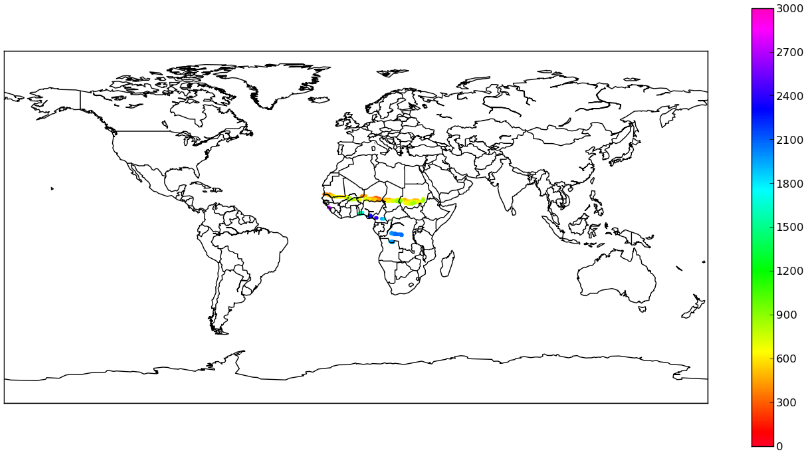

Figure 41: NPP

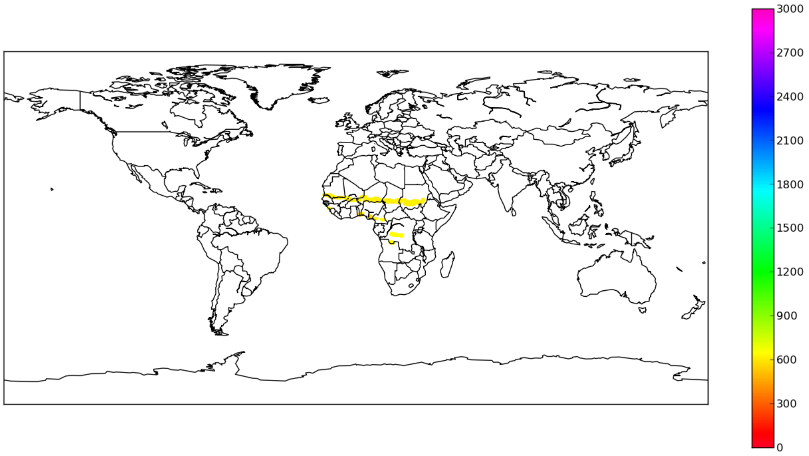

Figure 42: Predict

spatial k fold cross validation with random forest model can make good prediction in the setting of building global models. However, if the number of data is not that large, spatial leave one out cross validation with random forest model can be a good choice.

Furthermore, Figure 43 to Figure 46 shows performance of different models on four

model name	RMSE	MAE
Hierarchical clustering based model	383	304
Advanced hierarchical clustering based model(type 1)	356	273
Advanced hierarchical clustering based model(type 2)	316	235

Table 6: Error of models

normal 11-fold cross validation			
	r^2	RMSE	MAE
OLS regression	0.618	490	391
decision tree	0.863	294	198
rotation forest	0.740	405	326
gradient Boosting regressor	0.854	304	206
random forest	0.871	285	193

vertical spatial cross validation			
	r^2	RMSE	MAE
OLS regression	0.547	535	428
decision tree	0.639	477	352
rotation forest	0.649	471	387
gradient Boosting regressor	0.574	518	372
random forest	0.702	434	321

Table 7: result of 11 fold cross validation

continents in four types of cross validation. In those figures, r^2 is not included since variances of data in different continents are different and it can influence the value of r^2. Moreover, performance of OLS regression models on different continents are almost the same. In addition, tree based models excluded rotation forest have good performance in Africa and South America and the best model: random forest has the best performance in Africa.

normal leave one out cross validation			
	r^2	RMSE	MAE
OLS regression	0.618	490	391
decision tree	0.863	294	197
gradient Boosting regressor	0.857	301	205
random forest	0.872	284	192

spatial leave one out cross validation			
	r^2	RMSE	MAE
OLS regression	0.587	511	408
decision tree	0.723	418	307
gradient Boosting regressor	0.701	434	315
random forest	0.764	385	286

Table 8: result of leave one out cross validation

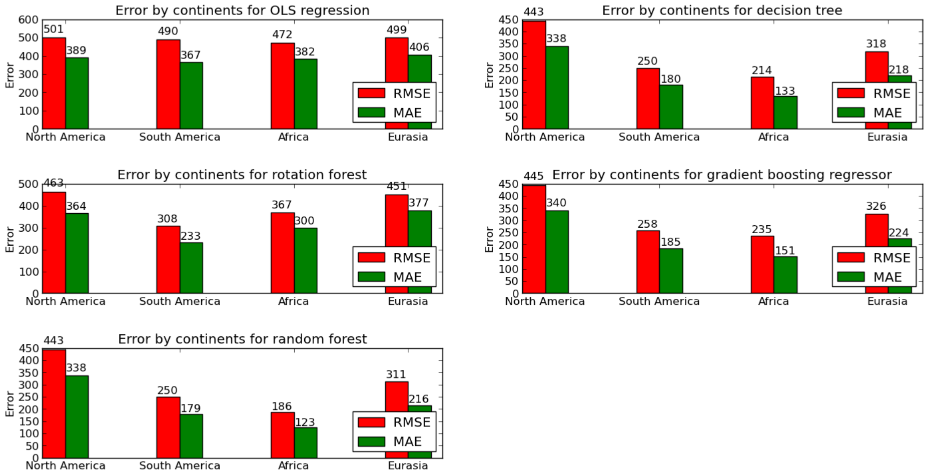

Figure 43: 11 fold cross validation

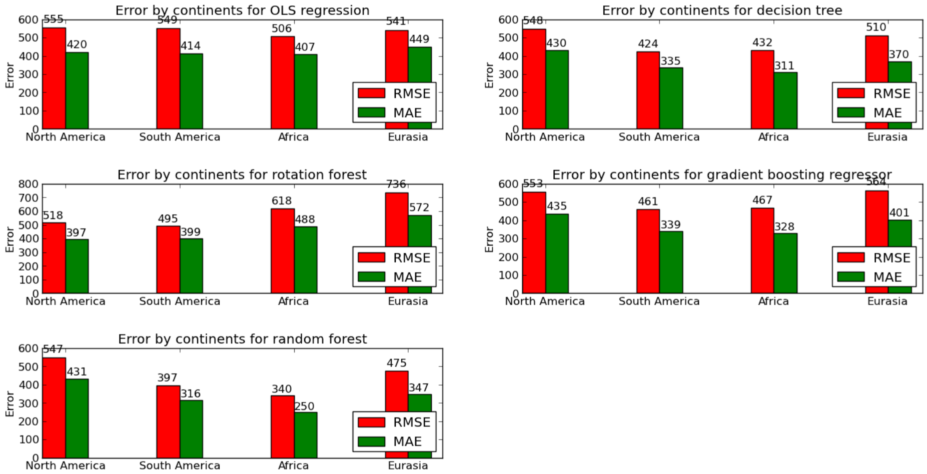

Figure 44: vertical spatial cross validation

5.4 discussion

Table 9 and Table 10 shows error for all models. However, Table 9 is like training error since we tune parameter which is number of clusters involved in training data by minimising RMSE of Africa data. Table 10 shows the real test error since the parameter namely the number of clusters in training data are tuned on validation data. After obtaining the best parameter on validation data, we apply the model

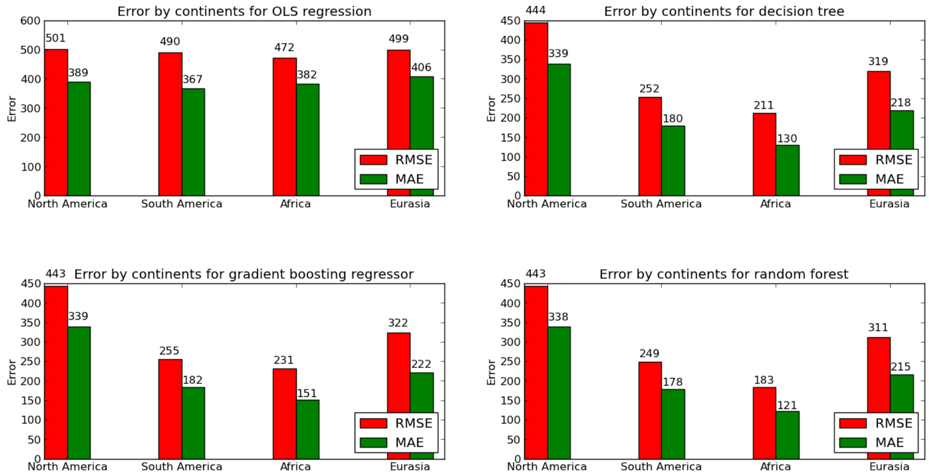

Figure 45: leave one out cross validation

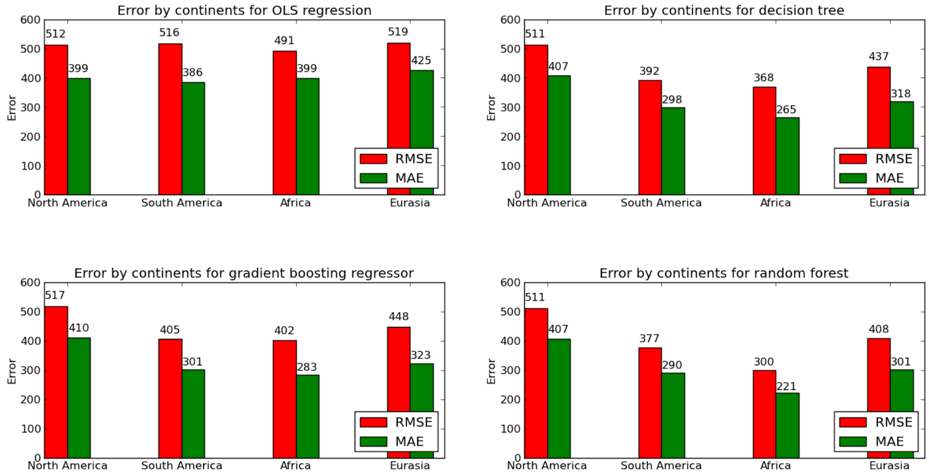

Figure 46: spatial leave one out cross validation

with fine tuned parameters on testing data. By comparing two tables, error of three clustering based models in table 10 are larger than them in table 9. Thus, tuning parameters on the whole Africa data can cause overfit indeed. However, table 10 still illustrate that AHCM is still the best model since RMSE is 398. Its RMSE decreases 22% compared with global models. Thus we can conclude that AHCM can build good predictive models.

model name	RMSE	MAE
Global models	512	416
Baseline models: Type 1	723	567
Baseline models: Type 2	541	422
Hierarchical clustering based model	383	304
Modified advanced hierarchical clustering based model	356	273
Advanced hierarchical clustering based model	316	235

Table 9: Error of all models

model name	RMSE	MAE
Global models	512	416
Baseline models: Type 1	723	567
Baseline models: Type 2	541	422
Hierarchical clustering based model	483	391
Modified advanced hierarchical clustering based model	435	325
Advanced hierarchical clustering based model	398	279

Table 10: Error of all models

6 Case study

According to experiment results described in previous sections, advanced hierarchical clustering based model has the best performance. Thus, we use the same parameter settings as AHCM for fossil data. Thus, in the first step, we cluster the present day data and fossil data to be 10 clusters. For fossil data, there are 5 cluster labels. Cluster 2, 4 and 5 are combined with the present day data and they are clustered again separately. Then we can clearly discover which cluster in fossil data are merged with which cluster in Africa data. Thus we can use AHCM for that cluster of Africa data for making prediction on its corresponding cluster in fossil data. For example, a sub cluster of cluster 2 of fossil data can be merged with sub cluster 1 of cluster 5 in Africa data so the best AHCM built in the experiment for that cluster of Africa data is applied to the sub cluster of cluster 2 of fossil data.

This step is repeated until NPP of all fossil data are predicted.

Figure 48 to Figure 51 show prediction of NPP over time period. Figure 47 shows NPP in present data. In present day, the average NPP in Turkana Basin area is around 600 to 800. When time period starts from 0.01 to 2 Ma, the mean NPP is 1021. When time is from 2 to 3 Ma, the average NPP is 981. Thus the trend is when time changes from 0.01 to 3 Ma, the environment in Turkana Basin area becomes dry. However, when time is from 3 to 4 Ma, the mean NPP is 1123 and when time is from 4 to 7.8 Ma, the mean NPP is 1104. So from 3 Ma, the environment in Turkana Basin area starts become a little bit humid. Then from 4 to 7.8 Ma, the environment remains almost the same humidity. In addition, environment in time period of 3 Ma to 7.8 Ma is more humid than the environment in present day.

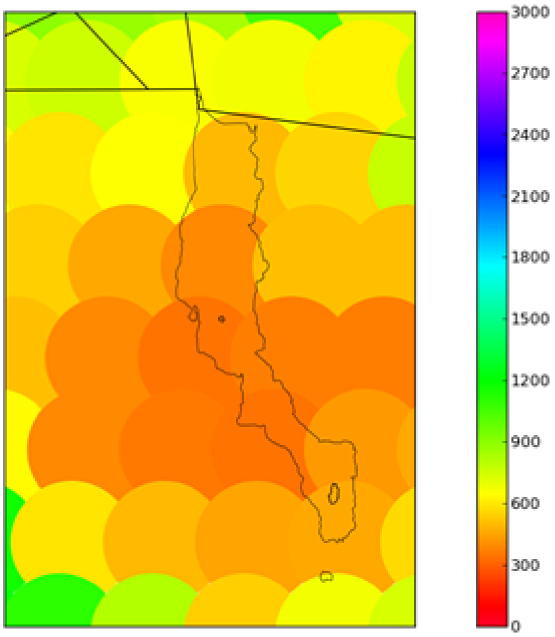

Figure 47: NPP at present day

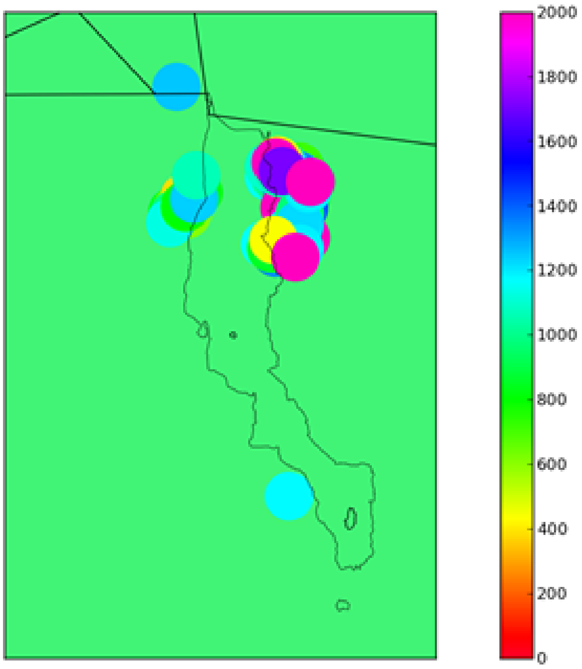

Figure 48: NPP from 0.01 to 2 Ma

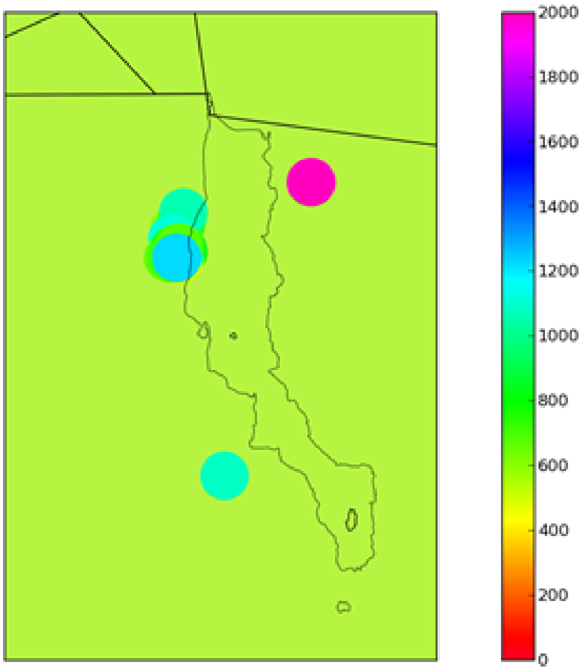

Figure 49: NPP from 2 to 3 Ma

e

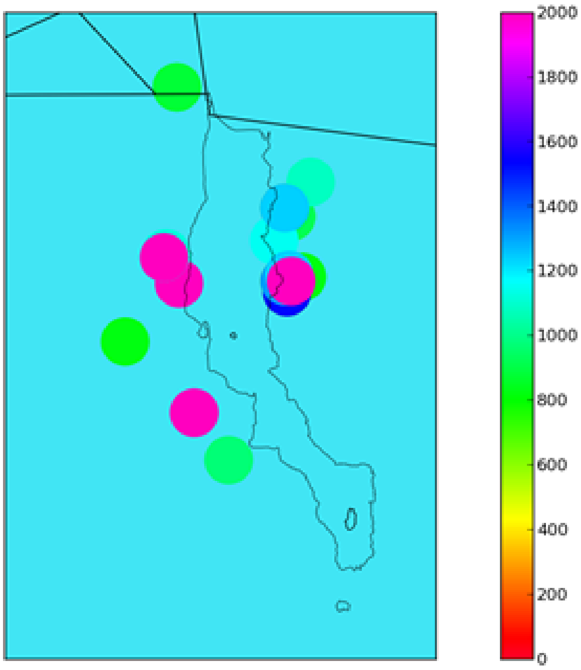

Figure 50: NPP from 3 to 4 Ma

7 Discussion and conclusions

In this paper, we propose local models and test global models. We also propose a spatial cross validation scheme that is vertical spatial cross validation and it is for evaluating performance of models. For VSCV, some data points that are near testing data are discarded for reducing influence of spatial autocorrelation for obtaining true accuracy of predictive models. For local models, less training data that matches testing data closely are also discovered. In addition, number of species for a data point can influence performance of local models. For savannas area in Africa, local models that are built on data points with number of species smaller than 4 or number of species larger than 32 can have higher accuracy. For forest area in Africa, the higher the number of species of datapoints is, the better the performance of models built on them. However, the relationship between number of species of datapoints and performance of local models are opposite for desert area in Africa. Namely, the smaller the number of species of datapoints is, the better the performance of models built on them.

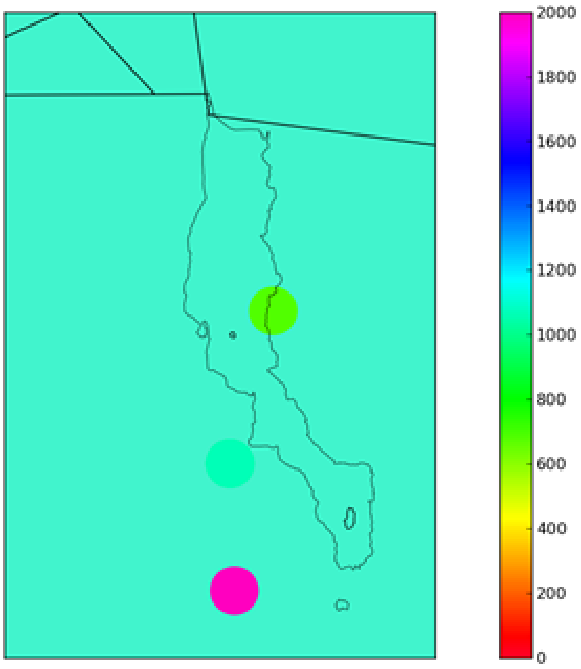

Figure 51: NPP from 4 to 7.8 Ma

To analyze performance of the best global models and the best local models for each clusters precisely, RMSE of both of the best global models and the best local models for each cluster in Africa are calculated. Figure 52 illustrates the result. In the figure, the blue line with black dots represent the RMSE for clusters for the best global models and the green line with blue dots shows the RMSE for clusters for the best local models. For cluster 2, 5 and 6, the difference of accuracy of global models and local models is small. Thus, we can conclude that if testing data are merged in the same cluster as 2, 5 and 6 in Africa, both global models and local models works well. But if the number of testing data is large, we still recommend local models. In addition, if testing data are merged in cluster 3, 4, 7 or 9, we recommend the best local models since performance of the best performance is much better than the best global models. Finally, it testing data are merged in the cluster 1, the best global models are recommended since RMSE of the best global models are much smaller than that of the best local models.

As considered in paper [H$^+$06], in our study, each prediction have the same cost. In order to avoid predictions that has very large error compared to real value, different costs should be added in the process of making predictions. For example, if we define desert as NPP is around 400, prediction of NPP of a data points in desert to be 100 or 500 do not change the fact that the environment of the data points is desert. But if the prediction of NPP of those data points is 1000, this prediction is totally wrong since the prediction show that this data point is in environment like savannas or forest. Thus those types of prediction need to be avoided. Therefore, this can be a future research of our study.

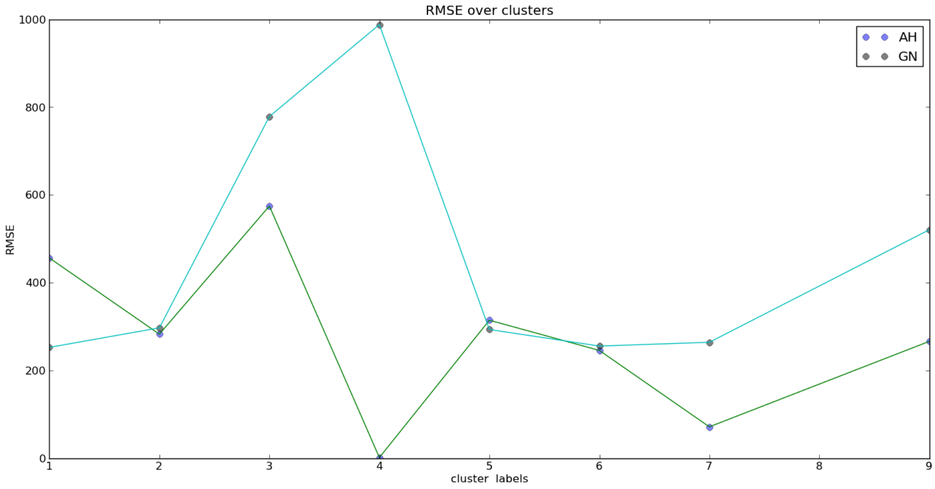

Figure 52: error over clusters

References

BHG+17 Barnosky, A. D., Hadly, E. A., Gonzalez, P., Head, J., Polly, P. D., Lawing, A. M., Eronen, J. T., Ackerly, D. D., Alex, K., Biber, E. et al., Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. *Science*, 355,6325(2017), page eaah4787.

Dar09 Darwin, C., *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life.* John Murray, 2009.

EPL+10a Eronen, J., Puolamäki, K., Liu, L., Lintulaakso, K., Damuth, J., Janis, C. and Fortelius, M., Precipitation and large herbivorous mammals i: estimates from present-day communities. *Evolutionary Ecology Research*, 12,2(2010), pages 217–233.

EPL$^+$10b Eronen, J., Puolamäki, K., Liu, L., Lintulaakso, K., Damuth, J., Janis, C. and Fortelius, M., Precipitation and large herbivorous mammals ii: application to fossil data. *Evolutionary Ecology Research*, 12,2(2010), pages 235–248.

FŽK$^+$16 Fortelius, M., Žliobaitė, I., Kaya, F., Bibi, F., Bobe, R., Leakey, L., Leakey, M., Patterson, D., Rannikko, J. and Werdelin, L., An ecometric analysis of the fossil mammal record of the turkana basin. *Phil. Trans. R. Soc. B*, 371,1698(2016), page 20150232.

GTFŽ17 Galbrun, E., Tang, H., Fortelius, M. and Žliobaitė, I., Computational biomes: the ecometrics of large mammal teeth.

H$^+$06 Hand, D. J. et al., Classifier technology and the illusion of progress. *Statistical science*, 21,1(2006), pages 1–14.

JWHT14 James, G., Witten, D., Hastie, T. and Tibshirani, R., *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014.

LPE$^+$12 Liu, L., Puolamäki, K., Eronen, J. T., Ataabadi, M. M., Hernesniemi, E. and Fortelius, M., Dental functional traits of mammals resolve productivity in terrestrial ecosystems past and present. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20120211.

LRPB13 Le Rest, K., Pinaud, D. and Bretagnolle, V., Accounting for spatial autocorrelation from model selection to statistical inference: Application to a national survey of a diurnal raptor. *Ecological Informatics*, 14, pages 17–24.

LRPM$^+$14 Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J. and Bretagnolle, V., Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23,7(2014), pages 811–820.

Mec17 Mechenich, M., Best practices for ecometric analysis: a case study correlating climate conditions and herbivore teeth in africa.

PPNH17 Pohjankukka, J., Pahikkala, T., Nevalainen, P. and Heikkonen, J., Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31,10(2017), pages 2001–2019.

PY10 Pan, S. J. and Yang, Q., A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22,10(2010), pages 1345–1359.

Val84 Valiant, L. G., A theory of the learnable. *Communications of the ACM*, 27,11(1984), pages 1134–1142.

Ž16 Žliobaitė, Indrė and Rinne, Janne and Tóth, Anikó B and Mechenich, Michael and Liu, Liping and Behrensmeyer, Anna K and Fortelius, Mikael, Herbivore teeth predict climatic limits in kenyan ecosystems. *Proceedings of the National Academy of Sciences*, page 201609409.

Zli16 Zliobaite, Indre and Tatti, Nikolaj, A note on adjusting R^2 for using with cross-validation. *arXiv preprint arXiv:1605.01703.*

ŽPEF17 Žliobaitė, I., Puolamäki, K., Eronen, J. T. and Fortelius, M., A survey of computational methods for fossil data analysis. *Evolutionary Ecology Research*, 18,5(2017), pages 477–502.