OIZOM
REDEFINING RESOURCES

WHITE PAPER

Collocation Study of Sensor-based CAAQMS with Reference Systems

By Ayyan Karmakar, Kruti Davda

November 2020

Introduction

Clean air is fundamentally essential for human health and well-being. With increasing population, urbanization and industrialization, air pollution is emerging as a significant problem. Availability of accurate and reliable air quality monitoring data is crucial to develop and evaluate pollution control strategies. It enables us to understand the trend and extent of pollution by identifying pollution sources and hotspots, thereby indicating where more efforts are needed. Unlike conventional continuous monitoring stations, sensor-based air quality monitors provide high quality real-time high spatial resolution monitoring data at economical cost.

Today, a wide and diverse range of low-cost air quality sensors are available, however, there is a general lack of uniform and regulated sensor manufacturing, monitoring methodology and data quality checks and assessments (QA / QC). Also, studies have found that sensor-based systems provide very accurate results in laboratory conditions while in ambient applications under varying atmospheric composition and meteorological conditions accuracy decreases. The rapid expansion of sensor-based air quality monitors prompted environmental regulation agencies around the world to evaluate sensor technology. The collocation of sensor-based air quality monitors with reference stations is a tested and evaluated method recommended by the United States Environmental Protection Agency (USEPA) and The European Committee for Standardization.

Collocation is the process of operating sensor-based air quality monitors and reference continuous air quality monitors at the same time and place under real-world situations for a defined evaluation period. It is carried out in similar conditions in which sensor-based systems are to be deployed. Collocation study improves sensor robustness and measurement precision and accuracy. Long-term sensor performance and sensor response is evaluated and compared with reference values to improve data quality.

This paper describes the step-by-step procedure of carrying out sensor collocation and also provides information on how to compare and interpret the results of collocation study and how to improve data quality from those results.

Need Of Collocation

After manufacturing, the precision of sensors is evaluated by testing sensors multiple times with reference “clean” air of “zero” concentration, containing no pollutants. Such testing is followed by testing the sensor multiple times with a sample of air having a known concentration of the pollutant. Comparing the data at “zero” and higher concentrations allow the determination of how well the sensor repeats itself under various conditions. Multiple data points are generated in order to determine measurement bias, the deviation of the measured value from the “true” value of pollutant concentration. Here, the “true” pollutant concentration is established by collocating sensors with a reference monitor.
Collocation is the process of operating a reference monitor and sensor-based air quality monitor at the same time and location, under real-world conditions for a defined evaluation period. Based on the results of collocation studies, correction factors can be developed to better simulate the dependency of meteorological parameters on various air quality parameters. Also, long-term performance evaluation of sensor-based systems can be carried to under

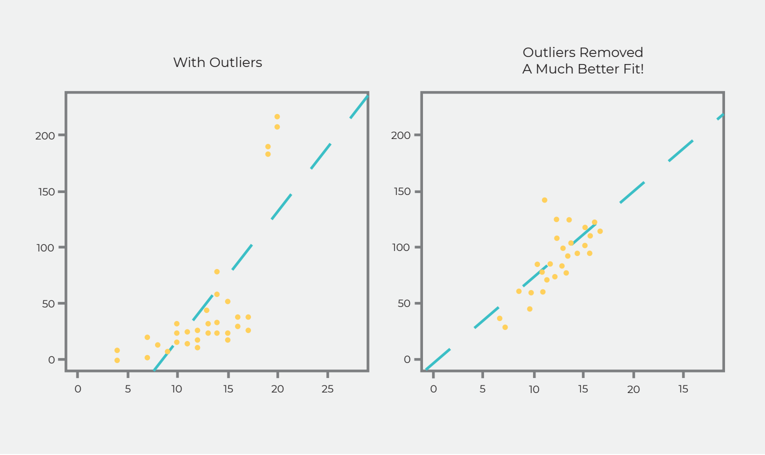
stand how sensors respond with changing air composition.

Agenda Of Collocation

The primary agenda and goal of collocation are to carry-out data review and data validation. Data review is the technical evaluation of the data collected by sensors and data validation is the process of evaluating collected data against the values collected from reference stations. The following factors are considered during data review and validation.

Outliers

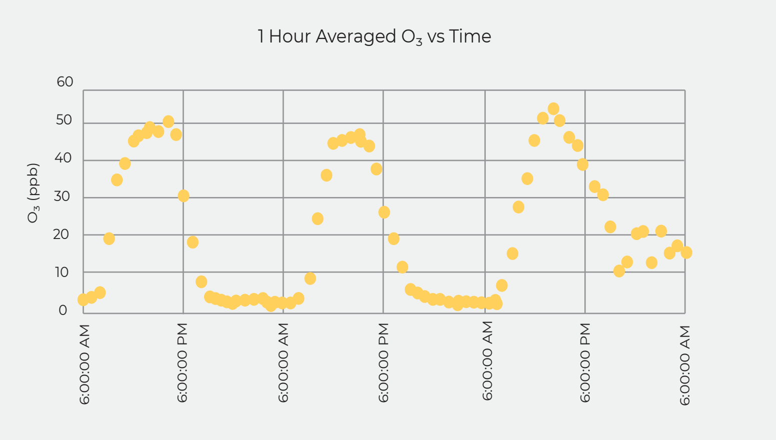
Outliers are the data points which significantly differ from other observations. They are unusual values in the dataset, which is abnormally higher or lower than nearby data points. There are many reasons for outlier presence in the datasets. If reference values also record such abnormal values at the same time, then it is not an outlier but a sudden spike in pollutant concentration.



Therefore, comparing the sensor values with reference values helps in determining real outliers, which can be later removed by various statistical techniques.

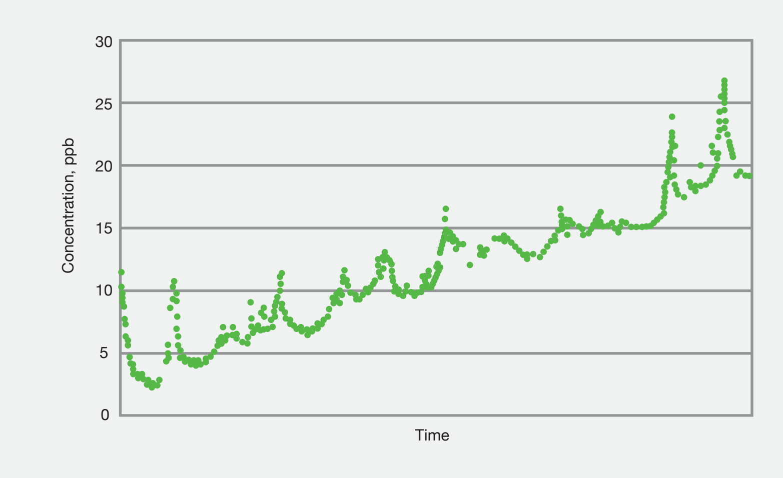
Pattern Identification

Pollutants such as ozone showcase a distinct concentration pattern. Ozone concentration increases throughout the day, peaks in the late afternoon and again decreases until the next morning. Patterns could be diurnal, weekly, monthly or seasonal. Absence of expected patterns and / or presence of unexpected patterns indicate problems with the sensors. Such patterns can be identified based on the reference data and required corrections can be developed to improve the data accuracy of sensors.



Data Drift/ Data Shift

Drift or shift in data refers to a gradual change in sensor's response characteristics over time. The shift can be positive or negative, which may lead to wrong conclusions. Drifts may occur due to a variety of reasons. One of the ways to reduce the drift is to calibrate the sensor frequently so that the instrument only drifts a small amount between each recalibration.



Collocation Procedure

It is crucial that all the sensors are collocated in a uniform manner in order to improve the quality and comparability of monitoring data. The United States Environmental Protection Agency (USEPA) has provided step-by-step guidelines to carry-out collocation with reference monitors.

Step 1: Planning

Planning is a preliminary survey to determine how the collocation shall be carried out. Selecting the reference station, power source, time duration, etc are thoroughly checked before execution. Proper data transmission and storage capacity are required as faulty data transmission may jeopardize the study. Number of monitors (Sensor-Based devices) to be collocated is decided and preferably at least 3 monitors are placed so any malfunctioning monitor can be identified easily. Moreover, as the pollutant concentration tends to vary seasonally ideally the collocation study should be planned to curb the seasonal variation.

Step 2: Measurements

For accurate measurements, the location of the sensor has to be selected. The sensor should be placed within 10 meters of the reference station and inlets for both the monitors should be kept at the same height to ensure uniformity. The frequency of data collection is set in the sensor to match with the reference station. Exact intervals are kept (for eg 5 minutes) or sometimes sensor data is recorded every minute and then averaged at 5 minutes of the time interval. Minimum 1000 data points are recorded to analyze the data and perform correlation tests. Sometimes the collocation period may extend up to three months based on the scope of the study.

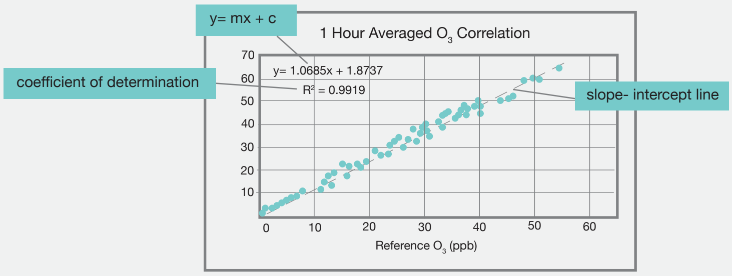
Step 3 (a): Data Review And Comparison

A superficial review can be carried out to identify outliers, interference, drift, or shift in the data by data visualization tools. For a more detailed analysis of correlation, various statistical tools are used. USEPA provides Excel compatible software called MAT(Macro Analysis Tool) to compare the datasets of both sensor-based and reference monitors. Various other software carries out these analyses and can be custom made by private companies, government agencies, researchers, etc.

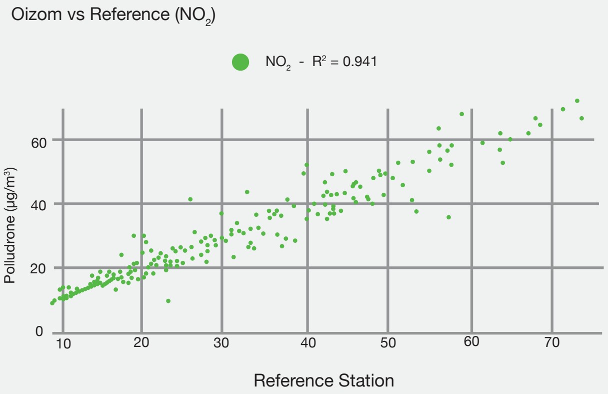
Correlation graphs are plotted between the sensor and reference monitors datasets against the X and Y axis as shown in the figure. A line going through the data points is known as the 'slope-intercept'. The slope intercept equation is denoted as:

y = mx + c

Where, 'y' denotes sensor data, 'x' denotes reference data. m denotes the slope of the line and deviation in data. As the value of m deviates more from '1' it shows the over or underestimation of sensors. While the intercept c shows the data recorded by the sensor when the reference monitor had the value 'zero', which shows initial bias in data.



The R2 value also known as the coefficient of determination is a statistical measure that shows the closeness of the data to the slope-intercept. It ranges from 0-1; value closer to 1 is interpreted as a stronger relationship between sensor and reference monitor datasets i.e. the accuracy and precision of the sensor is much better.



Step 3 (b): Data Quality Checks And Quality Assesment

Various statistical tests and analyses are carried out to ensure the highest level of data quality. Statistical parameters such as precision, accuracy, bias, coefficient of variance, standard deviation, presence of autocorrelation etc. are used for that purpose.

Precision is a measure of how close repeated measurements are to each other. In the context of air pollution data, if the sensor is to measure the same pollutant concentration in the same conditions, if the sensor records similar results in each measurement, the sensor has high precision. Also, precision is the description of random errors and a measure of statistical variability. Precision can be quantified by calculating standard deviation or variance of the data.

Standard deviation is a statistical measure of precision to describe how spread-out numbers are. It is calculated by taking the square root of variance.

Variance is the average of the squared differences from the mean.

Accuracy is the measure of how close a measured value is to the actual (true) value. In the context of collocation study, the measured value is the concentration value measured by sensors and reference station values are taken as actual (true) value. If the difference of measured value (sensor concentration) and the actual value (reference concentration) is less, the sensor is more accurate. High accuracy requires high precision and high trueness of data. Root Mean Square Error is a statistical analysis that interprets the accuracy in percentage error.

RMSE - Root mean square error indicates the accuracy of the prediction by the calibration model. It is calculated as the square root of mean squared terms of the difference between the individual predicted value(pi) and individual reference value(yi). The number of data points taken for averaging is denoted by n.

$$\text{RMSE} = \sqrt{\frac{\sum_{i} (p_i - y_i)^2}{n}}$$

Bias is a systematic (built-in) error in data measurement which makes all the measurements wrong by a certain amount. It is a persistent error in the measurement process that causes all measured values to be too high or too low by a certain amount, compared to the true (reference) value. This is the intercept value (b) in the slope-intercept equation y = mx + b.

Data quality may differ due to outliers, interferences, malfunctioning of sensors, etc and compromise the quality of 'averaged data'. Data completeness denotes the number of usable data records collected by the sensor compared to the total amount of data expected or collected by the reference monitor. It is denoted in percentage viz. If out of 1000 data points 950 were recorded by the sensor then the data completeness of 95% is said to be achieved. Similarly, if only 18 out of 24 hours of usable data is collected then data completeness is only 75%. Higher data completeness ensures a better average that represents the whole measurement period.

Autocorrelation is the similarity between consecutive data points in a time series. Autocorrelation is very common in time series data. Autocorrelation incorporates serious errors in data due to incorrect entry in the dataset. Autocorrelation tests help identify if the time series is autocorrelated. If the dataset has presence of autocorrelation, it can be removed by various statistical techniques.

Step 4: Corrections And Deployment

The results of correlation analysis are used to correct, calibrate, and update the sensor-based CAAQMS. Models and machine learning algorithms are corrected and retrained based on the R2 value, Root Mean Square Error (RMSE), and other statistical test results. It is updated manually or using cloud-based software and deployed in the field.

Oizom's Offerings

All of the sensors are calibrated at OIZOM's factory before dispatching them. OIZOM equipment is calibrated at the NABL (National Accreditation Board for Testing and Calibration Laboratories) accredited laboratory to ensure data accuracy. Also, to maintain and ensure the highest level of data quality, the monitoring devices are calibrated against a reference station before installation to test their performance in ambient conditions. The sensor is collocated in similar weather conditions in which it is to be deployed. Around 1000 data-points collected at the same time interval is sufficient to calibrate the system.

Also, OIZOM provides spot calibration, through which the equipment is calibrated over the cloud, reducing onsite manpower and maintenance cost. It is a periodic calibration performed quarterly or bi-annually. OIZOM uses recently calibrated sensor-based systems or a third party system trusted by the client as a reference. The reference system should be a continuous monitoring system. Around 150 data-points are collected at the same time interval to re-calibrate the system on the spot.

Oizom's Case Studies

OIZOM's sensor-based real-time air quality monitoring system POLLUDRONE was collocated with Continuous Air Quality Monitoring Station (CAAQMS) of Maharashtra Pollution Control Board (MPCB) located at Bandra. 15-min average data of reference stations were used for collocation and calibration of POLLUDRONE. The collocation study was carried out for 2-weeks (26th March 2019 to 9th April 2019). Prior to collocation, POLLUDRONE was calibrated in the laboratory against calibration gas mixtures and after the collocation, continuous air-quality data monitored by both the systems were compared for a period of two weeks (10th April 2019 to 24th April 2019).

For 1-hr averaged data, the overall R2 for most ofthe parameters is more than 0.9 with CO being best correlated with an R2 of 0.97. The lower correlation of $PM_{2.5}$ is due to its cross-sensitivity with PM_{10} measurement. While for 24-hr averaged data, the overall R2 for most of the parameters is more than 0.95 with NO being best correlated, with an R2 of 0.986.

This performance meets the recommended precision andaccuracy error metrics from the US EPA Air Sensor Guidebook for personal exposure (Tier III: Supplemental monitoring and Tier IV: Personal Exposure) monitoring.

A similar collocation study was carried out at Horiba Germany Office from 30th October 2019 to 24th November 2019. 5-min averaged data from the reference station was compared with Oizom's Polludrone measurements. Post calibration, continuous air-quality data monitored by both the systems were compared from 25th to 29th November 2019

Overall, it was concluded that with careful data management and calibration with the use of advanced machine learning models, the sensor-based system can significantly improve the ability to resolve spatial heterogeneity in air pollutant concentrations giving accuracy between 88% (for NO, NO_2 ,SO_2 ,O_3) to 78% (for particulate matter) in real-time.

Parameter	R2 (1hr average)	RMSE (1 hr average)	R2 (24hr average)	RMSE (25 hr average µg/m³)
$PM_{2.5}$	0.75	4.96	0.96	1.35
PM_{10}	0.77	15.95	0.95	4.73
CO	0.97	0.039	0.95	0.023
NO	0.93	10.85	0.98	1.23
NO_2	0.94	3.68	0.97	0.94
SO_2	0.92	0.64	0.98	0.42
O_3	0.94	4.84	0.95	1.58

Parameter	$PM_{2.5}$	PM_{10}	NO	NO_2	O_3
R2 (1 hr avg)	0.852	0.772	0.94	0.65	0.64

References

[1] How to Evaluate Low-Cost Sensors by Collocation with Federal Reference Method Monitors, National Exposure Research Laboratory Office of Research and Development, USEPA.

[2] Air Sensor Guidebook, National Exposure Research Laboratory Office of Research and Development, USEPA.

[3] United States Environmental Protection Agency. “FRMs/FEMs and Sensors: Complementary Approaches for Determining Ambient Air Quality.”, December 2019.

[4] New Paradigm for Air Pollution Monitoring, 2014-2018 Progress ReportAir and Energy Research ProgramRon Williams-US EPA.

[5] Reinventing air quality monitoring: Potential of low cost alternative monitoring methods, Centre for Science and Environment (CSE), India.

[6] Cordero, J. M., Borge, R., & Narros, A. (2018). Using statistical methods to carry out in field calibrations of low cost air quality sensors. Sensors and Actuators B: Chemical, 267.

[7] Munir, S., Mayfield, M., Coca, D. et al. Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—a case study in Sheffield. Environ Monit Assess 191, 94 (2019). https://doi.org/10.1007/s10661-019-7231-8.

[8] Low cost sensor systems for air quality assessment: Possibilities and challenges, Eionet Report -ETC/ACM 2018/21, European Topic Centre on Air Pollution and Climate Change Mitigation(2018).

About the Authors

Ayyan Karmakar

With an experience of more than 10 years promoting various Environmental Technologies, Ayyan Karmakar currently leads marketing at Oizom. He is an industry professional with core Environmental Engineering skills with a spirit of continuous learning.

Kruti Davda

With experience in environmental engineering and research, Kruti Davda currently leads environmental analysis at Oizom, where she puts her mind to gross air pollution data and extrude insights through environmental research and analysis.

OIZOM™
REDEFINING RESOURCES

Accurate And Affordable Air Quality Monitoring Solutions