Ali Hakimi

Geospatial neighborhood to enhance machine learning for dust storm susceptibility studies in the Middle East

Sand and dust storms (SDS) represent a significant and widespread obstacle to sustainable development, affecting its economic, social, and environmental dimensions. SDS pose severe challenges to achieving several Sustainable Development Goals (SDGs), including the eradication of poverty, the achievement of zero hunger, the improvement of health and well-being, the provision of clean water and sanitation, the promotion of affordable and clean energy, the creation of decent work and economic growth, the development of sustainable cities and communities, the reduction of climate change, the conservation of life below water, the conservation of life on land, and the establishment of partnerships for the goals. The Middle East and Central Asia alone contribute to about 30% of global dust emissions. A variety of factors such as topography, vegetation cover, soil moisture, soil type, precipitation, evapotranspiration, land cover, humidity, vertical air motion, and wind speed influence dust emission. By developing prediction models, based on the influencing factors, one can simulate and determine control strategies. There are several approaches to modeling and predicting SDS, including numerical modeling and simulation, spatial analysis, and machine learning. Numerical modeling is challenging due to atmospheric dynamics and identifying SDS sources and destinations. Spatial analysis relies on expert opinion and visualization, while machine learning is powerful but overlooks dynamic spatial patterns. Thus, a robust, expert-independent approach that can identify and interpret spatial relationships is needed.

This study aims to develop advanced machine learning techniques to predict dust storms. To address the neighborhood effect three techniques were proposed and evaluated:

1) Feature Creation: new features were created by integrating spatial statistical factors from neighboring features and distances to less influential features. This enhanced the performance of Global ML models by incorporating spatial statistical parameters.

2) Spatially Weighted Machine Learning (SWML): The Global ML model transformed into a SWML model by assigning spatial weights to observations and defining spatial parameters for tuning, such as bandwidth and weighted bootstrapping. This method is more efficient and interpretable than traditional ML. The combination of global ML and SWML enables the capture of expert opinion and the analysis of both global and local data behavior.

3) Combined regression and ML: A simple linear model was implemented, and its residuals were processed using SWML. This approach enabled the evaluation of a range of minimum and maximum evaluation metrics, thereby providing an effective analysis method for the dust storm dataset.

The findings indicate that the extraction of features and the creation of spatially statistical predictors could enhance the process of machine learning (ML) in identifying spatial relations. Additionally, SWML models are more reliable while they depend more on spatial parameters than on dataset size and traditional hyperparameters. These two factors have contributed to the

achievement of more sustainable results and evaluation metrics with feature distribution maps, which is a crucial aspect in introducing ML models to spatial relations.

Keywords: Physical Geography and Ecosystem analysis, Machine Learning, Spatial Neighborhood, Geographically Weighted Random Forest, XGBoost
Advisor: **Ali Mansourian**
Master degree project 30 credits in GIS and Remote Sensing, 2024
Department of Physical Geography and Ecosystem Science, Lund University. Student thesis series INES nr 659