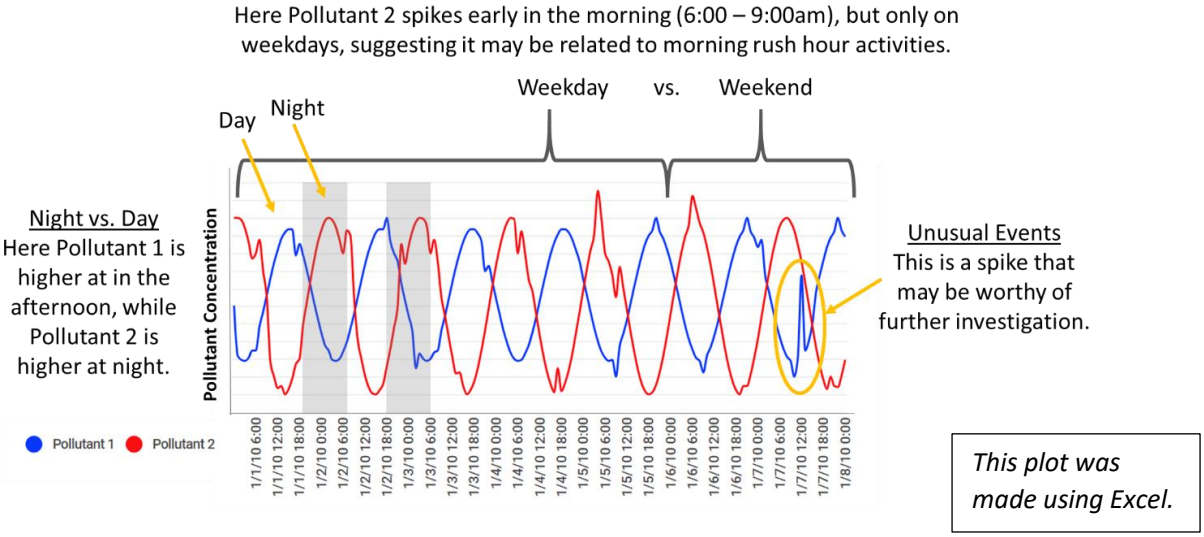
Low-Cost Sensor Data Analysis Guide

Guiding Questions

Low-cost sensors collect large amounts of data. Some sensors distributed through the US EPA STAR Grant program have been running continuously for over a year, recording data every minute (that's over 500,000 rows of data!). For this reason data analysis tools and software can be very helpful. In this guide we provide some brief instructions to help community scientists interact with the data they are collecting as well as some questions to help guide their analysis.

How do pollutant levels vary over time?

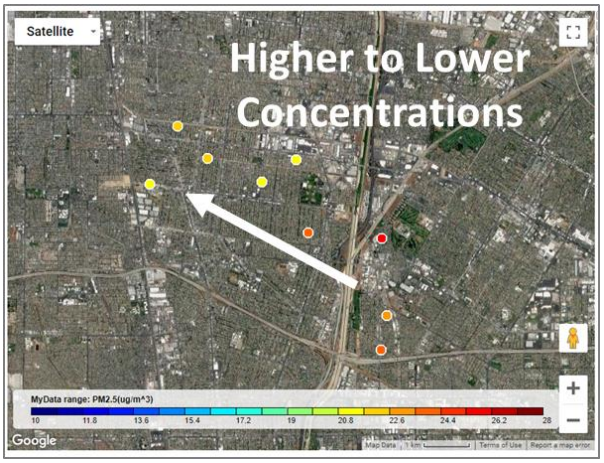
- Do you see any patterns in the data from day to day?
- Do you see any obvious differences between weekdays and weekends?
- You can also try comparing different periods, for example: the time of day, the morning and evening rush hour periods, or even seasons.



This plot was made using Excel.

Do you see any spatial trends?

- If you have data from multiple sensors available, how do the sites compare? Is one consistently higher or lower?
- Does one site experience more frequent spikes or elevations in pollution levels?

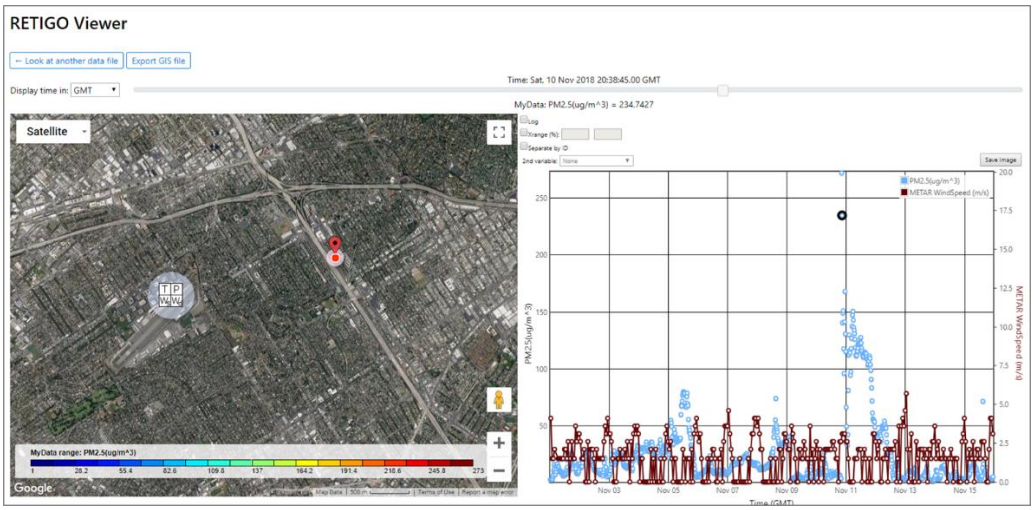


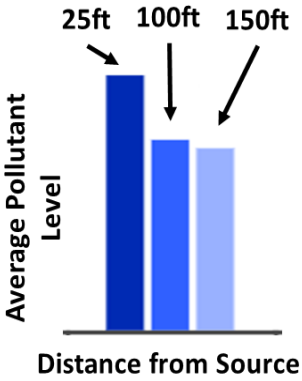
For the most part all of the sites are similar, however there are regular spikes in pollutant concentration at Site 1

In the two plots above, RETIGO was used to create the map and Excel was used to make the time series plot.

Are there impacts from potential sources?

- Do higher levels of pollution seem to line up with certain activities (e.g., high traffic times)?
- Or do you see elevated levels at a site closer to a potential source of pollution than you do at a site further away? Another way to look at this is: how does the data vary at different distances away from the source?
- Are there any relationships between wind speed or direction and pollutant levels? (looking at the example below





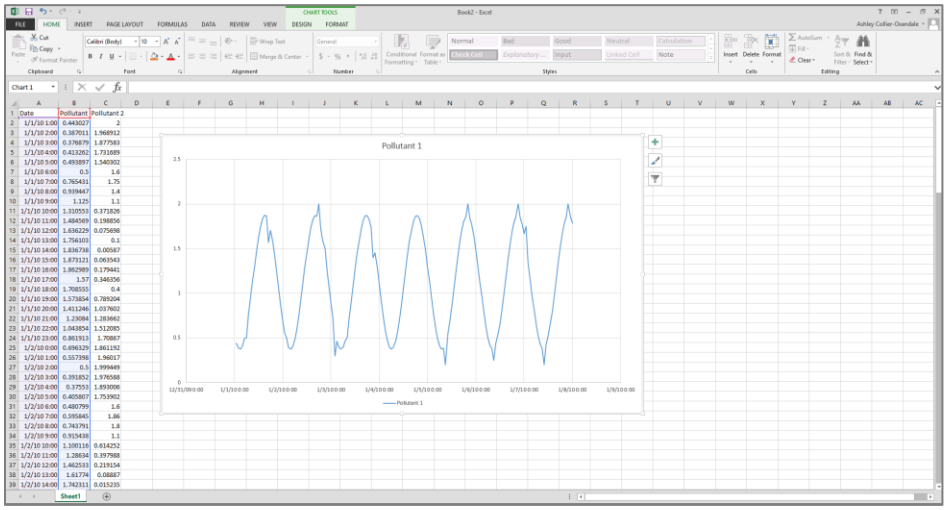
In the two plots above, RETIGO was used to create the map and time series and Excel was used to make the bar chart. For the RETIGO figure, wind speed was added in an effort to understand under what conditions enhancements in PM2.5 occur.

Resources Available

Excel

- Using Excel will provide the most freedom in terms of how the data can be plotted and what statistics may be calculated, however, using Excel may be a challenge for those with little experience with data analysis activities.
- That being said, there are many online resources and tutorials available to help with the use of Excel.

Excel Screenshot

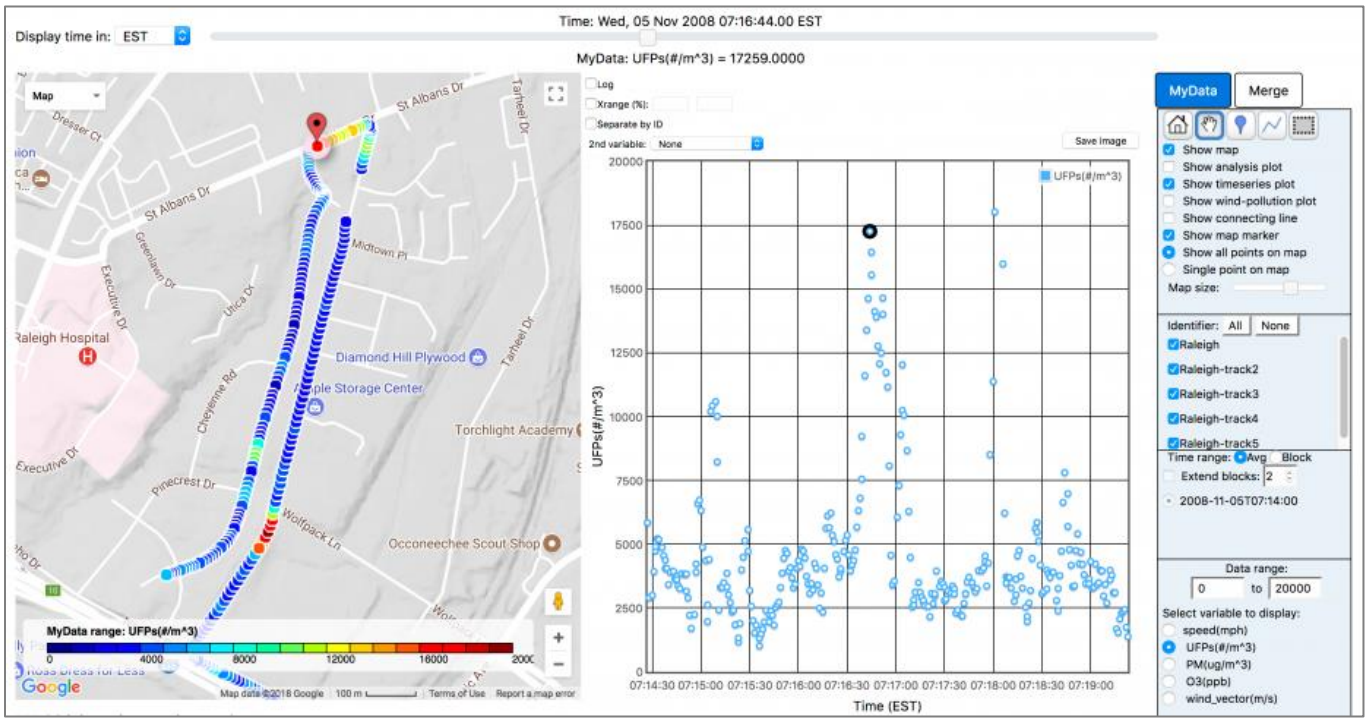


US EPA's RETIGO (REal TIme Geospatial Data Viewer) Tool

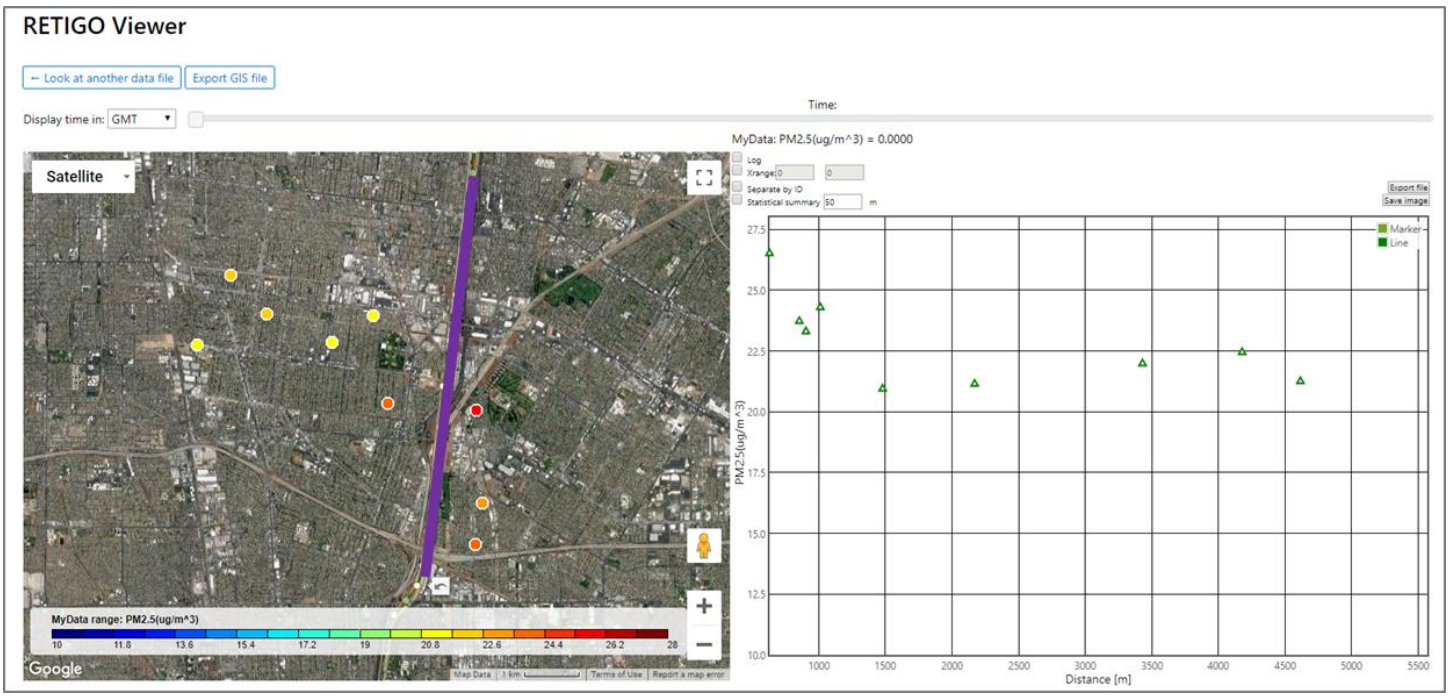
Contact for help: retigo@epa.gov

- This is a free, web-based tool that allows users to upload their air quality data and explore this data in maps and plots.
- This tool has a user friendly interface helpful for analyzing data, it can also produce nice visuals.
- Additionally the tool helps users look at how pollution levels change with respect to a particular point, line, or area that they can define. Users may also add in data from reference monitoring sites or wind speed/direction data which they can qualitatively compare to their sensor data.
- When uploading data, users have the option to share their dataset; meaning a single member of a community group could upload the data from all of the sensors in their network and attach key words (making the datasets easy to find), then any member of this community could more easily access the data and assist with analysis.
- The tool does not process sensor data to compare against the National Ambient Air Quality Standards (e.g., averaging PM to 24 hour intervals).
- More about RETIGO: https://www.epa.gov/hesc/real-time-geospatial-data-viewer-retigo

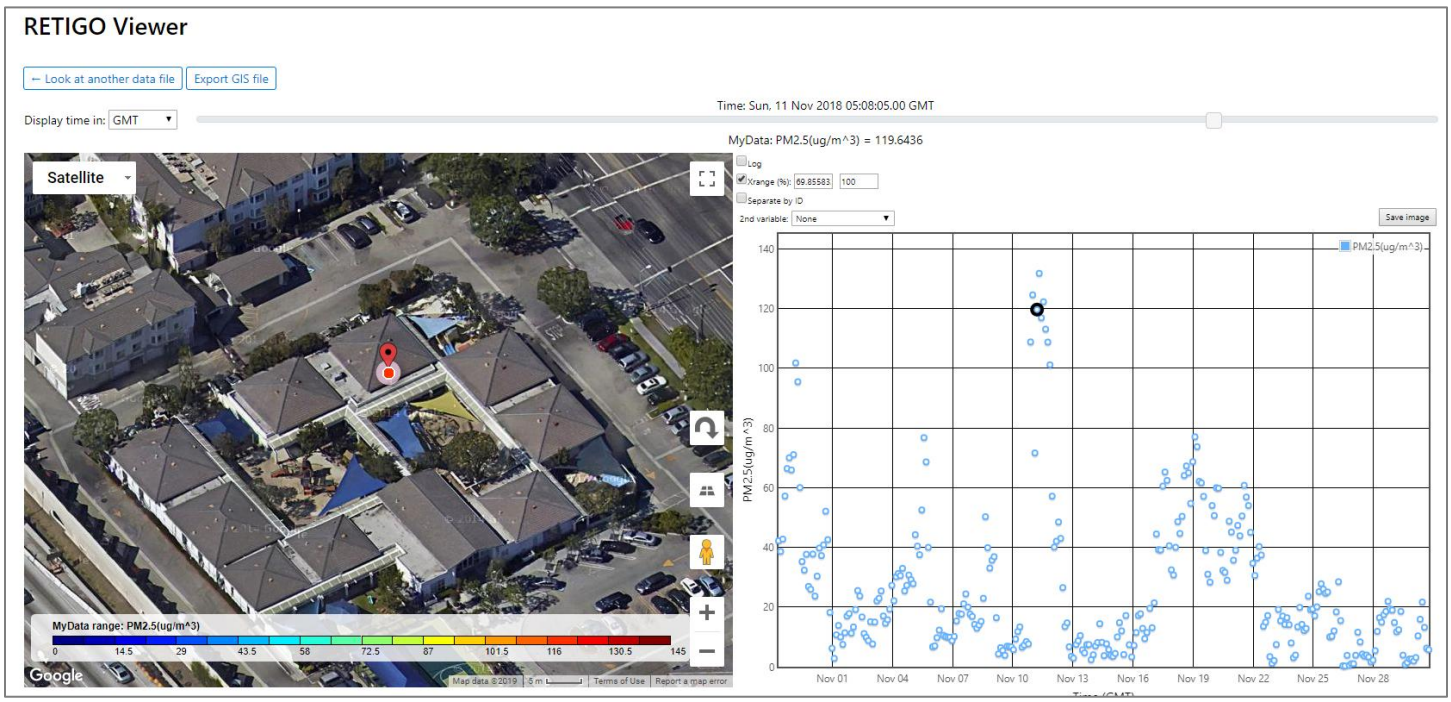
RETIGO Screenshot (source: US EPA, example of mobile data)



RETIGO Screenshot (source: AQ-SPEC, example of sensor network data with added analysis on the right illustrating pollutant levels in relation to the road highlighted in purple; in this example pollutant levels seem to be higher closer to the road)



RETIGO Screenshot (source: AQ-SPEC, example of sensor data from a single site with an interactive time series to the right)



A Few Quick Notes Regarding Data Quality

- Low-cost sensors are still an emerging technology, which means they may not be as reliable or as accurate as conventional monitoring instruments and methods.
- It's always good to consider whether your sensor data is "realistic", for example data that remains at the same level for a long time likely indicates an issue with the sensor. Similarly, very high or even very low levels of pollution indicated by a sensor may be reflecting local air quality trends, however, it is also possible that the sensor is malfunctioning.
- At this point it is not appropriate to compare low-cost sensor data with health-based regulatory standards, due to the previously discussed issues with accuracy and reliability. The data used to determine whether or not regulatory standards are being met is not only collected with higher-cost and higher-quality instrumentation, but also this instrumentation must be sited according to very intentional and specific criteria, the instruments undergo strict and routine maintenance, and the data is evaluated according to specific protocols to ensure important decisions about public health are only made using the best and most reliable data. Given the current challenges with low-cost sensors, it is possible for sensor data to suggest there are high levels of pollution when in reality there are not. However, the reverse is also true, it is possible that sensors may miss important air quality issues. For this reason, it is vital that low-cost sensors are used along with higher quality instruments for verification and validation purposes.
- Another thing to keep in mind is that sensor manufacturers typically calibrate their sensors or apply correction factors to improve the accuracy of their sensors. In some cases, the user can adjust the calibration models if they wish. It is also possible for users to calculate and apply their own correction factors, if there is interest in improving the accuracy of sensors being used.
- Despite these apparent limitations, there is still a lot we can learn from low-cost sensors. For example, comparing levels across a network of sensors can help to highlight "hot-spots" that may have been previously unknown, comparing trends across sensors can also highlight anomalies potentially caused by local sources, and a better understanding of our local air quality can give us information that might help us to reduce our exposure.

Read more about sensor calibration in the following resources:

The Air Sensor Guidebook, is a great resource in general and calibration specifically is discussed in Appendix C. (Available here: https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=519616&Lab=NERL)

The guide: "How to Evaluate Low-Cost Sensors by Collocation with Federal Reference Method Monitors", which accompanies the Macro Tool described above also provides great information.
(Available here: https://www.epa.gov/air-research/instruction-guide-and-macro-analysis-tool-evaluating-low-cost-air-sensors-collocation)

The following journal article provides an interesting and more advanced discussion around the use of complex algorithms in sensor calibration: Air Quality Sensors and Data Adjustment Algorithms: *When Is It No Longer a Measurement?* Gayle S. W. Hagler, Ronald Williams, Vasileios Papapostolou, and Andrea Polidori, Environmental Science & Technology. 2018, 52 (10), 5530-5531, DOI: 10.1021/acs.est.8b01826

If you have questions or comments, please contact the AQ-SPEC group at South Coast AQMD at: info.aq-spec@aqmd.gov

South Coast AQMD