Optimizing the microbial Global Population Structure (mGPS)

Author: Andrew Bergman

Supervisor: Eran Elhaik

BINP37, 15 CTS

Lund University

Contents

Abstract

The microbial Global Population Structure (mGPS) was developed to predict geographical coordinates based on the pan-metagenomic signatures present across the globe. This project aims to improve the predictive ability of mGPS on the Subway dataset and a soil dataset. mGPS performance on the Subway dataset was not significantly improved after applying feature-engineering and BORUTA, except for two cities; Bogota and Brisbane, further analyses are required to cement why that is. After interpolating and oversampling the soil dataset, mGPS increased predictions within 100km of their sampling sites by 1010.67%, the mean error was reduced by 84.75%. The increased predictive ability of mGPS serves as an indicator that sparsely populated datasets can be processed and trained and achieve a good effect. Further investigation is required to cement why BORUTA affected the Subway datasets differently and why the families dataset yielded worse performance than the species dataset overall. This work can be extended by applying interpolation and oversampling to the Subway dataset.

1 Introduction

1.1 mGPS

The microbial Global Population Structure (mGPS) is a machine-learning model that predicts geographic coordinates using the relative sequence abundance (RSA) of microorganisms sampled at train stations in major cities across the globe. mGPS is at its core an extreme gradient boosted decision tree (XGBoost). XGBoost is considered a state-of-the-art machine-learning model due to its gradient boosting (Chen, 2016). Gradient boosting means that several weaker models are layered on sequentially, each model attempts to correct the errors of the previous models (Chen, 2016). mGPS can utilize any dataset that is geo-tagged with the sampling sites' longitudes and latitudes, although the XGBoost model at mGPS's core also relies on two additional features. Sample city, country and continent are three examples of compatible features with mGPS. mGPS is composed of four layers. First, two logistic regression models that classifies samples into continents and cities/countries (depending on the dataset), followed by two linear regression models predicting longitude and latitude. Each layer outputs its predicted probabilities, these are used as additional features in the subsequent layers. Each layer aims to reduce *the loss function,* which is the log-loss for classification tasks and the squared error for regression tasks. Machine-learning models are often applied to high-dimensional data, but having too many features can result in noise and increased computational load (Guyon & Elisseeff, 2003). To mitigate this, feature selection is conducted using recursive feature elimination (RFE) (Zhang et al., 2024), which involves iteratively testing features for significance in a model fit and subsequently removing the least significant predictors (Guyon et al., 2002). When assessing the performance of mGPS, 5-fold cross-validation and leave-one-out-cross-validation are two common methods to choose from (Calomino et al., 2024). In this project, 5-fold cross-validation is used for model evalution. After mGPS has predicted geographical coordinates for the test data, the *pull_land* function is called (Zhang et al., 2024). *Pull_land* finds any samples that are in bodies of water, identifies the nearest land boundary and pulls the prediction to that boundary. This is a function that is called upon when utilizing mGPS for predictions that are known to be on land. When training mGPS, its hyper-parameters are tuned by grid-searching. These hyper-parameters are 1) the number of rounds of boosting iterations n=300 and n=600. The number of rounds reflect the number of times the decision trees are trained and updated in each boosting iteration. If the number of trees is too high, the model risks to become overfit to the training data. If the number of trees is too low, the model can become underfit, not capturing patterns in the data. 2) Learning rate (lr = 0.05, lr = 0.1). The learning rate is a coefficient that weighs the contribution of each decision tree to the final coordinate prediction. A low learning rate means that each tree will have a low impact on the overall prediction, a high learning rate means that each tree contributes more to the final prediction. 3) Max tree depth (d = 3, d = 6, d = 9). Deeper decision trees identify complex data patterns more easily, however, they can lead to model overfitting, and shallow decision trees can yield an undertrained model. 4) The fraction of features sampled for each tree, f = 0.6, f = 0.8. The features are split among different decision trees to counteract overfitting of any singular decision tree. In each decision tree, mGPS performance is evaluated at 60% and 80% of features.

When mGPS has generated predictions, the distance between the actual coordinates and the predicted coordinates are evaluated using Haversine's distance. mGPS predicts 62%, 74%, and 84% of samples within 250 km, 500 km, and 1000 km of the true coordinates, respectively, for the subway dataset (Zhang et al., 2024). The city classification accuracy is 92%, with a mean sensitivity of 78% and a specificity of 99%. Using the soil dataset, 61% and 71% of predictions were within 100km and 500km of the nearest sampling country border, the country classification accuracy is 88%. The sensitivity is 75% and the specificity is 99% for the soil dataset.

1.2 Datasets

1.2.1 Subway dataset

mGPS was initially trained on the subway dataset. MetaSUB is an international organization utilizing metagenomic data to understand and improve urban environments, they emphasize the study of microbial ecosystems. Among their datasets they have generated samples from subway stations, soil samples and aquatic samples (Ryon et al., 2022). The term *meta-pangenome* describes the genetic profile in any given environment (Ma & Ravel, 2020). At each subway station, the swabbed surfaces contain a mixture of metagenomic signatures. The subway dataset contains 4070 samples of 3757 microorganisms, 40 cities are represented around the globe (figure 1B). Using this information, the relative sequence abundance can be easily discerned by dividing the prevalence of the various microorganisms by the total abundance at the sampling site.

1.2.2 Soil dataset

The soil dataset contains 223 observations of 511 operational taxonomic units (OTUs) tagged with longitude and latitude (figure 1A4). It is smaller than the subway dataset, which leads to poorer performance of mGPS (Zhang et al., 2024). This is a common limitation of machine-learning models that can suffer from both excessive, and insufficient data. Too little data can lead to poorly trained models (Domingos, 2012), while large, high-dimensional datasets can suffer 'the curse of dimensionality' (Bellman, 1961).The low sample density is further exaggerated by significant geographical clustering of the samples, with vast areas void of samples in between clusters. Asia is poorly sampled, so is most of Europe (figure 1A4). Australia is the sampling country with the highest sampling density, however, all of the samples are tightly clustered in one area, not necessarily representing the meta-pangenomic signature of Australia, but rather a small subset of its microbial diversity (figure 1A4).

1.3 Introduction to BORUTA

Feature selection is explored by replacing Recursive Feature Elimination (RFE) with BORUTA, a wrapper model employing random forest classifier. BORUTA assesses each predictor's impact on city classification by comparing it to a version of the same predictor with randomly shuffled values, thus nullifying the possible relationship between it and the response variable (Kursa & Rudnicki, 2010). A Z-score is calculated, providing a metric describing the feature's importance in relationship to its randomly shuffled counterpart. Features are subsequently selected based on their importance relative to the response variable. The reasoning for choosing BORUTA is that it can thoroughly classify features as either important or unimportant (Kursa & Rudnicki, 2010), leading to possibly more informative features than RFE can provide.

1.4 Introduction to feature Engineering

Feature engineering involves creating additional information from already available features to enhance their informational value and/or to reduce the dataset dimensionality (Kumar & Patra, 2021). This approach can establish relationships between novel predictor variables and the response variables, which may be more informative than the original features. As previously mentioned, the subway dataset's metagenomic data includes count data for the species at various locations. Higher taxonomic categories can perhaps be utilized to improve the distribution of feature values which in turn may yield increased mGPS performance.

1.5 Introduction to interpolation

Sample interpolation entails generating pseudo-samples using previously existing datapoints to generate feature values (Gruven et al., 2016). While these methods may lead to overfitting of mGPS, it also provides an understanding of how datasets can be improved upon to achieve better performing machine-learning models.

1.6 Synthetic Minority Oversampling Technique (SMOTE)

Class imbalance occurs when any categorical variable is represented to a lesser extent than others (Chawla et al., 2002). The same study uncovered that combining oversampling the minority class and undersampling of the majority class improves the performance of several classification algorithms. When a machine-learning model is trained on an imbalanced dataset, it can lead to biased accuracy results. Consider the following example: If you train a ML model on mammogram images containing 99% healthy cells and 1% cancer cells, the ML model can correctly classify all healthy cells and incorrectly classifies all cancer cells and still achieve 99% accuracy. To address this issue, synthetic samples of the minority class can be generated via SMOTE (Chawla, 2002).

1.7 Aim

The aim of this study is to optimize the predictive capabilities of mGPS through feature engineering and feature selection in the subway dataset and to address the small sample size in the soil dataset through interpolation and oversampling techniques.

2. Methods

2.1 Employing BORUTA

BORUTA is a wrapper method that calculates the importance of each feature by comparing a random-forest classifier's accuracy using the feature to the accuracy using a corresponding shadow feature. The shadow feature is created by randomly reshuffling the feature values, thereby removing any meaningful relationship to the response variable. The difference in classification accuracy is converted to a measure of feature importance (Z-score). BORUTA was implemented using the *BORUTA* package in R. The feature selection process is initiated by generating a correlation matrix for all operational taxonomic units in the dataset. Next, if any feature correlation exceeds 98%, one of them is removed. The maximum number of iterations was capped at 10,000 with 50 decision trees per iteration.

2.2 Feature engineering the subway dataset

Feature engineering was conducted by aggregating the microorganisms by their highest taxonomic categorisation, the family level. The relative sequence abundance was subsequently calculated by dividing the family count by the total number of families in each sample.

2.3 Interpolating the soil dataset

When performing interpolation, we treated every feature individually. For every feature, 80% of samples were used for training and 20% of data for testing (5-fold cross-validation). Six different interpolation techniques were assessed for every feature, the highest performing interpolation technique (assessed by R-squared) was noted. The dataset was subsequently grid interpolated (10000 grid points covering earth) and filtered. Any interpolated sample without original datapoints within a 25 km radius were removed. Geographical data interpolation of data is aligned with the Tobler's first law of geography, stating that geographically proximal data points should be more similar than distant ones (Tobler, 1970). The following interpolation methods were considered when grid-interpolating the features.

2.3.1 Inverse distance-weighted interpolation (IDW)

IDW was implemented using the R *gstat* package. IDW utilizes k-nearest neighbours (k = 3) and weighs their influence on the interpolated sample as a function of the distance between them. The weights are inversely proportional to the Euclidean distance between the coordinates (equation I). The power coefficient *p* determines the degree to which the distance between points affect the weights. The feature value at the interpolated coordinate is calculated using the mean of the weighted values of the nearby samples (equation 2).

Equation 1. The weights of each original point considered for interpolation is determined by the distance and the power coefficient *p*.

$$w_i = \frac{1}{d_i^p}$$

Equation 2. The value assigned to any given feature of the interpolated coordinates is the mean of the weighted values.

$$Z = \frac{\sum_{i=1}^{N} w_i Z_i}{\sum_{i=1}^{N} w_i}$$

2.3.2 Generalized additive models (GAMs)

GAMs are generalized linear models that allow for non-linear relationships between the predictor and response variables. GAMs are able to capture complex patterns that do not present themselves in linear fashion, but they also require fine tuning to avoid over- or underfitting the model. GAMs facilitate non-linear relationships using splines, which are fitted using *penalized likelihood estimates,* a balance between curve complexity (smoothness) and goodness of fit (Brown et al., 1991).

Equation 3. g is the link function, β_0 is the intercept, f_i are splines applied to each predictor variable.

$$g(E|[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

2.3.3 Thin plate spline interpolation (TPS)

TPS optimizes the bending energy of a surface modelled as a thin metal sheet that is fitted over the known data points (Bookstein, 1999). This results in a smooth surface with smooth feature gradients. Since TPS splines are fit to all the data points, they are sensitive to outliers which can skew the feature gradients (Bookstein, 1999). TPS does not necessarily create biologically relevant gradients, as microorganisms are not necessarily distributed in a smooth surface. They are transmitted in various ways, such as human movement, which can lead to patchy abundance patterns. TPS are not meant to fit data which(Fang, 2023).

Equation 4. r_i is the distance from the data point to the interpolated coordinate (x, y). a and w are determined when fitting the model.

$$Z(x,y) = a_1 + a_2x + \sum_{i=1}^{N} w_i U(r_i)$$

2.3.4 Multi-level B-spline approximation (MBA)

MBAs employ basic splines (B-splines) in three layers that model the data at different resolutions. The difference between B-splines and other splines is that B-splines isolate each polynomial function, which leads to any one spline not influencing the others (Lee et al., 1997). There are three layers of splines in MBAs, 1) Coarse level: A low resolution B-spline that captures general trends. 2) Intermediate level: Fitted to predict regional variation. 3) Fine level: Captures local variation (Lee et al., 1997).

2.4 Synthetic oversampling in soil dataset

To address the imbalanced soil dataset Synthetic Minority Class Oversampling Technique (SMOTE) is employed. SMOTE addresses the imbalance by creating pseudo-samples, where each synthetic sample is generated by interpolating the feature values of K-nearest neighbours. We opted for k = 3, as higher values of k would result in remote pseudo-samples interpolating from samples that are too far apart. This would not represent realistic feature distributions.

3. Results

3.1 BORUTA yields similar predictive performance to RFE using the species and families Subway datasets

BORUTA was performed using a random forest with using a max tree depth of 50 and running 10,000 iterations. BORUTA identified 247 features as important using, while RFE determined a set 200 features to be most informative (Zhang et al., 2024). BORUTA did not yield any significant improvements over RFE, although I required significantly more computational time. However, a contradiction arose, that BORUTA yielded higher importance scores for features in the families dataset compared to the species dataset (Figure S1). Although the feature importance was increased, it did not lead to improved mGPS accuracy. To assess the experiment’s impact on the core XGBoost model, the *pull_land* function was disabled to isolate the effect of BORUTA. Applying BORUTA improved mGPS accuracy in the families dataset. The mean distance error was reduced from 1144 km to 992 km, and the median distance error decreased from 78 km to 71 km. BORUTA yields increased accuracy in terms of predictions located within 100 km, 250km, 500km, 1000km, 1500km and 3000km. However, when applying BORUTA to the species dataset, the performance was reduced. Using RFE with the species dataset yielded a mean distance error of 888 km and a median distance error of 72 km, compared to 894 km and 103 km, respectively. The positive impact of RFE is also reflected in predictions within 100km, 250km, 500km, 1000km, 1500km and 3000km (table 1). Since BORUTA proved beneficial on the families dataset, and some sites showed improved predictions, we briefly analysed the cause of the contradicting informativeness scores. We found that the sites where the families dataset performed best (e.g., Brisbane, Bogotá, Figure 1) had higher levels of the important features compared to the original dataset (Figure S1, S2).

Table I. The percentage of predictions within 100km, 250km, 500km, 1000km, 1500km and 300km for the different datasets and means of feature selection.

Feat. Sel.	**Dataset**	**<100km**	**<250km**	**<500km**	**<1000km**	**<1500km**	**<3000km**
RFE	Original	55.38	65.99	72.60	79.58	84.15	90.61
RFE	Families	52.7	61.64	67.54	74.76	80.02	87.78
BORUTA	Original	49.48	62.80	70.95	79.28	85.45	92.03
BORUTA	Families	54.91	64.59	70.44	77.71	81.96	90.63

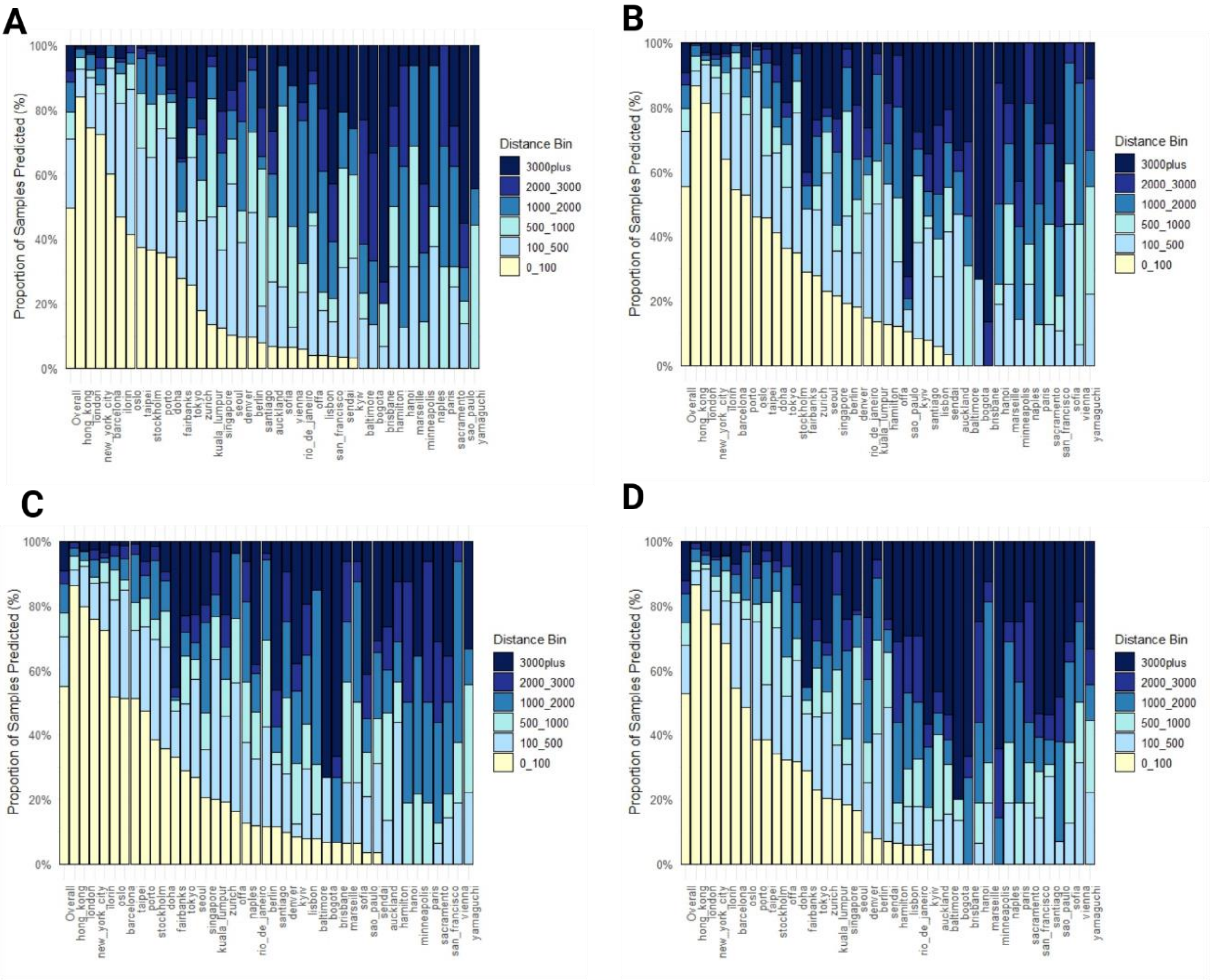

Figure 1. Cumulative fractions of predictions within 100km, 250km, 500km, 1000km, 1500km and 3000km for each city in the subway dataset. The different sub-figures illustrate the distributions of predictions for **A)** BORUTA Species dataset **B)** RFE species dataset **C)** BORUTA families dataset **D)** RFE Families dataset.

3.2 Interpolating samples to the soil dataset

3.2.1 Different features require different interpolation methods

The soil dataset is sparse, leading to mGPS struggling to learn the signatures of the clustered sampling sites. Some sampling sites are highly populated, such as Australia (n = 85 samples), while other regions are sparsely populated (figure 4A). Sparse and clustered samples generally lead to poor predictive performance, this is especially true in parts of Europe and Asia. Initially, individual interpolation methods were assessed, however, no singular method could properly interpolate all features. We then applied six interpolation methods to each individual feature, which proved fruitful (figure 2). The number of interpolated points is influenced by the number of samples originally present in their proximity, which further increases the class imbalance. To mitigate this, SMOTE was applied after interpolating the features. The final interpolation utilized GAM for 212 features, IDW for 213 features, nearest neighbour interpolation for 23 features, and TPS for 63 features. The mean R-squared was 0.371, 0.487, 0.413 and 0.501, respectively (table SII).

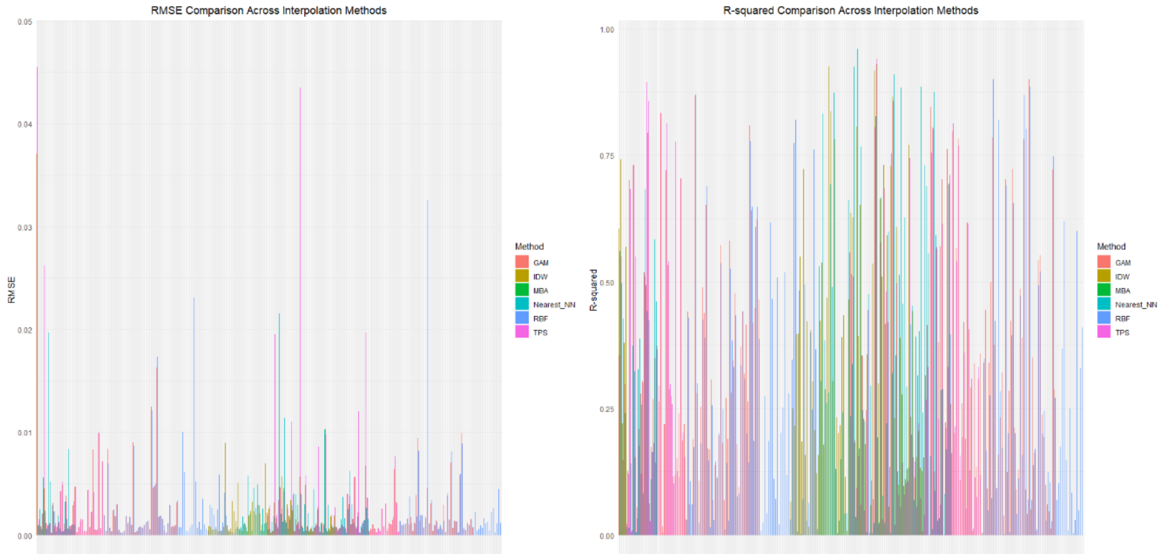

Figure 2. This bar chart illustrates the RMSE and R-squared for the different features in the soil dataset. The highest performing interpolation method is displayed, showcasing the variety of interpolation methods used to yield the best possible interpolation. The x-axis contains the features included in the original dataset.

3.2.2 SMOTE at 42 samples per country proved to provide informative data points

Initially we experimented with UBL's default SMOTE parameters, those being *balanced* and *extreme*. Comparing the two presets, 'balanced' yielded better predictive ability (figure 3A-B). The downside of using the balanced preset is that it applies oversampling of the minority classes and undersampling of the majority class. In an already sparse dataset, the balanced preset discards valuable information. We proceeded to experiment with SMOTE settings, finding that oversampling to 42 samples per country without any undersampling yielded the best results (figure 3C). The oversampling was assessed by plugging the dataset into a random-forest classifier and followingly assessing the country classification accuracy.

The improvement from the original dataset to the SMOTE balanced preset and to 42 samples per country was further supported by assessing the clustering of the features in each dataset. For each dataset, the features corresponding to OTUs were analysed using principal component analysis (PCA).The silhouette scores of each dataset were calculated, supporting that oversampling to 42 samples per country yields the least overlapping feature sets. The silhouette scores for the original dataset, the balanced SMOTE dataset and the 42 samples per country dataset are -0.009, 0.128 and 0.136, respectively. (Figure 3D-F).

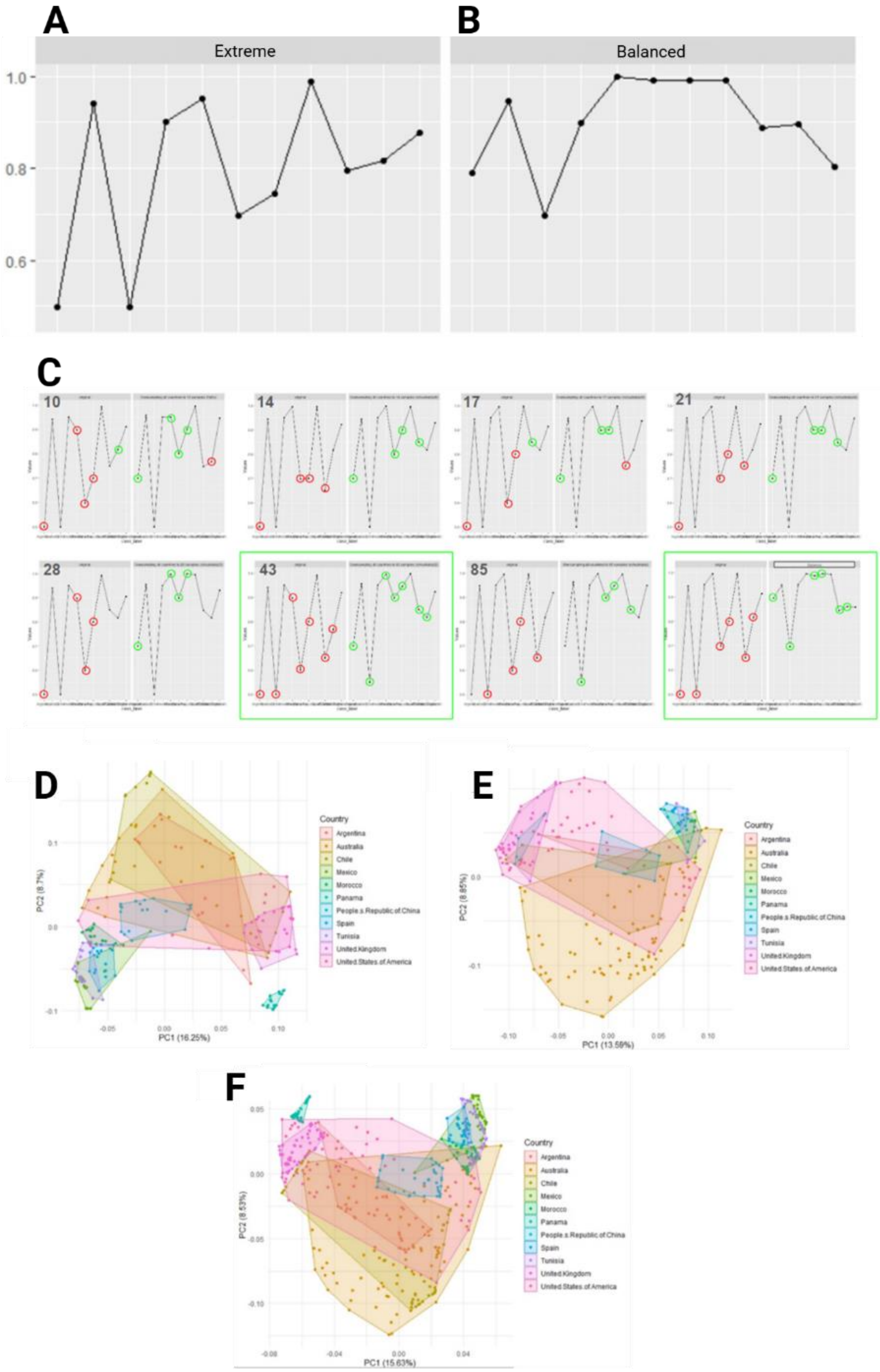

Figure 3. A-C) SMOTE optimization for soil dataset Two different SMOTE configurations were used, UBL's SMOTE contains two pre-configured methods: **A)** balanced and **B)**

extreme. Extreme SMOTE entails oversampling the minority classes to the number of samples present in the majority class (Australia, n = 85). Balanced SMOTE entails oversampling the minority classes and undersampling the majority class to yield a balanced dataset while minimizing the creation of synthetic data points. SMOTE was evaluated by plugging the SMOTEd datasets into a random-forest classifier and assessing the accuracy of country classification. **C)** The classification accuracies of countries using different oversampling thresholds. **D-F)** Principal component analyses (PCAs) of **D)** the original dataset, **E)** balanced SMOTEd dataset and **F)** SMOTEd to 42 samples per country. Principal component analyses display how datapoints cluster together, effectively providing a degree of similarity of the datapoints.

3.3 Interpolation and SMOTE increases mGPS predictive performance

mGPS's accuracy increased in the interpolated and SMOTE dataset. It was further improved when a second round of interpolation was applied before SMOTE (Figure 5A). It can also be observed that the interpolated dataset provides more accurate mGPS predictions, although the class imbalances are exaggerated. This could be due to the bias introduced by class imbalance. On the original dataset, mGPS had a mean distance error of 2981.73 km. This is significantly reduced to 597.71 km after applying the interpolation and SMOTE. The two-round interpolation followed by SMOTE yields the best performance metrics, with 44.76% of samples predicted within 100 km and a mean error of 454.81 km (Table 3).The SMOTEd and interpolated dataset increases the fraction of predictions within 100km of their sampling sites by 1010.67%, the mean error was also reduced by 84.75%. When observing the country-wise predictions, interpolation followed by SMOTE yields significantly better predictions compared to the original dataset, this is true for all countries represented in the original dataset, as shown in figure 5. mGPS’s accuracy when trained on the different datasets is illustrated in figure 6, where each red circle corresponds to a prediction and each blue circle corresponds to an original datapoint. Some outliers are hard to catch, these happen to be isolated samples, which renders them non-eligible for neither SMOTE nor any interpolation method.

3.4 Comparing oversampled datasets to previous metrics

The initial mGPS project's soil prediction quantification was handled differently from the approach in this project. Predictions were considered 0 km from their sampling sites if correctly assigned to their countries of origin. If a prediction falls outside its country's borders, the distance is calculated from the predicted coordinates to the nearest border of the sampling country. Our results show 96% of predictions within 100 km, compared to 86% in the previously reported results (Figure 7C). mGPS's performance on clustered data was also assessed using k-means clustering (with k ranging from 2 to 4). The target variable for mGPS was recalibrated to the cluster level. The modified dataset performs better than the original, but none of the datasets yielded particularly good results. This is likely because each cluster is represented by fewer observations compared to the country-level target variable used in all other experiments.

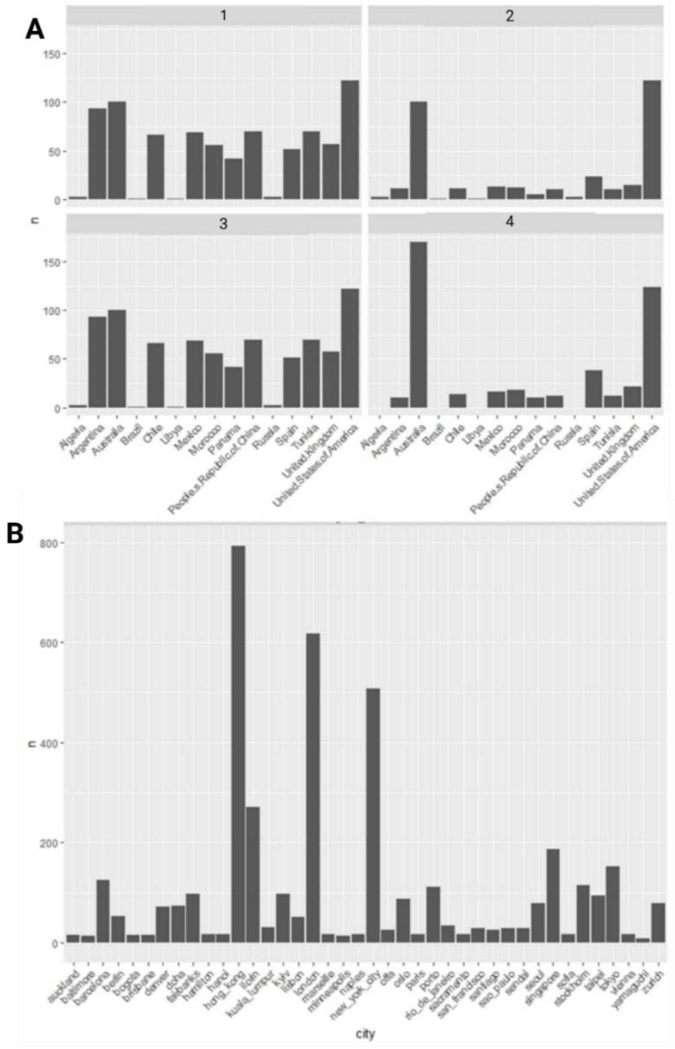

Figure 4. Distribution of samples in the datasets used in this study. **A)** Distribution of countries in the soil datasets. Each panel corresponds to a dataset. *1:* The product of employing two rounds of interpolation followed by SMOTE. *2: T*he product of interpolating the soil dataset using the most compatible interpolation methods *3:* The product of employing interpolation, then SMOTE. *4:* The original soil dataset. **B)** Distribution of cities in the Subway dataset.

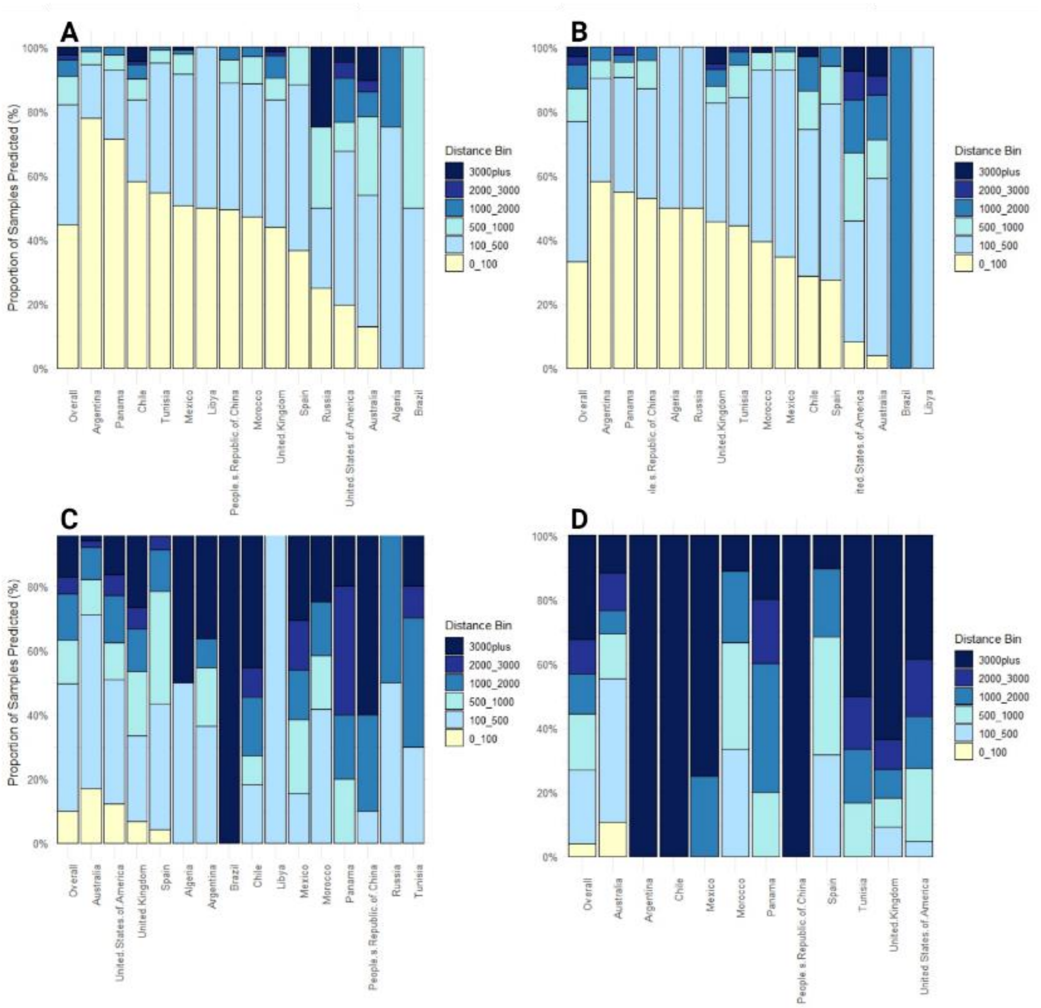

Figure 5. Cumulative fraction of predictions for countries represented in the soil dataset, the interpolated dataset, and the interpolated-SMOTEd dataset**. A)** Two-round interpolation + SMOTE, **B)** Interpolation + SMOTE, **C)** SMOTE, **D)** Original. The discrepancies in the countries present in the bar chart are due to interpolations and SMOTE yielding new country classifications.

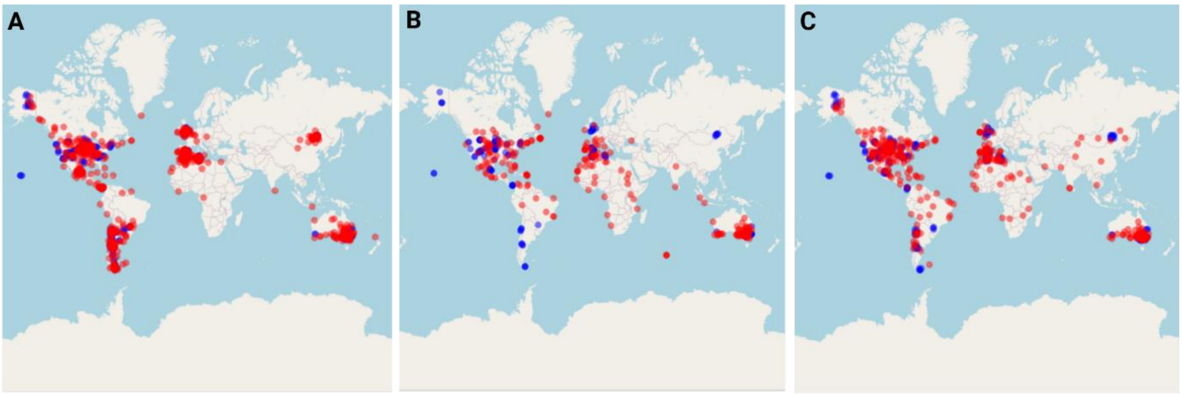

Figure 6. mGPS soil predictions using the interpolated-SMOTEd dataset. The interpolated-SMOTEd dataset provides the highest predictive accuracy, followed by the interpolated dataset. The poorest performance is observed in the original dataset. Blue circles represent actual data points, while red circles represent predicted points. **A)** The interpolated-SMOTEd dataset, **B)** The original dataset, **C)** The interpolated dataset.

Table III. Performance metrics using different soil datasets. *Interpolation* is abbreviated as *I* and *SMOTE* is abbreviated as *S*.

Dataset	<100km	<500km	<1000km	<2000km	<3000km	>3000km	Mean error
I + S	33.16%	60.72%	76.80%	87.03%	91.89%	97.00%	597.71
S	10.05%	31.36%	49.70%	63.01%	73.66%	82.84%	1700.43
I*2 + S	44.76%	68.01%	82.07%	90.90%	94.11%	97.42%	454.81
Original	4.03%	14.79%	26.90%	44.39%	52.91%	67.71%	2981.73

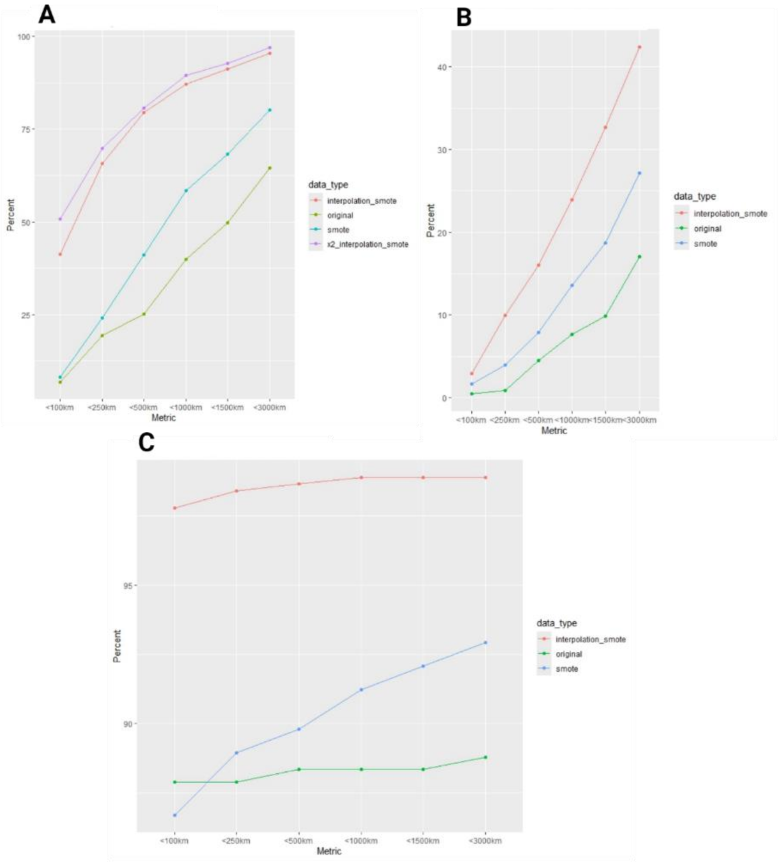

Figure 7. A) Fraction of predictions for distances between 0–3000 km, showcasing the differences between the original dataset, the interpolated dataset (with SMOTE), interpolation combined with SMOTE, and an iteration of 25 km interpolation. Raw metrics. **B)** K-means clustering with 2–4 clusters per country, assessed using Leave-One-Out Cross-Validation (LOOCV). **C)** Cumulative predictions for the same distance metrics as in the previous mGPS implementation.

4. Discussion

Rationale behind families aggregation

The Subway dataset contains 3757 species, which were aggregated at the family level to assess whether this aggregation would affect the distribution of the data. This idea originated from an observation made in the initial mGPS paper, where it was found that the most informative features are those that are globally dispersed, as opposed to features that are unique to specific regions (Zhang et al., 2024). We have not thoroughly investigated why the families dataset performs worse than the species dataset, but we have collected some data. Figure S1 illustrates the most important features of the species and families datasets. The families dataset produced features with higher peak importance than the species dataset. We then investigated the occurrence of features, ranked by importance. We found that in the cities where the families dataset performed better, the most important features of the families dataset occur more frequently. This may provide a good starting point for further investigation into whether the Subway dataset can be more selectively feature-engineered to improve mGPS predictions.

4.1 Comparing BORUTA and RFE in the different datasets

BORUTA is a feature selection method that determines the importance of features in a random forest classifier. The BORUTA algorithm is exhaustive and yielded more predictive features than RFE. BORUTA improved mGPS performance on the families dataset but reduced performance on the species dataset. We have not thoroughly investigated the reasons for this difference due to time constraints. Further investigation into why BORUTA improved performance on one dataset but not the other may provide valuable insights for future experimentation with feature selection.

Although BORUTA provided more features than RFE, it did not fully converge after

10,000 iterations, leaving some features unclassified as either important or unimportant. Interestingly, the kurtosis and skewness of the features were reduced in the families dataset (FIGURE S), which should improve the performance of the layered regression models in mGPS.

4.2 Interpolating and oversampling the soil dataset

Interpolation is applied before SMOTE because class imbalances are further exaggerated by interpolation. The balanced SMOTE yields results similar to the 42 samples per country approach. We opted for 42 samples per country because this approach avoids discarding data and only requires oversampling. mGPS's predictive ability is improved by combining interpolation and SMOTE, as it can produce more accurate representations of feature combinations that distinguish regions from one another. However, a downside to heavy interpolation and oversampling is that mGPS may become overtrained on the dataset.

The results demonstrate the ability of mGPS to utilize sparse datasets with this approach, improving classification accuracy.

MBA was not used for any feature during interpolation, as it failed to converge when applied to the data. This is surprising, given that MBA is more flexible than TPS. However, no further investigation into this issue was conducted due to time constraints. Future work should explore the effect of additional interpolation methods.

4.4 Distance calculations

Zhang et al. (2024) used a different method of calculating the distance between data points in the soil dataset than the method used in this study. They calculated the distance from the predicted coordinates to the nearest border of the country of origin. If a predicted coordinate was located within its origin country's borders, the distance was considered to be 0 km. This approach was adopted because the goal of applying mGPS to the soil dataset was to assess its ability to identify the country of origin for soil samples. However, the improvements presented in this study motivate discarding the previous distance calculation in favour of using raw predicted distances.

Acknowledgements

This thesis would not have been possible without the guidance and support of my supervisor Eran Elhaik. I would also like to extend my thanks to Yali Zhang, the main author of the mGPS paper (Zhang et al., 2024) for providing valuable context.

References

Chen T, Guestrin C. XGBoost. The Journal of Machine Learning Research [Internet]. 2016 Aug 8;785–94. Available from: https://doi.org/10.1145/2939672.2939785

XGBoost loss function for regression and classification

Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research [Internet]. 2003 Mar 1; Available from: https://dl.acm.org/doi/10.5555/944919.944968

Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning [Internet]. 2002 Jan 1;46(1/3):389–422. Available from: https://doi.org/10.1023/a:1012487302797

Calomino C, Bianco MG, Oliva G, Laganà F, Pullano SA, Quattrone A. Comparative Analysis of Cross-Validation Methods on PPMI Dataset. IEEE [Internet]. 2024 Jun 26; Available from: https://doi.org/10.1109/memea60663.2024.10596885

Ryon KA, Tierney BT, Frolova A, Kahles A, Desnues C, Ouzounis C, et al. A history of the MetaSUB consortium: Tracking urban microbes around the globe. iScience [Internet]. 2022 Oct 20;25(11):104993. Available from: https://doi.org/10.1016/j.isci.2022.104993

Zhang Y, McCarthy L, Ruff SE, Elhaik E. Microbiome Geographic Population Structure (MGPS) detects Fine-Scale geography. Genome Biology and Evolution [Internet]. 2024 Oct 7; Available from: https://doi.org/10.1093/gbe/evae209

Domingos P. A few useful things to know about machine learning. Communications of the ACM [Internet]. 2012 Sep 25;55(10):78–87. Available from: https://doi.org/10.1145/2347736.2347755

Kursa MB, Rudnicki WR. Feature Selection with theBORUTAPackage. Journal of Statistical Software [Internet]. 2010 Jan 1;36(11). Available from: https://doi.org/10.18637/jss.v036.i11

Kumar V, Patra SK. Feature engineering for machine learning and deep learning assisted wireless communication. In: Studies in computational intelligence [Internet]. 2021. p. 77–95. Available from: https://doi.org/10.1007/978-3-030-70542-8_4

Guven O, Eftekhar A, Kindt W, Constandinou TG. Computationally efficient real-time interpolation algorithm for non-uniform sampled biosignals. Healthcare Technology Letters [Internet]. 2016 Mar 11;3(2):105–10. Available from: https://doi.org/10.1049/htl.2015.0031

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research [Internet]. 2002 Jun 1;16:321–57. Available from: https://doi.org/10.1613/jair.953

Tobler WR. A computer movie simulating urban growth in the Detroit region. Economic Geography [Internet]. 1970 Jun 1;46:234. Available from: https://doi.org/10.2307/143141

Brown RA, Hastie TJ, Tibshirani RJ. Generalized additive models. Biometrics [Internet]. 1991 Jun 1;47(2):785. Available from: https://doi.org/10.2307/2532174

Bookstein FL. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence [Internet]. 1989 Jun 1;11(6):567–85. Available from: https://doi.org/10.1109/34.24792

Fang L. Smooth digital terrain modelling in irregular domain using adaptive finite element thin plate spline smoother. arXiv (Cornell University) [Internet]. 2023 Jan 1; Available from: https://arxiv.org/abs/2302.12974

Lee S, Wolberg G, Shin SY. Scattered data interpolation with multilevel B-splines. IEEE Transactions on Visualization and Computer Graphics [Internet]. 1997 Jan 1;3(3):228–44. Available from: https://doi.org/10.1109/2945.620490