Australia's National
Science Agency

Sydney Particle Study

Educational Datasets Teachers Guide

Year 7-10

Contents

1 Dataset Overview

CSIRO Educational Datasets

CSIRO Educational datasets have been derived from CSIRO research data and adapted for classroom use. They are delivered in three different levels; Novice, Expert and Programmer.

Novice level data has been simplified for the classroom. Potentially confusing outliers and partial entries have been removed from the data for the novice level, to make analysis and comprehension easier. Data labels have been modified to make them easier to understand.

Expert level data retains outliers and partial entries and has not always had the labels adjusted. This may mean that students are required to undertake research into subject language to fully understand what they are looking at. Both Novice and Expert level data contains a limited number of rows to ensure that they can be opened in spreadsheet packages.

Programmer level datasets and activities are intended to be used with more advanced tools and programming languages. This level provides the data in an unmodified format, allowing students to organise and analyse it independently.

Dataset Description

These datasets contain records of weather conditions and concentrations of different air pollutants, recorded at Westmead in Sydney in 2011. Many of the referenced government bodies are New South Wales agencies, since this project was based in NSW and much of the monitoring and management of air quality in Australia is done at a state level. However, the standards used in NSW are based on federal guidelines set by the Australian Department of the Environment and Energy. Each group of records covers approximately a month of recordings, with one set of recordings taken during Summer and another taken during Autumn in 2011.

The dataset was recorded with a system of sensors which measure the concentration of various substances in the air, such as particulates (small particles of dust, soot, pollen etc), Sulphur Dioxide (SO_2), Nitrogen Dioxide (NO_2) and others. In addition to these measurements, the system records weather conditions such as temperature, wind speed and direction, humidity, and solar intensity. The values recorded in these datasets are hourly averages, not individual measurements. This means that very brief highs or lows will not be readily apparent in the recorded data.

For pollutants like the ones measured here, the NSW Environmental Protection Agency (EPA) specifies maximum concentrations based on their known impact on human health. For most pollutants, these concentrations are specified as averages over a day, rather than a single peak value. For example, the maximum concentration of Sulphur Dioxide is an average of 0.08 parts per million (ppm) over a given day. However, the permitted maximum average concentration over a single hour is 0.2ppm, and over a year it is 0.02ppm. The EPA considers the standards for Sulphur Dioxide to be met if the hourly and daily limits are only exceeded once per year.

For a link to the original data in the CSIRO Data Access Portal, see Appendix A.

Understanding this Dataset

This section relates to understanding this specific dataset. For more general information on understanding and interpreting datasets, see the Educational Datasets Companion document.

The values in each column are the average of the values recorded over each hour, so it is possible that individual recorded values may have been much higher, or much lower, than the final average.

The pollutant concentration standards set by the EPA vary considerably depending on the pollutant and are periodically updated to reflect new information about the toxicity of pollutants. The standards for some pollutants are over different time periods to others, for example the limit for Carbon Monoxide (CO) is an average of 9.0ppm over 8 hours, and the limit for Ozone is 0.1ppm over 1 hour and 0.08ppm over 4 hours.

The most important thing to know about the standards is that they never refer to instantaneous readings – so sensors may record a level much higher than the standard over a short period of time without a breach if the average over the stated period is lower than the given value.

This means that while the maximum permitted limit for Ozone is 0.1 parts per million over one hour, you could see a reading of 0.2 or 0.3 over a few minutes that does not breach the standard, if the average of all readings for the whole hour is less than 0.1. In class, this is an opportunity to explore the difference between instantaneous and average values, and to consider the health impact of a short-term increase over a long-term increase in concentrations.

For each of the pollutants, the allowed average over a long period of time (for example a year) is lower than the allowed average over a short time. For example, for PM_{10} (particles under 10µm in diameter), concentration over an hour is permitted to average 50 micrograms per cubic metre (µg/m^3), but over a year the concentration is only allowed to average 25µg/m3. For PM_{10} if the hourly maximum value is exceeded at any time during the year, the standard is considered to have been breached. Alternately, Carbon Monoxide's maximum value can be exceeded for one day a year without it being considered a breach.

Restrictions on particulates are more rigid because of their potential health risk. The limits for $PM_{2.5}$ (particles smaller than 2.5µm) concentrations are significantly lower than for PM_{10}. This is because these particles are small enough to be absorbed into the bloodstream through the lungs, where they can cause and aggravate a range of health concerns including asthma and cancer. There is no 'safe' level of particulate pollution that has no health risk, so the allowable limits are a balance between what is achievable in an industrial society and the impact on health.

The NEPH column in the data (measured in bsp) is from a device called a nephelometer, which measures the concentration of particulates in a gas or a liquid by measuring how light is scattered in the sample. With more particulates in the air, more light is scattered as it bounces off particles.

Research Findings

The initial research undertaken with this data aimed to improve understanding of the particle sources that the population of the Sydney region are exposed to and provide a qualitative model of those sources. The major focus of this study was on particle pollution and ozone pollution, with a specific focus on fine particles, less than 2.5 microns in diameter (PM2.5).

The study found that organic matter, including particles from car exhaust, made up the largest percentage of all fine particles recorded: 57% during the autumn observations and 34% during the summer observations. The summer observations also recorded a very high number of particles (34%) caused by sea salt emissions from breaking waves. Additionally, it was found that the release of Volatile Organic Compounds from vegetation was a major source of secondary organic particles in summer, while wood heaters were the dominant source during autumn.

These results were used by the NSW Office of Environment and Heritage to build more advanced particle models, with the goal of using these more advanced models to help guide NSW air policy.

For more information about this research and a link to the reported findings, see Appendix A.

Learning Goals

As with any lesson resources, there are any number of ways this dataset could be brought into the classroom, depending on your approach and personal style. Here you'll find some potential overarching learning goals, most of which address general data literacy, understanding and representation to guide you in introducing this dataset to your students.

Understanding this dataset

Students examine simple ways of exploring datasets to understand them and discuss the positives and negatives of using a specific dataset. In achieving this learning goal, some activities might include:

- Sorting the data. Different trends become more obvious once the data has been sorted in certain ways. For example, how quickly do readings drop off after the maximum values? How quickly do they rise again after reaching minimum values? What does this tell us about the source of pollutants?
- Averaging. Knowing that the data in this dataset is all averages of readings, does that limit what we can find out at all? What does the average not tell us? Is it possible for two wildly different sets of readings to have the same average value?
- Mean vs Median. When taking the mean and the median of a dataset, it's possible to get two different results. What does this mean? Why are they different? Which one is a better indicator of the centre of the dataset? In this case, which value is more useful to us?
- Graphing. What kinds of graphs can we use to represent this data? Are there any subsets of the data that might be useful to compare on a graph?

Accurately report findings made from data

Students examine how to best represent their findings from the dataset. How can we display this data so that it can be easily read and understood? Representing the whole dataset in a single table can make it difficult to identify trends and link related concepts. Using statistical tools, such as using the average, range, median, mode or percentages can help give the audience a better idea of what the data tells us, but some of these values are more useful than others, depending on context. If you're packing for a trip, the range of temperatures for each day is more important than the median temperature for the whole trip. Knowing that the temperature will get as high as 27°C and as low as -2°C is more important than knowing that the median temperature will be 13°C, as it gives you a much better idea of what to pack.

With this dataset, consider if it is useful to represent pollutants as a percentage of total air volume on a pie chart. Even if carbon monoxide content exceeded the 8-hour limit by 10 times, 99.99% of the pie chart would represent clean air, giving no indication of how safe the air is to breathe. It's important to consider the purpose of a visualisation, in terms of the story it presents the viewer.

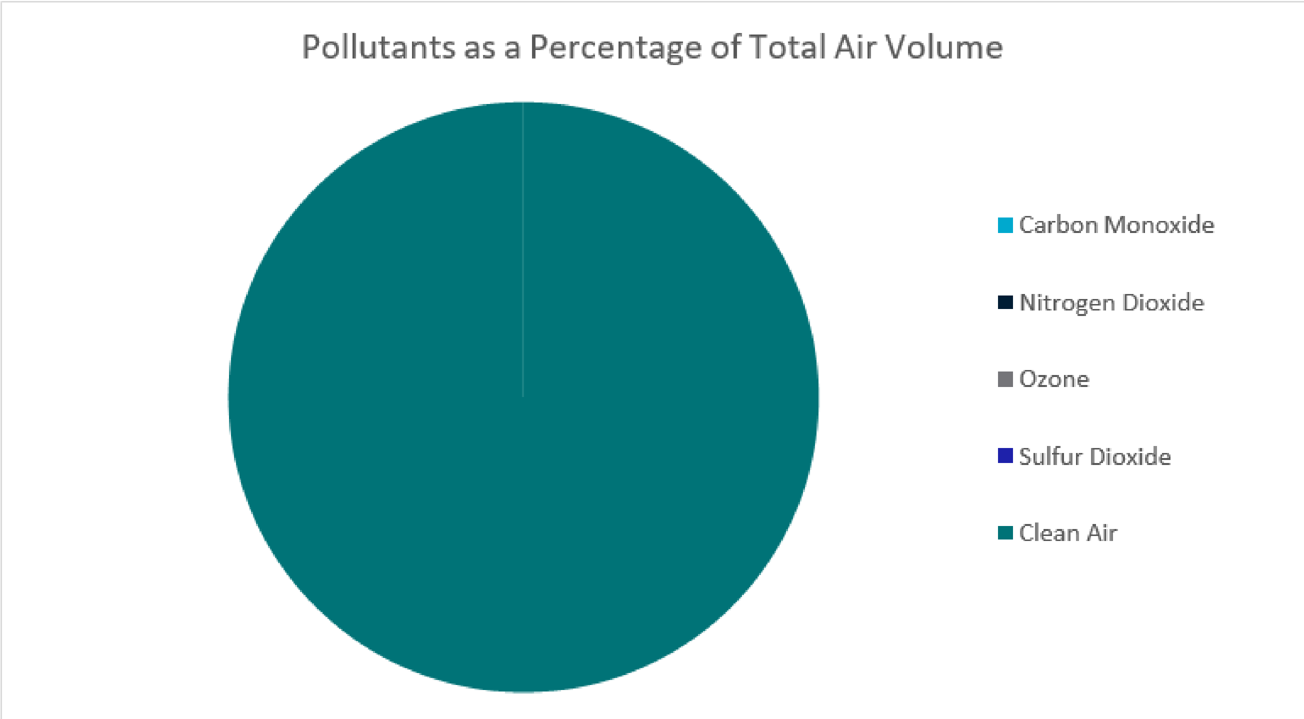

Figure 1 - Pie chart indicating pollutants as a percentage of total air volume. Listed pollutants are ten times over allowable limits. The only segment of the pie chart that is visible represents clean air, making it difficult to get information from this visualisation.

What else can we do to make sure that the findings we're reporting don't skew the data or misrepresent it? Examine ideas such as considering the whole data set, not just favourable sections of it, or ensuring that when using two graphs to compare data, they use the same start and end points, the same scales, and the same display ranges.

Alternately, you could reverse this lesson idea and ask students to find a way to misrepresent the dataset to distort the truth, without lying about the data. How could you display it so that it looks like it's saying something that it isn't? This could generate discussion about misuse of data in the media and advertising, or ethics in scientific research, and the importance of representing data accurately.

Understanding Outliers

Outliers in the context of data refers to data points that fall well outside of the other values observed. They can be legitimate variations in the thing you're measuring, or can be measurement errors, where the reading was not taken correctly for a variety of reasons.

With this lesson goal, students examine the dataset, identifying the average, median and range. Once that is done, students can then identify any outliers, using methods like looking for sharp spikes in pollutant values, or surprisingly high or low values, remove them from the dataset and recalculate their average, median and range, compare the values, and discuss which ones have changed and why, as well as discussing where the outlier values may have come from.

For more information on Outliers, see the Educational Datasets Companion document.

Identifying the Right Visualisation

As the idiom goes, a picture is worth a thousand words and there are lots of ways we can take data and make it visual. Some of the more common methods of creating visualisations are pie charts, line graphs and bar graphs. Depending on the data set, other visualisations may be appropriate to give the audience a better illustration of the data and the trends and patterns it contains.

For this dataset, since we only have readings from a single location, maps are not a relevant way of displaying this data. Most graphs are acceptable, especially line graphs, which can indicate spikes in the pollutant values, and with the right ranges can indicate where the values have moved beyond the safe threshold.

While students can generate visualisations using digital resources, there is also the opportunity with smaller datasets to create these visualisations by hand, using printed maps or sketches.

For more examples of data visualisations, see the Educational Datasets Companion document.

Spreadsheet and Numeric Skills

Spreadsheets and numerical skills are embedded across the curriculum, and this data offers an opportunity for students to put their skills to work on real-world scientific measurements. While a wide range of mathematical skills and spreadsheet skills can be applied, some key examples are:

- Sorting data. Sorting the data along different values can reveal different trends. What different ways can we sort it? How useful are each of these different methods?
- Developing spreadsheet formulae. Look for places in the data that an automatically calculated total or average might be useful. We might need to check the average pollutant level across the set time periods addressed in the standards. Formulas could be used to determine the minimum and maximum value of a list of data, to identify spikes in pollutants, or the difference between each day's minimum and maximum values, both in numerical form and as a percentage.
- Graphing. Consider the different types of charts that your spreadsheet software can make. How can we modify the settings of a graph to display data appropriately? What is an appropriate title? What labels and value ranges should be used for its axes? Students could construct graphs on paper, to build manual graphing skills.
- Conditional formatting. Create a set of rules so that when the value for a given pollutant breaches the standard, the cells change colour to indicate that the pollutant is at a dangerous level and highlight the readings that have indicated that it is dangerous.
- What-if calculations. Students can use the real-world data to make predictions. Examine what happens if the pollutant values increase by 1% across all readings. Does that put any into the dangerous range? What if they increase by 10%? Or 50%?
- Non-digital numerical skills. Students can manually take averages of sets of readings, examine other statistical quantifiers such as median and range or identify the standard deviation. Alternate goals for this dataset could be utilising algebraic skills to determine the maximum possible value an unknown reading could be without breaching the standards.

Programming

Many of the files in this dataset can be opened and manipulated in a variety of programming languages. CSV files are very easy for most programming languages to work with, since they are simple text files which use commas to split data points. Python has a specific module (csv) that adds additional functionality when working with these files.

Teaching programming with this dataset gives students an opportunity to practice skills relating to reading and writing data to and from files directly and incorporates string manipulation so they can directly access specific pieces of data. Students can investigate data structures such as lists, dictionaries and objects, assessing their usefulness in storing this data, and utilise control structures to perform calculations on the data, or organise it in a manner appropriate for output.

Depending on the prior understanding students have of programming principles, this can lead to activities ranging from calculating averages automatically and outputting them to the screen, to searching for potential outliers and removing them from the dataset before outputting it as a separate file, to creating interactive visualisation tools for the dataset.

Subject Links

This dataset can be linked to the Australian curriculum learning areas of Mathematics, Science, Technologies (Digital Technologies), Humanities and Social Sciences (Geography) and Health and Physical Education (Personal, Social and Community Health)

2 Lesson Materials

Required Understanding

A list of the existing skills students will require to work effectively with each level of this dataset can be found in the table below. This dataset can also be used as a tool to develop these skills.

For this dataset, the activities can be performed using the Novice and the Expert level datasets, but the Expert file presents additional difficulties for more advanced students. In the Expert version of the file, headings are not as intuitive or easy to read, and sections with missing data have been left as they were in the original dataset.

Spreadsheet Novice

- Spreadsheet software and the relevant key terminology, such as cell, row, column, sheet, data, cell reference and cell range
- Developing spreadsheet formulas
- Creating charts in spreadsheet software packages
- Basic mathematical statistical concepts, such as averages, range and median values.

Spreadsheet Expert

- Spreadsheet software, including appropriate formatting skills and relevant key terminology, such as cell, row, column, sheet, data, cell reference and cell range.
- Developing spreadsheet formulas
- Creating charts in spreadsheet software packages
- Basic statistical concepts, such as averages, range and median values.

Programmer

- Basic understanding of commands for a specific programming language
- Understanding of data structures and file input/output
- Understanding of programming control structures, such as sequence, selection and repetition
- Basic statistical concepts, such as averages, range and median values.

Content Engager

Use these resources to introduce students to the topic of air quality.

- CSIRO Blog – Could the Aussie BBQ go up in smoke?
- YouTube – What is Particulate Matter?
- YouTube – Aerosols: Airborne particles in Earth's atmosphere
- SBS News – Revealed: The most polluted suburbs in Australia

Some questions that you can use to start discussion about this topic and activate students' prior knowledge include:

- Why do we care about air quality?
- Do you, personally, ever notice differences in air quality?
- Why do we measure different substances in the air?
- What are particulates and how might they affect us?
- What might be the impact of different sized particulates?
- How is air quality measured?

Introductory Description

To introduce students to this dataset, consider reading the following paragraph to them, or something similar.

'Today we're going to be looking at data that was measured for a scientific study on air quality. It took place in NSW and looks at the different things in our air that could affect our health. The data files we will be looking at are measurements of different pollutants that were found in the air, and each reading is the average of pollutants over an hour. The study collects readings from two months, with one month of measurements taken during summer and another month of measurements taken during autumn. Both sets of measurements were taken during 2011.'

Thinking Time

Once students have an idea of the dataset's content, give them 5 minutes to brainstorm questions they'd like to try and answer using this data. Try not to lead students too much during this time. There is a high chance that students will develop questions which cannot be answered by the data. This creates an opportunity to explore why those questions cannot be answered.

Activities

Spreadsheet Novice and Expert

1. What kinds of questions do you need to ask about the dataset and how it was collected in order to help make sense of this data?

   What do the labels mean? Are these levels safe? What kind of area was it collected in? What are the substances being measured? What creates these substances? What are their health effects?

2. Consider one type of particle, like PM10, and find the average concentration across all readings. What does the average value tell you about human risk? What doesn't it tell you?

   If the average is below the NEPC risk level that's good, but it doesn't tell you how often the standards have been breached.

3. How else can you explore the concentration of PM10 to make sense of it?

   Sorting it to see how many of the values are above the maximum, graphing a single day, graphing each day, graphing the whole column.

4. Consider the default graph your spreadsheet creates if you graph the whole column. What can be easily determined from this graph? What can't be determined from this graph? How can it be improved to make it easier to understand the data?

   The default line graph will try and plot all the readings on a single graph, making the labels hard to read. It's easy to see that there is an overall pattern with this graph, but hard to understand the details of that pattern. Grouping the data into days then graphing the individual days would make the information much clearer.

5. How might you represent all the data in a way that is useful and meaningful? (consider ways to summarise and visualise the data)

   – What is the best graph type for one column of this data?
   – Since weather data and pollutants are measured on different scales, is it possible to graph them together? How can you represent ozone and solar intensity on the same graph?
   – Can you put all this data on a single graph?
   – Explore different ways of visualising the dataset
   – Consider the strengths and weaknesses of each.
   – Which graph types are appropriate, which are not?
   – What characteristics of the dataset does each valid graph type highlight?

   Line graphs are very useful for this specific dataset. The weather data and pollutant data can be graphed together, using independent scales. This will allow comparison of the patterns without the exact values skewing the range of the graph. You can put all this data on a single graph, but it will be difficult to read. Experiment with the different graph types available and list the strengths and weaknesses of each one. What does it reveal about the data? What doesn't it show?

6. What do we learn by comparing the Autumn data with the Summer data? How could you effectively represent that much information?

   Summarise the data using statistical tools. Graph daily averages against each other, comparing the hourly averages across a given week of each variable.

7. Explore the values by sorting them - how you can find the lowest and the highest concentration of each pollutant? Examine each pollutant independently. How much higher than the other values is the highest value? How much lower is the lowest? How do the top and bottom of the range compare to the average? What does this say about how safe the air is to breathe?

   Sort each column by value, ascending and descending to find the maximum and minimum values (Or alternately, use the max and min formula). Once you've got the highest and lowest values for each pollutant, you'll need something to compare it to. The median or mean might be a good place to start, as would the second highest/lowest values.

8. Why might the average daily value not be a useful indicator of air quality if the range is large?

   Spikes of pollutants may dramatically exceed the standards for short periods of time and be exceptionally dangerous without changing the overall daily averages much.

9. Are there values that are significantly different from the surrounding values? What impact do these outliers have on the averages? What might cause the outliers?

   The particle reading in the NEPH column demonstrates a huge spike on the 26th of February, around 2/3pm. As it has a gradual build and fall, it's possible that it's a legitimate reading caused by a particle source being in the area for a significant amount of time. Alternately, it could be a recording or measuring error.

10. How would you graph the values for a single pollutant? What is the right graph type? How could you graph the entire set of values in a readable way?

    Line graphs are very useful here, but there are multiple correct answers. If you can see the rise and fall of

    particles in the graph, it's likely to be useful. A line graph that displays all the values in the data is going to be hard to read. Consider graphing each day or week as an individual line.

11. Given that there are a lot of data points, what are the advantages and disadvantages ofusing sampling to reduce the dataset size? How would this change a visualisation of the data?

    If you take every tenth value, there is a risk you miss the highest or lowest values. How fast do the values change? What details might you lose using sampling? What is the lowest sampling rate that still gives you an accurate picture of the data? (i.e. every second, third, tenth, hundredth value?)

12. In the readings from Autumn, there are missing records at 2am daily. How can we account for the missing data when analysing the dataset? Make an estimate of the missing values and enter it into the dataset.

    We could estimate what the value could be using the values around it. This is called interpolation, and there are many methods for achieving it.

13. Consider the default graph your spreadsheet program creates for this data. What features does the graph need to have to be fully readable? How can you best label each value? Include the date/time stamps, title the graph, etc?

    Coming up with appropriate, relevant and informative names for the labels is important here. Don't forget to include a legend to indicate to the viewer what the colours mean.

14. Make graphs for each day and compare them. Are the default graphs comparable? Why or why not?

    By default, the spreadsheet graphs the data range, rather than starting at zero. This means the scaling on each day will be different if the range is different. Even graphs that both start at 0 are incomparable if one goes to 30 and one to 80 in the same size area. Discuss the dangers of default settings in misrepresenting data.

15. Do the maximum values for different pollutants occur at the same time each day? How can you find out?

    The most time-consuming way would be to sort each column individually and check the times. A much easier and faster way is to graph the pollutants so you can look for the spikes at a glance. Another method would be using an Excel formula to return the time linked to the maximum value for each pollutant.

16. Do any of the values look correlated? Can you see this clearly if you graph two values together? What if you use the correlation function?

    The graph is an easy way to indicate correlation, if the graph of two different values seem to follow the same pattern in the way they rise and fall, they are correlated. To use the correlation function effectively, you may need to investigate the appropriate correlation formula for your spreadsheet package.

17. Why are the allowable standards for the pollutants higher over the short term than over the longer term? The limit for CO is an average of 9.0ppb over 8 hours, and the limit for Ozone is 0.1 over 1 hour and 0.08 over 4 hours. Why is the concentration allowed to be higher for short periods of time?

    This is largely because the standards are focused around the amount you inhale in total. If it is over a very short period of time, you're not inhaling that much of the pollutant overall.

Programmer

Write a program using your chosen programming language to perform the following tasks:

1. Calculate summary statistics for each column of data.
2. Create an algorithm to estimate the missing 2am values from the Autumn data and insert them into the file.
3. Detect if a pollutant has breached the standard.
4. Print a report that lists each pollutant and details how close the pollutant was to breaching the standard as both a percentage of the standard's limitations and as a measurement.
5. Graph data for each pollutant separately.

Open Inquiry

In addition to the activities listed above, this dataset can be used for student-centred open inquiry projects. Using open inquiry, students generate research questions and design investigations to answer those questions. Students can use this dataset to support their independent research and investigation in a range of areas.

Examples of inquiry questions that could be explored using this data include:

- What air pollutants are prominent in the local area and how do they compare to pollutant levels found in this data?
- What are the major sources of air pollution in the local area?
- What are the notable air pollutants found in this dataset, what are their likely sources and what impact do they have on local residents?
- What patterns can be identified in pollutants found in the dataset and what do those patterns reveal about weather conditions and pollutant sources?

Assessment

Assessment items for this dataset could include:

- A labelled graph, or series of graphs with the default settings customized to make sure that scales start at zero, labels are accurate and readable, and legends are clear, readable, and in English (for example, using 'Particles smaller than 10µm' rather than PM10).
- Graphs comparing Summer and Autumn readings for key measurements, and a discussion of the differences between the recorded values.
- A poster about the health impacts of different pollutants, with a visualization of key pollutant values, showing standards as well as measured values.
- A spreadsheet with formulae to calculate summary statistics.
- Code to extract pollutant values from the file, calculate and report summary statistics.
- Code to detect when pollutants exceed the standards (bearing in mind that the standards are calculated differently for each pollutant).

Appendix A References

Educational Dataset:
Sydney Particle Study

Original Dataset:
Keywood, Melita; Selleck, Paul; Galbally, Ian; Lawson, Sarah; Powell, Jennifer; Cheng, Min; Gillett, Rob; Ward, Jason; Harnwell, James; Dunne, Erin; Boast, Kate; Reisen, Fabienne; Molloy, Suzie; Griffiths, Alan; Chambers, Scott; Crumeyrolle, Suzanne; Zhang, Chuanfu; Zeng, Jianrong; Fedele, Rosie (2016): Sydney Particle Study 1 - Aerosol and gas data collection. v3. CSIRO. Data Collection. https://doi.org/10.4225/08/57903B83D6A5D

Keywood, Melita; Selleck, Paul; Galbally, Ian; Lawson, Sarah; Powell, Jennifer; Cheng, Min; Gillett, Rob; Ward, Jason; Harnwell, James; Dunne, Erin; Boast, Kate; Reisen, Fabienne; Molloy, Suzie; Griffiths, Alan; Chambers, Scott; Humphries, Ruhi; Guerette, Elise-Andree; Cohen, David (2016): Sydney Particle Study 2 - Aerosol and gas data collection. v1. CSIRO. Data Collection. https://doi.org/10.4225/08/5791B5528BD63

Published Papers:
NSW Office of Environment & Heritage – Sydney Particle Study

Supporting Information:
- CSIRO – Air Quality
- NSW Office of Environment & Heritage – Glossary of Air Quality Terms
- NSW Office of Environment & Heritage – Standards and Goals for MeasuringAir Pollution
- Australian Department of Energy and the Environment – Air Quality

Australian Curriculum:

As Australia's national science agency and innovation catalyst,
CSIRO is solving the greatest challenges through innovative science and technology.
CSIRO. Unlocking a better future for everyone.

Contact us
1300 363 400
+61 3 9545 2176
csiroenquiries@csiro.au csiro.au

For further information
CSIRO Education and Outreach
1300 136 376
+61 2 9490 5588
education@csiro.au csiro.au/education

Cover Image Credit
CSIRO Marine Research