OPENAIR

Operational Network of
Air quality Impact Resources

Best Practice Guide

BP407 | Manage and analyse

Data interpretation: analytics

Introduction

To support activities that create impact, raw sensor data needs to be interpreted. The first two steps in this data interpretation process involve correction and harmonisation, and quality control. The final step is data analysis, which is the interrogation and utilisation of corrected, harmonised, and quality-controlled data to produce insights, or support operational outcomes. These, in turn, support data-driven actions, project outcomes, and impact creation. This chapter aims to help you with some of the main considerations for analysis of data produced by low-cost air quality sensors. It provides practical advice relating to the following key methods of data analysis:

- statistical analysis
- temporal interpolation
- spatial aggregation
- spatial interpolation
- heterogenous data synthesis
- atmospheric pollution dispersal models
- digital twins
- AI and machine learning applications.

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 fourth 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 quality control) in the order shown in Figure 1.

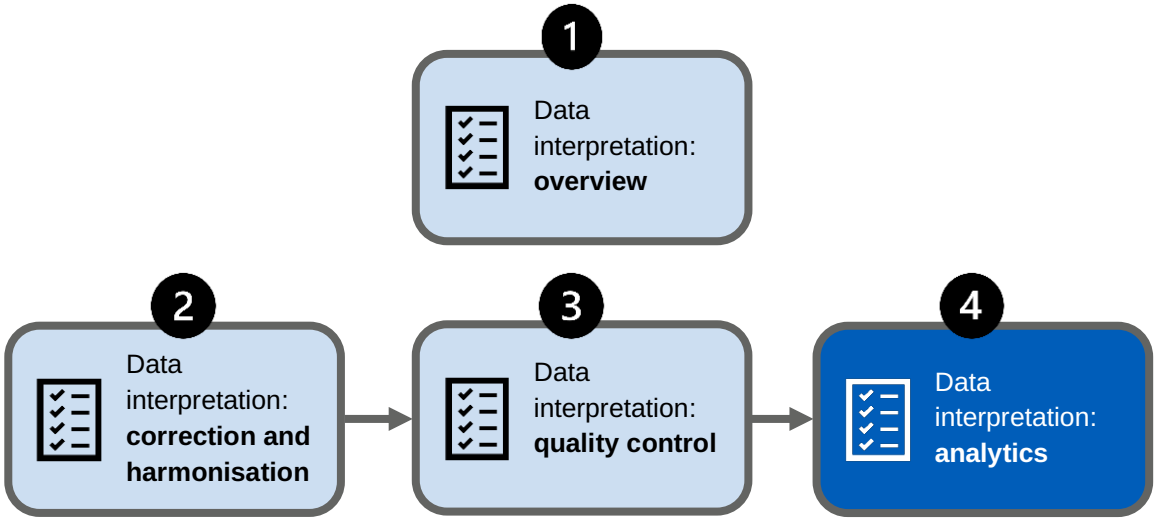


Figure 1. OPENAIR data interpretation Best Practice Guide chapters

Key messages

The key messages of this chapter are:

- Data analysis can require a range of approaches, of varying complexity. While analysis is always necessary to some degree, the specifics of your data use case will dictate your requirements in terms of effort, resources, and expertise.
- Data analysis can be applied manually (to exported static data sets), or it can be built into the functionality of an IoT or data platform with varying degrees of sophistication. It is a good idea to consider your data analysis requirements as part of your technology procurement decision-making process.
- Data analysis platforms with specific advanced functionality are often integrated with a data platform, as part of a more advanced data architecture. For further guidance, see the OPENAIR Best Practice Guide chapter *IoT reference architecture for smart air quality monitoring*.

Statistical analysis

Statistical analysis involves interrogating a set of data to identify patterns and trends, around which insights may be established. These insights can then support evidence-based actions.

What is it used for?

Statistical analysis is a broad term that covers a diverse range of techniques of varying complexity. Some common techniques that are useful for analysing air quality data from low-cost sensors are:

- descriptive statistics (e.g. to calculate mean, median, standard deviation, etc.)
- regression analysis (e.g. to examine the relationship between two or more variables)
- T-test (e.g. to compare the means of two samples).

Practical approaches

Statistical analysis may be conducted on static data sets that are exported from your database (a form of 'manual' analysis), or incorporated into your IoT or data platforms and applied as a rolling real-time function (a form of 'automated' analysis, e.g. a 24-hour rolling average). If you are working manually, then it is recommended that you use specialist software to support statistical analysis, such as Microsoft Excel, R Studio, SPSS, or MatLab.

VISUALISING DATA TO SUPPORT STATISTICAL DATA ANALYSIS

There are several basic data visualisations that can support manual data analysis by:

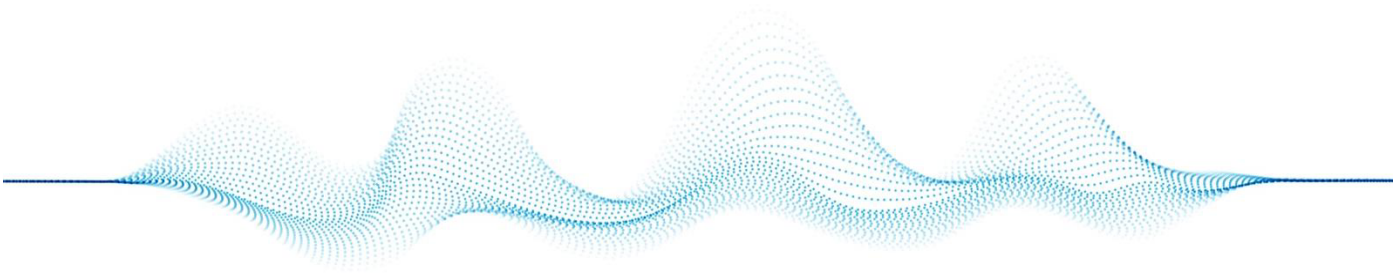
- translating large amounts of information into more manageable formats that are easier to interpret
- providing an initial understanding of a data set to support the development of a more thorough data analytics strategy
- summarising the outcomes of your statistical analysis.

Summary or 'Pivot' table	• A simple overview of data, often based on statistical summaries (e.g. mean, median, or standard deviation) • Allows comparisons to be made between categories of data • An efficient way to represent data that allows interrogation of processed data sets (rather than individual data points)
Scatter plot	• A graphical summary of data that displays the relationship between two or more categories (e.g. $PM_{2.5}$* concentration and time) • Supports visual evaluation of trends • Can be used to calculate regression parameters, such as slope
Box and whisker plot	• A graphical summary for groups of data that shows the group's position and distribution, relative to a larger data set • *Example*: for a group of data containing one full day of $PM_{2.5}$ readings from a device, the plot would show the mean, minimum, maximum, median, and 1st and 3rd quartiles for that day
Histograms	• A graphical summary of data that is useful for comparing levels of pollutants, and condensing a data series into an easily interpreted visual (by taking many data points and grouping them into ranges or 'bins')

* $PM_{2.5}$ refers to airborne particulate matter (solids or liquids) that have a diameter of 2.5 micrometres or less. (NSW Health, 2020)

Temporal interpolation

Interpolation is the process of inferring unknown values from adjacent data. Temporal interpolation infers values for a fixed point in time, using data from before and after that point.



What is it used for?

There are two main uses of this method of analysing data using temporal interpolation:

1. To compare data from multiple devices at a single point in time

Temporal interpolation allows you to directly compare data from multiple sensing devices for a fixed point in time. This is helpful because devices in a network do not synchronise their transmission of data packets (which means they have random, unaligned timestamps). Temporal interpolation allows you to infer a value associated with each device for a fixed point in time. This provides a critical foundation for data verification, and makes a wider variety of insights possible from statistical analysis (e.g. exploring the correlation of data between two or more locations).

2. To fill gaps in your data set

Temporal interpolation is also valuable for filling in the gaps in a data set. This is a common issue with low-cost sensing, where data packets are often lost due to marginal data communications. As such, temporal interpolation is often a foundational step in data processing, and one that supports higher-order analytics activities.

Practical approaches

Sensor data that is exported from an IoT or data platform tends to have raw timestamped values that will require manual temporal interpolation. A variety of data management and analytics platforms can be used for this (e.g. Microsoft Excel).

Many IoT and data platforms incorporate temporal interpolation as a standard automatic function (e.g. many graphing tools will provide interpolated X-axis values as you scroll over them). This function is worth checking for when you are procuring a platform solution for your air quality monitoring project. Any advanced analytics or automated data verification functionality will be reliant upon inbuilt temporal interpolation.

INTEGRATING SENSOR DATA WITH YOUR GIS PLATFORM

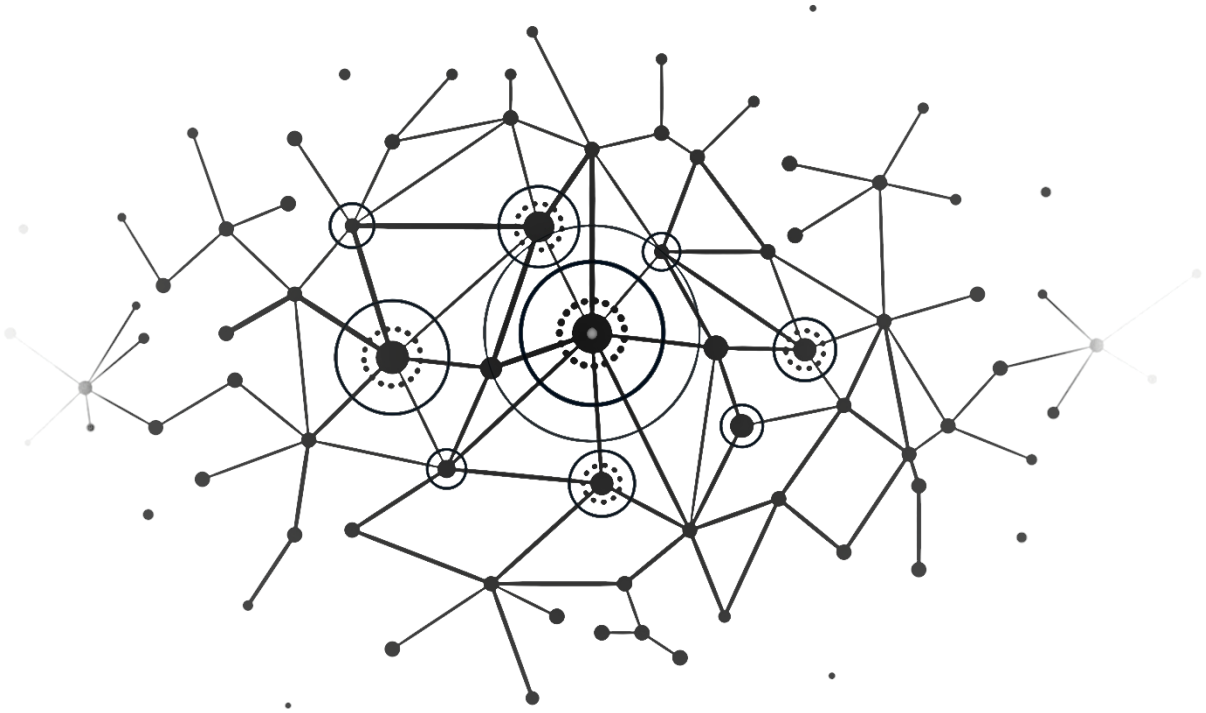
Any sensor data that is associated with geospatial co-ordinates (longitude and latitude) can be entered into a geospatial model. This lets you explore its spatial relationship with other data. The most common types of geospatial models are Geographical Information Systems (GIS), which are used by most local governments at an enterprise level.

Most major GIS providers now operate cloud-based services that are capable of ingesting and modelling live data streams, whether they are from your own sensor network, or from a third-party source (such as the Bureau of Meteorology). Speak with your IT department about your current GIS service package, and explore options for integrating your sensor data. A variety of advanced modelling options may be available.

One simple but useful integration involves connecting a GIS platform with your IoT platform, but only synchronising device metadata (e.g. type of sensor, co-ordinates, date of deployment, asset cost and lifetime, etc.). This creates an asset registry within your enterprise GIS system, allowing all staff to pull up a map layer that shows what sensing devices you have, and where they are deployed.

Spatial aggregation

Spatial aggregation is where data from multiple devices in a given area is combined. This provides various average values for air quality in that area.



What is it used for?

Spatial aggregation is useful for two main reasons:

1. Building a representative view of air quality for an area

Regulatory ambient air quality monitoring involves the collection of data in highly specific contexts that minimise interference from localised phenomena. This allows for the capture of representative air quality measurements that can be associated with the surrounding area. An individual low-cost sensing device deployed in the urban environment, however, *can* pick up on localised pollution sources. It will also be subject to complex interplays of environmental factors relating to air movement, heat, humidity, and air chemistry. It can thus be difficult (or impossible) to build a more representative understanding of air quality for a wider area, based on one low-cost sensing device alone.

On the other hand, one of the key advantages of low-cost sensors is that you can deploy many of them across a study area, allowing you to capture a diversity of conditions. By aggregating data from multiple devices with different locations, you can calculate a more representative indication of pollutant levels for an area. This effectively lets you smooth over the various localised extremes and outliers, and focus on average conditions for the network. This also allows you to compare your low-cost sensor data more easily with data from a nearby regulatory sensor. Note that a degree of caution is advisable with this approach, as you will still be comparing data generated using highly different methodologies.

2. Decoupling sensor data from specific addresses to support safe data sharing

Spatial data aggregation associates air quality data with an area, rather than with single spatial co-ordinates. This is valuable in the context of data privacy and ethics, particularly when it comes to matters of data sharing. You may think that most air quality data does not carry privacy or ethical concerns, but that is not always the case. Consider the following scenarios:

- You are running a citizen science project with local schools. Children take sensors home and run them outside. Data appears on a central dashboard. Anyone with access to that dashboard now has a map of where all the children live. Who has access to the dashboard? What if the project includes an exploration of open data release?
- A local government is using a small network of low-cost particulate sensors to study woodsmoke in a valley. Woodsmoke and air quality are highly political hot-button issues for the community, with strong feelings on both sides. The local government wishes to publish the collected data, but is concerned about associating it with specific streets or addresses. What if people jump to (possibly incorrect) conclusions about pollution sources, leading to community tensions?

In both of these scenarios, spatial data aggregation is a viable solution that protects individuals and groups, and may be a prerequisite for open data release. For the school example, values could be generated for suburbs, based on data from multiple children. For the woodsmoke example, general hotspots could be identified, without specifying individual addresses.

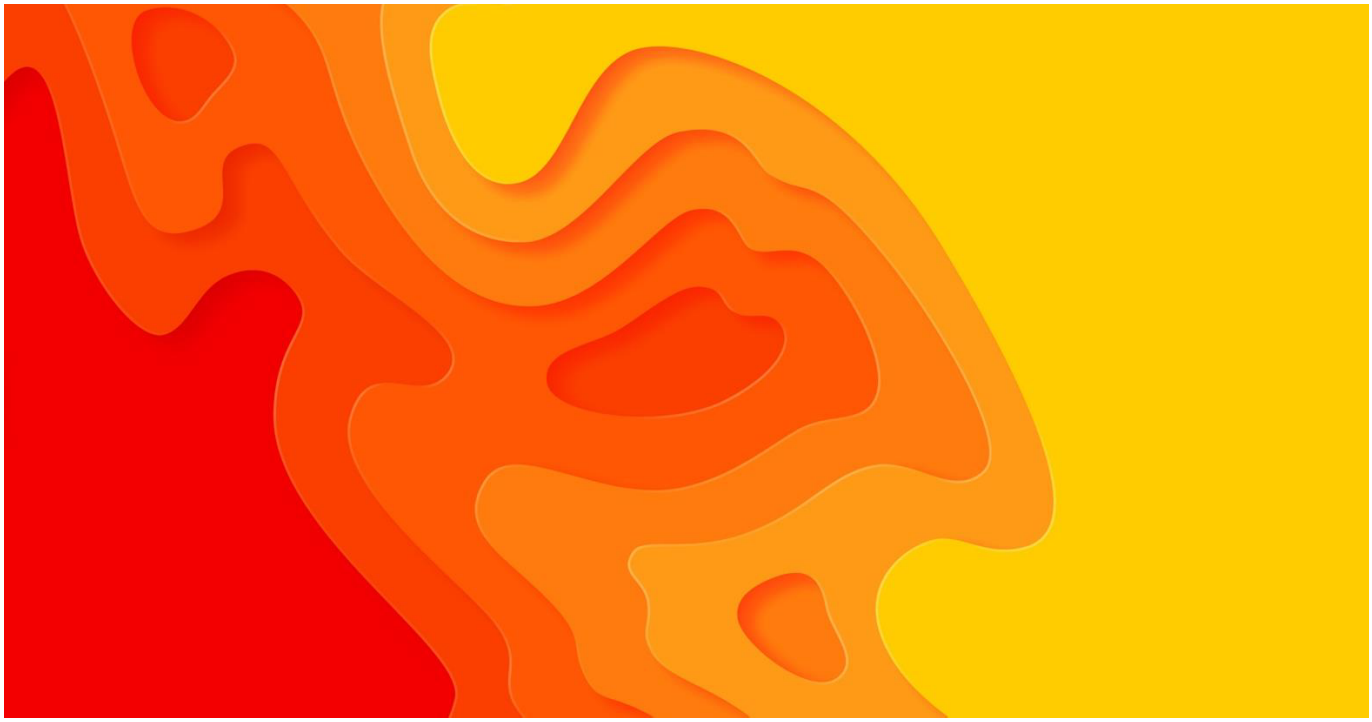
Practical approaches

Spatial aggregation of sensor data can vary greatly in sophistication. At its simplest, it can be done using basic spreadsheet software to combine and summarise data from multiple devices in a network.

More sophisticated approaches require the use of geospatial modelling platforms (such as Geographical Information Systems, or GIS), which can add in things like a weighting factor for device data, based on the device's spatial position relative to other devices in the group. This can help to correct for the effects of uneven distribution of sensing devices on a map.

Spatial interpolation

Interpolation is the process of inferring unknown values from adjacent data. Spatial interpolation infers values for a point on a map where there is no data, using data from other points in the area where there are sensors. More sophisticated forms of spatial interpolation can factor in a range of other spatial variables (such as topography, vegetation, or ground cover) to provide more accurate inferences.



What is it used for?

The most common output of spatial interpolation is a heat map (also known as an ‘isopleth’ or ‘contour’ map). This is a visual representation of a variable (which could be temperature, air quality, or any other metric) across an area. It allows you to quickly understand complex spatial patterns at a glance, as well as to see what the inferred value is for any point on the map (regardless of whether there is a sensor deployed there).

Heat maps are powerful visual tools that can also incorporate a temporal dimension (e.g. a dynamic sequence of heat maps, capturing shifting conditions over a day or year). They can help you determine pollution sources and hotspots, pollution dispersal, and vulnerable communities and receptor sites (e.g. childcare centres that fall within hotspots). Finally, spatial interpolation can support air quality alerts related to pollutant threshold breaches. These alerts can be locally specific, down to the suburb or even street level.

Practical approaches

Beware of poor-quality heat maps

Many commercially available IoT and data visualisation platforms include heat maps as part of their suite of functions. However, many of these will not provide you with accurate outputs if you use them with air quality and urban heat data. You should be cautious about the use of heat maps, particularly if you plan to base policy or services on them.

Why are accurate heat maps for air quality and urban heat so difficult to achieve?

Environmental variables like ambient heat and air quality are extremely complex, and are thus difficult to spatially interpolate. They vary relative to other spatial variables (such as topography, vegetation, or ground cover). In order to accurately spatially interpolate sensor data, a heat map model must understand sensor data *relative* to these other variables, and then correct an interpolated value for another location, based on what those variables are at that location. Most basic commercial heat maps do not do this.

Is it possible to generate an accurate heat map for air quality and urban heat?

An accurate heat map needs to be generated within a GIS platform, or other advanced spatial model. This ensures that you have the capability to work with multiple spatial variables, and access to the right third-party geospatial data sets (e.g. elevation, 3D building models, etc.). Internal cross-verification of heat map accuracy is advised (where the model 'infers' values for a point where you have sensor data, and checks to see how well it predicted these). The leading approaches are now incorporating machine learning to adapt models based on these types of feedback loops.

If this sounds like a complex process, it is – thus very difficult to do well.

Note that accurate heat mapping of air quality data is not a standard function for GIS platforms, and will likely need to be developed as a custom delivery for your project.

Heterogenous data synthesis

Heterogenous data synthesis refers to the modelling of one variable, using multiple 'heterogenous' data sources as an input. Harmonisation of these data sources may be quite complex, due to fundamental differences between them.

Heterogenous data refers to two or more data sets that relate to the same environmental variable (e.g. $PM_{2.5}$), but have fundamental differences. This means that they cannot be directly compared without a degree of complex modelling.

Examples of fundamental differences between air quality data sources include:

- data quality
- fundamental design of a sensor (e.g. a sensor that directly measures the mass of particulate pollution differs from a sensor that measures light diffusion by particulate pollution)
- design of a device (e.g. a particulate monitoring device that features a heated air intake will perform differently to one that does not)
- sampling interval
- reporting interval (or available averaging time)
- how data has been corrected, interpreted, or abstracted.

What is it used for?

In the context of local air quality monitoring, the production of heterogenous data sets for a single variable commonly occurs in the following circumstances/scenarios:

- **Hybrid device networks.** You are running a monitoring network comprising more than one type of low-cost sensing device. For example, you might run many low-cost devices, and a smaller number of higher-performance devices in key locations.
- **Regulatory data inclusion.** You are running a network of low-cost sensing devices, and there is a regulatory ambient air quality monitoring station nearby. You can ingest a live data stream from the regulatory station into your own platform.
- **Fixed + mobile**. You are gathering data from sensing devices deployed at *fixed* locations for an extended period, and from *mobile* monitoring activities (e.g. vehicle-mounted devices, or wearables).

In any of these scenarios, heterogenous data synthesis can be used to bring together data from different sources and with different fundamental attributes, and build up a single picture of air quality for an area.

Practical approaches

Heterogenous data synthesis requires a relatively complex model that treats raw data from all sources as inputs, accounts for fundamental differences, and generates a single type of data as a derived output. Depending on the capabilities of your chosen analytics platform, it may be possible to do this modelling in real time. Alternatively, you may need to run a model using static data sets that are exported from your live data environment.

Atmospheric pollution dispersion models

An atmospheric pollution dispersion model predicts where (and how fast) air pollution disperses from a source.

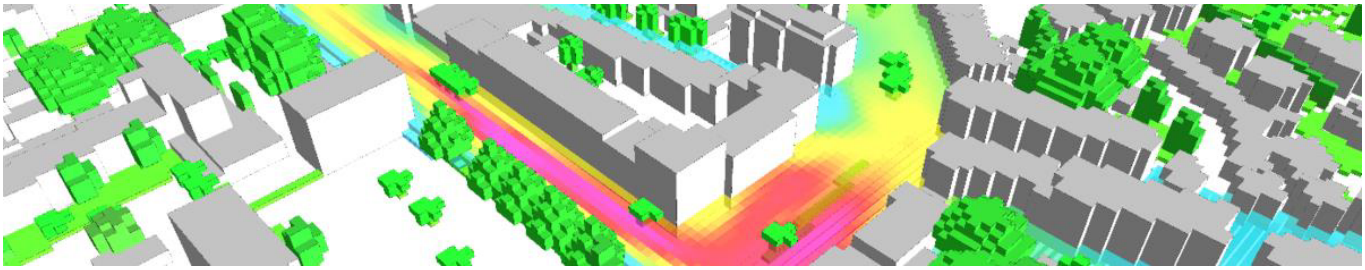


Image source: Creative Commons

What is it used for?

Dispersion models can be applied at various scales. At larger scales, they can be used to predict the behaviour of smoke plumes from bushfires or industry. At smaller scales, they can model the localised dispersal of vehicle emissions, and other urban pollution sources.

A fundamental benefit of dispersion modelling in the low-cost sensing context is that it reduces reliance on blanket sensor coverage. Data from a few devices can be leveraged to generate insights that benefit everyone in a community, rather than being restricted to a handful of deployment locations.

A dispersion model can be used as an advanced, multi-variable approach to spatial interpolation (a method discussed earlier in this chapter). This allows for inference of pollution concentrations for a location that lacks sensor data, based on other information (such as wind speed and direction, rainfall, and topology). This is one way to produce more reliable and accurate heat maps for air quality.

When applied to live data streams, a dispersion model can be used as part of a 'nowcasting' model, which supports very short-term predictions for the following 2 to 6 hours. These capabilities can support geographically targeted alerts, making dispersion models critical tools for public health services. They also play a vital role in generating regional air quality metrics, where fewer sensors often mean a greater reliance on models to fill in the gaps in sensing networks.

Practical approaches

Most commonly used air pollution dispersion models are based on something called 'gaussian plume dispersion.' A Gaussian model can be used to calculate dispersion of pollution from a single point source (e.g. a chimney stack). There are free online resources available, such as air pollution dispersion models for chimney stacks, which may be useful for investigating some local air quality issues.

Gaussian modelling can be incorporated into a more complex geospatial model that generates 2D or 3D dispersion heat maps for an entire area. Such models are not standard for smart city analytics platforms, and a specialised stand-alone platform is likely to be required.

Digital twins

A digital twin is a virtual representation of a place (or area) that is connected to multiple real-time data sources, and runs over an extended and ongoing period of time. It supports the investigation of complex relationships between multiple systems and processes, and the simulation of 'what if' scenarios.

A digital twin includes a three-dimensional model of the physical environment. It applies complex, multi-variable simulation of environmental conditions and related processes to produce system-level insights.

A digital twin simulation will often bring together the outputs of multiple parallel models for specific sub-systems (such as weather, pollution dispersion, traffic congestion, crowd movement, energy demand, or urban planning). Machine learning may be used to improve a digital twin over time.

What is it used for?

A digital twin allows you to simulate highly complex relationships between multiple, interconnected urban systems. The potential insights and impacts that digital twins might be able to support are only just emerging, but already hold great promise. Some possible scenarios in three different sectors are described here.

Transport planning

A digital twin could help to answer important transport-related questions, such as what happens to air quality in a residential precinct if a nearby highway is widened; what impact a change of bus routes

might have on the air quality in an inner-city building canyon; or what the optimal design might be for a new low-emission zone.

Public health modelling

If a digital twin were to incorporate a health impact model, an air pollution model, and an urban planning model, then it could predict health outcomes based on planning decisions.

Green infrastructure planning

If urban planning and health impact modelling are combined with digital twin models that relate to air pollution mitigation by trees, then you could plan future green infrastructure with respect to air pollution mitigation.

Practical approaches

Digital twins that can work with local air quality data as part of a broader, complex system do exist as commercial options, and at the time of writing can be procured through a standard procurement process. However, these systems are still relatively new and evolving rapidly, so a degree of caution is advised if they are intended to inform core business operations or policy.

GET AIR QUALITY INCLUDED IN YOUR ORGANISATION'S DIGITAL TWIN AGENDA

Many local governments are choosing to invest in enterprise-scale digital twins, and it is quite possible that your own organisation is already considering options in this space.

It is worth exploring if this is the case, and – if there are discussions about setting an agenda for that digital twin – engaging with relevant colleagues about your air quality project. Not all digital twin platforms have air quality modelling as part of their core functionality. You may be able to get air quality included as a priority focus for a future digital twin, and influence the technology procurement process.

AI and machine learning applications

Artificial intelligence (AI) and machine learning (ML) are advanced data analytics capabilities that can augment and enhance an existing data analytics model. They can make use of low-cost sensor data to produce a range of outputs.

Artificial intelligence is an umbrella term that includes a broad range of decision-making and task-oriented functions based on data inputs. Machine learning is a subfield of AI that is increasingly seen in smart city applications. The most common application of machine learning is to improve the performance and accuracy of a data analysis model over time, through the continued ingestion and accrual of data.

What is it used for?

Machine learning can be applied to air quality sensor data in several different ways.

One concrete example is that machine learning can continually improve the accuracy of a near-term forecasting model for air quality. The model might predict future sensor readings based on all past readings, past weather conditions, and a near-term weather forecast.

A machine learning program compares that prediction to actual sensor data, as it arrives. If there is a difference between the prediction and reality, the machine learning program will alter small aspects of how the predictive model works, and then check its predictive accuracy against new data. If the prediction is more accurate, the machine learning program will tweak the model further, in the same direction. If the prediction is poorer, the program will reverse the previous alteration and try something different. In this way, the model iteratively improves, or evolves, over time.

Practical approaches

Currently, AI and ML applications for low-cost sensing are advanced, custom capabilities that tend to be possible only through partnerships with research institutions and a small number of commercial platform providers. Case studies are limited, and the use of these technologies is in its infancy, but the field is rapidly growing and changing.

THE FUTURE OF AI – EXPECT RAPID CHANGE
Artificial intelligence has seen major progress recently, and large language models such as ChatGPT have exploded in both popularity and their range of applications. It seems highly likely that these technologies will intersect with low-cost sensing in the coming years. They have the potential not only to significantly improve advanced analytics capabilities for air quality data, but also to make these capabilities accessible to local governments running sensor networks.

YOUR DATA INTERPRETATION JOURNEY: CHECKPOINT 3

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

- you have selected and applied a data analysis approach (or more than one, if relevant) that meets the needs of your planned data use case
- you are incorporating early insights from your data analysis into your ongoing project strategy (e.g. identifying a need for additional data, or a change to the way you are collecting data) to enable analytics outcomes that deliver impact.

References

NSW Health. (2020). *Particulate matter (PM10 and PM2.5)*. NSW Government. https://www.health.nsw.gov.au/environment/air/Pages/particulate-matter.aspx

Additional resources

***David Carslaw, Jack Davison and Karl Ropkins* | *Openair: open source tools for air quality data analysis*†**

An advanced set of open-source data analytics resources, including tools for importing, manipulating, and undertaking analysis of air pollution data. Although focused on UK air quality monitoring stations, it has useful tools for estimating model statistics and generating quantile plots. (Please note that the Openair software package is not related to the OPENAIR project that produced this document).

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: quality control

This Best Practice Guide chapter provides guidance for the quality control of data produced by smart low-cost air quality sensors. Data quality control helps to isolate trusted data that can then be used to support chosen activities. This chapter explores approaches for cleaning static data sets to prepare them for analysis, and approaches for operational verification and quality control of live data streams.

IoT reference architecture for smart air quality monitoring

This Best Practice Guide chapter introduces the OPENAIR reference architecture for smart air quality monitoring. The reference architecture is a framework that identifies the various components and data flows that make up a complete technical solution for smart air quality monitoring. It is a generic reference that can help local governments to design and implement their own technical solutions.

† Despite sharing a name, this UK-based resource has no affiliation with the Australian OPENAIR project.

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: 20231107 BP407 Data interpretation: analytics Version 1 Final