MacroX
Know Faster, Invest Better

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

Georgia Tech Practicum 2022.
Project #1 Air quality and Economic Progress in India

The goal of this project is to conduct a time series and cross-sectional analysis of the changes in air quality vs. changes in economic outcomes in more than 50 cities across India. This work supports our larger agenda of helping researchers model trade-offs involved in achieving various sustainable development goals (SDG).

The project consists of four major steps

I. Initial Step

Getting high-quality time-series data for big polluting gases such as NO2, SO2, CO, O3, Black Carbon, and Particulate matter (2.5, and 10) is not easy. Openweathermap.org is an excellent API that provides these and other data from Nov 2020 onwards. However, to conduct in-depth social research, one typically needs longer data. Next, we will discuss how to backfill the time series by estimating emissions history.

1. In this step – you can use the data provided for 54 Indian cities in the google cloud storage (GCS) bucket, plot it and get a feel for it.
2. **<Here>** is the initial onboarding document with instructions on accessing the data as well as starting and stopping VMs. **Before you start and stop VMs, please do drop a slack message to our engineering team and explicitly get permission.**
3. *Bonus: If you find other trustworthy APIs with longer time series, please do let us know.*

II. Estimating Emissions History

In this step, you will treat Openweathermap data as a "source of truth." You will use the official Indian govt. data, satellite estimates, and traffic data to generate an estimate of the Openweathermap data from 2018 to 2020.

1. Estimate the initial models for the relationship between Openweathermap and other data (govt., Satellite, Traffic) from 2020 to 2021. Keep 2022 for validation. These are daily data so you should have a reasonable sample size.
2. To estimate the initial relationship

Openweathermap ~ Govt. Data + Satellite Data + Traffic + Other factors + E (t) [1]

you can use any ML technique you like. ElasticNet, as well as Random Forest and XGBoost in particular, may be useful.

a. Do a good write-up and justify your model selection.
b. You can also create a pseudo air quality index and check the results against AQI data.

3. Using the same relationship generate your estimates for the Openweathermap data from 2018 to 2020. In other words, generate the estimate:

   Openweathermap* ~ f(Govt. Data + Satellite Data + Traffic + Other factors) [2]

   for the years 2018-2020 where these data do not exist.
4. **Cross Section.** Some cities like Delhi have 40 stations which help cover a lot of areas in the city. Estimate the *with-in-city* pollution variation based on reasonable factors such as population density, factory locations, etc.
   a. **Check** if other cities that do not have as many stations would benefit from having more stations.
   b. **In your opinion** does the API estimate for other cities appear to capture the intra-city variation? Check out an article by MSR researchers on finding more granular "hotspots" within Chicago.
5. **Traffic and Other Factors**
   a. Google mobility data can proxy for traffic. Check the "Resource" part at the end of document.
   b. *Extra: You can find other sources of traffic and congestion and even think of other factors. Please document the data sources and WHY you think these factors are reasonable.*
6. **Missing Data**
   a. Some of the X variables will have a large number of missing values – such as Satellite data on cloud days. You will have to think about reasonable (and defensible) ways of filling it in.
   b. Map missing data using packages such as Naniar and think about whether these data are missing at random or if there is a pattern to this.
   c. These missing data are the reason we are doing the project and it is so valuable.

III. Changes in Emissions vs. Economic Growth

In this step, you will use the various cities' unemployment rates time series provided here as a reasonable proxy for economic growth. Correlate the changes in air quality to the changes in the economy at the city level.

1. Estimate the
   a. Change in Unemployment Rate ~ Changes in Emissions (NO2, SO2, Particulate matter, Black Carbon, etc.). [3]
   b. Does any variable matter in particular?

c. You will have to decide the appropriate way to aggregate or summarize the emissions data to match the monthly economic data.
d. Does the change in emissions of any particular gas or emission correlate more with economic changes?
e. *Extra: You can check if a "fixed effects" model or "random effects" model for the economic and emission relationship by city works best*
f. *Extra: You can think about the "health costs or impact" vs. the "economic changes" for each gas.*
g. *You can attempt to find other economic metrics at the city level that may be interesting*

2. Classify cities in a 2 by 2 matrix for each quarter – positive and negative economic changes vs. positive and negative emissions.
    a. *Extra: Can you predict city rankings over time? Feel free to use any ML model.*
3. COVID Analysis: What was the impact of COVID on the air quality as well as the economy? Which cities were the most resilient?

IV. Dynamic Visualization

In this step, you will create a code-based interactive "dashboard."

1) You can use Shiny, Dash, or any other code-based and interactive visualization to help citizens and scientists understand your findings.
2) Create an accompanying academic style paper in latex on overleaf.com or Quarto.org. Feel free to use any top-conference – such as ICWSM - templates.

Resources

1. **GitHub Repository**
2. **GCS Bucket for data**
    a. **Satellite Emissions Data**
    b. **Economic Data**
    c. Openweathermap **Vendor API data**
    d. **Mobility: Google dataset location for your queries**
        i. **[bigquery-public-data.covid19_google_mobility.mobility_report]**
        You can get the dataset using a simple query like:

```
ds <- bq_dataset("publicdata", "covid19_google_mobility.mobility_report")
tb <- bq_dataset_query(ds,
  query = "SELECT *
           FROM `bigquery-public data.covid19_google_mobility.mobility_report`
   WHERE country_region = 'India'
   ORDER BY date ",
  billing = "YOUR PROJECT")
```

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

Note that it is for national mobility. You can change it to appropriate cities.

ii. **Please feel free to find other data available publicly.**

e. **Govt. Data** The excel file shows the # stations for each of the 54 Indian cities as well which gases have been downloaded.

About MacroXStudio

We are a growth stage fintech startup and use lots of alternative data and ML.
At MacroXStudio, our goal is to measure the *entire* world economy faster and better than anyone else and use that information to generate superior returns for investors as well as help society. Our cloud-native platform uses many alternative data sources (like satellite, search, twitter, credit card, supply chain, etc.) and trained machine learning models to measure the entire world economy in real-time at the city level – a technique known as nowcasting. The platform is currently live for 50 cities and 3 countries. We are now scaling the platform all over the world. We are currently in semi-stealth mode and expect to be more public later this year.

Know faster. MacroX's real-time information or nowcast is 1 to 3 months ahead of the government data, since a typical government growth estimate – for instance for the first quarter ending in March 2022, would only be released in May or June 2022. Additionally, our city-level nowcast is more granular than the state-level data provided by the government. In developing countries – such as the ones in Sub-Saharan Africa, the current government data may be too low quality to be reliable. In contrast, our accurate and real-time information is commercially and socially valuable.

Invest better. Using our platform, we could track the COVID recession and recovery's economic impact in real-time. Such faster information can be extremely useful in doing tactical asset allocation or trading – in face of the coming recession, it may lead us to trim the stock portfolio right away rather than wait for 2 to 3 months, and as the positive effects of the stimulus kick in, identify the bounce-back as a buying opportunity.

Social aspect. For us, investments are not just financial. We will open-source parts of our toolkit to enable the community to research alternative data, and we will also contribute to and maintain datasets for real-time UN sustainable development goals (SDG) metrics. We typically spend 20% of our time contributing to this aspect. Do something for the SDG you care about.