Open Data and Air Quality Toolkit

Prepared by:

AIR SOUTH ASIA

Clean up Nepal

Supported by:

UKaid from the British people	Data for Development in Nepal	development initiatives	The Asia Foundation

August 2018

What is this toolkit and who is it intended for?

This toolkit is intended as a resource for people interested in working on air pollution and want to utilize open data. The toolkit provides a list of resources and examples from different parts of the world. Sections of the toolkit are taken from different sources, and attributed in individual sections. This toolkit was developed with support from the Data for Development Program implemented by The Asia Foundation and Development Initiatives, and supported by DFID. For further information, please contact Pallavi Pant (pallavipnt@gmail.com).

TABLE OF CONTENTS

1. What is open data and why does it matter?

Open data is defined as data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike[1]. Data must be both *legally* and *technically* open, i.e., data must be placed in the public domain with minimal restrictions, and be available in machine-readable formats so that data can be used with freely available software. Additionally, when talking about open data, the focus is on data that is non-personal and identifiable.

Key features of open data include:

- ***Availability and access***: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.

- ***Reuse and redistribution***: the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.

- ***Universal participation***: everyone must be able to use, reuse and redistribute — there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

Open data is truly useful only if it is used for other products and services, and therefore, it is important to build . It is interesting to consider that if data is open, it can lead to several benefits including better services and stimulation for innovation, as well as provide societal benefits.

There are many good examples of open data being used in a variety of applications, and select examples are listed below:

[1] http://opendatahandbook.org/guide/en/what-is-open-data/

- The Open Streets Map (OSM) is the map of the world created by people and free to use under an open license. OSM is used by scientists, civil society groups and others everyday for a variety of geospatial applications.

- In the UK, open data on transport from Transport for London has led to a variety of applications including applications for journey planning, routes and fares and disruptions in the system.
- In India, an open data initiative (Open Works) on city budgets and operations in Bengaluru, Karnataka is trying to catalyze citizen participation in governance, and push for transparency in decision-making. (take a look at the mapped data)
- In the context of air quality, AIR Louisville in USA brings together the city's public health department and a health management company and a research institute, and using a smart inhaler which combined with weather and traffic data is helping in identification of trigger points allowing for better management of air quality in the city.

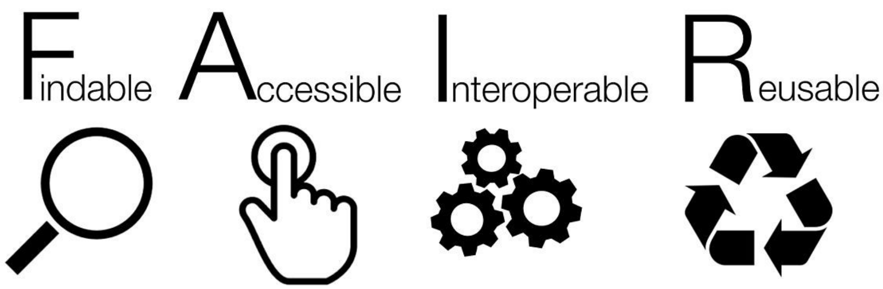


Image: https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

"Government-collected data, when shared openly and in a timely manner, can bridge this gap by galvanizing communities, guiding effective data-driven mitigation policies and external, independent analyses, and helping answer remaining public health and economic policy-relevant science questions. It can also open up entirely new economic potentials, similar to how access to weather data spurred public and private innovations a few decades ago."

- Hasenkopf et al.,2016, Clean Air Journal

An overview from Dr. Christa Hasenkopf, OpenAQ on the role of open data in air pollution: download slides

Credits

This section of the toolkit is derived from the Open Government Data Toolkit.

Additional Resources

Open Data Toolkit- http://opendatatoolkit.worldbank.org/en/essentials.html

- This toolkit also includes training modules for data producers as well as users.

Open Data Handbook- http://opendatahandbook.org

2. Air Quality Basics

Ambient (outdoor) as well as indoor air pollution are key environmental health risk factors contributing to the disease burden. Sources of air pollution include transportation, industries, energy production and use, construction, dust, brick kilns and waste burning among others. Air pollutants can either be released directly into the atmosphere (primary emissions) or can form as a result of chemical interactions (secondary pollutants).

Key air pollutants of interest are described below:

(i) Carbon Monoxide

Carbon monoxide (CO) a colorless and odorless gas, which at high levels can be harmful to humans by impairing the amount of oxygen transported in the bloodstream to critical organs. Although high concentrations of CO are more of a concern indoors, emissions outdoors, particularly in developing countries can be high. New evidence also reveals that long-term exposure to low concentrations is also associated with a wide range of health effects. The main sources of ambient CO include motor vehicle exhaust and machinery that burn fossil fuels.

(ii) Nitrogen dioxide

Nitrogen dioxide, mainly emitted by power generation, industrial and traffic sources, is an important constituent of particulate matter and ozone. There is growing evidence that independently, it can increase symptoms of bronchitis and asthma, as well as lead to respiratory infections and reduced lung function and growth. Evidence also suggests that NO2 may be responsible for a large disease burden, with exposure linked to premature mortality and morbidity from cardiovascular and respiratory diseases.

(iii) Ozone

Ozone is a gas made up of 3 oxygen atoms; it has the chemical formula O3. Ozone found at ground level, where people live and breathe, is formed by chemical reactions between nitrogen oxides and volatile organic compounds in the presence of sunlight. For this project, ozone levels are measured in units of parts per billion (ppb) by volume. Ozone concentrations, averaged over the summer season in each region when ozone levels tend to be highest, are used to represent the exposures experienced by human populations in those regions.

(iv) Particulate Matter (PM)

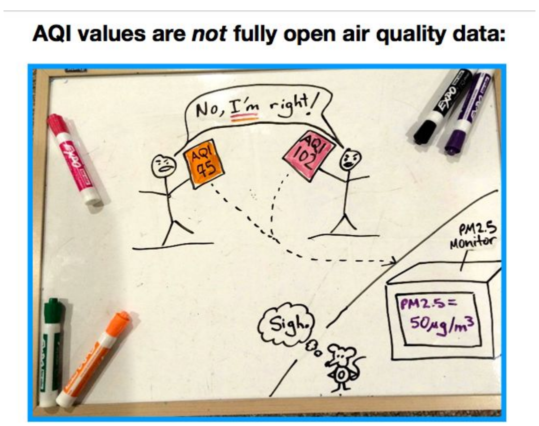
Particulate matter (PM) are inhalable and respirable particles composed of sulphate, nitrates, ammonia, sodium chloride, black carbon, mineral dust and water. Particles with a diameter of less than 10 microns (PM10), including fine particles less than 2.5 microns (PM2.5) pose the greatest risks to health, as they are capable of penetrating peoples' lungs and entering their bloodstream. Sources of PM include combustion engines (both diesel and petrol), solid-fuel (coal, lignite, heavy oil and biomass) combustion for energy production in households and industry, as well as other industrial activities (building, mining, manufacture of cement, ceramic and bricks, and smelting).

2.1 Impacts of air pollution

There is a large body of research on the impact of air pollutants on human health and studies have conclusively shown that long-term exposure to PM increases mortality and morbidity and can lead to respiratory as well as cardiovascular diseases.

2.2 Air Quality Index

The Air Quality Index (AQI) is an index for reporting daily air quality. It tells how clean or polluted the air is, and what associated health effects might be a concern for people. The AQI focuses on health effects people may experience within a few hours or days after breathing polluted air. Countries and regions often calculate AQI with different metrics since their standards are different, rendering it difficult to compare AQIs across platforms, even for data from the same stations.



Credit: OpenAQ (link)

Typically, the AQI values are roughly in the following range:

Index Value	Name	Color	Advisory
0 to 50	Good	Green	None
51 to 100	Moderate	Yellow	Unusually sensitive individuals should consider limiting prolonged outdoor exertion
101 to 150	Unhealthy for Sensitive Groups	Orange	Children, active adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion
151 to 200	Unhealthy	Red	Children, active adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else should limit prolonged outdoor exertion
201 to 300	Very Unhealthy	Purple	Children, active adults, and people with respiratory disease, such as asthma, should avoid outdoor exertion; everyone else should limit outdoor exertion
301-500	Hazardous	Maroon	Everyone should avoid all physical activity outdoors.

☛ **What is the difference between AQI and pollutant concentrations?**

2.3 Measurement and monitoring of air pollution

Fixed-site (ambient) monitoring: Such monitoring is often conducted as part of long-term monitoring networks focused on measurement of air pollutants. Air quality instruments are operated at a particular site over long intervals, and data can be used for understanding temporal patterns and undertaking trend analysis. There are a number of monitor types including manual instruments and real-time air quality monitoring equipment. Costs can vary based on the type of technology being used in the system. Location of air quality monitors depends on the purpose of the monitoring program, and typically, monitoring is conducted across a range of sites including urban, rural and source-impacted sites (e.g., roadside, industrial, commercial etc.).

‘Each of London’s 33 boroughs has a minimum of one and up to half a dozen analysers,’ he explains. These monitoring sites form part of the London Air Quality Network (LAQN) and vary in size from *one the size of a post box* on Oxford Street to the *huge shipping container* opposite Madame Tussauds on Marylebone road. Some of the sites measure a single pollutant, such as NO2 or particulate matter (PM), in a location where boroughs are keen to closely monitor air pollution, and others measure the entire range of legislated air pollutants. The large site on Marylebone road and another in South Kensington also take additional measurements solely for research purposes.

Link to article

Low-cost Sensors: Low-cost sensors offer an opportunity to generate high-resolution data at a lower cost, and with fewer deployment and access limitations (Snyder et al. 2013, EST). A number of community projects have been launched worldwide for crowd-funded air pollution measurements, and while such sensors can be useful in designing citizen science projects and generating novel data, there are still a number of uncertainties associated with the accuracy of measurements using low-cost sensors (Lewis and Edwards 2016, Nature). So far, such monitors have not been proven to provide long-term, accurate data without systematic calibration (Lewis and Edwards 2016, Nature; Rai et al. 2017, STOTEN). Efforts are underway to improve precision among such sensors, and latest analyses are supporting the case for deployment of well-designed low-cost sensors for measurement of air pollution at the city level. If designed carefully, such networks can provide valuable data on spatial variability of pollutants, and help in identification of hyperlocal pollution hotspots. Additionally, such instruments can be used as indicative instruments to define and design more efficient regulatory monitoring networks.

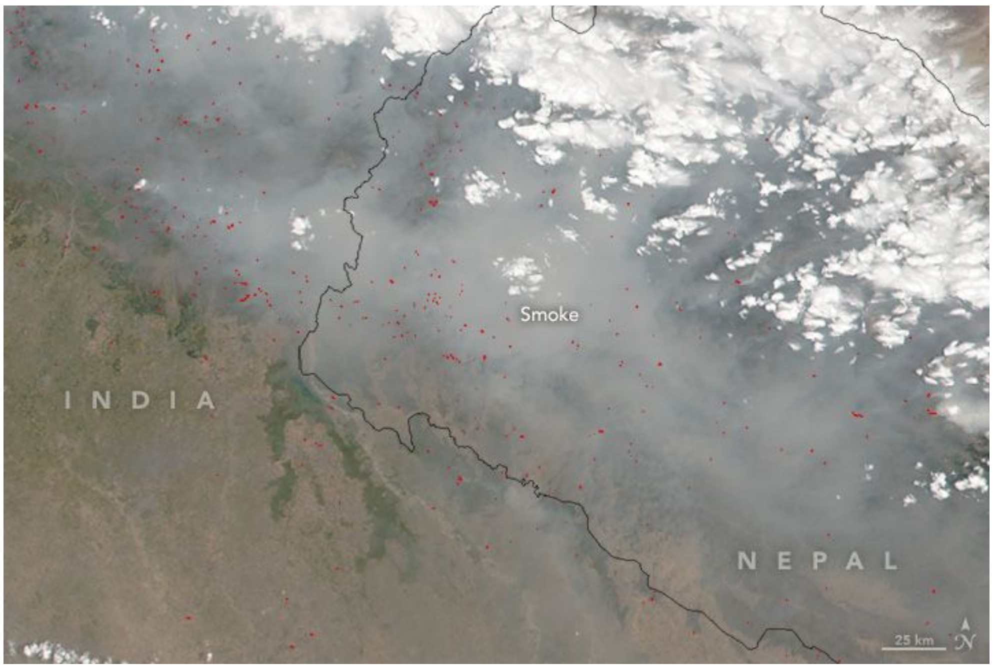
Low-cost sensors are currently available for a range of air pollutants including particulate matter (PM), and gases (E.g., nitrogen oxides, ozone and carbon monoxide), and the technology is rapidly evolving leading to improvements at a very rapid pace.

Satellite Monitoring: Data on aerosol optical depth (AOD) is typically collected via satellite overpass measurements (e.g., MODIS- https://modis.gsfc.nasa.gov/) and can be used for AQ applications such as forecasting, tracking pollution sources and plumes, and as input/evaluation for AQ models (Duncan et al. 2014, AE).

The most common use of satellite data has been the use of aerosol optical depth (AOD) data which can be used for estimation of PM concentrations. AOD is a dimensionless number that is related to the amount of aerosol in the vertical column of atmosphere over the observation location, and indicates how much direct sunlight is prevented from reaching the ground by these aerosol particles (more about AOD here). A value of 0.01 corresponds to an extremely clean atmosphere, and a value of 0.4 would correspond to a very hazy condition.

NASA WorldView , an open source project, provides near-real-time data on a daily basis, well suited to event analysis, in a clean, intuitive interface and all data, software and services are freely available for public use.

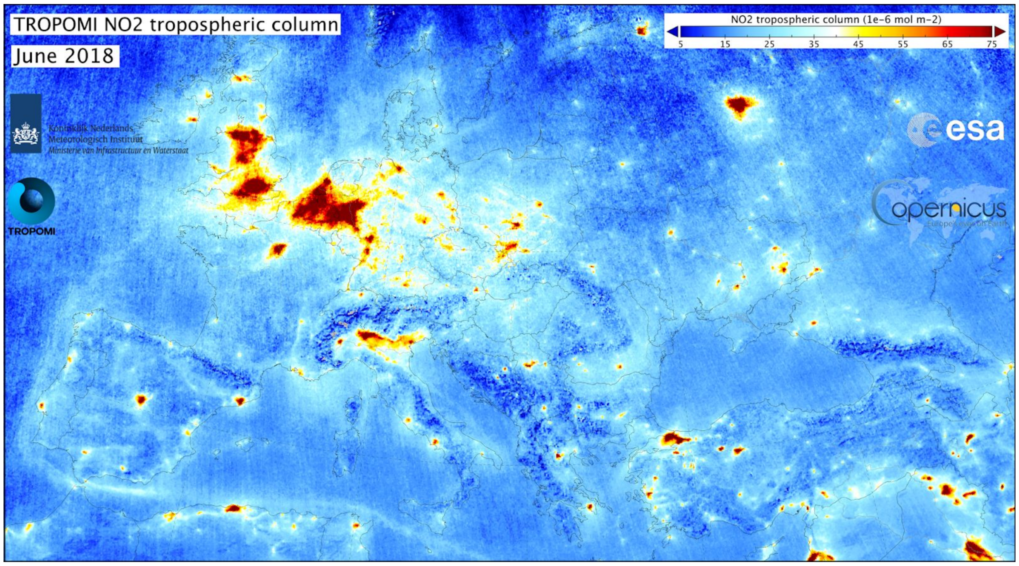
There are three main ways to use satellite data for policy applications: for qualitative applications (e.g., to detect trends in air pollution or understand spatial patterns), for quantitative applications (to quantify changes over time), and for more advanced analysis (e.g., to derive ground-level air pollution concentrations which can then be used for modelling and health studies). Satellite data have been used for wide-ranging studies, including the estimation of health effects related to air pollution as part of the Global Burden of Disease study, and for detection of agricultural burning and wildfires in India(example 1 and 2).



Smoke and fires in Nepal on April 28, 2016 (link) [*Red outlines indicate hot spots where VIIRS detected warm surface temperatures associated with the fires*.]

MODIS (Moderate Resolution Imaging Spectroradiometer) is a key instrument aboard the Terra (originally known as EOS AM-1) and Aqua (originally known as EOS PM-1) satellites. Terra MODIS and Aqua MODIS are viewing the entire Earth's surface every 1 to 2 days, acquiring data in 36 spectral bands, or groups of wavelengths (see MODIS Technical Specifications). These data will improve our understanding of global dynamics and processes occurring on the land, in the oceans, and in the lower atmosphere. MODIS is playing a vital role in the development of validated, global, interactive Earth system models able to predict global change accurately enough to assist policy makers in making sound decisions concerning the protection of our environment.

Sentinel 5-P was launched in October 2017 and will perform atmospheric monitoring. The Tropomi instrument on board is the most advanced multispectral imaging spectrometer to date, and this allows collection of highly detailed and accurate atmospheric data.



Example of data from TROPOMI (link)

Visible Infrared Imaging Radiometer Suite (VIIRS) collects visible and infrared imagery and radiometric measurements of the land, atmosphere, cryosphere, and oceans. The data is used to measure cloud and aerosol properties, ocean color, sea and land surface temperature, ice motion and temperature, fires, and Earth's albedo.

Let's check out an animation for AOD data over the years.

☛ How many fixed-site monitors are operating in the Kathmandu Valley?
☛ Are there any low-cost monitor networks in Nepal?

2.4 Source Apportionment of Air Pollutants

The term, source apportionment (SA) describes techniques used to quantify the contribution of different sources to atmospheric particulate matter (PM) concentrations. Source apportionment techniques are used widely for quantitative estimation of the contribution of different sources to ambient PM concentrations and can be implemented in many different ways, receptor modelling being one of the methods. There are three key approaches for source apportionment including use of emission inventories and dispersion models, receptor models and monitoring data. Essentially, receptor models (RMs) form a subset of source apportionment techniques and apportion the pollutant concentrations based on the measured ambient air data and the knowledge about composition of the contributing sources.

2.4.1 Receptor Models

Receptor models (RMs) form a subset of source apportionment techniques and apportion the pollutant concentrations based on the measured ambient air data and the knowledge about composition of the contributing sources. Such models provide relevant information for development of air pollution management and control programs, validation of dispersion models and are particularly helpful in cases where complete emissions inventories are not available. With the assumption that the relative concentrations of chemical species are preserved between sources and receptors, receptor models use the principle of mass conservation for apportionment of PM mass to different air pollution sources. Chemical receptor modelling methods utilize the chemical composition of airborne particles for identification and apportionment of sources of PM in the atmosphere. Each source has a characteristic emission profile, and differences among the profiles can be used for quantitative apportionment of mass to different emission sources.

2.4.2 Dispersion Models

Dispersion modeling uses mathematical formulations to characterize the atmospheric processes that disperse a pollutant emitted by a source. Based on emissions and meteorological inputs, a dispersion model can be used to predict concentrations at selected downwind receptor locations. Commonly used models include AERMOD and CALINE3.

2.5 How to access air quality data?

There are a number of websites, apps and analog sources for accessing air quality data. Popular websites for AQ data include:

- **AQICN**

AQICN curates AQI data from more than 60 countries around the world. AQI information can be accessed here. Currently, data from the US Embassy and Phora Durbar stations is available on the website. However, historical data cannot be downloaded from this website.

#Let’s check the current AQI data for major cities in Africa on the AQICN website.

- **OpenAQ**

OpenAQ fights air inequality through open data, open-source tools, and a global, grassroots community. *Because data need a collaborative community for impact.* OpenAQ’s community has collected air quality measurements from 8,613 locations in 67 countries, and data are aggregated from 110 government level and research-grade sources. OpenAQ's mission is to *enable previously impossible science, impact policy, and empower the public to fight air pollution*. Historical data are available for download and analysis. There is a also a vibrant community which can be useful if you have specific questions or are looking for collaborators.

Let’s download data for an AQ station from Africa from the OpenAQ website.

- **State of Global Air**

In addition to the real-time data, long-term trends can also be useful in determining changing patterns of air quality in a city/region/country. Annual country-level PM2.5 data (and associated mortality/morbidity) data are currently available on the State of Global Air (SoGA) website run by the Health Effects Institute. Data are available for download and analysis.

#Let’s take a look at data for Africa on the SoGA website.

- **Resource Watch**

A World Resources Institute (WRI) product, the ResourceWatch combines a wide range of datasets including fires, power plants, tree cover, urban population and air quality among others.

#Let’s take a look at fire data for Ghana on the website.

Credits

This informational material is from the following websites:
USEPA (link)
NASA MODIS (link)
JPSS VIIRS (link)
HAQAST (link)

Additional Resources

- Primer on Air Quality Management (link)
- Primer on Pollution Source Apportionment (link)
- Air Pollution Monitoring 101 (link)
- *Global AOD data*: AOD data can be downloaded from NASA Earth Observations (NEO) free of cost here. This data comes from Terra/MODIS and Aqua/MODIS.
- *HAQAST (Health and Air Quality Applied Sciences Team):* HAQAST is a collaborative team that works in partnership with public health and air quality agencies to use NASA data and tools for public benefit.
- *NASA ARSET (Applied Remote Sensing Training) program*: In-person and online trainings are offered all year round. Keep an eye out for the next training if you are interested!
- Training module on VIIRS (link)
- Air Sensor Toolbox for Citizen Scientists, Researchers and Developers (link)
- A brief history of air quality sensors

3. Data visualization and analysis

3.1 Plotly

Plotly creates leading open source tools for composing, editing, and sharing interactive data visualization via the Web. A number of online graphing tools are available, in addition to libraries for software such as R, Python and Matlab. Plotly has a graphical user interface for importing and analyzing data and the graphs can be downloaded for future use or embedded (e.g., on a website).

With a free community account, you can build public charts & dashboards and export charts to PNG or JPG versions. Using the Chart Studio, you can create, design and share plots using your data without any coding!

😃 *This is a great way to create data dashboards without the need for advanced coding skills.*

There are other, more intensive methods which are open-source and free, and provide the user considerable freedom in analysis and visualization of data.

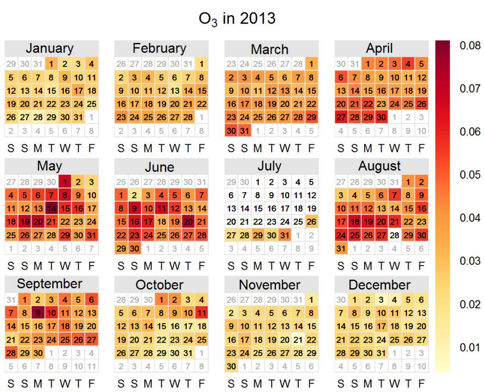
3.2 R

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. R *is available as Free Software* under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

Learn more about R here.

There are a number of packages that cater to analysis of air quality data with openair being the most common.

Openair is an R package primarily developed for the analysis of air pollution measurement data (learn more about the package here and here). Plots developed using openair have been used in a large number of research studies as well as reports and presentations. The openair manual (download here) provides detailed instructions on the use of different functions within the package. A sample exercise set and answer key is available here for practice.



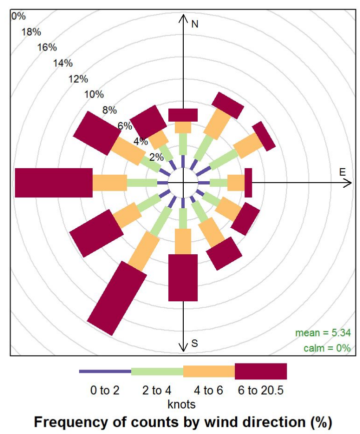


Image Credits: RPubs

3.3 Python

Python is an open-source programming language and is free to use and distribute. Similar to R, Python includes a large number of libraries, several of which are relevant for air quality applications.

The python-aqi 0.5.1 allows for conversion between AQI value and pollutant concentration (µg/m^3 or ppm).

3.4 GIS Mapping

Geographic Information System (GIS) provides a framework to gather, manage, analyze, visualize, and store spatial data. GIS allows us to delve further into spatial patterns, trends, and relations present to solve real world problems and deliver effective solutions. Today almost every field makes use of GIS to make maps to analyze, communicate and share information. GIS can be used to identify problems, monitor changes, manage and respond to events, perform forecasting, set priorities, and understand trends (link). GIS provides a powerful medium to tell stories and communicate effectively to public (examples). Simply, GIS provides a way to connect data with geography.

Today various powerful open source GIS software are available.

3.5 Examples

1. UrbanEmissions.info uses Python and Plotly to visualize data from ambient air quality monitoring stations across India (box plots and time series)
2. In Wales (UK), the government website uses openair for plotting air quality monitoring data (check out the website)
3. Gareth Kennedy used Python to create a movie on air quality in Beijing using a series of images (link to page).

Additional Resources

A number of free resources are available for learning R/Python etc. There are also a number of online communities where you can seek answers (and ideas). A select number of courses and resources are listed below:

- *Plotly*
  - Tutorial on Plotly, link
- *R*
  - Introduction to R (DataCamp), link
  - R Programming (Coursera), link
  - R Basics - R Programming Language Introduction (Udemy), link
  - Tutorial on openair, link

    - Introduction to openair, link
    - R for Data Science, link
- *Python*
    - Introduction to Python for Data Science (DataCamp), link
- *GIS Mapping*
    - QGIS- Learning resource | Guide for use
    - Open Source Geospatial Foundation (OSGeo)
    - GIS for air quality (link)
    - GIS in Public Health research (link)
    - Examples of maps created using QGIS

😃 Stack Exchange is one of the best places to look for answers if you are stuck. Chances are, someone had the same error/issue before you, and an idea/solution/tip might already exist.