Predicting Air Quality Index with Data Science and Machine Learning

GROLLEAU Florian
2A–Dual Diploma
Sciences Po Saint-Germain-en-Laye / CY TECH Engineering School

July 2023

Abstract

Since the massive urbanization and high-density human activities phenomena, such as industrialization and overconsumption of services and goods, the global Air Quality Index (AQI) has regressed. However, the global AQI does not reflect the very localized impact and changes on a rather small geographic area. For this reason, there is a need for more accurate metrics and qualitative analysis of the AQI over smaller regions than a worldwide analysis. Thanks to the emergence of several Data Science tools and Machine Learning models, we aim for better monitoring and prediction of local AQI.

Contents

1 What is AQI?

By the acknowledgement of global climatic changes, the question of the quality of the air, and mainly its troposphere layer, emerged. As a matter of fact, the troposphere represents 90 percent of the total mass of the atmosphere and contains almost the entire volume of water vapor (Olschewski, 2025). Moreover, according to the state of progress of the life sciences, every living being interacts within this critical layer of the atmosphere. For this reason, the quality of the air matters and, in the current position, constitutes a major challenge in the preservation of ecosystems and human societies. Among the several factors which affect air quality, human pollution is considered as the one affecting the most the values of air quality (Beig, 2010). Indeed, "It can pose a serious threat to human health if it exceeds the permissible limit" (WHO, 2000; USEPA, 2008). Every year, over 2 million premature deaths are reported due to air pollution, and "the effect of urban (outdoor/indoor) air pollution is caused by burning of solid fuels" (WHO, 2000, 2002, 2005). As the challenges for a healthier air quality arose, the need for a more rigorous metric designed for scientific monitoring emerged: the AQI. Since the official publication by the AEPA (American Environmental Protection Agency) of the AQI in 1976, its worldwide usage has significantly increased.

1.1 Definition and Purpose

Supporting the WHO air quality guidelines (updated around every 2–3 years), the AQI has been established as a standard for most countries in their aim of monitoring the air quality. Particularly, it has been assessed that "more than half of the air pollution-driven disease burden is borne by the population of developing countries" (Beig, 2010). This strengthens the prospects for improving air quality, as these countries are undergoing profound technological and economic transformations. A definition of the AQI has been provided by the Florida Department of Environmental Protection, which defines the AQI as a "tool for reporting daily air quality. It tells you how clean or polluted your air is and what associated health effects might be a concern for you. The AQI focuses on health effects you may experience within a few hours or days after breathing polluted air. It takes all the monitored pollutants and relates them to a single-scale value to indicate air quality." It is usually a scale of 0 to 500; the higher the AQI value, the greater the level of air pollution and the greater the health concern. However, there isn't a worldwide harmonized scale of standards; for example, Australia's environmental agency uses the NEPM (National Environment Protection Measure), which considers an AQI in the range of 67–99 as "fair," while in Europe, an index of 75–100 is considered as "high polluted air." Moreover, numerous countries use a unique mathematical equation to measure the AQI (such as Canada's Air Quality Health Index, Malaysia's Air Pollution Index, and Singapore's Pollutant Standards Index). Considering a more physics-oriented definition, the AQI "requires an air pollutant concentration over a precise averaging period, obtained from an air monitor or model. Taken together, concentration and time represent the dose of the air pollutant. Health effects corresponding to a given dose are established by epidemiological research. Its air quality index values are typically grouped into ranges. Each range is assigned a descriptor, a color code, and a standardized public health advisory" (Wikipedia). The pollutants that are largely prioritized in the monitoring process are ozone, particulates, sulfur dioxide, carbon monoxide, and nitrogen dioxide.