EE798 FISA: Assignments

Prof. Tushar Sandhan
sandhan@iitk.ac.in

6 June, 2023

Introduction

India's coal production scenario is dominated by open-pit mining. Future coal demand is expected to be very strong. However, environmental problems, such as worsening air quality brought on by the emission of particulate matter and other gaseous pollutants from diverse mining operations, would restrict the use of coal.

You will be analyzing real-world recent time-series data, for descriptive as well as inferential statistics. Data-set is given in the accompanying `.csv` file. You need to use any programming language or programming environment for statistical analysis. You have to submit your answers via detailed report.

All assignments are merged together, total marks are 40%. As weightage is high, we will only be giving approximate guidelines for data analysis and not formulating precise questions. You have to do descriptive, exploratory and inferential statistical data analysis for this work on your own. This evaluates how well can you analyze the data and infer from it via various tools. Quality of graphs, detailed report and data presentation skills in final report will be considered for the grading. Submission is single comprehensive `.pdf` report.

References

- [1] https://en.wikipedia.org/wiki/Air_pollution.
- [2] https://en.wikipedia.org/wiki/Interpolation
- [3] Appendix

1 COAL INDIA OPEN-PIT BLASTING

`Blasting from open-pit coal mines causing massive air pollution`

The two main air pollutants in NCL coal fields are suspended particulate matter (SPM) and reparable particulate matter (RPM). Air quality monitoring is regularly carried out at both dust-generating and non-generating locations in the vicinity in order to evaluate the particulate pollution in and around the opencast mining projects of the Singrauli coalfield. SPM and RPM concentrations are predominate at coal working surfaces, coal yards, coal handling facilities, and

D1 | Singrauli, Surya Kiran Bhawan Dudhichua PM10 (㎍/m3)

A	B	C	D
4612	18-02-2023 00:45	18-02-2023 01:00	186
4613	18-02-2023 01:00	18-02-2023 01:15	186
4614	18-02-2023 01:15	18-02-2023 01:30	186
4615	18-02-2023 01:30	18-02-2023 01:45	186
4616	18-02-2023 01:45	18-02-2023 02:00	NA
4617	18-02-2023 02:00	18-02-2023 02:15	NA
4618	18-02-2023 02:15	18-02-2023 02:30	NA
4619	18-02-2023 02:30	18-02-2023 02:45	NA
4620	18-02-2023 02:45	18-02-2023 03:00	171
4621	18-02-2023 03:00	18-02-2023 03:15	171
4622	18-02-2023 03:15	18-02-2023 03:30	171

Figure 1: PM10 data column from the dataset at every 15 minutes of interval.

haul roads used to transport coal, as well as close to drilling sites, in overburden, and on such haul roads. Air pollution [1] measurements available via multi-sensory system are PM10, PM2.5, SO2, NO2, NOx, CO, NH3, O3 and BENZENE.

Due to reasons like sensor failure, sensor-to-central-hub communication link failure, data packet loss etc., there will be some missing sensory data for certain duration of the time. Entire data is arranged in the tabular form as shown in Fig. 1. Entire sensory array link failure renders missing values in entire rows, whereas individual sensor mishap causes few entries missing from the column. Some missing data entries are also shown in Fig. 1 via `NA` letters.

How can you plot the time-series? Make a clever use of various graph plotting techniques to plot multi-variate time-series. How `NA` values are interfering with plotting? Can we just replace `NA` with 0 values? What is better strategy? Can we establish ARMA/ARIMA process for the dataset? Should it be per-row or per-column? Try to set the process for some columns and plot the corresponding time-series.

Can the below interpolation useful as compared to ARMA/ARIMA processes? Show by descriptive statistics (e.g. plotting various graphs and jointly comparing them).

Interpolation [2]

By cutting a slick curve through the time(t_i) $i = 1, 2, 3, ...$, we are able to estimate PM10 and other column of data for any given time(t). Interpolation is used when the intended time(t) falls between the greatest and smallest of the time

(a) Linear Interpolation:

This is basically like connecting two points in a dataset by drawing a line between them.

(b) Cubical Interpolation: It offers true continuity between the segments. As such it requires more than just the two endpoints of the segment but also the two points on either side of them.

(c) Spline Interpolation:

Low-degree polynomials are used in each of the intervals in spline interpolation, which is similar to polynomial interpolation in that it selects the polynomial parts to fit together smoothly. The outcome is a function known as a spline.

2 Statistical inference

A particular method of analysing a set of air pollution data points gathered over a period of time is called a 'time series analysis.' Instead of just capturing the data points intermittently or arbitrarily, time series analyzers record the data points at regular intervals over a predetermined length of time. Blasting time in coal India is 13:45 pm to 14:45 pm major effect on air pollution. see figure [2].

Can you validate this information from actual observed data? Can you derive combined weighted combination of air polluting factors to obtain a single time-series data, which should capture the pollution effect of blasting? How can you detect the blasting time from this time-series? Can you plot the histogram of this blast trigger times across all months of data. What kind of distribution it is following? Can you infer from QQ plot whether is it Normal distribution or not? Can you find the probability of open-pit blast happening during 14:15 to 14:30?

3 Problem setting and prediction

You have to setup your own problem statement for analyzing the information from data. It should follow minimum 3 of the below statistical inference techniques.

- (a) Classification: time series data classify mainly two categories (1) Stock time series data refers to measuring characteristics at a specific moment, much like a static image of the data as it was. (2) Flow time series data means measuring the activity of the attributes over a certain period, which is generally part of the total whole and makes up a portion of the results. It identifies and assigns categories to the air pollution data?

- (b) Curve fitting: Plots the air pollution data along a curve to study the relationships of variables within the data. Plots the all air pollution data on same curve? Also explore non-parametric curve fitting or fitting data via parametric distributions instead of deterministic curves.

- (c) Descriptive analysis: Identifies patterns in time series data at the time of coal India open-pit blasting effect, in coal india blasting effect time is 13:45 pm to 14:45 pm , find trends like cycles, or seasonal variation. Do Descriptive analysis can be categorized into four types which are measures of frequency, central tendency, dispersion or variation, and position of air pollution data?

- (d) Explanatory analysis: Attempts to understand the air pollution data and the relationships within it, as well as cause and effect air pollution coal India at the time of blasting?

- (e) Exploratory analysis: Highlights the main characteristics of the time series air pollution, usually in a visual format.

- (f) Forecasting: Predicts future data. This type is based on historical trends of air pollution data set. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points. Analyse the time series methods used for forecasting are Autoregression (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA)?

- (g) Intervention analysis: Intervention analysis in time series refers to the analysis of how the mean level of a series changes after an intervention.

- (h) Segmentation: Splits the data into segments before and after the time of blasting to show the underlying properties of the source information. It is like categorizing the data into various unifying parts for appropriate inference.

4 Appendix

Introduction about air pollution data-set
The air pollution data set is obtained from the Singrauli Coalfield Pollution Control Board for coal India's (Singrauli Coalfield). The pollution is monitored during open-pit blasting.

There are 13 columns overall in the air pollution data collection of pollutants that are available at intervals of 15 minutes. Coal India open-pit blasting effects on air pollution at time 13:45pm to 14:25pm.

- column (A) Indicates serial no of the data set.

- column (B) and (C) Indicates date and time from and to for 15 minutes of interval.

- column (D) Indicates PM10 pollutant of the data.

- column (E) Indicates PM2.5 pollutant of the data.

- column (F) Indicates NO pollutant of the data.

- column (G) Indicates NO2 pollutant of the data.

- column (H) Indicates NOX of the data.

- column (I) Indicates CO pollutant of the data.

- column (J) Indicates SO2 pollutant of the data.

- column (K) Indicates NH3 pollutant of the data.

- column (L) Indicates Ozone pollutant of the data.

- column (M) Indicates Benzene pollutant of the data.

■