OPENAIR

Operational Network of
Air quality Impact Resources

Best Practice Guide

BP406 | Manage and analyse

Data interpretation: quality control

Introduction

In order to support activities that create impact, corrected and harmonised sensor data must be subjected to a quality control process. This involves verifying and cleaning data to ensure it can be trusted, and is useable for analysis. This chapter outlines some of the main considerations for quality control of data produced by smart low-cost air quality sensors. It will explain and present practical advice relating to the following key topics:

- **field testing** (where the 'normal' operation of a device, and the quality of the data that it produces, is verified prior to commencement of your main data collection activities)
- **data cleaning** (where data anomalies, outliers, and other unusable data are detected and removed as part of your main data collection activities)
- **data verification** (where data is verified against external references, or through internal cross-verification, to improve user trust in its ability to support impact).

Who is this resource for?

This resource is for local governments and other organisations undertaking similar projects. It is intended for staff engaged with the design and delivery of air quality monitoring projects, including project managers, environmental officers, smart city leads, and planners. It is also a useful reference for senior management who wish to understand the complexities and challenges related to this kind of project.

How to use this resource

This OPENAIR Best Practice Guide chapter is the third in a series of four chapters on the topic of data interpretation. It is recommended you read the overview chapter first, then refer to the other chapters on data interpretation (correction and harmonisation, and analytics) in the order shown in Figure 1.

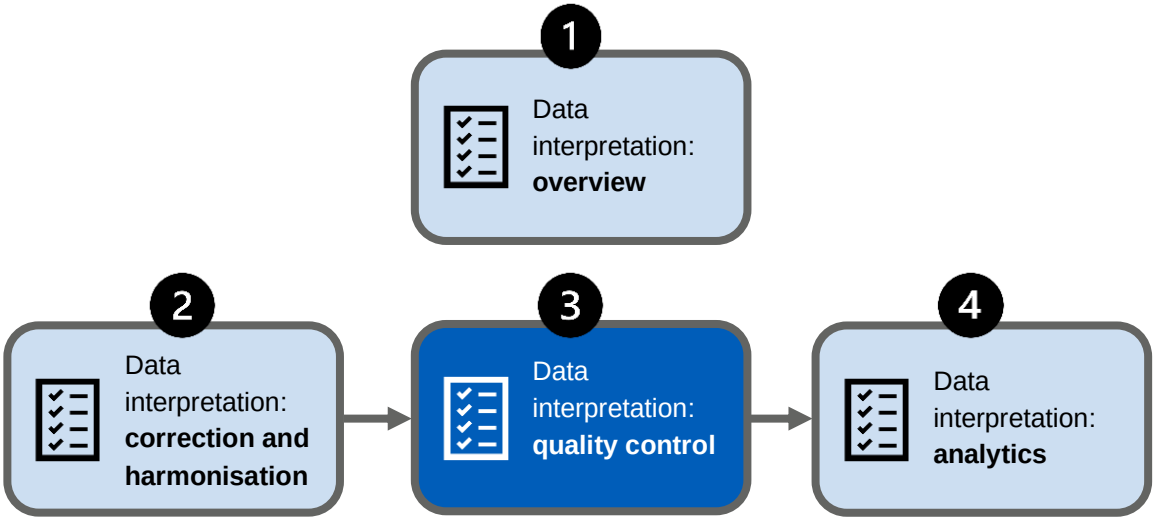


Figure 1. OPENAIR data interpretation Best Practice Guide chapters

Key messages

The key messages of this chapter are:

- Data quality control can involve a range of approaches, of varying complexity. While quality control is always necessary to some degree, the specifics of your data use case will dictate your requirements in terms of effort, resources, and expertise. As such, effective planning for data quality control demands that you clearly understand the needs of your data use case.
- Data quality control can be applied manually (to exported static data sets), or built into the functionality of an IoT or data platform with varying degrees of sophistication. It is a good idea to consider your data quality control requirements as part of your technology procurement decision-making process.

Field testing (initial device and data verification)

Following deployment of a device, field testing is required to verify metadata completeness and accuracy, confirm correct operation of a device, and establish an initial period of stable data quality. Only when all of these steps have been cleared can the data from a device be considered verified and useable. **This is a foundational process that is part of device deployment operations, and occurs *before* normal data collection operations commence**.

The initial device and data verification process includes three stages, as described in Figure 2.

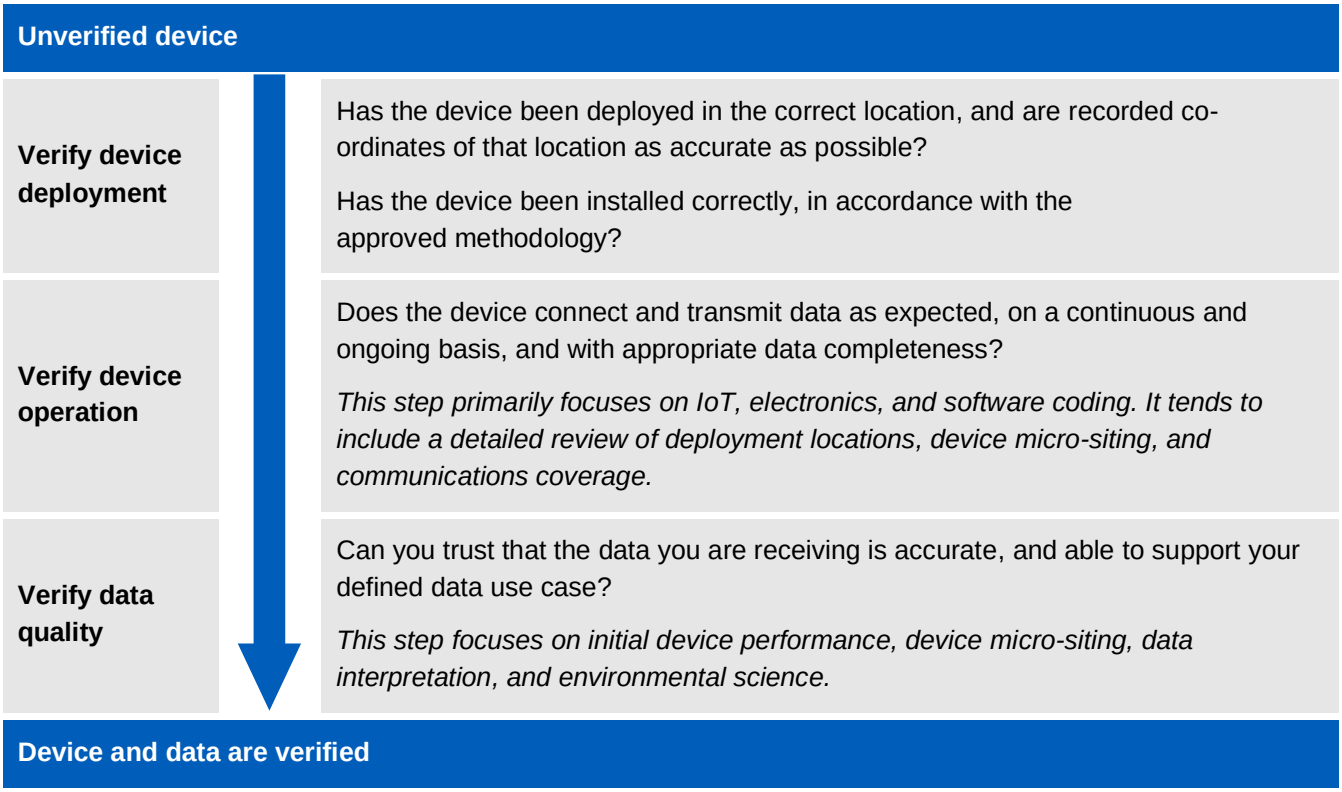


Figure 2. Overview of the device and data verification process

Detailed guidance on how to conduct field testing for initial verification of newly deployed devices can be found in the OPENAIR supplementary resource *Air quality sensing device activation and deployment checklist*, and the Best Practice Guide chapter *Air quality sensing device activation and deployment*.

Data cleaning

Almost any data set or live data stream will have anomalous data points, outliers, and other inconsistencies that need to be addressed before you can make use of the data. Data cleaning is the process of detecting and either removing or fixing such anomalies. There is no prescribed way to do this, however it is important to establish a consistent approach. This section will guide you through some of the key approaches you might take.

Manual data cleaning

Manual data cleaning applies to static data sets. In the context of a s**mart** low-cost sensing network, a static data set is produced by downloading a set of historical data from your IoT platform, defined by a start and end date. Before such a data set can be analysed, it first needs to be cleaned.

Manual data cleaning can be undertaken using a range of software packages, including open-source software (such as R Studio), or fully licenced software that is available to your organisation. Manual techniques can include filtering large sets to identify outliers, or visual identification using scatter plots. Alternatively, rules can be applied, such as a modified version of the Grubbs test that detects outliers in data. Websites and online tutorials/guides can provide useful tips for data analysis and interpretation.

Automated data cleaning

Automated data cleaning is applied to live data streams within your IoT, data, or analytics platform. It involves the automatic application of threshold-based actions to remove or correct data anomalies. This can deliver clean data sets for analysis, removing much of the hard work associated with manual cleaning. Automated data cleaning is also critical for:

- use cases where live data is streamed directly to third party users (e.g. via a public data portal)
- use cases where live operations or services are reliant upon the accuracy of live data (e.g. health alerts triggered by high pollution concentrations)
- use cases where live data is integrated into a live model or advanced analytics application (such as machine learning).

NEVER DELETE ORIGINAL DATA

Original (or 'raw') data is vital for understanding how data has been processed, and will be requested by anyone who wants to use that data for more advanced analytics applications. You should *never* delete it.

With manual data cleaning, ensure that you save a copy of the original data set, and delete anomalous points in a separate data file.

For automated data cleaning and quality control, ensure that your database is storing raw data alongside any corrected and cleaned data.

Main steps for data cleaning

1. Detect and remove anomalous data points

Some data points may sit outside of what can be considered 'possible' for the parameter in question. Anomalous points can arise due to errors in how a device is configured and calibrated, or in how data is structured and interpreted. Table 1 describes two common examples.

ALWAYS DISCOUNT DATA FROM YOUR TEST PERIOD

When a device is first deployed, you should run it for at least a week to test for issues, and verify that it is operating correctly and reliably. Make a note of the device deployment date (and, ideally, the date that you verified its operation), and discount/remove from your analysis any data from *before* that verification date.

This principle also extends to situations where a device develops an issue after it has been functioning correctly for a time. Once the issue has been addressed, it is advisable to implement a new test period and device verification date, and to discount/remove any data associated with the entire troubleshooting period.

Table 1. Common examples of anomalous data points from smart low-cost sensors

What to look for	Explanation	What to do
Negative values for pollution concentrations	When you are measuring very low concentrations of a pollutant, you are reliant upon the sensor to be calibrated so that its understanding of 'zero' matches true zero for that pollutant. If this **calibration** is slightly off, a negative value can be produced. Furthermore, consider that many smart low-cost sensors have a notable **error range**. This means that a small negative value that falls within that error range is possible (and even expected) under conditions where true pollutant concentrations are zero, or near-zero.	If you find many negative values from one device, it may indicate that you need to **recalibrate** that device. You may **formulate a bias correction** for all historical data from that device (e.g. if you consistently see a gas concentration of -2.0 as your lowest recorded value, then you might add 2.0 to *all* values in your set, adjusting negatives to zero). **Speak with your IoT platform provider** about automatically applying a correction to future data.
Extreme outliers	There are several possible causes for extreme outliers (defined as very large or small values, relative to those expected). Some sensing devices may have **firmware bugs** that cause the production of extreme outliers under certain conditions. Extreme outliers may be the product of a **misplaced decimal point** applied in the decoding of raw data. A review of the decoder may be required. There may be a discrepancy between the **structure of a data packet**, and what the decoder module for that data packet expects to receive. This can result in strange outcomes, such as temperature data being interpreted as a particulate concentration.	Discrepancies like this can occur following device firmware updates, or updates to decoder software within your IoT platform. **Speak with your device vendor.** Ask them about firmware bugs that may be associated with a recent update, and check for possible changes to the structure of data packets sent by the device. **Speak with your IoT platform provider.** Ask them about software updates that might coincide with data anomalies, and request a review of the device decoder relative to the current data packet structure.

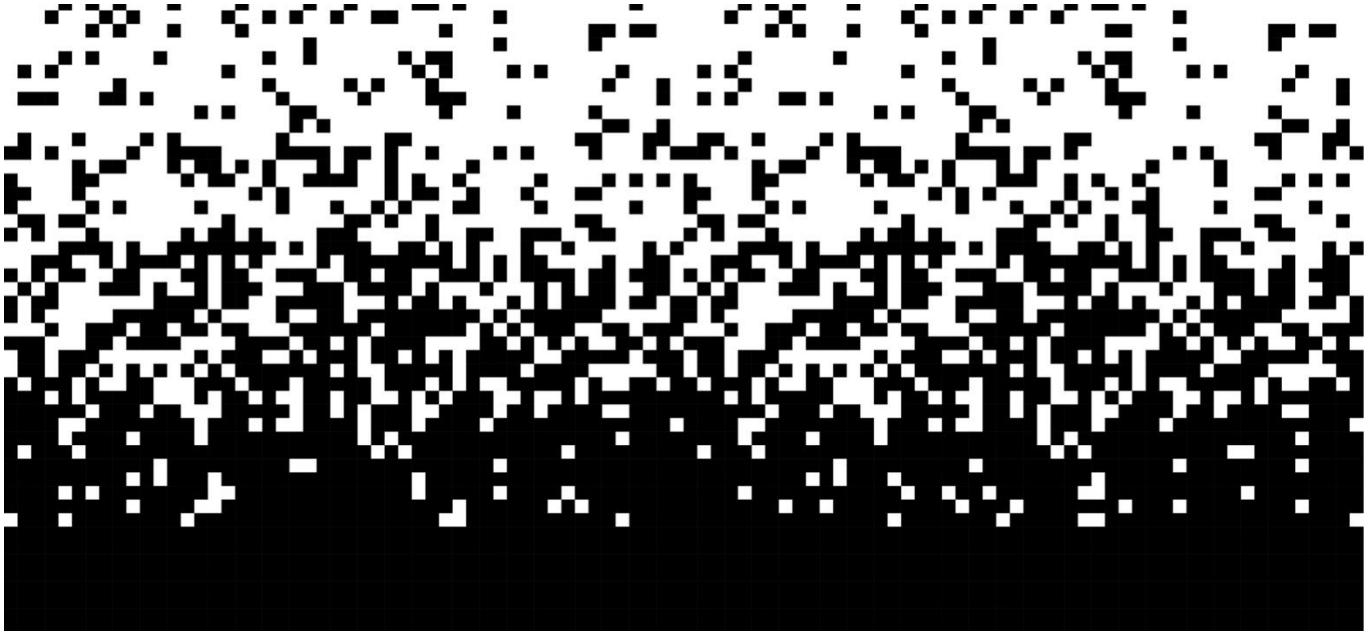
2. Check for abnormal data trends

Certain trends in a data set should be cause for concern, indicating that you are probably not capturing an accurate record of environmental conditions. Table 2 describes two common examples.

Table 2. Common examples of abnormal data trends from smart low-cost sensors

What to look for	Explanation	What to do
A flatline trend	Flatline trends are where a data set shows little or no variation in values over time, producing a flat line (counter to what the deployment context implies). An example is an outdoor ambient temperature sensing device. We know there must be day/night temperature fluctuations on a 24-hour cycle. If no fluctuations are apparent, there is cause for suspicion. For air pollutants, such an assessment may be more subtle. However, assuming you are measuring a pollutant that relates to variable activity (e.g. traffic), a flatline trend would still be suspicious.	Most abnormal trends are likely to result from a problem with a device, rather than a decoding issue in your IoT platform. Share the abnormal data with your device vendor, and request support. If you identify an abnormal trend from a particular device, you should discount all data from that device until you are able to either: • verify that the trend is, in fact, correct. • diagnose and troubleshoot the problem that is causing the abnormality.
A steady rising or falling trend over a period of days, weeks, or months	This indicates a possible device calibration error, or sensor malfunction.	

3. Check data completeness



Data completeness refers to the amount of useable data that is obtained from a sensor over a period of time, compared to the total amount of data expected over that same period (expressed as a percentage).

***Example*: A one-month period has a data completeness of 50%. This means that only half of the data expected for that period is present.**

A data use case should define a lower threshold for data completeness (e.g. 75%). This tends to relate to having a sample size that is able to support statistically significant data analysis. Put simply, if you only have partial data, it may not support reliable conclusions.

There are two practical approaches to checking data completeness:

1. **Visual inspection.** Use spreadsheet software or a data visualisation tool (which may be present within your IoT platform) to visually inspect your data record for a defined period. This won't give you a precise data completeness metric, but it may be adequate for an initial rough assessment if you have concerns.
2. **Data profiling.** Use a data analysis tool or platform to generate a data completeness value, and various other statistics and metadata about a data set. This might include the number, type, range, frequency, and distribution of values in the set.

Data completeness is a useful indicator of how well a data set can support your data use case; however, it is not definitive. Table 3 describes a few nuances to be aware of while you assess data completeness.

Table 3. Nuances to be aware of when assessing data completeness

Nuance	Description
The timing of pollution events being investigated	Imagine that you are exploring pollution exposure at a specific time of day, such as during a half-hour period around school drop-off and pick-up times. You might have a relatively small data gap that happens to fall during one of those half hours. Across the day, you may have high data completeness, however your data use case still cannot be served because you are missing the data from a critical period.
Investigating peak pollution events	You may be specifically interested in capturing short-lived peak pollution events. If a data gap coincides with a peak event, then the maximum concentration of pollution for your study period will not be recorded (resulting in a false, much lower peak value recorded for the period).
Distribution of available data	Consider a 24-hour period with 50% data completeness. It could be possible that the first 12 hours of that period have *all* data points present, and the second 12 hours have *no* data points present. Alternatively, it could also be possible that the sensing device failed to send *every other* data packet for the entire 24-hour period, resulting in an even distribution of data and gaps. The distribution of available data (or data gaps) is not captured by a data completeness metric, but may be highly relevant to your data use case.

FILLING IN DATA GAPS

A data gap can prevent you from querying an air quality value for a specific point in time. This can be overcome using temporal interpolation.

For instance, if you need to know what the $PM_{2.5}$ * concentration for location A was at 3pm last Tuesday, what should you do if you are missing data for that specific time?

Temporal interpolation enables you to infer or 'interpolate' values for data gaps, using data from either side of that gap. To do this accurately, a minimum data completeness is required. Relatively even data distribution is also important (e.g. you cannot reliably interpolate data for the second half of a day if you only have data from the first half).

Data profiling tools are required to support temporal interpolation.

* PM (particulate matter) refers to airborne solids or liquids. Its size is measured in micrometres and is indicated by the subscript. E.g. $PM_{2.5}$ has a diameter of 2.5 micrometres or less. (NSW Health, 2020)

For further guidance on data completeness, please refer to the OPENAIR Best Practice Guide chapter *Sensing device troubleshooting: common problems and how to fix them*, and the supplementary resource *Sensing device troubleshooting: extended guide*.

Verification of data quality during standard operations

During standard operations of an air quality monitoring network, it may be necessary to verify the quality of data that is being collected. This can improve user trust in the data and its ability to support impact. This activity is distinct from 'initial' data verification carried out during the period of field testing (immediately after a device is deployed).

The most likely scenario for verification of data quality during standard operations is a defined period of focused data collection, where a higher level of certainty about data quality is required to support intended data analysis. This can be done in two main ways:

1. Verify against **external high-performance references**
2. Verify using **internal cross-verification**

A regulatory ambient air quality monitoring station in Wagga Wagga (NSW) that is part of the NSW Government's state-wide monitoring network. Stations like this one produce high-quality reference data that can be used to verify data from low-cost sensors. Image source: Creative Commons

Verify against external high-performance references

Verification of air quality data against an external reference is commonly done in one of two ways:

1. Verify smart low-cost sensor data against regulatory air quality data

Data from a regulatory ambient air quality monitoring station (see the image on page 9) can be used to help verify data from smart low-cost sensors deployed in the same general area. However, the appropriateness of this approach will depend on what you are trying to measure with your own sensors (see Table 4).

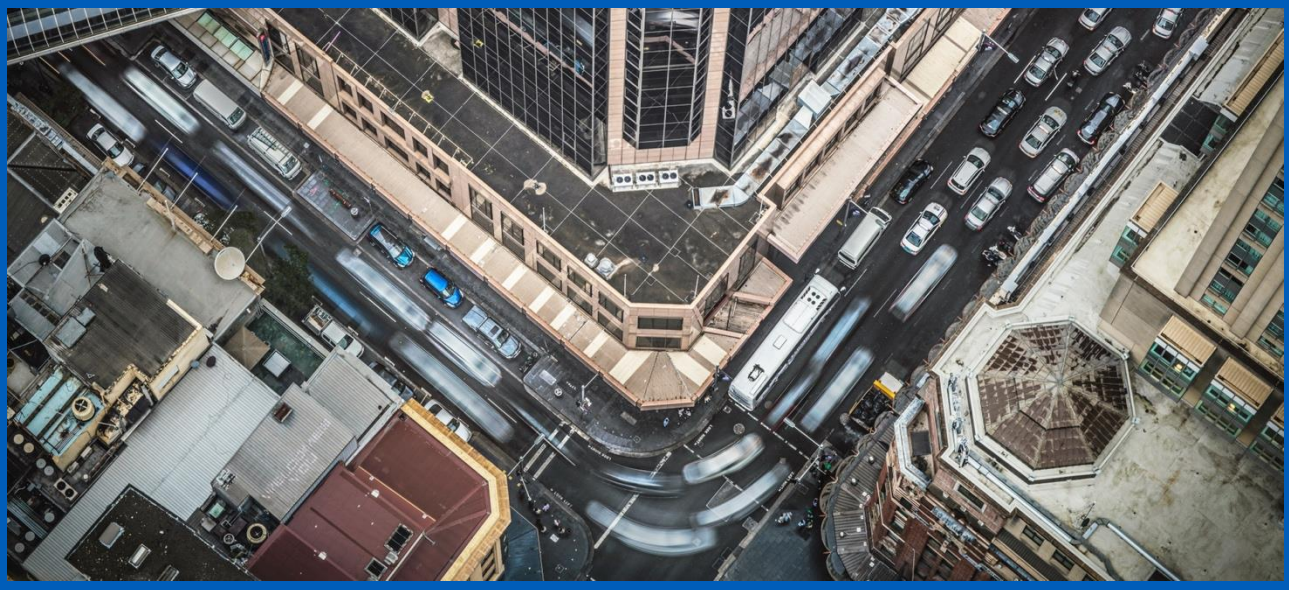
Table 4. Understanding when it is appropriate to verify low-cost sensor data against regulatory air quality data

Works well when…	**Does not work well when…**
Your sensors are positioned to capture representative readings for the wider area, and you *avoid* highly localised emissions and micro-climates. AND There is a regulatory station in close proximity to your sensors, in a location that is roughly comparable to the locations of your own sensors. AND You take an average of data from many devices in your network, and apply longer averaging periods (e.g. 4 hours or more). This reduces the 'noise' associated with individual device deployments to create values that are more representative of a wide area over time. It is then more appropriate to verify these values against regulatory data. IDEALLY, you should also co-locate your devices with your local regulatory station prior to deploying them. This ensures that your devices are calibrated to that station, and that you clearly understand how their performance differs from it.	You are using low-cost sensors to measure a highly localised phenomenon, such as roadside vehicle emissions, or coal dust around rail corridors. In such cases, data peaks and temporal trends may have little to no relationship with regulatory data, making direct comparison unhelpful.

2. Verify smart low-cost sensor data against high-performance portable equipment

Smart low-cost sensor data can be verified against high-quality data obtained from portable equipment. For a variety of reasons, such high-performance equipment is generally not suitable for leaving in place, meaning that data collection is generally restricted to a short, fixed period of time, usually of several hours (though readings may be repeated on multiple days). Data verification activities that use portable high-performance equipment should therefore be designed to meet the specific needs of a project, with the aim of verifying a particular observed phenomenon. See EPA Victoria's air monitoring network for more information.

AN EXAMPLE OF VERIFICATION USING HIGH PERFORMANCE PORTABLE SENSORS



Context

Three smart low-cost sensors are measuring pollution associated with peak hour traffic in an inner-city street canyon. The sensors show pollution peaks that regularly exceed recommended safe levels. This has implications for city planning policy, so there is strong desire to validate the findings.

Method

High-performance portable equipment is taken to the three locations where smart low-cost sensors are deployed, and reference data is captured during peak hour traffic on multiple days.

Results

Reference data shows peaks and trends that verify the smart low-cost sensor data. The smart low-cost sensor data captures longer-term trends (e.g. weekdays vs weekends; or the impact of COVID-19 lockdowns) that could not reasonably be captured with mobile reference equipment alone. By verifying just *some* of your smart low-cost sensor data against high-performance reference equipment, you can now place much greater trust in *all* of your smart low-cost sensor data over time.

VERIFYING AGAINST OTHER TYPES OF EXTERNAL DATA

More indirect external verification approaches can work for certain use cases. Other types of external data that relate to pollution-generating activities can be checked for correlation against trends in smart low-cost sensor data.

Example scenario:

A sensor deployed next to a railway line to monitor coal dust associated with coal trains reports periodic spikes in levels of PM_{10}.

Verification actions:

Acquire **train timetables** and compare these with the timing of reported elevations in PM_{10} from the sensor. If there is close correlation, then it supports validation of a hypothesis (in this example, that coal trains are a likely cause of local particulate pollution) and rules out other possible causes of observed elevations (e.g. diesel particulates from a nearby road).

Internal cross-verification

Internal cross-verification of air quality sensor data involves verifying the data received from one device (or group of devices) against data from another device (or group of devices) in your network. This is most useful for identifying stand-alone devices that develop unique issues (e.g. calibration drift) relative to the rest of your network. It is less useful in cases where there are systemic issues that impact all of your devices equally.

The more devices you have in your network, the more effective cross-verification will be. This is because you can weigh the reliability of one device against the reliability of many devices. However, this approach does have its limitations:

1. It assumes that the sensing devices you are using perform at a standard appropriate to the needs of your use case. Cross-verification can identify devices within a network that have issues (i.e. ones that deviate from the standard). However, if all devices perform poorly, then this approach will not work.
2. It is entirely possible for only one device in a network to record major pollution events that are not recorded by the other devices. Cross-verification identifies outliers, but an outlier can sometimes be an entirely accurate phenomenon. Keeping this in mind, internal cross-verification is a useful tool that should nevertheless be approached with a clear understanding of your specific device deployment context and local conditions. This will allow you to assess what you are seeing based on more than just the numbers.

AUTOMATING VERIFICATION OF DATA QUALITY

Verification of data quality during standard operations can be applied manually (to data that is exported from an IoT or data platform), or it can be automated into a platform and applied to live data streams. Automated verification is an advanced functionality that is likely to be restricted to more sophisticated, enterprise-scale smart city platforms.

Automated data verification using external references will be limited to the use of data sources that are accessible as live streams via an application programming interface (API). This external data can be used to support a dynamic verification function that is applied to incoming data from your own sensors.

Automated internal cross-verification is a platform function that can compare (and potentially correct) every data point received, relative to every other data point received. This may require the use of machine learning applications. It is contingent upon automated temporal interpolation, which allows for data from two or more devices to be compared for a single point in time. See the OPENAIR Best Practice Guide chapter *Data interpretation: analytics* for more on this method.

YOUR DATA INTERPRETATION JOURNEY: CHECKPOINT 2

At this point in your data interpretation journey, you should check that:

- your device network is installed, and all field testing and verification of data is complete, giving you a baseline of trusted and reliable data
- you have clean and appropriately complete data sets available and ready for analysis
- your data set quality has been verified, making it ready for analysis.

TOOLS FOR DATA CLEANING AND QUALITY CONTROL

Data cleaning and quality control can be undertaken using a range of different software packages, including open-source software (such as R Studio), or commercially available software (such as MatLab and Microsoft Excel).

References

NSW Health. (2020). *Particulate matter (PM10 and PM2.5)*. NSW Government. https://www.health.nsw.gov.au/environment/air/Pages/particulate-matter.aspx#:~:text=PM10%20(particles%20with%20a%20diameter,and%20cause%20serious%20health%20effects.

Associated OPENAIR resources

Best Practice Guide chapters

Data interpretation: overview

This Best Practice Guide chapter provides guidance for interpreting data produced by smart low-cost air quality sensors. It outlines the three main stages of the process (data correction and harmonisation; data quality control; and data analysis), explores the relationship between data interpretation and impact creation, and supports the planning of a data interpretation strategy.

Data interpretation: correction and harmonisation

This Best Practice Guide chapter provides guidance for correction and harmonisation of data produced by smart low-cost air quality sensors. It introduces several types of correction factor that may need to be applied to raw sensor data, and explores how data formatting and labelling should be harmonised with a project data schema to support effective data management and sharing.

Data interpretation: analytics

This Best Practice Guide chapter introduces common analytical approaches that can be applied to data produced by smart low-cost air quality sensors. These include statistical analysis; temporal interpolation; spatial aggregation and interpolation; complex geospatial system modelling; and AI and machine learning applications.

Air quality sensing device activation and deployment

This Best Practice Guide chapter provides guidance for activating and deploying smart low-cost air quality sensing devices.

Sensing device troubleshooting: common problems and how to fix them

This Best Practice Guide chapter introduces a framework of common problems that can arise with smart low-cost air quality sensors and the provision of useful data. It includes some practical information to help diagnose issues, fix them, and mitigate against reoccurrence.

Supplementary resources

Air quality sensing device activation and deployment checklist

This resource provides extended guidance for activating and deploying low-cost smart air quality sensing devices. The process begins with onboarding and configuration of devices, includes testing and installation, and ends with commissioning.

Sensing device troubleshooting: extended guide

This resource presents an extended, systematic list of problems that can arise with smart low-cost air quality sensors and the provision of useful data. It includes practical information to help diagnose, fix, and mitigate each type of issue.

Further information

For more information about this project, please contact:

Peter Runcie

Project Lead, NSW Smart Sensing Network (NSSN)
Email: peter@natirar.org.au

This Best Practice Guide section is part of a suite of resources designed to support local government action on air quality through the use of smart low-cost sensing technologies. It is the first Australian project of its kind. Visit www.openair.org.au for more information.

OPENAIR is made possible by the NSW Government's Smart Places Acceleration Program.

Document No: 20231030 BP406 Data interpretation: quality control Version 1 Final