BUAN 6312: Project Proposal – Group 6

Time Series Analysis Of Air Pollution Data In The United States

Prriyamvradha Parthasarathi, Premi Jawahar Vasagam, Mira Radhakrishnan, Sofia Rajan, Martin Navarro, Manoj Mareedu, Vyshnavi Gangineni, Siva Renuka Chowdary Nandigam

---

Introduction:

This project aims to conduct a comprehensive time series analysis of air pollution data in the United States. We will leverage the dataset available at https://data.world/data-society/us-air-pollution-data, which contains information on various air pollutants measured at different locations over time. This analysis aims to uncover patterns, trends, and potential correlations within the air pollution data.

Dataset Description:

The dataset was documented by the U.S. EPA from 2000 to 2016. It contains the following variables: state code, county code, site num, address, state, county, city, date local, NO2 units, O3 units, SO2 units, and CO units.

Objectives:

1. Analyze the time series patterns of various air pollutants such as Nitrogen Dioxide, Ozone, Sulfur Dioxide, and Carbon Monoxide across different geographical locations.
2. Assess and analyze each state's contribution to air pollution, understanding the variations and trends in pollutant levels over time.

To achieve these objectives, we will follow the following steps:

1. Data Preprocessing: In this step, we will examine the dataset for missing values, outliers, and other issues affecting the accuracy. Missing values are placed with mean imputation. Categorical variables are encoded, and feature scaling will be performed.
2. Exploratory Data Analysis: Here, the relationships between the variables and trends will be analyzed using visualizations and statistical analysis. We are studying the correlations between target and explanatory variables.
3. Feature Selection: In this step, we will use feature selection techniques such as correlation analysis, mutual information, and recursive feature elimination to identify the most critical variables that contribute to the pollution. We will then select a subset of the most important variables for use in the classification models.

Model Selection:

We will perform pooled OLS, random effect, and fixed effects models to test the objectives.

Model Evaluation:

We will do hypothesis testing to study the join significance between the variables and R-squared, Adjusted R-squared and Mean Squared Error (MSE) as metrics to evaluate the performance among the models to get better insights regarding air pollution in United States.