Climate Classification: A data-driven approach

Ji Luo[*] **Bilal Hussain**[†] **Sakina Patanwala**[‡]

Abstract

This research attempts to make a global climate classification via unsupervised clustering methods. Based on the idea that climate cannot fully explain the vegetation, this paper proposes an evaluation metric, harmonized homogeneity, for this type of clustering problem in which ground-truth labels can be hard to define but there are related label evidence. To improve the clustering results in terms of fitness with real-world vegetation, a merging algorithm is designed. The optimal result is interpreted and assigned climate type names. Furthermore, with the idea of supervised learning, we explore if there is a mapping between climate and vegetation, which is an important goal of climate classification.

1 Introduction

Climate and vegetation, directly and indirectly, influence each other, where the question arises if there is a mapping between climate and vegetation[1]. To discover the humidity-heat conditions for different vegetation types, which is a vital goal of climate classification, we are pursuing a purely data-controlled approach for climate classification. The methodology of our research is shown in Figure 1.

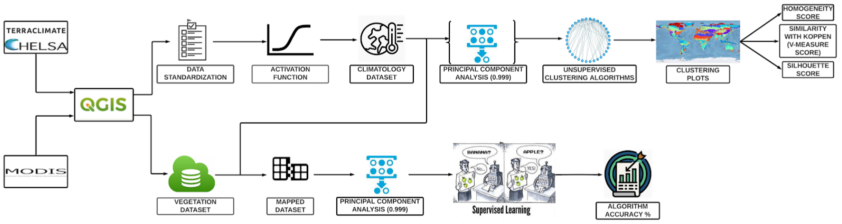

Figure 1: Methodology overview

The Köppen-Geiger system classifies the world into five climate zones based on temperature and precipitation, which are further divided into 30 sub-climate zones, which allow for different vegetation growth. This system categorizes climate zones throughout the world in an attempt to recognize local vegetation[4]. The Köppen-Geiger classic climate classification system is based on heuristic and empirical rules, which may appear arbitrary and subjective[2]. While these dating from the 1900s have been used, they have heuristic deficiencies, causing numerous alteration variants.[5]

[*] Department of Computing Science, Simon Fraser University, luojil@sfu.ca

[†] Department of Computing Science, Simon Fraser University, bha59@sfu.ca

[‡] Department of Computing Science, Simon Fraser University, spa170@sfu.ca

[4] https://www.nationalgeographic.org/encyclopedia/koppen-climate-classification-system/

[5] https://www.researchgate.net/publication/26640584_Updated_World_Map_of_the_Koppen-Geiger_Climate_Classification

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2 Dataset

We have researched two datasets for this project: climatology and vegetation. The former is continuous data while the latter is categorical. Climatologies are used for unsupervised clustering and as input features for supervised learning, while vegetation serves as a reference for clustering evaluation and target labels for supervised learning.

After downloading the raw dataset, we aligned them to be in the same coordinate referencing system, WGS 84, with the help of QGIS[6].

2.1 Climate dataset

The validating time period for the climate data set is 1981 to 2010 and is derived by incorporating potential evapotranspiration(PET)[7] from TerraClimate[8][4] with other key variables picked from CHELSA[9][5]. The dataset is downsampled to 0.5°C × 0.5°C resolution using Cubic method. The bounding box for the dataset is [55°S, 80°N], [180°W, 179.5°N], Antarctica excluded. This results in 270 × 719 resolution.

From the point of signal processing, precipitation (denoted as $\mathbf{p_1}$) and PET (denoted as $\mathbf{p_2}$) are each regarded as a discrete signal with a period of 12. This study introduced three novel features derived from the precipitation-PET cross-correlation to guide the algorithm to better distinguish between monsoonal (dry winter), perhumid and Mediterranean (dry summer) climates.

Appendix A[6] lists the climatic features, with their processing methodology and descriptions.

According to the "Standardization method" column in Appendix A, Z-score standardization is used to transform each feature to a mean value of 0 and a standard deviation of 1. After standardization, a hyperbolic tangent function is applied [7, 8]:

$$y(x) = \tanh\left(\frac{x-\mu}{\sigma}\right)$$

In the formula, μ is called the center and σ is the deviation. The choice of center and deviation is shown in the column of "Activation method" in Appendix A. Since the data has been standardized by Z-score, under "normal" cases, centers are set as 0 and deviations are set as 1, but for some features with particular meanings, Z-score standardization is skipped. These features are specially processed during activation.

This study introduces hyperbolic tangent function activation to simulate the simple judgment (or binning, both are hard thresholding) used in traditional climate classification. For example, the Köppen classification defines that areas with all monthly average temperatures higher than 18°C have a tropical climate, but such a decision boundary appears arbitrary and rigid from the perspective of data science. Logistic functions are used in data science to simulate this judgment, but it is more natural and smooth. Commonly used logistic functions include sigmoid, hyperbolic tangent, etc. Here, the hyperbolic tangent is selected to map all the features to the $[-1, 1]$ interval.

The finalized climate dataset contains 72 features and 61784 records (data for ocean areas is not used). About half of the features are dominated by temperature, and the other half by precipitation. Further, principle component analysis(PCA) is used to perform dimensionality reduction. We retain 99.9% of information and obtain a data set containing 43 features.

2.2 Vegetation dataset

The vegetation dataset, from MODIS(MCD12C1[10][9]), has its categories based on the International Geosphere-Biosphere Programme(IGBP). We acquired the data for 2019. The layer is downsam-

[6] https://qgis.org/en/site/

[7] Defined as the amount of evaporation that would occur if a sufficient water source were available. If the actual evapotranspiration is considered the net result of atmospheric demand for moisture from a surface and the ability of the surface to supply moisture, then PET is a measure of the demand side. Surface and air temperatures, insolation, and wind all affect this. A dry-land is a place where annual potential evaporation exceeds annual precipitation.[3]

[8] http://www.climatologylab.org/terraclimate.html

[9] https://chelsa-climate.org/

[10] https://lpdaac.usgs.gov/products/mcd12c1v006/

pled and clipped with QGIS using Mode method and has 14 vegetation types after preprocessing. Urban areas, unrelated to climate, are ignored by assigning them as water bodies. Croplands and Cropland/Natural Vegetation Mosaics are merged into one type called Cropland Mosaics.

3 Metrics for clustering

3.1 Problem setting

For unsupervised learning, the key issue is that, there are actually no ground-truth labels for climate classification. Although climate classification somewhat aims to fit vegetation types, the vegetation cannot be regarded as ground-truth labels. On the one hand, it is generally believed that the vegetation type is not only a function of climate[2], but is also influenced by factors like human activities and topography; on the other hand, the vegetation dataset used in this study contains only 14 vegetation types, while 14 climate types are generally considered insufficient to describe the variety of global climates (for example, the commonly used Köppen classification[10] contains 30 types). In other words, a vegetation type often contains multiple climate types, while within a climate type, there should be as few vegetation types as possible. But vegetation does provide some insight to evaluate the applicability of a climate classification result.

3.2 V-measure

Use V(vegetation) to denote the vegetation classification result for reference, and C(climate) to denote the climate classification result. Taking advantage of the concepts of homogeneity, completeness in V-measure[11], the mathematical expression of this setting is that, homogeneity of vegetation given climates, $h(V, C)$, is better to be high, but since there are fewer categories in V than in C, completeness of vegetation given climates, $c(V, C)$, should not be included in the evaluation.

The mathematical expressions of homogeneity, completeness, and V-measure are as follows:

$$h(V, U) = 1 - \frac{H(V|U)}{H(V)}$$

$$c(V, U) = h(U, V) = 1 - \frac{H(U|V)}{H(U)}$$

$$v(V, U) = \frac{2h(V, U)c(V, U)}{h(V, U) + c(V, U)}$$

In the formulas, $H(V|U)$ is the conditional entropy of V when U is given, which measures the homogeneity of V within each cluster given by the clustering algorithm:

$$H(V|U) = -\sum_{i=1}^{n_U} \sum_{j=1}^{n_V} \frac{N_{i,j}}{N} \log \frac{N_{i,j}}{N_i}$$

N is the total number of data points, N_i refers to the number of data points contained in the i-th cluster in U, and $N_{i,j}$ refers to the number of data points in the entire dataset belonging to both class j in V and cluster i in U.

Homogeneity, completeness and V-measure are all in the interval of $[0, 1]$. According to this indicator, the closer the three values are to 1, the closer U is to V, the better the clustering result will be. [11]

3.3 Harmonic homogeneity

However, it is not advisable to directly use the homogeneity in the V-measure. The reason is that the homogeneity index alone will favor a high number of clusters. The finer the classification, the fewer vegetation types will be included in each climate type on average, and thus, a higher homogeneity score will be obtained; in the extreme case, each data point is classified as one separate climate type, and thus within each climate type, there must be only one kind of vegetation. The homogeneity achieves full mark, but such a climate classification does not make sense.

Based on the homogeneity in the V-measure, this study introduces $H(C)$ in its numerator and denominator to antagonize the preference of homogeneity for more clusters. In this way, the harmonized homogeneity is proposed:

$$hh(V, C) = (1 + \alpha)(1 - \frac{H(V|C) + \alpha H(C)}{H(V) + \alpha H(C)})$$

$H(C)$ is the information entropy of the climate classification results. The finer the classification(with more clusters and more details), the higher the value. When α and $H(V|C)$ are constant, the value of harmonized homogeneity decreases with the increase of $H(C)$ to achieve the desired effect of suppressing the indicator's preference for more climate types.

α is the non-negative harmonization intensity coefficient. Larger α introduces stronger suppression effect. An α that is too large will produce a reversed inclination on the metric that prefers fewer climate types. For a specific problem, the optimal α should make the harmonized homogeneity value does not significantly change with the number of categories, so as to achieve the desired suppression effect without overcorrecting.

The coefficient $(1 + \alpha)$ is used to scale the value to $[0, 1]$, to be consistent with V-measure. When α is 0, the harmonized homogeneity goes back to the homogeneity as in the original V-measure.

This proposed metric is applicable to a series of clustering problems with class labels that are not ground-truth but provide some insight and can be a reference. In these problems, the number of referencing classes and derived clusters may differ.

Three metrics are employed in this study: harmonized homogeneity with vegetation(external indicator), silhouette coefficient(internal indicator), and V-measure with Köppen classification. Among them, the harmonized homogeneity is the most important. We believe that an applicable climate classification should be close to vegetation. The silhouette coefficient is used to evaluate the effectiveness of the clustering model on a certain dataset. A model achieving high silhouette value fits better with the input dataset. V-measure with Köppen classification is for reference only. It does not indicate the quality of a clustering: a good clustering is not necessarily consistent with Köppen.

Euclidean distance is used in the clustering algorithms to measure the similarity between observations.

3.4 Determination of the harmonization intensity coefficient α

Choose a series of cluster numbers K, a series of K-medoids models are trained on the climate dataset without PCA, and use a series of α to calculate harmonic homogeneity values. Take the number of categories K as the X axis, α as the Y axis, and the harmonized homogeneity as the Z axis, a mesh plot is made(Figure 2).

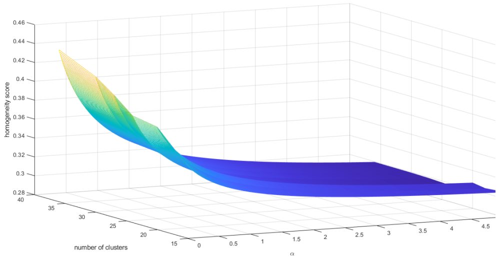

Figure 2: Cluster number-α-harmonized homogeneity figure

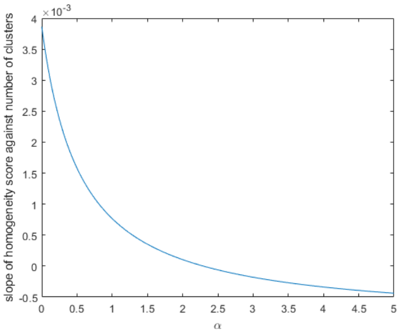

Figure 3: α-"slope" figure

It is easy to know from Figure 2 that when α is relatively small, the harmonized homogeneity increases significantly with the number of clusters, which is in line with the previous discussion that the homogeneity in V-measure has a preference for more categories. With the increase of α, the growing trend of harmonized homogeneity with the number of clusters gradually weakened, also in line with the expectation that the harmonized homogeneity restrains this preference.

In order to quantify this trend, linear fit is performed on cluster number-harmonized homogeneity curves with different α, and the slope is obtained. Figure 3 illustrates the slope of α-harmonic homogeneity curve(hereinafter referred to as “slope”) versus the number of clusters. It can be seen from Figure 3 that when α is near 2.29, the "slope" is close to 0. As α further increases, the "slope" decreases to a negative value, which means harmonized homogeneity grows a tendency to prefer fewer categories. The desired α is the zero point of this curve. Therefore, we determine that the optimal α is 2.29. The harmonized homogeneity mentioned later all takes $\alpha = 2.29$.

Some details are omitted for clarity of explanation. In fact, for the PCA dataset, we also trained a series of K-medoids models and calculated the "slope". The "slope" used to determine the optimal α is actually the arithmetic average of the "slope" on the non-PCA dataset and the one on the PCA dataset.

4 Postprocessing of clustering - merging

Among all the models we tried, the highest score of harmonized homogeneity is obtained by the SOM with 5×3 grid topology on the PCA dataset, which is 0.3302, still failing to exceed the Köppen’s 0.3346. Judging from the maps plotted from these classification results, the results with fewer clusters usually contain broad clusters that should be further divided, and the results with more clusters appear to be fractionated, expecting merging of clusters.

To this end, we hope to improve the classification results with the help of some post-processing methods. Human experts may consider manually merging some fractionated categories, or subdividing some categories that are too broad. Merging is often easier than subdividing. [12]

This research designs a merging algorithm to simulate such post-processing, which is otherwise done by human experts. Compared with human experts, the merging algorithm will be more objective, judging whether a certain pair of clusters should be merged based on some objective indicators.

The merging algorithm examines three indicators to determine whether two clusters should be merged: Are the two clusters close to each other in the feature space(i.e., are the two climate types similar)? After merging, how much will the internal evaluation metric of the cluster(Calinski-Harabasz metric, hereinafter referred to as C-H metric, is chosen here in place of silhouette due to its low computational complexity) and the external evaluation metric(harmonized homogeneity with vegetation types) change respectively? Higher similarity between the two clusters, together with the greater increase in the indicator values after merging, secures a higher priority for them to be merged. The three factors should be considered comprehensively.

Basically, such merging is at the cost of clustering validity in the sense of data science, in exchange for the fitness with vegetation types. The silhouette value of vegetation types on the climate dataset is actually negative(-0.2235). In this sense, the vegetation classification itself poorly fits the natural classification of climate data. The C-H values always decline after merging. The ideal merging should limit the decrease of C-H value as small as possible, and increase harmonized homogeneity as much as possible, so as to satisfy the rationality in terms of data science to the greatest extent and preserve fitness with the actual vegetation types.

The similarity between the two clusters is measured by the Euclidean distance between the cluster centroids. The C-H indicator is examined in terms of the ratio that it changes (i.e. its value before merging / its value after merging). Harmonized homogeneity is examined by the absolute value in its improvement after merging (i.e. its value after merging $-$ its value before merging). For the first two, after their values for all the cluster pairs are calculated, the values are converted into quantiles, and then respectively input into an activation function similar to ReLu: $y(x) = \max(0, 1 - 2x)$, $y(x) = \max(0, 2x - 1)$.

Basically, these two activation functions limit merging to be considered only when the distance between the two clusters is below the median of the distances between all cluster pairs, and also the C-H index ratio is higher than its median. The two activated values and the absolute value of harmonized homogeneity improvement are multiplied together to obtain an entry in the optimization matrix. Perform such calculation on all cluster pairs to obtain the entire optimization matrix. The flow is shown in Figure 4.

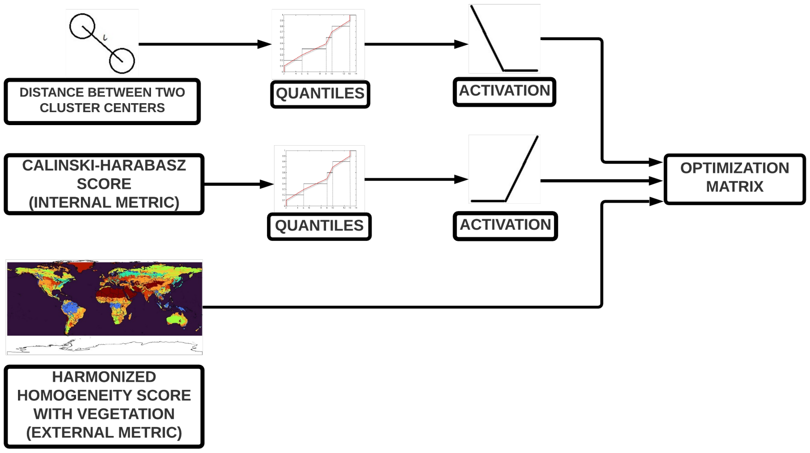
DISTANCE BETWEEN TWO CLUSTER CENTERS
QUANTILES
ACTIVATION
CALINSKI-HARABASZ SCORE (INTERNAL METRIC)
QUANTILES
ACTIVATION
OPTIMIZATION MATRIX
HARMONIZED HOMOGENEITY SCORE WITH VEGETATION (EXTERNAL METRIC)

Figure 4: Optimization matrix in the merging algorithm

The merging algorithm selects the maximum value from the optimization matrix each time, merges the two clusters it represents, and then sets the corresponding rows and columns in the optimization matrix to 0, so that in one round of the algorithm, each cluster can be merged at most once, until there is no value higher than the set threshold in the optimization matrix. The threshold is set to be 0.03 / the total number of categories before merging. The fewer clusters before merging, the higher the threshold. This setting is made due to the finding in our experiments that merging two large clusters may significantly improve the harmonized homogeneity score, but such merging is not reasonable in many cases. In order to suppress the inclination of the algorithm to give large clusters, the threshold is set in this way. The merging algorithm is summarized as follows.

Algorithm 1: The merging algorithm

```
Data: Optimization Matrix $M$, threshold $t$
repeat
    $(x, y) \leftarrow$ the index of the highest value in $M$;
    merge cluster $(x, y)$;
    row $x$, row $y$, column $x$, column $y$ of $M \leftarrow 0$;
until no value in $M > t$;
```

In the experiment, we found that one round of the merging algorithm is generally enough. Even though further merging may improve the harmonized homogeneity score, it will often give unreasonable large clusters, which should not be adopted.

Significant improvements in harmonized homogeneity score of the models are observed after one round of the proposed merging algorithm. After merging, 24 out of 54 models obtain a harmonized homogeneity score exceeding Köppen's 0.3346. The highest score, obtained by 6×6 hexagonal structure SOM on the non-PCA dataset after merging, is 0.3473, where 10 pairs of clusters are merged, thus giving 26 categories. On the other hand, its silhouette value is 0.1675, although significantly lower than before merging(0.2032), it is still much higher than Köppen's 0.0062. This is the finalized classification result. Judging from these statistics, our climate classification achieves the goal of both fitting vegetation types and exhibiting a data science perspective.

5 Results

5.1 Clustering - SOM

SOM can learn the topology of the data set and reflect it in the clustering results. Take the 6×6 hexagonal SOM on non-PCA data set as an example, which, after merging, becomes the final result.

Concluding from the cluster centroids given by the algorithm and the climate type map plotted, the position of each category in the feature space roughly conforms to Figure 5. The SOM model learned the two key dimensions of the climate data set: temperature and precipitation, and thanks to features like growing season length, it further understood the importance of humidity and heat synchronization. Mediterranean climate type 4 and 8, where the total amount of precipitation is considerable but concentrated in winter, are regarded by the model as special dry types of climate. Compared with perhumid and monsoonal climates with similar temperature and precipitation, their growing season is shorter. This conforms to climatology common sense and is a successful discovery by the model.

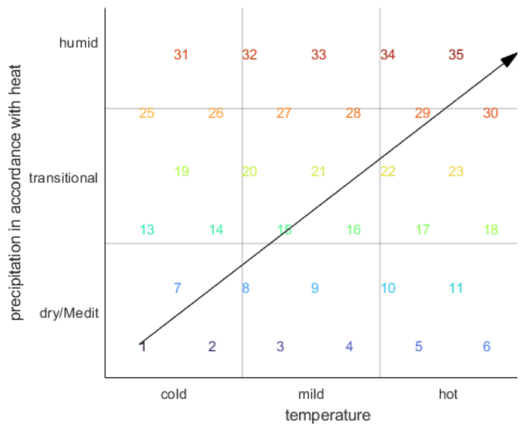

Figure 5: Interpretation of the results before merging

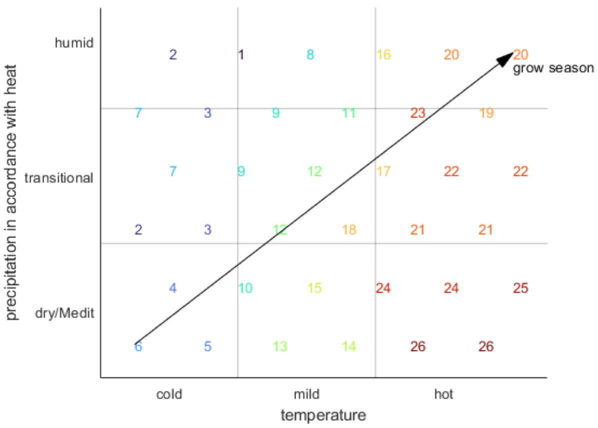

Figure 6: Interpretation of the final result after merging

After running the merging algorithm, the position of each category in the feature space is shown in 6. To cater to human perception, we reordered categories and reassigned color. From #1 to #26, the color gradually changes from dark blue to dark red, to visualize the climates' transition from wet and cold to dry and hot. The global climate classification map is drawn with this colormap (Appendix B).

Concluding from the figure 6, the vegetation types and the cluster centroids, for each category, we also assigned an interpret-able name, and plotted a map. The maps, together with the datasets and source code of this research, can be found in the Github repository[11].

5.2 Clustering - other models

By using unsupervised learning algorithms, we are modeling the underlying structure and distribution in the dataset in order to extract patterns hidden behind climate data. Unlike supervised learning, the answers to clustering problems are subjective. For these models, world maps are plotted to visualize clusters. Harmonized homogeneity and silhouette scores are used to validate the clustering performance, and the V-measure score is used for comparison with the Köppen-Geiger classification system.

Aside from SOM, Agglomerative Clustering, BIRCH Clustering, Gaussian Mixture Clustering, K-Means/K-Medoids Clustering and Mini Batch K-Means Clustering are also clustering algorithms implemented in this project. Among these models, K-Means on the non-PCA dataset is close to giving a better climate classification as the homogeneity score of 0.3216 is close to Köppen's 0.3346. The scores are shown in Figure 7.

[11] https://github.com/logic2333/climate-clustering

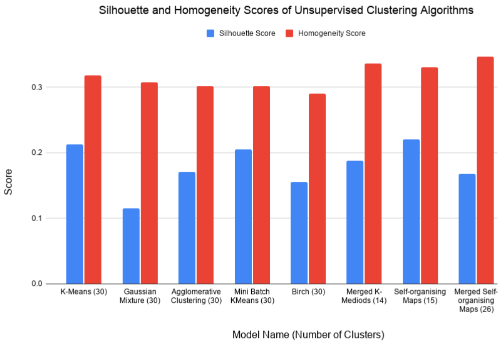

Figure 7: Performance of other clustering models

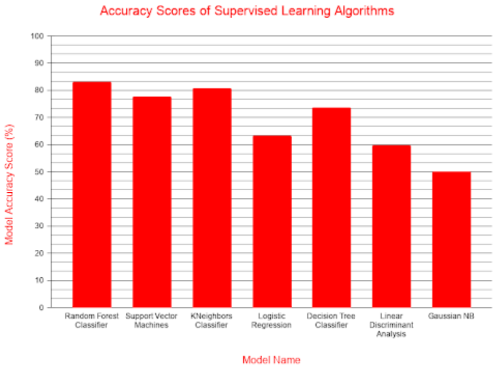

Figure 8: Performance of supervised learning models

5.3 Supervised learning

Supervised learning algorithms model the relationship and dependency between the input features and target labels so that one can envision the output values for the new data based on its interconnection links learned from the target column [13]. We practiced several supervised learning models on the mapped dataset, namely, Support Vector Machine(SVM), Logistic Regression, Linear Discriminant Analysis(LDA), K-Nearest Neighbour Classifier, Gaussian NB Classifier, Random Forest Classifier and Decision Tree Classifier. Random Forest Classifier fits a number of decision tree classifiers on various subsections of the dataset and uses the mean values to boost the accuracy between the input features and target labels. For this project, Random Forest Classifier is able to find the mapping between climate and vegetation without PCA with an accuracy of 84.09%, highest among all attempted supervised learning models. The scores are shown in Figure 8.

6 Conclusion and Future Work

6.1 Conclusion

Defining the "best" climate classification is intricate, but the advantage of these models is that they fit better with climatology data and are objective. On the other hand, it sacrifices simplicity and interpretability, comparing with traditional classification.[14] After all, whether the classification is precise or not should be judged by climatologists. However, data science provides some insight for climatologists on global climate.[15]

6.2 Future Work

Climatology data for the years 1991-2020 is still a work in progress. Due to computational restraints and time limit, we were unable to evaluate models like mean shift, affinity propagation and spectral clustering. Future work should attempt these models and may involve variables like altitudes on the newly collected data to see how the results describe global climate change.

References

[1] Tomislav Hengl et al. "Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential". In: *PeerJ* 6 (2018), e5457.

[2] Jakob Zscheischler, Miguel D Mahecha, and Stefan Harmeling. "Climate classifications: the value of unsupervised clustering". In: *Procedia Computer Science* 9 (2012), pp. 897–906.

[3] Wikipedia. *Potential evaporation*. URL: `https://en.wikipedia.org/wiki/Potential_evaporation`.

[4] John T Abatzoglou et al. "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015". In: *Scientific data* 5 (2018), p. 170191.

[5] D Karger et al. *Data from: Climatologies at high resolution for the earth's land surface areas, Dryad Digital Repository*. 2017.

[6] Dirk Nikolaus Karger and Niklaus E Zimmermann. *Climatologies at High Resolution for the Earth Land Surface Areas CHELSA V1. 2: Technical Specification*. 2019.

[7] Anil Jain, Karthik Nandakumar, and Arun Ross. "Score normalization in multimodal biometric systems". In: *Pattern recognition* 38.12 (2005), pp. 2270–2285.

[8] Barry L Kalman and Stan C Kwasny. "Why tanh: choosing a sigmoidal function". In: *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*. Vol. 4. IEEE. 1992, pp. 578–581.

[9] Sulla-Menashe D Friedl M. *MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V006*. NASA EOSDIS Land Processes DAAC, 2015.

[10] Markus Kottek et al. "World map of the Köppen-Geiger climate classification updated". In: *Meteorologische Zeitschrift* 15.3 (2006), pp. 259–263.

[11] Andrew Rosenberg and Julia Hirschberg. "V-measure: A conditional entropy-based external cluster evaluation measure". In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. 2007, pp. 410–420.

[12] Brian Everitt. *Cluster Analysis*. UK Wiley, 2011.

[13] David Fumo. *Types of Machine Learning Algorithms You Should Know*. URL: `https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861`.

[14] AJ Cannon and B van de Hurk. "Köppen versus the computer: comparing Köppen-Geiger and multivariate regression tree climate classifications in terms of climate homogeneity." In: *Hydrology & Earth System Sciences* 16.1 (2012).

[15] Benedikt Knüsel and Christoph Baumberger. "Understanding climate phenomena with data-driven models". In: *Studies in History and Philosophy of Science Part A* 84 (2020), pp. 46–56.

Appendices

A Climate variables specification

Group	Standardization Method	Activation Method	Feature	Description
Seasonality	independent	normal	Annual temperature range	The difference between annual average maximum temperature and the average annual minimum temperature
			Diurnal temperature range	The average difference between the daily maximum temperature and the daily minimum temperature
			Standard deviation of temperature	Standard deviation of monthly mean temperature series
			Standard deviation of precipitation	The standard deviation of the series of the proportion of monthly precipitation in total annual precipitation
Mean Temperature	no	use two centers: 0 and 10[12], the deviation σ is set to be the standard deviation of all the monthly average temperature data	Monthly average temperatures	A total of 12, sorted from low to high [13]. A total of 24 after two activations
	independent	normal	Mean annual temperature	
			Winter [14]Average temperature	
			Summer average temperature	
			Dry season [15]average temperature	
			Wet season average temperature	
Precipitation	combined	normal	monthly precipitation	A total of 12. Aligns with the months in the monthly mean temperature series
			Monthly average precipitation	

[12] 0°C is the freezing point, 10°C is commonly acknowledged as the lowest temperature that can support plant growth.

[13] The seasons in the northern and southern hemispheres are opposite, but this should not lead to different climate types. Sort the months according to the average temperature to eliminate the difference between the northern and southern hemispheres.

[14] Refers to the three consecutive months with the lowest average temperature. The same applies to summer.

[15] Refers to three consecutive months with the least precipitation. The same goes for wet season.

	no	0.25 as the center, the standard deviation of the four series combined as the deviation σ	Maximum monthly precipitation Minimum monthly precipitation Proportion of winter precipitation Proportion of summer precipitation Proportion of dry season precipitation Proportion of wet season precipitation	in total annual precipitation
Humidity	No	0 as the center, the standard deviation of all data in this feature group as the deviation σ	Humidity of each month	Each month's precipitation minus its potential evapotranspiration (PET). 12 features in total, aligns with the months in the monthly mean temperature series
Precipitation-PET correlation	independent	normal	Cross correlation at zero offset	$\frac{\mathbf{p_1 p_2}}{\sqrt{\|\|\mathbf{p_1}\|\|^2\|\|\mathbf{p_2}\|\|^2}}$
			Standard deviation of the cross-correlation series	Fix the time window of $\mathbf{p_1}$, slide $\mathbf{p_2}$, calculate the correlation between the two, and get the cross-correlation series. The standard deviation of the series
	No	No	Cosine of precipitation-PET phase difference	Phase difference is expressed as the offset of the maximum value in the cross-correlation sequence. Denote it as ϕ in months. $\cos{(\pi/6 * \phi)}$
Day count	independent	normal	Length of grow season Snow cover duration	according to TREELIM model[6]
Month count	No	6 as the center, 1 as the deviation	Average temperature higher than 10°C Average temperature lower than 0°C	
			Dry	i.e. humidity is negative

B Final result visualization

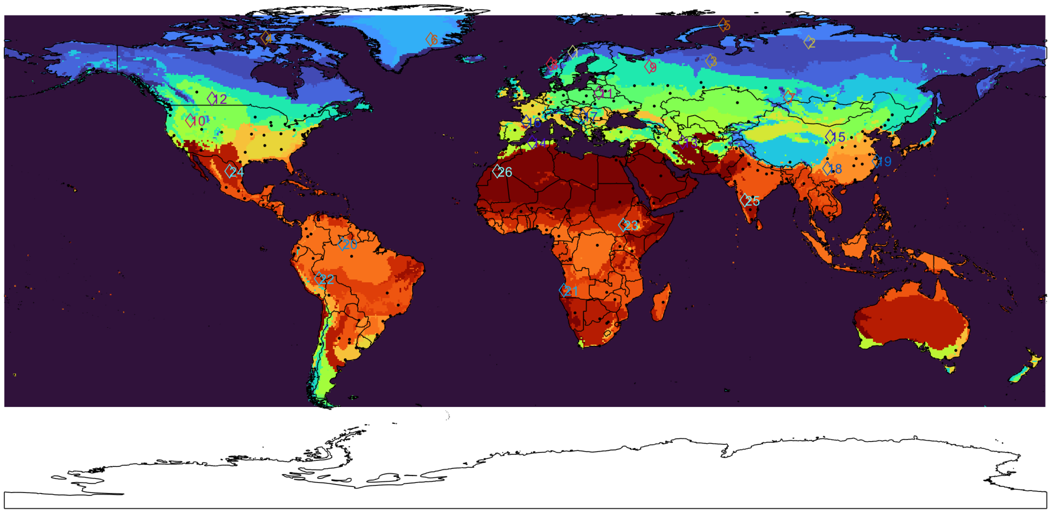

Figure 9: Final classification result

C Contributions

Task	Group Member
Climatology Data set Extract Transform Load	Ji Luo
Vegetation Data set Extract Transform Load	Bilal Hussain, Ji Luo, Sakina Patanwala
Mapped Data set	Bilal Hussain, Sakina Patanwala
Merging Algorithm	Ji Luo
Unsupervised Learning Algorithms (K-Medoids and Self-Organizing Map)	Ji Luo
Unsupervised Learning Algorithms (Agglomerative Clustering, BIRCH Clustering and Gaussian Mixture Clustering)	Sakina Patanwala
Unsupervised Learning Algorithms (K-Means Clustering and Mini Batch K-Means Clustering)	Bilal Hussain
Supervised Learning Algorithms (Logistic Regression, Random Forest Classifier and Support Vector Machines)	Sakina Patanwala
Supervised Learning Algorithms (Linear Discriminant Analysis, Decision Tree Classifier, Gaussian NB and K-Nearest Neighbour Classifier)	Bilal Hussain
Harmonized Homogeneity Metric, V-Measure (Similarity with Koppen) and Silhouette Score	Ji Luo, Bilal Hussain, Sakina Patanwala
Poster	Ji Luo, Bilal Hussain, Sakina Patanwala
Report	Ji Luo, Bilal Hussain, Sakina Patanwala