mGPS Algorithm Optimization

Course: Bioinformatics Research Project (BINP37), 15 credits

Student: Chandrashekar CR

(email: ch1131ch-s@student.lu.se)

Supervisor: Eran Elhaik

(email: eran.elhaik@biol.lu.se)

Lund University 2025

Abstract

mGPS (microbiome Geographic Population Structure) is a novel algorithm developed to analyze the geographical distribution of microbial communities. This report presents an enhanced version of mGPS with significant improvements in computational efficiency and predictive accuracy, enabling robust analysis of large-scale datasets. Optimization efforts focused on refining the algorithm's core functionality, improving data handling, and implementing ensemble-based learning architectures.

The core of this study involves the application of advanced machine learning models to improve geolocation predictions based on microbiome signatures. The Metasub dataset, comprising 4,070 samples collected from 40 cities across seven continents between 2016 and 2017, served as the basis for evaluation. Several neural and ensemble models were developed and benchmarked against the original mGPS framework.

Our refined pipeline introduces a hierarchical ensemble learning architecture incorporating XGBoost, CatBoost, TabPFN, neural networks, and GrowNets, with performance-gated meta-models at each geographic resolution. Class imbalance was addressed using SMOTE, and stratified 5-fold cross-validation was employed to ensure fair and robust evaluation. A critical advancement includes a corrected error calculation methodology that more accurately quantifies performance by accounting for hierarchical prediction dependencies.

The optimized mGPS framework offers superior predictive performance and sets a foundation for future applications in public health surveillance, forensic investigation, and ecological research.

1. Introduction

1.1 Geographical Prediction Using Microbial Signatures

The ability to predict the geographical origin of microbial communities has significant implications across multiple disciplines, including biosurveillance, forensic investigation, and public health monitoring [REF NEEDED]. Microorganisms present in environmental samples can serve as biological signatures of specific locations, reflecting local environmental conditions, human activity patterns, and ecological factors unique to particular geographical regions [REF NEEDED].

The microbial Global Population Structure (mGPS) algorithm represents an innovative approach to leverage these microbial signatures for geographical prediction [REF NEEDED]. By analyzing the relative sequence abundance (RSA) of microorganisms in samples, mGPS can infer the likely origin of a sample at multiple geographical scales—from continent to precise coordinates. The original implementation employed a hierarchical prediction model using XGBoost, where continent classification preceded city prediction, followed by latitude and longitude coordinates [REF NEEDED]. This hierarchical approach achieved notable success with 92% accuracy at the city level and a reported median error distance of 137 km [REF NEEDED].

1.2 Previous Work and State of the Art

Several studies have built upon the original mGPS framework, primarily utilizing XGBoost as the core algorithm [REF NEEDED]. These studies have implemented various improvements including smart analysis techniques and hyperparameter tuning to enhance prediction accuracy at the coordinate level [REF NEEDED]. The typical workflow involved a top-down hierarchical approach where:

- The initial feature space is reduced from thousands of microbial features to approximately 200 informative features through recursive feature elimination (RFE) [REF NEEDED]
- The reduced feature set is used to predict the continent classification.
- Continent prediction probabilities are augmented to the original feature space
- This augmented feature set predicts the city classification.
- Both continent and city predictions are incorporated to predict geographic coordinates.

This cascaded prediction approach has become the standard methodology for microbial-based geographical prediction systems [REF NEEDED].

1.3 Limitations in Existing Approaches

Despite its success, the original methodology contains notable limitations. First, the reported error metrics are interdependent—the distance error calculation benefits from the high accuracy of continent and city predictions, potentially obscuring the true predictive performance [REF NEEDED]. Second, the hierarchical nature of the model means that errors propagate through prediction levels, with early misclassifications affecting subsequent predictions [REF NEEDED].

These cascading errors significantly impact coordinate-level predictions, with misclassifications at the continent level leading to substantial geographical displacement in final predictions.

Additionally, previous approaches have not adequately addressed:

- The challenge of error analysis in hierarchical models where mistakes at early levels compound
- Dataset imbalances at different geographical scales
- The potential benefits of ensemble approaches beyond single-algorithm solutions
- Proper transformation of coordinate data for machine learning applications [REF NEEDED]

1.4 Dataset Characteristics

This research utilizes the MetaSUB dataset, a comprehensive global atlas of metagenomic samples systematically collected from mass-transit systems across major cities worldwide [REF-metasub-2020]. The dataset comprises 4,070 quality-controlled samples from 40 cities and 7 continents, each geo-tagged with precise latitude and longitude coordinates. Samples were collected from high-contact surfaces such as railings, benches, and ticket kiosks, following standardized protocols developed by the International Metagenomics and Metadesign of Subways and Urban Biomes (MetaSUB) Consortium [REF-metasub-2020]. The MetaSUB dataset provides an annotated, geospatial profile of urban microbial communities, including taxonomic profiles, functional characteristics, and antimicrobial resistance (AMR) markers. Taxonomic profiles were generated using KrakenUniq, with relative sequence abundances (RSA) calculated for each taxon. The dataset also includes AMR gene profiles quantified as reads per kilobase per million mapped reads (RPKM), and features a large number of novel genetic elements, such as 10,928 viruses, 1,302 bacteria, 2 archaea, and over 838,000 CRISPR arrays not found in reference databases [REF-metasub-2020].

A globally consistent "core" urban microbiome of 31 species was identified in 97% of samples, distinct from human commensal organisms, though its abundance varies by city.

The dataset reveals strong geographic variation in microbial signatures, with taxonomic diversity decreasing with increasing latitude, and widespread but variable AMR gene profiles across cities. Notably, a significant fraction of the urban microbiome remains uncharacterized, highlighting the presence of unique, city-specific DNA sequences.

For the mGPS analysis, the MetaSUB dataset underwent rigorous quality control, retaining only cities with at least eight samples. The resulting dataset's high biodiversity and standardized collection make it ideal for developing and evaluating machine learning models for microbial biogeography and AMR tracing [REF-metasub-2020]. The feature space consists of relative sequence abundances of microorganisms, initially containing more than 3,000 different organisms per sample.

Not all microorganisms contribute equally to location-specific signatures. Through recursive feature elimination (RFE) using random forest as a base model, the feature space is typically reduced to approximately 200-300 of the most informative microbial features [REF NEEDED]. This dimensionality reduction focuses the models on the microbial signatures most predictive of geographical origin. The mGPS tool, leveraging this dataset, achieved 92% accuracy in predicting source cities and enabled fine-scale localization and AMR transmission route tracing.

1.5 Proposed Enhancements

Our research addresses these limitations through several methodological improvements. Instead of relying on a single algorithm, we implement an ensemble approach combining multiple machine learning models: XGBoost, CatBoost, TabPFN, neural networks, and GrowNets. We introduce a threshold-filtering mechanism where only models achieving specified accuracy thresholds contribute to meta-model predictions at each geographical level. This selective ensemble approach helps mitigate error propagation through the hierarchical prediction chain. To address dataset imbalances, we employ Synthetic Minority Over-sampling Technique (SMOTE) particularly at the continent level, ensuring more balanced training representations. Our implementation also incorporates 5-fold cross-validation with stratification to maintain consistent class distributions across training and testing splits, reducing potential biases in model evaluation. A significant contribution of our work is the implementation of corrected error calculations that provide a more comprehensive understanding of model performance. By computing the expected coordinate error as a weighted sum over different correctness combinations (continent and city predictions), we offer a more nuanced view of prediction accuracy that accounts for the hierarchical nature of the geographical prediction task. Through these methodological enhancements, our primary objective is to develop a robust, ensemble-based framework for microbial geographical prediction that addresses the limitations of previous hierarchical models. Specifically, we aim to improve prediction accuracy, reduce error propagation, and provide a more realistic assessment of model performance across multiple geograph-

ical scales. Ultimately, our approach has the potential to advance the field of microbial forensics and environmental monitoring by enabling more accurate and reliable location predictions from metagenomic data. This study does not address temporal variation in microbial signatures, which may be explored in the future work. The following section details the materials and methods used to implement and evaluate our proposed framework.

2. Materials and Methods

2.1 Data collection

This study utilized the Metasub dataset [REF], comprising approximately 4,070 samples collected from subway stations across 40 cities worldwide between 2016 and 2017. The MetaSUB Consortium, established in 2015, conducted this large-scale global metagenomic sampling campaign primarily during the Global City Sampling Days on June 21st in both years [REF].

Samples were collected following standardized protocols [REF] from high-contact surfaces within mass-transit systems, including railings, benches, and ticket kiosks. Two main sampling methods were employed: Copan Liquid Amies Elution Swabs and Isohelix Buccal Mini Swabs, each paired with specific preservative media. DNA extraction was performed using commercial kits like MoBio PowerSoil DNA Isolation Kit and ZymoBIOMICS 96MagBead DNA Kit [REF].

16S rRNA sequencing was performed on the samples, with libraries prepared for Illumina NGS platforms and sequenced using Illumina HiSeq X Ten System, generating approximately 5-7 million 125bp paired-end reads per sample. The resulting data was processed to extract microbial population information. The metagenome dataset underwent quality control, filtering, and normalization according to established protocols [REF]. Raw reads were processed using the MetaSUB Core Analysis Pipeline (CAP) for taxonomic profiling, with taxonomic assignments generated using KrakenUniq based on the NCBI/RefSeq Microbial database.

The processed dataset, containing relative sequence abundances (RSAs computed after subsampling to 100,000 classified reads) and corresponding geo-tagged locations, was obtained from previous work on the mGPS algorithm [REF], which was designed to predict sample locations based on their microbial composition.

2.2 Preprocessing of the dataset

Dataset preprocessing involved multiple quality control steps, including removal of locations with insufficient samples (cities with fewer than eight samples were excluded) and correction of collection coordinates to improve geographical accuracy. These procedures followed the methodology described in the original mGPS paper [REF]. While the original implementation used R, we developed equivalent preprocessing workflows in Python to maintain methodological consistency.

Feature engineering included recursive feature elimination (RFE) to reduce the initial $\approx 3,000 microbial features to approximately 200-300 most informative features using Random Forest as$ $fold cross-validation to ensure robust feature ranking and selection.$

To address class imbalance issues, particularly at the continent level where regions

like Oceania and Africa were underrepresented, we employed Synthetic Minority Oversampling Technique (SMOTE) [REF] to generate synthetic samples for underrepresented geographic regions. This approach helped prevent model bias toward overrepresented continental regions like Europe and North America, thereby improving generalization capabilities across diverse geographic locations. The SMOTE algorithm was configured to create synthetic samples until class distributions achieved a threshold ratio of 1:3 between minority and majority classes [REF].

2.3 Neural Network Approaches for mGPS Optimization

2.3.1 Model 1: Separate Neural Networks

We developed a hierarchical approach using discrete neural networks for each prediction level (Figure X). Each network operated independently but in sequence:

- **Continent Classification Network:** A feed-forward neural network with three hidden layers (256, 128, and 64 neurons) employing ReLU activation functions. The output layer used softmax activation to generate probability distributions across 7 continents. This network was trained using categorical cross-entropy loss with Adam optimizer (learning rate: 0.001, $\beta 1 : 0.9, \beta 2 : 0.999$).

- **City Classification Network:** The second network accepted the continent classification probabilities as additional features alongside the original microbial features. It consisted of three hidden layers (128, 64, and 32 neurons) with ReLU activations and a softmax output layer for city classification. This network was trained using categorical cross-entropy loss with the same Adam optimizer settings.

- **Coordinate Regression Network:** The final network accepted city probability distributions and the original features to predict geographic coordinates. This network employed four hidden layers (256, 128, 64, and 32 neurons) with ReLU activations and a linear output layer for coordinate regression.

Each network was optimized independently, allowing for specialized tuning at each geographical level. This approach mirrors the original mGPS hierarchical structure but with neural networks replacing XGBoost at each level.

2.3.2 Model 2: Combined Neural Network

We developed an integrated neural network architecture (Figure Y) designed to simultaneously predict continent, city, and geographic coordinates. This model utilized shared lower-level feature representation layers followed by specialized branches for each prediction task:

- **Shared Layers:** Four dense layers (512, 256, 128, and 64 neurons) with ReLU activation functions extracted high-level features from microbial abundance data.
- **Continent Branch:** Two dense layers (64 and 32 neurons) followed by a softmax layer for continent classification.
- **City Branch:** Three dense layers (128, 64, and 32 neurons) that received both shared layer outputs and continent branch outputs, followed by a softmax layer for city prediction.
- **Coordinate Branch:** Three dense layers (128, 64, and 32 neurons) accepting inputs from all previous branches, ending with a linear output layer for coordinate prediction.

The model was trained using a multi-objective loss function:

$$\text{Loss} = \alpha \cdot \text{CrossEntropy}_{\text{continent}} + \beta \cdot \text{CrossEntropy}_{\text{city}} + \gamma \cdot \text{MSE}_{\text{coordinates}} \quad (1)$$

where $\alpha, \beta, and \gamma were weighting parameters (set to 0.4, 0.3, and 0.3 respectively). This allowed end-to-end training with backpropagation across all prediction levels, enabling the model to learn interdepende$ 0.3) were included after each hidden layer to prevent overfitting.

2.4 Ensemble Learning Approach

We implemented a comprehensive ensemble learning system combining multiple machine learning algorithms: XGBoost [REF], CatBoost [REF], LightGBM [REF], TabPFN [REF], Neural Networks, and GrowNet [REF]. This approach was specifically designed to enhance prediction capabilities for underrepresented geographic classes in the dataset.

The ensemble framework was structured as a hierarchical prediction system:

2.4.1 Continent Classification

Base models were trained on microbial sequence abundance features to generate probability distributions across continents. To optimize ensemble performance, we implemented a quality-based filtering criterion where only models achieving individual accuracy above 94% contributed to the meta-model. This threshold prevented lower-performing models from introducing prediction noise. The filtered probability outputs were combined using an XGBoost meta-learner to produce final continent predictions.

2.4.2 City Classification

For city-level prediction, we augmented the microbial abundance features with probability distributions from the continent classification stage. The same ensemble methodology

was applied, with a filtering threshold of 92% accuracy (established based on benchmark performance from previous mGPS studies [REF]). Filtered model outputs were again combined using an XGBoost meta-learner to generate city probability distributions.

2.4.3 Coordinate Regression

For geographic coordinate prediction, we employed model-specific strategies:

- **Tree-based models:** XGBoost, CatBoost, and LightGBM directly predicted raw latitude and longitude values. Latitude predictions were incorporated as additional features for longitude prediction to leverage geographic relationships.
- **Neural network models:** For TabPFN and other neural network approaches, we transformed geographic coordinates to Cartesian coordinates to address gradient vanishing issues [REF]. The transformations applied were:

$$x = \cos(\text{lat}) \times \cos(\text{lon}) \tag{2}$$

$$y = \cos(\text{lat}) \times \sin(\text{lon}) \tag{3}$$

$$z = \sin(\text{lat}) \tag{4}$$

Final coordinate predictions were generated by combining outputs from qualified regression models using an XGBoost meta-model. This hierarchical ensemble approach enabled more accurate location prediction by leveraging diverse algorithmic strengths while systematically filtering out lower-performing models at each geographic resolution level.