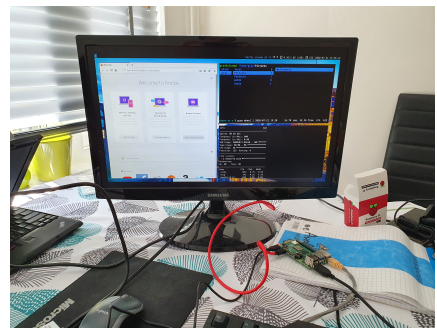
Geospatial Data Mining Using **R!**

Guanqun Cao
guanqun.cao@ieee.org

Why R? 2020 Conference
Geo + Finance Session
Sep 27, 2020

Technical Background

- Specialized in multimedia retrieval, representation learning from different sources.
- Now in auto industry working on data analytics for self-driving cars.
- Started using R less than a year ago
- Favorite packages: `ggplot2`, `rgdal`, `spBayes`.
- Development environment: Vim, Anaconda, Jupyter Notebook and Firefox.



Home setup

Air Pollution Problem

Impact

Analyzing the geo-spatial data improves our quality of life.

Causes

- Energy use and production: Gases and chemicals; smog, and haze.
- Natural causes; climate change.

Dispersion and Transport

Wind, turbulence, and air temperature.

Health Effects

Eye irritation, cardiovascular system, respiratory system

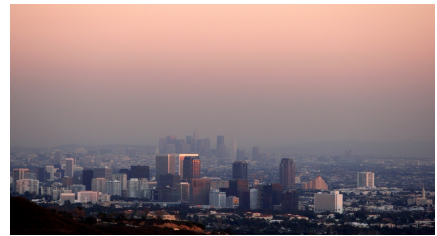


Image Source: Reuters news

Particles and Measurements

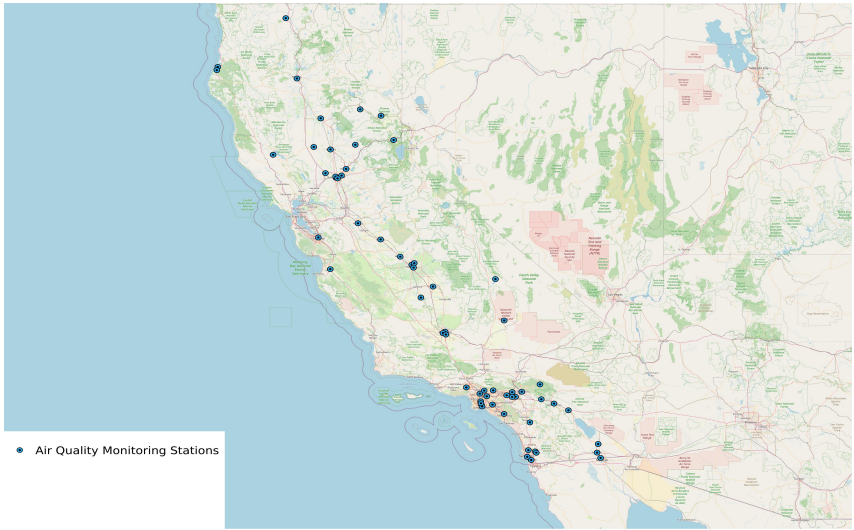
- Particulate matter: Coarse particles PM_{10} ($> 2.5\mu m$). Fine particles $PM_{2.5}$ ($< 2.5\mu m$).
- Air Quality Index (AQI): Report daily air quality in metropolitan areas for PM (PM_{10} and $PM_{2.5}$).
- Air quality modeling: Gaussian model, CMAQ Model

Pollution Levels Based AQI

AQI	Air Pollution Level	Health Implications	Cautionary Statement (for PM2.5)
0 - 50	Good	Air quality is considered satisfactory, and air pollution poses little or no risk	None
51 -100	Moderate	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
101-150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
151-200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects	Active children and adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else, especially children, should limit prolonged outdoor exertion
201-300	Very Unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
300+	Hazardous	Health alert: everyone may experience more serious health effects	Everyone should avoid all outdoor exertion

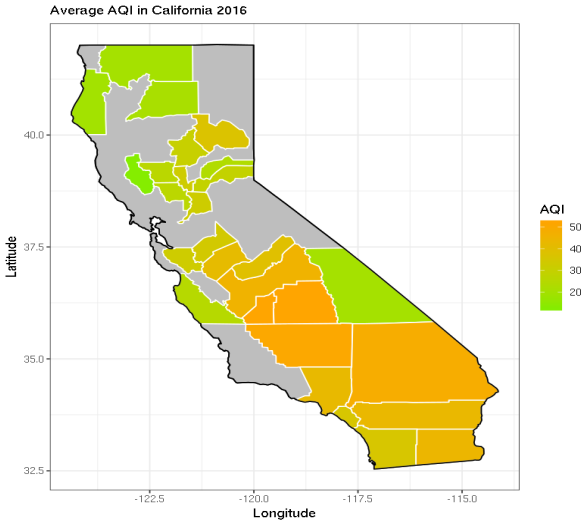
Source: https://aqicn.org/

62 Stations Monitor California Air Quality Shown in QGIS

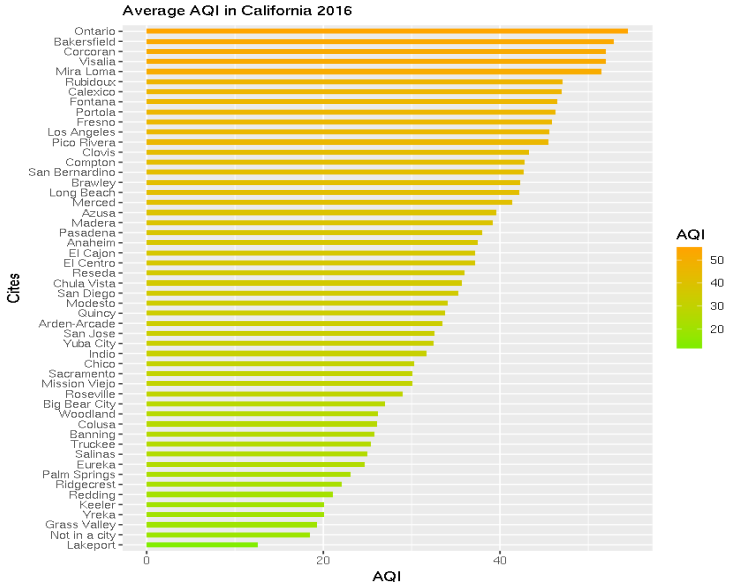


Data Source: `https://aqs.epa.gov/aqsweb/airdata/download_files.html`

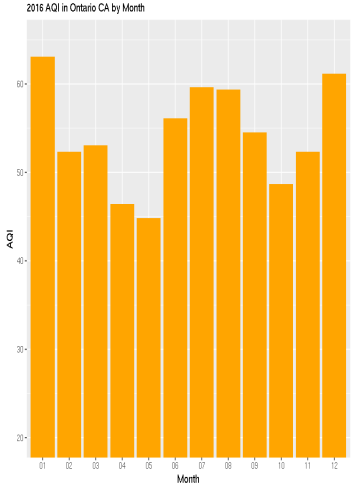
AQI by County

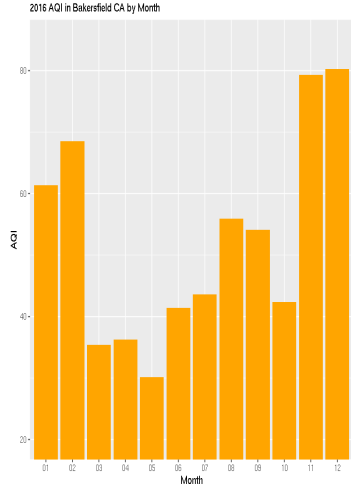


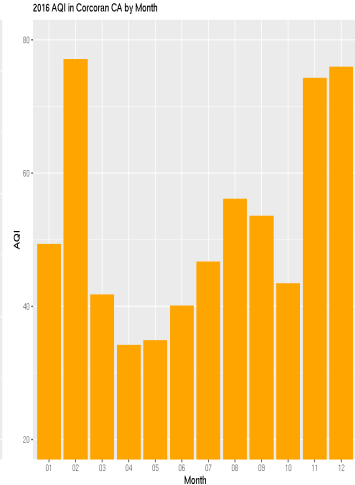
Average AQI in California



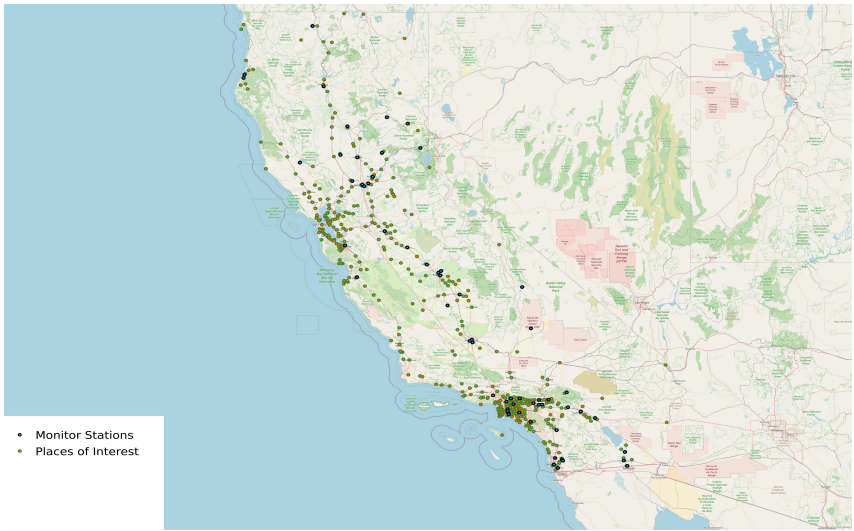
AQI by Month in Most Polluted Cities





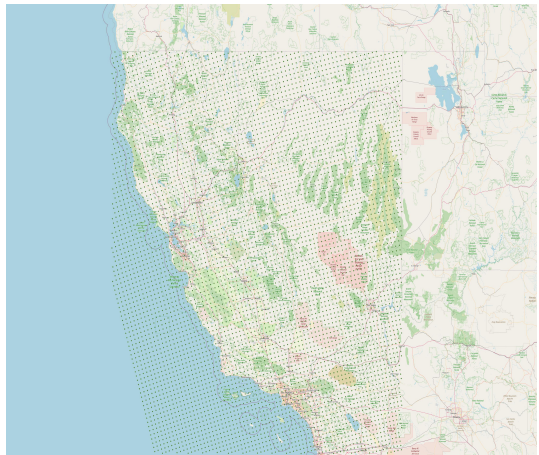


How do we measure AQI in every city of California?



CMAQ Model

- CMAQ stands for Community Multiscale Air Quality.
- It is an active open-source development project of the U.S. EPA that consists of a suite of programs for conducting air quality model simulations.
- It provides sound estimates of particles, ozone, toxics, etc.



CMAQ Grid Over California

Predict AQI by Spatial Bayesian Method

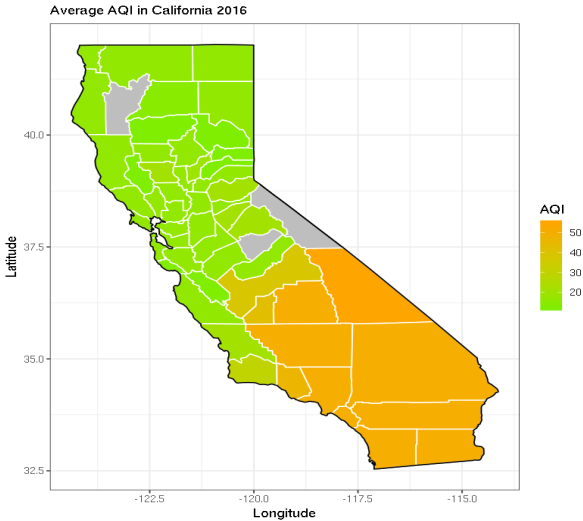
Our solution

- Couple observations from sparsely distributed stations with CMAQ model outputs.
- Use `spBayes` to model point-referenced data.
- Gaussian spatial process model:
  $y(s) = X(s)\beta + w(s) + \epsilon(s)$, where $w(s) \sim \text{GP}(0, \Sigma_w(d)), d = |s_i - s_j|$.

Under the hood

- California city data from Kaggle. Address info from Nominatim.
- Adopt CMAQ results from the closest 12 km-grid output.
- Convert longitude & latitude to UTM system.
- Our model use square root of AQI.
- MCMC takes a long time.

Predicted AQI by County



Number of Days Spent in Each AQI Level

References

1. Godish, T., Davis, W. T., & Fu, J. S. (2015). Air quality. 4th Edition. CRC Press.
2. Congdon, P. D. (2019). Bayesian Hierarchical Models: With Applications Using R. CRC Press.
3. Finley, A. O., & Banerjee, S. (2020). Bayesian spatially varying coefficient models in the spBayes R package. Environmental Modelling & Software, 125, 104608.

Take-Away Message

- R provides powerful open-source packages for geo-spatial data mining and visualization.
- Similar techniques can be applied to disease mapping, crime mapping, traffic accident mapping, etc.
- Still room for improvement for command-line users: In-line debugging.
- Feel free to contact:
  - guanqun.cao@ieee.org
  - https://github.com/cgq5/spatial-release
  - https://www.linkedin.com/in/guanquncao/