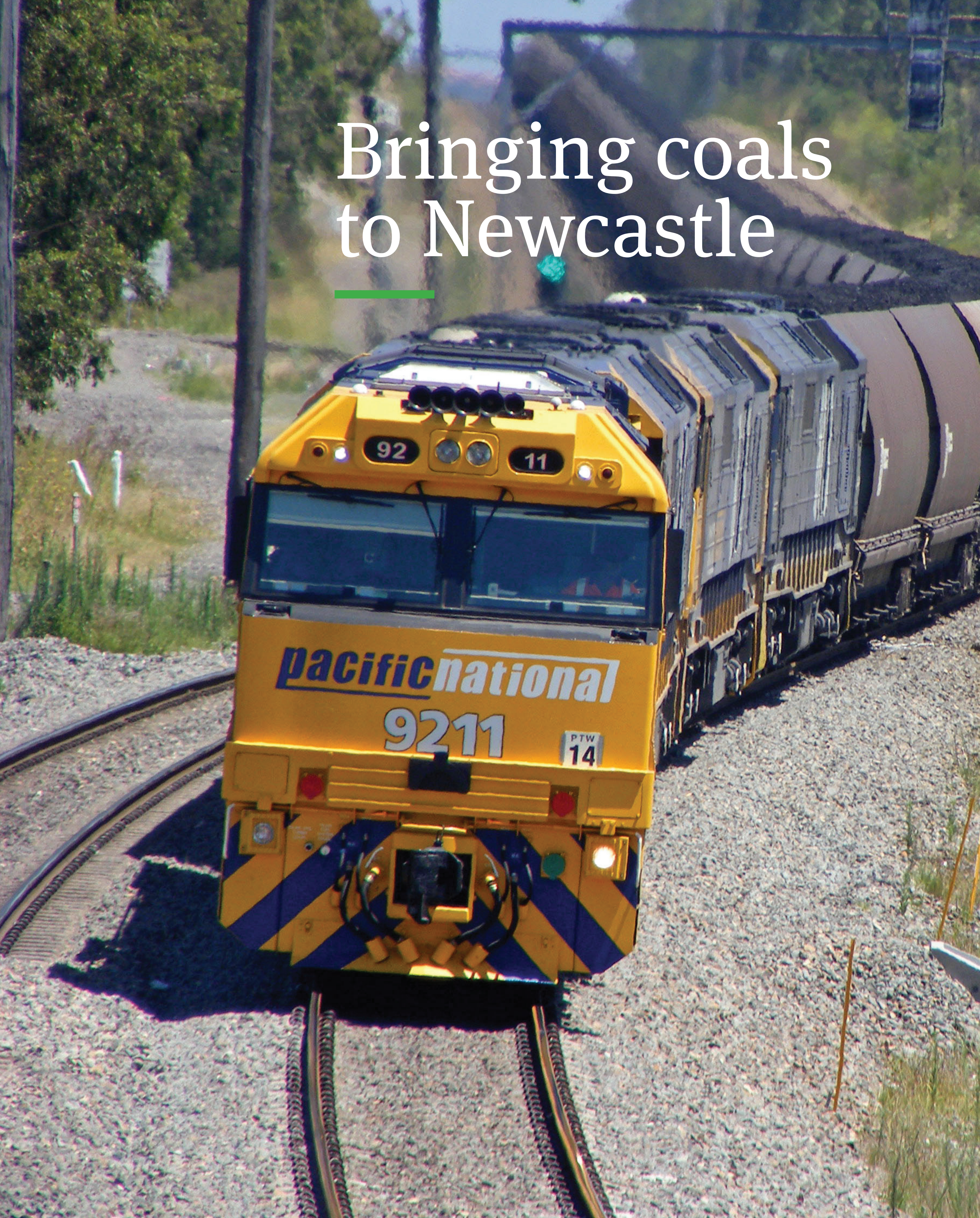
Bringing coals to Newcastle



Making effective public policy decisions is challenging at the best of times, but especially in the context of environmental regulation, which typically requires managing opposing interests and strong opinions from industry and private citizens. In this case study, **Louise Ryan**, **Matt Wand** and **Alan Malecki** show how statistical analysis can help resolve conflict and inform effective decision-making under uncertainty

Louise M. Ryan and **Matt P. Wand** are distinguished professors of statistics, and **Alan A. Malecki** is a doctoral student at University of Technology Sydney, Australia. All three are part of the Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers

Main image: Vince Wang/flickr.com

The city of Newcastle – just two hours' drive north of Sydney – is a key hub for Australia's coal industry. It boasts the world's largest coal export port by tons shipped per year. However, the city's strong industrial base belies the fact that it is also an area of stunning natural beauty. Newcastle has beautiful beaches and is a gateway to the Hunter Valley, a renowned wine-producing region and holiday destination for tourists and locals alike.

Citizens in Newcastle and the Hunter Valley region are understandably anxious about the impact of large-scale coal mining on their environment. Many report that their homes and cars are often coated in a fine layer of soot and they are concerned how this might affect their health. The citizens believe that coal transportation is a major cause of this pollution, as the railway tracks used to transport coal to the port pass directly through residential areas.

A citizens' action group in Newcastle has advocated for regulation to enforce the covering of wagons on coal trains, believing that this will reduce the level of coal dust particulates in the air. However, the coal industry has resisted the move on the grounds that it would be expensive and have little benefit.

To help resolve the issue, the New South Wales Environment Protection Authority (EPA), in conjunction with the Australian Rail Track Corporation (ARTC), conducted a study to collect data on airborne particulate levels in the Hunter Valley.

As happens all too often when statisticians are not involved in the design phase of a project, the study had some weaknesses. Most notably, particulate data were collected from only a single site, with a view to assessing whether or not levels increased when loaded coal trains were passing. It would have been much better – expense and operational issues aside – to design a study with multiple monitors over a network of sites in the region.

The environmental consulting company which designed the study also conducted a statistical analysis, and subsequently this was criticised as having some limitations. And so, in late 2013, we were asked by the EPA to assess the analysis. Our initial review confirmed that the statistical analysis had some shortcomings, and subsequently we were asked by the EPA to reanalyse the data. Our first reanalysis,[1] conducted in early 2014, found clear evidence that loaded and unloaded coal trains, as well as freight trains, were associated with increased airborne particulate levels. We hypothesised that this was due to the diesel-powered locomotives pulling these trains. To explore this possibility, the ARTC provided us with additional

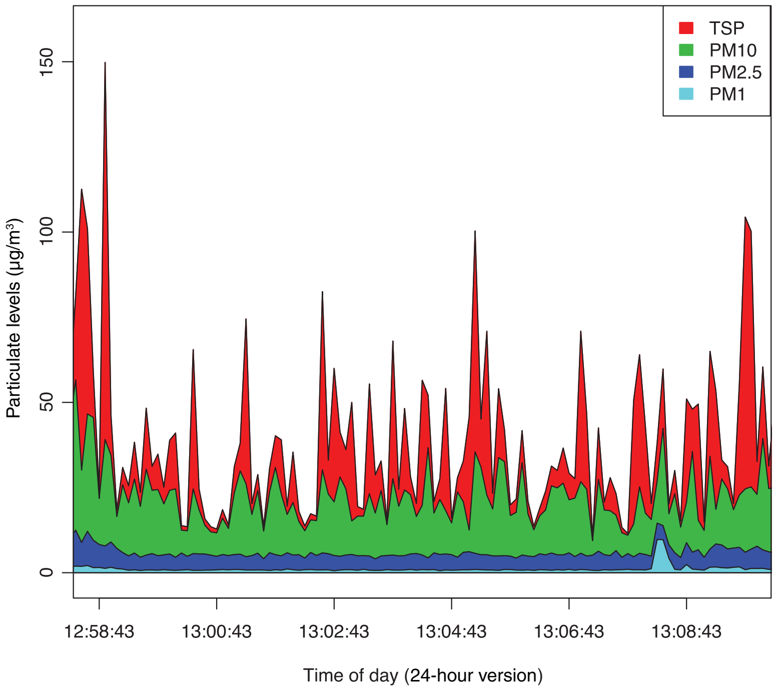


FIGURE 1 Recorded levels of PM1, PM2.5, PM10 and TSP during a 10-minute period on 30 November 2012

data on the number of locomotives pulling each train. Data on rainfall (precipitation) were also made available to us at this time. We presented our second reanalysis in 2015.[2]

The data

To collect information on particulate levels, an Osiris air pollution monitor was placed adjacent to the rail tracks in Metford, a town on the outskirts of Newcastle. This particular location was near a busy railway junction with four active tracks, two being used by north- and southbound passenger trains and the other two by north- and southbound coal and freight trains. Southbound coal trains were "loaded" since they were on their way to deliver coal to the port in Newcastle. In contrast, northbound coal trains were empty, as they were returning to the mines after unloading at the port.

The monitor was set up to record particulate measurements (in micrograms per cubic metre) every 6 seconds in three different size groups – particulate matter (PM) less than 1, 2.5 and 10 micrometres in diameter – and counts were also aggregated as total suspended particulate (TSP). Information was generated every time a train passed, including an indicator of which type of train it was, how fast it was travelling, how long it took to pass the monitor and the number of locomotives on the train. Data were obtained from the Australian Bureau of Meteorology on wind speed and direction, as well as precipitation.

While the monitoring system was set up to measure particulate levels every 6 seconds throughout each day over a 61-day study period, there were a number of periods of missing data, especially at the beginning of the study when the system was being set up. We analysed data from the subset of 55 days that had at least 1000 monitoring measurements, corresponding to a total of just over 600 000 sets of observations.

To illustrate the nature of the data, Figure 1 shows the recorded levels of PM1, PM2.5, PM10 and TSP during a 10-minute period on 30 November 2012. We see that the largest particles (PM10) were the most common, and that their values closely tracked TSP levels. Figure 1 also shows that the data are very jumpy and prone to extremes. Notice how TSP levels are predominantly in the range from 30 to 60, but can jump as high as 150 on occasion. For this reason, we considered a natural logarithmic transformation and found this to be more appropriate for subsequent analysis.

In our analysis, we were struck by just how many trains there were. Over the course of the study period, a total of 5601 trains passed by the monitoring station, with a median of 137 trains per day. Table 1 shows the median numbers of trains, as well as the upper and lower quartiles, by different train types, along with the duration (in seconds) of each train passing and the average train speed. There were a few instances where train type could not be determined (these were recorded as "Unknown"). It is interesting to see that while passenger trains are the most common, they travel quickly and pass by the monitor in just a few seconds. At the other extreme, loaded coal trains travel slowly and are very long. Hence they can take several minutes to pass the monitor.

Further exploration (see Figure 2) showed that trains are much more likely to pass the monitoring station during the daytime and peak commuting hours of 5am to 9pm, than nighttime hours, underscoring the importance of adjusting for time of day in our analysis. The bar plots in Figure 2 show that while this pattern holds for all train types, a relatively higher proportion of loaded and unloaded coal trains travel at night.

Data on concurrent wind speed and direction were available for approximately 75% of the particulate observations. Precipitation measurements were supplied for each day over the study period at the nearby towns of Cessnock and Maitland. At Cessnock, rain was recorded as a cumulative measurement (in millimetres) staggered at 30-minute intervals, while Maitland rain was only reported at the total daily level.

TABLE 1 Information about trains passing during the study period. Columns show median value, along with lower and upper quartiles (LQ, UQ).

Train type	Trains per day Median (LQ,UQ)	Duration (s) Median (LQ,UQ)	Speed (m/s) Median (LQ,UQ)
Empty coal	40 (27, 47)	72 (66, 78)	19.8 (18.5, 20.7)
Freight	6 (2, 8)	66 (36, 180)	18.2 (14.7, 28.5)
Loaded coal	36 (24, 41)	102 (96, 120)	13.9 (11.7, 15.1)
Passenger	60 (55, 92)	6 (6, 6)	24.9 (24.0, 28.9)
Unknown	3 (2, 3)	36 (30, 42)	19.1 (15.7, 21.0)

Analysis and modelling strategy

The objectives, as posed to us by the EPA, were to determine: (1) whether trains operating on the Hunter Valley rail network were associated with elevated concentrations of particulate matter; and if so, (2) whether loaded coal trains were associated with higher concentrations than unloaded coal trains or other trains on the network.

If there is an increased effect due to passing trains, then determining whether the loaded coal trains were associated with an increase in particulate levels would aid in the implementation of policies. Possible outcomes include the idea of covering loaded coal wagons before leaving the mine or, alternatively, washing out the unloaded coal wagons before they depart the port back to the mine.

The consulting company that had initially reported on the study did a fairly naïve analysis, computing average particulate levels associated with each train passing, and then comparing these to averages computed during periods when no train was passing. This resulted in several problems, including inadequate control for factors such as time of day. Their analysis also ignored the fact that the data were subject to autocorrelation, meaning that observations measured close together in time were likely to have similar values. This phenomenon violates the independence assumption required for many commonly used statistical models. Ignoring autocorrelation exaggerates the precision of an analysis and can lead to incorrect conclusions.

We based our analysis on regression modelling of the individual data, without aggregation and with appropriate adjustment for autocorrelation due to the time series structure of the data. Our analysis involved no loss of information due to aggregation and also permitted the use of covariates – predictor variables – indicating train type, the number of locomotives pulling each train, wind speed, and precipitation, as well as additional variables reflecting time of day, day of week and other temporal effects. (See the box on page 36 for a more detailed description of the statistical methods we used.)

It was critical that we communicate our analysis strategy effectively to the EPA and especially to the citizens of Newcastle. We explained that the big advantage of regression analysis is that it allows for simultaneous adjustment with respect to various confounding factors that may otherwise bias the analysis. We explained that without careful adjustment for time of day, the fact that loaded and unloaded coal trains were relatively more common during the night, compared with passenger trains, could distort the results. Regression analysis can easily account for this temporal effect.

Figure 3 shows some actual data, collected over a 6-hour period with plotting symbols colour-coded to indicate the presence of various train types. The speed and length of the passenger trains are evident in the absence of long trails of red dots in contrast to other train types which pass the monitor for longer periods of time. The effect of time of day is highlighted by the distinct shift at 17:30 of log-transformed values from 4 to 3. Figure 3 also reveals a strong tendency for observations to "track" together in time (this is a reflection of autocorrelation

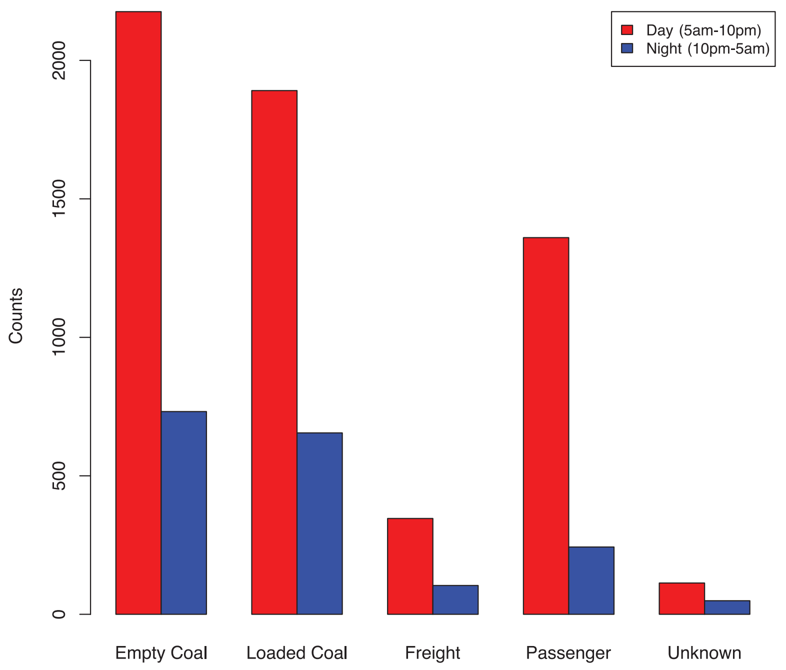


FIGURE 2 Number of trains of various types passing at different times of day

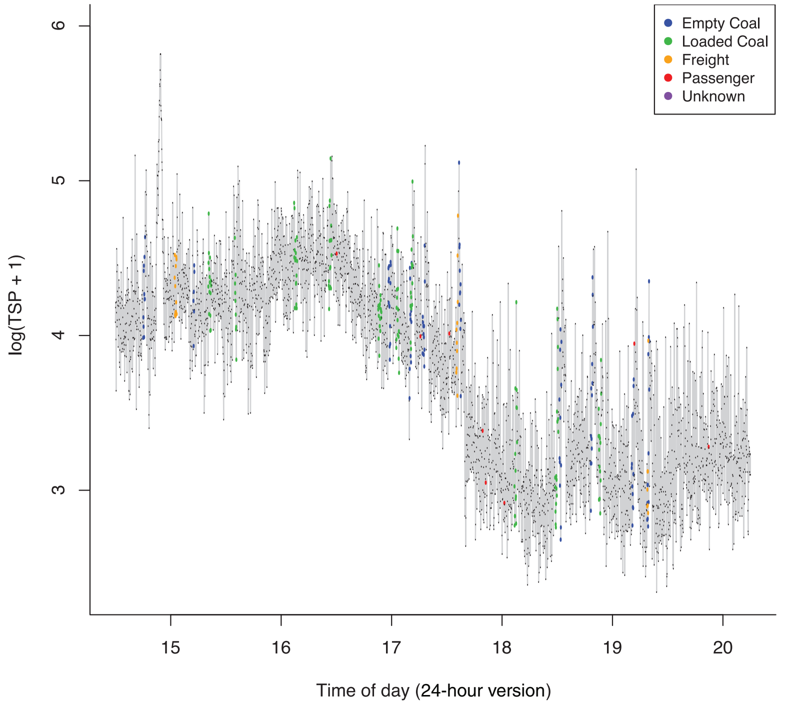


FIGURE 3 TSP data during a 6-hour period on 9 December 2012. Plotting symbols are colour-coded to indicate the presence of various train types

in the data). Through the explanation of this figure and its key features, we were able to address why regression analysis was the most effective method.

Results

Table 2 shows the results of our main regression analysis on logged TSP levels. All variables listed in the table represent indicators of whether or not that particular condition applied at various time points. For example, the variable "Freight train passing" took the value 1 whenever a freight train was passing the monitor, 0 otherwise. Although not shown here, models for PM10, PM2.5 and PM1 show similar patterns, namely that particulate levels are higher on average when any kind of train is passing, as well as during the 5-minute period after a train has passed. The magnitude of increase is similar for freight, loaded coal and unloaded coal trains, and roughly half that magnitude when passenger trains are passing. There was certainly no indication that loaded coal trains were associated with higher particulate levels than other train types.

Particulate levels were significantly lower if it had rained in Maitland on the previous day. Indeed, this rain indicator variable was the strongest predictor in the analysis. As seen in Table 2, TSP levels were reduced by 0.303 (on the logged scale) if it had rained the previous day in Maitland. We explored interactions between train types and rainfall, but they were not significant. We also investigated other measures of rainfall, including current day rainfall levels and rainfall at Cessnock. However, the dominant effect remained Maitland's previous day measure.

Although not shown in Table 2, all models included smooth spline terms to allow for daytime effects as well as day-to-day variation (again, see the box for details). Also not shown are any of the extensive explorations we undertook to assess the impact on particulate levels of the number of locomotives on each train. Overall we found virtually no effect. This finding to some extent dispelled our original hypothesis that diesel emissions were the culprit. However, the ARTC itself had low confidence in the accuracy of the locomotive data, so these results had to be interpreted with caution.

TABLE 2 Fitted regression model based on logged TSP values

Variable	Estimate	Standard error	*t*-value	*p*-value
Intercept	3.508	0.080	43.795	<0.001
Freight train passing	0.100	0.037	2.729	0.006
Freight passed within 5 min	0.100	0.030	3.316	0.001
Empty coal train passing	0.090	0.017	5.339	<0.001
Empty coal passed within 5 min	0.113	0.012	9.675	<0.001
Loaded coal train passing	0.073	0.017	4.196	<0.001
Loaded coal passed within 5 min	0.081	0.015	5.387	<0.001
Passenger train passing	0.049	0.018	2.694	0.007
Passenger passed within 5 min	0.044	0.010	4.542	<0.001
Unknown train type passing	0.124	0.041	3.044	0.002
Unknown passed within 5 min	0.074	0.044	1.710	0.087
Rain in Maitland on previous day	−0.303	0.119	−2.548	0.011

The statistical methods

To allow for particulate levels to vary over time, we used an advanced version of linear regression analysis, the so-called generalised additive model,[3] which allows for the flexible modelling of continuous functions using splines and other kinds of smooth functions. Analysis was conducted using the ***gam*** function from the mgcv package[4] in R.[5] While the ***gam*** function has an option to adjust for autocorrelation, these models did not run due to the large size of our data. Consequently, we used bootstrapping[6] to adjust the standard errors computed for our models. The bootstrap works by resampling replicate data from the original data set, running the generalised additive model on each replicated data set, and then averaging the results. In our case we used a specific variant, the so-called blocked bootstrap, which has been developed for use with autocorrelated data. It maintains the dependence structure of the data, in our setting, by resampling days.

The large size of our data set meant that each model took several minutes to run and bootstrapping took several hours. In order to explore the data and to identify suitable models, we first ran analyses using ordinary linear regression with polynomial (non-linear) terms for time of day and day of study. Our final analyses were then repeated using the computationally more intensive but accurate ***gam*** function with bootstrapping to produce reliable standard error estimates. In general, we found that linear regression models gave qualitatively similar results to the generalised additive model analyses (aside from standard error estimates) and hence they provided a useful practical approach to exploratory analysis.

Discussion

The case study presented here provides a good illustration of how modern statistical science can help to bring clarity to a complex real-world situation. Clear logic and sound methodology are needed in situations where government agencies have to make decisions in the face of considerable uncertainty and many strong and sometimes opposing opinions. The analysis presented in this article has armed the EPA and citizens with useful information, aiding their discussions with industry on how to reduce the environmental impact of coal mining and coal transport in the Hunter Valley. In particular, our analysis suggests that there would be little to gain by covering the cars on the loaded coal trains. The similarity between the particulate levels associated with the various train types, as well as the strong impact of rainfall on the previous day, suggest that the stirring up of existing dust on the tracks is a major source of the increased particulate levels associated with passing trains. Anecdotal and visual evidence (see main photo, page 32) at the monitor location suggest that tracks used by the unloaded coal trains are the ones with the most dust. This further suggests that money might then be better spent on cleaning the coal cars after they have unloaded at the port, rather than covering them on their way to the port. The EPA is presently exploring these and other alternative avenues.

Our analysis illustrates the importance of thinking not only about the right technical approaches, but also more pragmatic issues such as how to effectively manage large and complex data sets, as well as how to communicate complex ideas to a professional but non-statistical audience as well as a lay

Clear logic and sound methodology are needed where government agencies have to make decisions in the face of uncertainty

audience. A simple simulation accompanied by a thorough explanation of statistical methods supported by visual inspection, with graphs and images, was key to communicating our work to the less statistically informed members of the public. Engaging with practical problems that help the community is also a source of great satisfaction to the statistician.

In any real-world study, it is inevitable that available data will involve lots of problems with missing values, data inaccuracies and less than ideal study designs. The statistician's role in such settings is to come up with a sound and reliable solution, keeping these limitations in mind. In practice, there will rarely be time to pursue what might be considered the ideal solution since there will generally be time pressure to produce a report. A good perspective to keep in mind is the 80–20 rule: that is, finding a solution with which one is about 80% happy, but knowing that there will be time down the line to pursue more nuanced approaches. In this particular case, we have emerged from the project with some exciting ideas for future statistical research. ■

Acknowledgements

The findings described here are drawn from two reports submitted by to the New South Wales Environmental Protection Authority, the first by Ryan and Wand in February 2014 and a follow-up report submitted in August 2015 by Ryan and Malecki. The authors thank the late Professor Peter Hall for advice on the blocked bootstrap.

References

1. Ryan, L. M. and Wand, M. P. (2014) Re-analysis of ARTC data on particulate emissions from coal trains. Technical Report, University of Technology Sydney.

2. Ryan, L. M. and Malecki, A. A. (2015) Additional analysis of ARTC data on particulate emissions in the rail corridor. Technical Report, University of Technology Sydney.

3. Wood, S. N. (2006) *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press.

4. Wood, S. N. (2015) Package "mgcv". R package version 1.7.

5. R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org

6. Efron, B. and Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. New York: Chapman & Hall.