Dr Luke Knibbs *BSc PhD*
l.knibbs@uq.edu.au
07 3365 5409

1 July 2013

To: Mr Matthew Reilly
Director, Climate & Atmospheric Science
Office of Environment and Heritage
NSW Department of Premier and Cabinet

Re: Peer review of Australian Rail Track Corporation (ARTC) Pollution Reduction Program (PRP) 4.2 - Particulate Emissions from Coal Trains.

Dear Matthew,

Thank you for giving me the opportunity to conduct an independent peer review of the ARTC's recent report detailed above, which was performed on their behalf by Katestone Environmental Pty. Ltd.

I was engaged to assess whether the project was conducted with appropriate scientific rigour and that the results are robust and within the specifications set by the EPA.

I have read through the final report and the study specifications in detail. I have identified a number of methodological and analytical issues that affect the scientific rigour of the study, the robustness of its conclusions, and the ability of the study to meet its objectives and specifications. These are listed in the attached report.

Some of these issues are fairly minor in nature, and their impact on the findings may be negligible. However, there is a major error in the statistical analysis of the data that is very likely to obscure the true statistical significance of the comparisons undertaken. This error affects many of the conclusions drawn, and is likely to underestimate the number of statistically significant differences observed in the data.

In my opinion, this major error needs to be rectified for the study to be able to address its objectives. I have listed how this might be accomplished in the attached report.

Yours sincerely,

Dr Luke Knibbs

General comments

The project sought to determine whether:

(a) Trains operating on the Hunter Valley rail network are associated with elevated particulate matter concentrations; and

(b) Loaded coal trains operating on the Hunter Valley rail network have a stronger association with elevated particulate matter concentrations than unloaded coal trains or other trains on the network (and by inference contributing to ambient rail corridor particulate levels).

To me, these general objectives imply assessment of multiple particulate matter (PM) sources, namely; (1) emissions from the train's power source (e.g. diesel engine), (2) resuspension of settled particles on or near the track, and (3) fugitive emissions from the train's cargo (where applicable). This is not mentioned in the report, and I think it should be stated at the beginning to make it absolutely clear *what* was being measured.

All PM sources are effectively treated as one and not differentiated. While the study was limited by its specifications in this regard, it makes it very difficult to determine what, if anything, was the cause of the measured concentrations.

It also means that the results are confounded. For example, an unloaded train which emits no fugitive dust but which happens to be a high emitter of diesel exhaust could register much higher PM readings than a fully loaded train with more efficient combustion and moderate fugitive emissions of coal dust. In such a case, it could be concluded that loaded trains do not contribute excess PM compared to unloaded trains. This would be numerically correct, but based on data confounded by the effects of diesel exhaust.

One would hope that the sheer number of trains sampled would offset this to an extent, but it is not possible to determine this from the report. This issue is discussed further in the following section in point 12 under the minor issues sub-heading.

Specific comments

Major issues

Section 5.2 ('Assessment of contribution by train type') described the statistical analyses used. It was assumed that if the 95% confidence intervals around the mean of PM concentrations for an individual train type overlapped those of another type, then no statistically significant difference existed between the 2 means. This did not sound correct to me, and I'm unaware of such a method to determine statistical significance.

However, since I'm not a statistician I consulted a colleague who is and who has extensive experience in air quality research (A/Prof Adrian Barnett at QUT). He confirmed that the analysis used in the report is erroneous.

By making this assumption, potentially significant differences are undetected and the true number of such differences is likely to be underestimated. This affects many of the conclusions drawn (at least those reporting significance), and makes it impossible to determine if the study met its objectives or not.

It's also impossible to say whether this error has a profound or comparatively minor effect on the conclusions until the analysis is repeated using an appropriate method.

Regarding how this might be rectified, I quote verbatim from A/Prof Barnett:

"An ideal statistical model would be a regression model using the six second averages as the dependent variable and train type as the independent variable. Statistical significance could then be compared by comparing the regression coefficients for the train types using p-values and 95% confidence intervals. Using the six second data (rather than the average during a train's passing) would account for the fact that coal trains have more data than passenger trains, hence the averages for coal trains should be more accurate. Averaging the results to a per train passing figure ignores this difference, which is potentially very important as the coal trains could have five times the amount of data (Section 5.1). Statistically significant differences would be more likely for coal trains as they had more data."

And:

"Wind direction could be analysed using a circular variable to identify the peak direction. An interaction between a circular variable for wind direction and train type could be added to the regression model suggested above. This would test whether wind direction influenced the pollution levels of the trains."

Minor issues

(1) Page viii and elsewhere: The phrase 'Total Suspended Particulates' (TSP) is used to refer to measurements made by the Osiris instrument. As the name suggests, TSP refers to all airborne PM and is not defined by a size threshold like PM_{10} or $PM_{2.5}$. The Osiris instrument is only capable of measuring particles up to 20 μm, so it can't measure TSP. The report correctly points this out on page viii, but the ongoing use of TSP to describe the Osiris measurements is misleading and it is impossible to compare with any previous or future measurements of *actual* TSP.

(2) Page 4: The Osiris instrument has a heated inlet (60° C). As the aerosol being measured when trains passed would've contained diesel combustion products, this raises questions about the evaporative loss of PM samples containing volatile components, leading to an underestimation of concentrations. I don't expect that the authors would've done a detailed calibration to assess this, but they should at least find some information to indicate whether or not this was likely to be an issue (from the manufacturer or the literature) and have a statement to this effect in the report.

(3) Page 4: Light-scattering photometers like the Osiris typically report measurements that are substantially different from those recorded by compliance monitors. This is noted in the report, and it is stated that the emphasis in the study was on *relative* differences, which should be unaffected by the measurement bias. However, there is no indication of the likely magnitude of the bias. This makes is very difficult to compare the results to any previous or future work. Again, I don't think that a detailed calibration is required, but a sense of what the size of the bias is should be included.

(4) Page 6: The statement "TSP and the finer particle size fractions will remain suspended for many tens of metres downwind of the emission source. Hence, the distance from each of the tracks to the Osiris monitor is unlikely to result in substantially different concentrations of TSP, PM10 and PM2.5." is unsubstantiated. It may not be the case for *actual* TSP.

(5) Page 7. 30% data loss over 2 months is obviously less than ideal. It is stated that the loss has not biased the data in any way, but the assessment used to determine this was fairly cursory. A more detailed assessment might give additional credibility to this statement.

(6) Page 9: Was it universally the case that coal trains on the DC track were unloaded? Was any validation of the train data performed?

(7) Page 10: Section 4.4.2 – was any validation of the algorithm undertaken based on comparisons with observation as per the data synchronisation?

(8) Page 13: The number of 6-sec PM averages was not increased for passenger trains because of their much shorter pass-by time. Presumably these trains could also lead to suspension/entrainment of particles after they passed. Were any such trends observed? If not, then it gives the above decision more credibility. If so, then perhaps these should be analysed similarly to the non-passenger trains.

(9) Page 15: See comments under 'Major issues' heading above.

(10) Page 16: Section 6.1 – rather than simply analysing by train type, the overall weight of the train, length or engine capacity/type would be more telling and help to identify the source of emissions. I understand that this is probably secondary to main objectives, but it would be very useful in addressing objective (b) in my opinion.

(11) Pages 16-21: See comments under 'Major issues' heading above.

(12) Page 21: As the project was concerned with differences between loaded and unloaded coal trains, I think a good opportunity to probe the role of loading might've been overlooked. For example, the ratios of fine particles to TSP could potentially be used to assess what the likely predominant sources of PM were (e.g. combustion vs. coal dust or resuspension). Perhaps this is not relevant to this report, where the focus is purely on reporting measured concentrations, but it really would've made the results more compelling (analytical errors notwithstanding).

(13) Page 23: I think more should be made of the effects of the dry weather in the main measurement period. It is intuitive that wet weather reduces resuspension and increases 'washout' so the difference between the pilot study and the main study should be explained in that context.

(14) Page 26: The 'limitations' section should be expanded to reflect my comments above.

(15) Page 27: See comments under 'Major issues' heading above. The same applies to the executive summary and all other sections that report the results of statistical analyses.

(16) Page 51: Appendix A – looking at the data points in question, some appear to be part of a distinct peak, rather than due to just noise. It's stated that the overwhelming majority of these data were collected when no trains were passing. This raises the question of whether the sub-deflection limit data that are also part of the peak should be excluded if it's thought that they're due to instrument error. If this is not the case, and the data are legitimate, then even the deflection limit peaks should be left in place, as they are better than no measurement at all.

(17) Page 57: See comments under 'Major issues' heading above.

Summary

There is major flaw in the statistical analysis used, which needs to be rectified for the study to meet its objectives. It's not possible to say to what extent this will alter the findings of the report. It has the potential to underestimate the importance of differences in PM concentrations between different train types.

There are a number of minor issues which should be considered when amending the report to improve its scientific rigour. These are less likely to alter the substantive findings compared to the analytical error described above.