Microclimate Comparison for Human Health and Agriculture

2024 INBRE Summer Fellowship Research Plan
University of Idaho

Student: Leif Huender, Student Researcher, North Idaho College
Preceptor: Dr. John Shovic, Research Faculty, University of Idaho
Research Mentor: Dr. Mary Everett, Postdoctoral Fellow, University of Idaho

Abstract

During the summer of 2024, North Idaho College student Leif Huender will be working with the University of Idaho computer science department on the Coeur d'Alene campus. With research mentor Dr. Mary Everett, Leif will be exploring microclimate comparisons to predict malaria outbreak and agricultural yield outcomes. This project will first look at dimension-reducing similarity score metrics between microclimates areas and seasons, and use scored similarities based on a variety of metrics to predict outcomes. More direct machine learning models will also be used on the sequence data to make the same predictions. Model performance will be compared for accuracy, data and power consumption, and model explainability. Results will be presented at ICUR in the summer and potentially the International Conference on Precision Agriculture, pending abstract acceptance. Additional conferences and publication opportunities will be sought for the student.

Introduction

Understanding climates worldwide is an important task for a variety of concerns relating to health and human safety. Accurate forecasts help create well-informed management decisions for mitigating adverse climate impacts on societies. However, creating accurate forecasts is not a trivial task and much time and effort is devoted to understanding how weather patterns impact an area.

The University of Idaho department of computer science on the Coeur d'Alene campus is interested in understanding and forecasting microclimate behavior given the following assumptions:

1. Not all microclimates have information for a long period of time. Weather information and outcome information for an area of interest may be limited to a few seasons.

2. Model explainability is an important factor along with model explainability. "Black box" models may find success limited success with data outside testing, and may limit trustworthiness of the model. However, accurate models are the ultimate goal.

3. Focus is placed upon similarity to microclimates in the same and other areas in the same and other seasons. This is noted because prior management decisions on times and areas with similar behavior may be useful in determining future management decisions.

With these assumptions, this research project aims to focus on two microclimate outcomes with human health impacts:

1. Malaria Outbreaks: Climate conditions are hypothesized to impact the severity of malaria outbreaks, as weather patterns such as temperature and rainfall can greatly impact mosquito development and disease transmission [1]. This focus will compare climate data from at-risk areas to determine if similarities can be quantified between seasons to predict malaria outbreak severity.

2. Agricultural Production: Climate is a huge factor in agriculture production and climate change issues have made proactive management of agriculture all the more important as weather patterns become more predictable. Efficient and sustainable agriculture is important to meet the predicted increased food demand on the world population. This focus will compare growing season microclimates to predict agricultural yield.

Research Questions

1. What is the best method or combination of methods to quantify similarity between two microclimates?

For this research question, a variety of methods will be employed to quantify similarity between two historical microclimate sequences, including distance metrics and methods derived from information theory. Similarity scores will be taken between different areas and different years.

2. Can similarity comparison be used to predict microclimate outcomes?

Once similarity scores are obtained, this research questions will look at their ability to be used in a variety of models to predict outcomes.

3. What is the best combination of information to use to compare similarity and determine outcomes between two microclimates?

This research question will look at which combinations of the available weather information are most useful in determining outcomes.

4. How do similarity comparison-based methods perform compared to direct machine learning methods when it comes to predicting outcomes?
These methods will be compared based on their performance on prediction outcomes, explainability of models, data and computation needs, and ability to relate additional historical information.

5. For either direct machine learning methods or similarity-based methods, can early (incomplete) season information be used to predict outcomes?
In general, is will be assumed that complete data for season is available for similarity use score and modeling. However, in real-world applications, stakeholders will likely want to know how early season indicators related to outcomes. For experiments relating to this research question, data will be reduced.

Datasets

Malaria Outbreaks: For malaria outbreak prediction task, data from the World Health Organization (WHO) will be used as microclimate comparison outcomes, in particular vector species prevalence in the areas of North New Delhi, South New Delhi, Gujarat, and Tamil Nadu in India. Climate data for these areas is obtained through the OpenWeather website, including information on temperature, humidity, light, and rainfall. This information is available at a high resolution through the years of interest when the data was collected, primarily in the 1980s.

Agricultural Production: For the agricultural yield prediction task, data from the United States Department of Agriculture will be used for pecan yield in four counties in New Mexico: Chaves, Doña Ana, Otero, and Sierra for several years in the 2000s. Again, OpenWeather data will be used to obtain microclimate data for these sites.

Leif has already begun work with the University of Idaho during the spring semester, and has obtained both the microclimate and outcomes data for both tasks. Leif has been able to format it in a manner conducive to his experiments.

Methodology and Experiments

Note: *Research may not be conducted in the order indicated here*. However, these tasks will comprise the bulk of the research project and the order below is one logical flow of task completion.

Phase I – Research

One of the first steps to any project is to research the state of the art practices in the task at hand. Leif has already conducted much of the work in this face, which is a literature review of various climate comparison methods and seasonal prediction tools.

Phase II - Microclimate Similarity Quantification

This phase involves similarity score metric calculations for seasons between areas and years. The following similarity score in the geometric, correlation, and information theory categories will be calculated across full seasons and years for individual microclimate data streams (just temperature, just humidity, etc). The manifold comparison technique(s) will use two or more datastreams together.

Geometric:
Euclidean Distance: Also called the Pythagorean distance, it is the distance between two points via a straight line.
Manhattan Distance: The Manhattan distance between two points uses right angles rather than a diagonal line.

Cosine Similarity: The dot product of a two vectors (in this case, datasets) divided by the product of the length of the vectors. It essentially gives a size measure of the angle between the vectors of information.

Correlation:
Pearson Correlation: This is a straightforward linear correlation measurement between two variables.
Spearman Correlation: Similar to Pearson correlation, but uses data rankings.
Kendall Tau Correlation: Also uses data ranking to measure dependence.

Information Theory:
KL Divergence: Uses principles from information theory and cross-entropy to compare two distributions.

Manifold Comparison:
Earth Mover's Distance: compares multi-dimensional data distributions, based on ideas of filling gaps in multi-dimensional space.
MTop-Divergence: Also compares multi-dimensional data distributions, based on Cross-Barcode [2].
Riemann Manifold Learning: Maps high dimensional data into low dimensional spaces [3].

Other manifold comparison metrics may be used as they are uncovered during research.

Phase III: Outcome Prediction with Similarity Comparison
Models will be created using the similarity scores from phase 1 to predict outcomes. In particular, input to the models will be similarity scores on features (from one or more metric calculations), along with outcomes from the reference year, to predict the outcome for the year of interest. The following are options for models to conduct the score prediction (though may not be the ultimate choices).

- Linear Regression (and other regression models):
- Support Vector Machines
- Bayesian Models
- Decision Trees/Random Forests
- Small neural networks

Performance of each model will be evaluated based on accuracy of outcomes, data and power consumption, and explainability.

Phase IV: Feature Analysis
Experimental results from Phase III will be repeated with different combinations of features and similarity scores based on primary results in directions that appear promising.

Phase V: Direct Machine Learning Models
Machine Learning Models will be trained from the direct sequence data to predict outcomes. Options for these models include:

- Markov Chains
- Transformers
- LSTM Networks

These direct models (without similarity score pre-processing) will be evaluated and compared to the Phase III and Phase IV models based on accuracy, data and power consumption, explainability, and ability to tie in historical data.

Phase VI: Early Season Data
A subset of promising experiments from Phases III-V will be repeated with a smaller amount of seasonal data to determine if outcomes can be predictor by early indicators.

Mentoring Plan

The student, Leif Huender, will be working directly with postdoctoral fellow Mary Everett on the University of Idaho – CDA campus in the same office block. Mary will be available for the duration of the project. Check-ins with the student will be conducted on progress each week, and Mary will be responsible for connecting the student with resources he needs to be successful.

Additionally, Mary plans to supervise the project and assist where needed, especially with running experiments with less well-supported metrics. Mary will also assist in creating publications and presentations and provide feedback to the student.

Project Outcomes

The outcomes for the student on the research project are as follows:

- Student will understand the research process, from question formulation to experiment development to results tabulation to communication of findings.
- Student will develop greater understanding of statistical processing and machine learning models.
- Student will develop skills in data management and microclimate analysis.
- Student will gain practice in writing research publications.
- Student will gain confidence in communicating research findings.
- Student will have the opportunity to work alongside other members of a research team, including undergraduates, graduate students, and research staff/faculty.

Publication and Communication Plan

There will be several opportunities for the student to present and communicate their research throughout the summer and beyond:

- Idaho Conference on Undergraduate Research [required]: The student will present a poster at ICUR in Boise during July 2024.
- International Conference on Precision Agriculture: At the moment, it is planned that the student will submit a poster abstract to this conference in Kansas, which several other members of the computer science team are attending. This is subject to abstract acceptance. The conference takes place in late July of 2024.
- University of Idaho Conference on Undergraduate Research: This conference is an opportunity for the student to present their summer project, and will take place likely in April of 2025.
- Potential Conference on Human Health: A conference will be sought for the student to present their findings in particular for malaria outbreak data.

References

[1] Fernando, S.D.Climate Change and Malaria - A Complex Relationship. *UN Chronicle*. https://www.un.org/en/chronicle/article/climate-change-and-malaria-complex-relationship
[2] Barannikov, S., Trofimov, I., Sotnikov, G., Trimbach, E., Korotin, A., Filippov, A., & Burnaev, E. (2021). Manifold Topology Divergence: a Framework for Comparing Data Manifolds. *Advances in neural information processing systems*, *34*, 7294-7305.

[3] Lin, T., & Zha, H. (2008). Riemannian manifold learning. *IEEE transactions on pattern analysis and machine intelligence*, *30*(5), 796-809.