Review of Pollution Reduction Program 4.2 Particulate Emissions from Coal Trains

by Distinguished Professor Louise Ryan
on behalf of accessUTS Pty Ltd

for NSW Environment Protection Authority

18 September 2013

Project No: 2013001201

accessUTS Pty Limited
PO Box 123
Broadway NSW 2007 Australia
Tel +61 2 9514 1916 Fax +61 2 9514 1433

ABN 55 098 424 312

TABLE OF CONTENTS

1. SUMMARY

This is an interesting study. While use of a single monitoring site is not ideal, the time series design appears adequate to address the questions of interest. Unfortunately, there are some serious limitations with the data analysis that has been used. I recommend that the study be reanalysed using regression analysis methodologies, taking into consideration the possibility of autocorrelation.

2. INTRODUCTION AND BACKGROUND

Katestone Environmental Pty Ltd was commissioned by Australian Rail Track Corporation to conduct a monitoring study of particulate matter, with a view to determining whether

1. Trains operating on the Hunter Valley rail network are associated with elevated particulate matter concentrations; and
2. Whether trains loaded with coal have a stronger association compared with unloaded coal trains or other trains on the network.

To achieve these objectives, a continuous particulate monitoring station was installed to measure particulate levels in the rail corridor adjacent to tracks carrying different types of train. Specific details of the monitoring design were informed by a previously conducted pilot study. In addition to continuous monitoring of particulate levels, the time of each train pass-by was recorded, along with other details about each corresponding train. Additional data were available regarding wind speed and rainfall.

After the final report was submitted, a number of independent reviews were undertaken to evaluate the reliability of the results. Several of the reviews raised concerns regarding the statistical methods used to analyse the data. I was engaged as an independent, expert statistical reviewer, charged with assessing the suitability and reliability of the statistical methods used to analyse the study data. I purposely did not read any of the independent reviews before undertaking my own evaluation. On concluding my evaluation, I then read the reviews and compared these assessments with my own. In particular, I found that the opinion of a well-qualified statistical reviewer concurred closely with my own.

I will begin by offering some comments regarding the study design itself. After critiquing the statistical analysis methods, I conclude with some recommendations.

3. EVALUATION OF THE STUDY DESIGN

I have some concern about the use of a single monitoring site. While I understand the reasons articulated in section 3.4 for why the Mayfield site was discontinued for the main study, it would

have been better, I believe, to identify at least one more monitoring site. This would have provided more confidence regarding the generalizability of any results. That being said, the study should definitely provide valuable insight into the nature of particulate levels associated with different types of trains.

While I am not qualified to comment on the suitability or reliability of the monitoring instrument itself, the Osiris, I think the decision to choose a monitoring system with the potential to provide a resolution as fine as 2 minutes was excellent. Similar comments apply with regard to specifics of the instruments used to monitor wind speed and wind direction. I note that the report does not provide much information at all about where data related to precipitation were obtained.

With data available from only a single site, the resulting data will be a time series, along with time-varying covariates indicating presence/absence of trains, characteristics of those trains, as well as meteorological conditions. While multiple monitoring stations would have been helpful, a continuous time series at this single site should be adequate to answer the questions of interest.

4. EVALUATION OF THE STATISTICAL ANALYSIS STRATEGY

The report does a good job at describing some of the underlying data capture issues (section 4 of the report). The tables showing data capture rates by date, as well as by hour of day and day of week are very helpful. However, I think it would have been helpful to have more detail in this section. It is clear that capture rates settled into a good level on or around December 22nd and continued through the end of the monitoring period, with an exception of a few days around New Year. While some explanations are offered, it is not clear however why the earlier data capture rates were so consistently low. I think it would be useful to consider some sensitivity analysis that considers if and how the results might change if data from the less reliable periods were excluded.

The report describes an assumption that coal trains heading towards the port (UC track) were loaded, while coal trains headed in the other direction were all unloaded. Is this assumption completely valid? It makes sense, but was there any empirical evaluation of this? Table 5 reports 168 trains of "unknown" characteristics, from among the 11,245 train pass-bys evaluated. What was the reason for these unknowns?

There is virtually no description of the algorithm used to determine train type. Since train type is such a key issue for the overall study objectives, this is an absolutely critical issue. Significantly more information needs to be provided regarding how train type determinations were made. Similarly, how were the determinations reported in Table 7 made?

Section 4.5 describes how particulate levels were averaged over the duration of train pass-bys. For non-passenger trains, it appears that the averaging period was increased to allow for potential entrainment in the air of particulates after a train has passed the monitor. While the report makes

a coherent argument for the approach, there are a number of serious concerns from a statistical perspective. First of all, the data are treated different according to train type. This is highly problematic when a fundamental objective of the study is to compare particulate levels by train type! While the concept of entrainment make sense and the concern that lingering effects could be missed is valid, averaging over an extended period could actually end up in diluting real effects. It is also problematic that because freight trains are slower, the averages calculated for these kinds of trains use much more data than averages calculated for passenger trains. In my opinion, it is not necessary to average the data at all. The individual measurement (6 second averages) could be analysed directly, taking into account possible autocorrelation between points measured near each other in time. I will discuss this concept later in the recommendations section of my report.

More generally, the averaging method used to compute concentrations by train type (see section 5.1) introduces several problems. First of all, it is clear that the resulting data points are likely to be averaged over variable numbers of 6 second periods, ranging from one or two periods for a fast passenger train, to 30 6 second periods for a coal train, to who knows how long for periods when there are no trains. This differential averaging will lead to heteroscedasticity in the data (the standard deviation of data points will vary by presence/absence of a train, as well as by train type. This is likely to induce some anomalies in the data analysis and in particular it is likely that hypothesis tests may be affected. I also think that it does not make sense to simply average all the measurement made in each non-train period. For comparisons with times when trains are passing, it would make sense to compare with non-train times that are relatively nearby, time wise. With the current approach, it is possible that biases are being introduced

Missing data issues are always inevitable and indeed they come up a few times in the report. (e.g. 5.1). However, it is important to make sure that the missing data are not informative. I don't expect this to be the case, but some more nuanced consideration of missing data issues would be appropriate.

There are serious limitations with the reported statistical analyses (see page 15 of the report) which basically correspond to the computation of summary statistics (averages, medians, standard deviations, upper and lower 5^{th} percentiles, maxima) and upper and lower 95% confidence intervals within the various groups of interest (No train, loaded coal train, unloaded coal train and passenger train). Group comparisons were evaluated according to whether or not group-specific confidence intervals overlapped, as well as more informally using graphical tools. Unfortunately there are a number of problems with the analysis methods used:

- They do not take account of the fact that the data points may have been averaged over different numbers of time periods.

- The averaging approach that has been used may introduce some confounding, since it is likely that the non-train time periods will not be representative of the time periods when trains are passing. For instance, it may be that the non-train periods include night-time.
- They do not take account of the fact that even with averaging, there is likely to be residual serial correlation in the data (observations measured more closely in time are likely to be correlated).
- The approach used to test hypotheses is incorrect. The report indicates that hypothesis tests are conducted by assessing whether or not there is overlap in the confidence intervals associated with the relevant groups. This does not constitute a valid hypothesis test for comparing two groups. It works in the much simpler single sample setting where one can ask, for example, whether a confidence interval covers zero. But this does not extend to two sample comparisons.

Furthermore, in the results section and the various graphs referred to from there, the link with the averaging periods discussed in section 4.5 is really unclear! For example, Table 8 shows that there were 916 instances where loaded coal trains passed, corresponding to 41,102 6 second averages. However, we can't determine from the report how these 41,102 numbers were further averaged. From above, it appears that these numbers were not analysed directly.

- The very extreme maximal values reported in the tables suggest to me that perhaps the analysis would benefit from a log transformation. This should at least be explored. Figure A1 also suggests that a log-transformation might be a sensible strategy for the data.
- The various subgroup analyses (e.g. wind speed, train speed) etc. suffer from all the same problems as described above for the main analysis. There are also some inefficiencies associated with breaking out the data into so many different subsets.
- In my opinion, the data are crying out for reanalysis based on regression analysis. This would allow for rigorous comparisons of interest (loaded coal train vs. no-train etc.), while adjusting for effects of wind speed, train speed etc. A regression analysis would also have the advantage of not needing to average the data, but rather could treat each 6 second measurement as the data point. Of course it would be critical to adjust for potential autocorrelation. A regression approach would also allow for a more sophisticated exploration of entrainment effects.
- Finally, I note that the multiple pass-by occurrences seem to be associated with relatively high levels. This makes sense for a number of reasons. First of all, chances are that coal trains are involved in many of the multiple pass-bys. Also, it would seem likely that multiple pass-bys would cause significant air turbulence which would further stir up particulate levels.

5. RECOMMENDATIONS

As indicated above, there are some serious limitations with the data analysis that has been used. I recommend that the study be re-analysed using regression analysis methodologies, taking into consideration the possibility of autocorrelation. The outcome variable for the analysis should be the individual particulate measurements (that is, without any averaging). Such an analysis will allow for incorporation of covariates indicating train type, wind speed, precipitation, as well as additional variables reflecting time of day, day of week and other temporal effects. I also recommend that preliminary exploratory analysis be conducted to see if a log transformation of the data would improve the modelling process. Such an approach would more appropriate address the questions of interest, whilst accounting for possible confounding effects.

Louise Ryan

Distinguished Professor of Statistics, UTS