MacroX
Know Faster, Invest Better

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

Project: 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 deeply understand the trade-offs that achieving various sustainable development goals (SDG) involve.

The project consists of four major steps

I. Initial Step

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

(i) In this step – you can use the data provided for 54 Indian cities in the GCS bucket, plot it and get a feel for it.

(ii) <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.**

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, NASA estimates, and traffic data to generate an estimate of the Openweathermap data from 2018 to 2020.

(i) We suggest estimating the initial models for the relationship between OpenweatherMap and other data (govt., Nasa, Traffic) from 2020 to 2021 and keeping 2022 for validation. These are daily data so you should have a reasonable sample size.

(ii) To estimate the initial relationship Openweathermap ~ Govt. Data + NASA Data + Traffic + Other factors + E (t) you can use any ML technique you like. We suggest ElasticNet, as well as Random Forest and XGBoost in particular.

a. Do a good write-up and justify your model selection.

b. You can also create a pseudo airquality index and check the results again AQI data.

(iii) Using the same relationship generate your estimates for Openweather map from 2018 to 2020.

(iv) Missing Data

a. Some of the X variables will have a large number of missing values – such as NASA data on cloud days. You will have to think about reasonable (and defensible) ways of filling it in.
b. We also suggesting mapping missing data using Naniar package and having an idea of whether these data are missing at random or 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. You will now correlate the changes in air quality to the changes in the economy at the city level.

(i) Estimate the Change in Unemployment Rate ~ Changes in Emissions (NO2, SO2, Particulate matter, Black Carbon etc.). Does any variable matter in particular?
   a. Extra: You can check if a "fixed effects" model or "random effects model" works best

(ii) Classify cities in a 2 by 2 matrix for each quarter - +ve and -ve economic changes vs. +ve and -ve emissions.
   a. Extra: can you predict city rankings? Feel free to use any ML model.

(iii) 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 "dashboard" – we suggest using Shiny or Dash to illustrate your findings and you can create an academic style paper to accompany it (one per group) in latex on overleaf.com or Quarto.org

Resources

(i) GCS Bucket for data
(ii) Onboarding, Git links etc.
(iii) Academic References
(iv) Jupyter Notebook for fetching and plotting Indian Govt. Data, API data, as well as NASA data

You will especially focus on the COVID period as a natural experiment and observe which states had the most change in air quality with the least change in economic progress.

Idea

MacroX
Know Faster, Invest Better

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

Project: Air Quality vs. Economic Progress at the city level in India

AIR QUALITY DATA

Step 0: Openweathermaps. Com API is source of truth but only available from 2020.

- According to us other APIs have data from 2022
- If you find better APIs – please tell us
- Other air quality indices may have relevant information, however they tend to be composite. You can also use the overall composite air quality as a check for your estimates. [AQI]
  [SOURABH provides top relevant link]

Step 1: Backfill from 2018-2020 API [ML]
API_hat ~ Govt + Traffic + Other
Justify your approach.

- Which technique for missing values?
- Which model to relate other factors and NASA to govt. data?
- Which other factors (be creative in finding other factors)

Use the API values after 2020 as source of truth and hence a Holdout sample for your model.
Govt. data is heterogenous [only provides cities which have data from 2018?]

- Traffic API?
- [mobility for sure]
- Some scaling libraries and information [SOURABH – conversion factor AND AJINKYA for scaling functions]

Step 2: [ML/Stats]
Estimate
Econ Changes (provided) ~ Airquality by City (state proxy for city econ)
City vs. State mapping [AJINKYA]
"City-level" unemployment series (MacroX) (each city in the same state will have the same time series)

AIR QUALITY ~ ECON RELATIONSHIP

Step 3
Report and Visualize

(i) Create a live viz – we suggest dash or shiny
(ii) Classify by dEcon progress vs. dAirquality
  a. Which cities are doing the best (+ve econ progress and +ve air quality) vs. worst

b. **What do you think is responsible? (feel free to support your thesis with diff data or reasoning)**
c. **Can you predict city ranking each year? (any ML technique is fine)**
d. **How was COVID?**

Econ Activity ~ Air Quality Total (SO2, NO, PMI levels). (i)

Econ Activity ~ Non-Vehicular Pollution (i A)

Create live tracker.
What is the cross-sectional picture?
Which Indian cities did best and which ones did worst during COVID?

Data Sources

Economic Data
India State Unemployment rate
This will act as a proxy for each city's unemployment rate change since most economic activity is concentrated in cities.
You can also map additional indicators like google mobility for workplace and see how they relate to the unemployment rate.
You can also find other indicators from the World Bank or SDG goals here

Air Quality
NASA
API
Official

Offical_AirQuality_ est ~ Official + NASA + API + other sources (ii)

Official_AirQuality_Vehicle_Est ~ mobility, traffic, congestion (iii)

The official datasets are not clean and you will spend most of the time cleaning the data and justifying your choices.
NASA data has many missing values as do many government data.

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

You can follow some of the work done by our engineers to clean data for all cities and match the govt. data time series to API>
You can also use the NASA data.

You can use ML techniques to fill in the data and justify your choices.
Some ML packages -

India City coordinates and states
Table here

Questions for AD and SS

1) Can we find more X variables related to air quality
    a. google mobility, traffic congestion,
2) City

The goal of this project is to map emissions levels as measured by Satellite data in 100 cities vs. the actual on-ground measurement. Then we will see how these pollution levels are changing vs. WHO and UN Norm and with economic progress.

You will loosely follow this paper in Atmospheric Environment by Just et al. on applying machine learning to evaluate spatiotemporal models for fine particulate matter using satellite data.

MacroXStudio will provide the satellite data and the latitude and longitude and the bounding box coordinates for the relevant cities. You will find the relevant on-ground data for the 100 cities across by the local meteorological agencies.

You will use various ML techniques like XGBoost, Random Forest, Elastic net etc. to help understand which features reduce the RMSE in mapping the noisy satellite measurement to the more accurate on-ground measurement.

Team

1. We strongly suggest a team of 3 to 4 people. It will help if at least one person has interest in meteorology or has experience reading academic papers. If you have never read an academic paper – it is fine but you should definitely have interest in the environment and machine learning.

2. It helps to clarify upfront who will handle administrative things like organizing regular meetings etc.

Meetings

1. You will meet on a bi-weekly basis with the team over zoom. We will also try to make at least one on-campus appearance.
2. You will be assigned a slack channel and can ask questions and receive answers there.

Templates and Engineering Support

1. MacroXStudio will provide basic templates for how we want the code to be written, as well as how to professionally document your research process in a Markdown file.
2. MacroXStudio will provide the cloud computing resources as well as handle the complex engineering aspects of Airflow scheduling etc.

Deliverables:

1. **Scripts**
    a. Fit the ML models to the data.
    b. To create the table that will host the data. Think what schema you will use to support regular updates.
    c. regular upload script.
2. **Relevant Visualizations** We really care about informative plots since that communicates information to the rest of the team and helps drive insight generation.
3. **Short write up and slides. Make sure to cover**
    a. **Outline** of the initial data quality check
    b. **Time series models.**
    c. Seasonal Adjustment libraries used.

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 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.

MacroX
Know Faster, Invest Better

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

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 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 do research on alternative data, and will also contribute 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.

Work with world-class experts. You will be guided by leadership with work experience at firms such as MSFT, Deutsche Bank and Bridgewater, and research at schools such as HBS. Access our investors and advisors at senior finance and technology positions such as the president of RStudio and a global head of PE, as well as ones with Unicorn exits. Check out a talk by our CEO on applying AI to asset management. **We believe in publishing research in top journals and conferences and have access to many unique data sets.**

Modern tech-stack. You will be expected and empowered to set our research agenda and evolve our technology stack as we scale. You will program in python primarily (and some R), work on the latest tech and ML stack – GCP, data lakehouses, Airflow, Kubernetes, Redis, Data Version Control etc. You will leverage the latest open-source software and help develop web-based, mobile friendly, low-code AI applications for the platform.

Important Dates
Student NDA – July 13 [Can send this now].
Data available – July 25.
[3-4 person] On-campus practicums are always available.
Class begins – Aug 22. Registrar.

Revised Idea

MacroX
Know Faster, Invest Better

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

Replicate and create the air quality metrics for cities.
Chart economic progress vs. social goals and narratives.
For now – air quality vs. economic output for cities. Which cities do well?

Technical Appendix

Basic idea
Y_actual ~ X_satellite + Other variables.

Y_actual	X_satellite	Other Variables	Comment
US Cities- EPA AQS.			
India **API here** **World EPAs (map below)** **SAFAR (Mumbai, Delhi, Pune, Ahmedabad)**			Most comprehensive world-wide source is aqicn.org it seems. Combines 8,000 other EPAs Private start-ups and concerns have sprung up. Collaborative and commercial here. More data and links here

MacroX
Know Faster, Invest Better

MacroXStudio Inc. 981 Mission Street, San Francisco, 94103

https://aqicn.org/api/

Papers on Indian Emissions.
NOX via Satellite. [2008]
Emissions ~ seasonal variation in boundary layer, meteorological conditions, biomass burning seasonality.

Air Pollution @ mega cities.
Objective evaluation of stubble emission of North India and quantifying its impact on air quality of Delhi Beig et al.

Data @ SAFAR + World Meteorological organization.

Particulate pollution ~ meteorology + transport
The relative share of $PM_{2.5}$ emissions by different sectors for Delhi National capital Region is discussed elsewhere (Singh, 2014). **Transport sector is the most dominant sector contributing 39.1%** in total $PM_{2.5}$ emission. Industrial sector is the second most dominating factor contributing 22.3%. The contributions from power sector, biofuel sector, resuspended dust and others are found to be 3%,5.7%, 18% and 11.7% respectively. The model results under SAFAR project were also routinely validated for Delhi region for normal case as well as for extreme events (Beig et al., 2019; Marrapu et al., 2014; Srinivas et al., 2016b