DATA STATISTICS

ANALYSIS MATHEMATICS

AUSTRALIA INFORMATION

STRATEGIC INFORMATION CONSULTANTS

Review of CSIRO Report on Burrup Peninsula Rock Art Monitoring

May 2017

Project: DER/2

Review of CSIRO Report on Burrup Peninsula Rock Art Monitoring

May 2017

Client: Department of Environment Regulation

Project: DER/2

Consultants: Dr John Henstridge
Dr Kathy Haskard
Marzieh Mehryar

Data Analysis Australia Pty Ltd
97 Broadway
Nedlands, Western Australia 6009
(PO Box 3258 Broadway, Nedlands 6009)
Website: www.daa.com.au
Phone: (08) 9468 2533
Facsimile: (08) 9386 3202
Email: daa@daa.com.au
A.C.N. 009 304 956
A.B.N. 68 009 304 956

Executive Summary

Data Analysis Australia has reviewed a draft report by the CSIRO, *Burrup Peninsula Aboriginal Petroglyphs: Colour Change & Spectral Mineralogy 2004–2016 by* Noel Duffy, Erick Ramanaidou, David Alexander & Deborah Lau (the "Draft Report"). The Draft Report covers the data collection and analysis with respect to a monitoring project running since 2004 focusing on possible effects on the rock art due to industrial developments. As such it is the latest in a number of reports by the same CSIRO group that have presented earlier data from the monitoring program.

Data Analysis Australia examined a number of these earlier reports in 2016 as part of a project evaluating a possible third party statistical analysis. That evaluation highlighted a number of shortcomings with both the data and the statistical analysis and made a number of recommendations. We understand that the contract for the 2016 monitoring program required the CSIRO to address the recommendations of our 2016 review. Our understanding is that the Draft Report aimed to address these shortcomings.

Our review found that a considerable amount of work has been done to address some of the concerns. In particular there have been substantial improvements to the statistical analysis of colour changes using linear mixed models and greater care has been taken to highlight the problems associated with the BYK spectrophotometer used in the early years of the monitoring program. There also appears to have been action taken to better manage the data, both to make it available for analysis and to preserve it for future years.

However significant work remains if the 2016 Recommendations are to be addressed. The progress towards meeting the 2016 Recommendations can be summarised as follows:

1. *The historical data collected by the CSIRO should be systematically archived and held by DER …***Largely met**.
2. *The CSIRO should be asked to revisit the cross calibration issues with the BYK and KM spectrophotometers …* **Not addressed.**
3. *An analysis similar to that of Black and Diffey should be conducted using verified ASD estimates of* L^*, a^*, b^*, … **Partially met**.
4. *Future work by the CSIRO should be based upon an agreed analysis plan certified by a competent statistician ….* **Not met**.
5. *Consideration should be given to expanding the number of measured sites and in doing so, improving the balance of the design to include more effective controls …* **Not met**.
6. *To maintain scientific rigour, future data collection should follow a fully documented and detailed protocol …* **Not met**.

If the work to meet these recommendations cannot be completed for this Draft Report it should at the very least be highlighted as work in progress so the reader is not given to think that the Draft Report is complete or its conclusions final. In particular:

- The use of the BYK data is highly problematic and while we understand that efforts have been made to find possible ways in which it might have been corrupted, it is a reasonable statement that little if any scientific weight can be given to it. This needs to be made more prominent, and indeed the right solution is probably to assign the BYK data to a historical note.
- This raises the importance of the ASD spectrograph data and its derived colour measures that appear to have been collected with reasonable consistency with one instrument since 2004, although we have some concerns about the 2004 component. The Draft Report gives prominence to the spectra from this instrument and to the spectral parameters (near infrared spectral features of direct relevance to certain mineralogical components) but little attention has been given to the ASD colour measurements.
- The presentation of the ASD spectral parameters is not particularly helpful (numerous small barely readable plots) and no statistical analysis is conducted. There is little purpose presenting them as done in the Draft Report and no purpose if they are not going to be statistically analysed.
    - Our initial examination found that they contained some statistically significant trends, but it is not clear whether these are relevant to the preservation of the rock art.
- The application of linear mixed models to the ASD colour data is a marked improvement on the previous CSIRO reports and should be commended. However the presentation of the models is far from clear and it appears that the most basic test (of whether the contrast between engravings and their background is changing over time at a different rate for sites close to the industrial developments than for the northern control sites) was not included. More detailed analysis and clearer reporting of that analysis is required if the Draft Report is to be convincing.

As a result of these shortcomings the Draft Report could be considered misleading in its statement that "the data is scarcely unequivocal and there are reservations on the conclusions of the statistical analysis". In our opinion such a statement may lead to an impression that the data is not capable of giving clarity, whereas a more thorough statistical analysis may be able to resolve the question more completely.

It is unavoidable that in reviewing the Draft Report one is forced to review the design of the monitoring program. It is unfortunate that, for whatever the reasons, this was not based upon firmer statistical principles. More sites should have been monitored, especially more control sites and the number of replicate measurements taken at each point seems excessive (or unnecessary). Furthermore, as there are concerns that the measurement process is damaging the engravings, a fractional design is indicated where not all spots were measured each year. It is not possible to fix the historically collected data but moving forward consideration should be given to redesigning the monitoring scheme.

Table of Contents

1. Introduction

The Indigenous rock art on the Burrup Peninsula in the north-west of Western Australia is recognised as unique in the world, being the oldest extant artistic expression. The area is also a key economic hub for Australia, bringing together mining, natural gas industries and a major port. The potential for the industrial development to affect the rock art has been long recognised and a monitoring program has been in place to assess what the effects might be on the rock art. Part of this has been an annual data collection and reporting by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) using photometric methods to measure possible changes to the rock art[1].

In recent years questions were raised about the statistical analysis of this data, resulting in a draft paper[2] (hereinafter termed the Black & Diffey Draft Paper) suggesting that significant changes had taken place, in sharp contrast to the findings in the CSIRO reports. In 2016, the Department of Environment Regulation of Western Australia (DER or the Department) engaged Data Analysis Australia to review the statistical issues raised in the Black & Diffey Draft Paper, making reference to the data itself and the CSIRO reports.

The 2016 review by Data Analysis Australia was primarily focused on the Black & Diffey Draft Paper, with the CSIRO work and reports examined only when it became evident that the Black and Diffey work may depend upon problematic data. The result of that review (hereinafter termed the 2016 DAA Review) was to recommend that the Black & Diffey Draft Paper should not be published while there exists such uncertainties about the data.

The 2016 DAA Review made a number of specific recommendations and most of these related to the CSIRO's management and analysis of the data. We understand that the CSIRO accepted these recommendations and undertook to address the issues they raised. The draft report by the CSIRO titled *Burrup Peninsula Aboriginal Petroglyphs: Colour Change & Spectral Mineralogy 2004–2016* (hereinafter termed the Draft Report) is their first opportunity to provide data and analyses that meet these recommendations.

The Department has engaged Data Analysis Australia to review the Draft Report. Data Analysis Australia's role is that of statistical experts with relevant experience. We understand that the CSIRO work has been criticised in several non-statistical respects, such as the physical limitations of certain instruments. We do not attempt to review the Draft Report in regard to such issues.

---

[1] Each of the CSIRO reports covers the data over a period from 2004 and is generally published in the year following the latest data.

[2] Black, J and Diffey, S., 2016, Reanalysis of the Colour and Mineralogy Changes from 2004 to 2014 on Burrup Peninsula Rock Art Sites, unpublished.

1.1 Structure of This Report

An outline of the findings from the 2016 DAA Review is given in Chapter 1 along with an overview of the data collection process current used for the monitoring program. A detailed review of the statistical content of the Draft Report is presented in Chapter 2. In Chapter 3 more general issues are discussed, including limitations of the data as a result of the design decisions made at the start of monitoring in 2004. Chapter 4 summarises our findings, particularly with respect to compliance with the recommendations made in 2016.

1.2 Recommendations from 2016 DAA Review

For completeness and to provide context, the recommendations of the 2016 DAA Review are as follows:

1. The historical data collected by the CSIRO should be systematically archived and held by DER, with consistent naming conventions, both to provide a baseline record and to facilitate comparisons with future data. The archival data format should enable ready access to the data via standard statistical software such as R.
2. The CSIRO should be asked to revisit the cross calibration issues with the BYK and KM spectrophotometers, both to ensure that the historical data is properly understood and to confirm whether or not the historical BYK data is capable of comparison with current and future measurement instruments.
3. An analysis similar to that of Black and Diffey should be conducted using verified ASD estimates of L^*, a^*, b^*, ideally using the original ASD spectra rather than the averaged spectra.
4. The publication of the Black and Diffey paper should ideally wait until the problems with the BYK data are resolved or should use the ASD data.
5. Future work by the CSIRO should be based upon an agreed analysis plan certified by a competent statistician. Since each year the CSIRO reports have covered the full data set since 2004, it would be appropriate for the next published report to incorporate this improved analysis and in doing so, make it clear that it should replace the analyses in their previous reports.
6. Consideration should be given to expanding the number of measured sites and in doing so, improving the balance of the design to include more effective controls, if feasible.
7. To maintain scientific rigour, future data collection should follow a fully documented and detailed protocol, and ensure that departures are documented.

Note that the above recommendations summarise the findings of the 2016 DAA Review, with further details provided in the full 2016 DAA Review. Note also that Recommendation 4 is not relevant to the current review.

In addition to commenting on the degree to which the Draft Report meets these 2016 Recommendations, we also make a number of fresh recommendations (that might be termed the 2017 Recommendations).

1.3 Information and Data Provided

We were provided with:

- The draft CSIRO report *Burrup Peninsula Aboriginal Petroglyphs: Colour Change & Spectral Mineralogy 2004–2016 by* Noel Duffy, Erick Ramanaidou, David Alexander & Deborah Lau (the "Draft Report").
- Various data sets provided by the CSIRO.

1.4 Overview of the Data Collection

This review has focused on the Draft Report itself but, where necessary, refers to the data and the code used in the analysis.

It is appropriate here to briefly describe the data collection framework – we do so here as the Draft Report and its terminology are not always clear on some methodological details.

- An attempt was made to define areas that might be affected by the industrial developments and those what would not be – in statistical parlance *intervention* and *control* areas. The control area was on islands on the north of the peninsula.
- A total of 10 *sites* with engravings were monitored, most since 2004 but 3 were introduced only in 2014. Two sites were in the control area (monitored in all 13 years), and the remainder in the treatment area.
- The logic was that if the treatment sites changed over time in a manner different from the control sites then this might be *prima face* evidence that the treatment sites were affected by the industrial areas.
- At each site measurements were taken at a number of *spots*. The numbering or labelling of the spots was only meaningful within each site – the first spot of one site was not linked to the first spot at another site.
- Each spot was, somewhat confusingly, made up of two measurement *points*, one which was on the engraving at that site and one that was close by but not on the engraving, that is on the background. (The Draft Report does not use the terminology of a point, but we do so here to reduce ambiguity.) These points were defined when the monitoring at each site was first done and ideally the measurements should attempt to measure at exactly the same points every year, no simple task since it is not appropriate to mark the rocks in any way.
- The *type* of a point refers to whether it is on the engraving or the background. It is the colour contrast between the engraving and background that makes the engravings visible to the human eye and this contrast is essentially measured by the difference in colour variables between the two types of point at each spot.
- At each point, measurements were made with up to three instruments:
    - The BYK spectrophotometer (BYK) that measured colour;
    - The Konica Minolta spectrophotometer (KM) that similarly measured colour; and

  - The Analytical Spectral Devices spectroscope (ASD) that measured the full spectrum covering the visible and near infrared.

- At each point, each instrument recorded several measurements, a standard approach to both achieve greater accuracy and to understand the nature of the variation in measurements. It appears that sometimes several measurements were made taking the instrument measuring head off the point between measurements so that each measurement replicated the full process of placing the head on the point as accurately as possible (we call these *replicate* measurements) while at other times measurements were repeated without taking the head off the point between measurements (we call these *repeat* measurements). The Draft Report p. 2 states that *repeat* measurements (as defined here) were made from 2013, while p. xiii suggests this was only from 2015 – possibly a typographical error.

- Colour was measured in terms of the L^*, a^*, b^* framework where L^* is essentially a measure of overall reflectance or "lightness", a^* a measure of the red-green contrast and b^* a measure of the blue-yellow contrast. The BYK and KM instruments measured directly in this space while the L^*, a^*, b^* values could be calculated from the visible part of the ASD spectra.

- At a number of places in the Draft Report differences in colour measurements are reported as ΔE^* (sometimes also referred to as ΔE and ΔE^*Lab in the Draft Report) defined by $\Delta E^* = \sqrt{\Delta L^{*2} + \Delta a^{*2} + \Delta b^{*2}}$. We understand that this is a standard method that approximates the manner in which the human eye perceives differences in colour.

Recommendation 1. The Draft Report should include a succinct description of the measurement framework similar to that given above (with corrections as appropriate).

2. Review of the Draft Report

2.1 Colour Measurement (Chapter 4)

The bulk of Chapter 4 of the Draft Report presents summaries of the data collected since 2004. This comprises of:

- Table 4 to Table 13 which give average L^*, a^*, b^* values collected at each point using the BYK and KM instruments. The transition between the BYK instrument and KM instrument is made clear (black and red text, respectively). Year-by-year differences are presented for every point and every year except for 2009 where the change in instrument rendered this inappropriate.

- Table 14 presents average colour changes (year by year) for each site in each year, following the black-red convention of the previous tables to distinguish between the BYK and KM instruments. The table also includes comparisons between 2004 and 2016 – since these times are measured with different instruments, a cross calibrated version of the BYK measurements of 2004 was used.

- Table 15 and Figure 16 to Figure 25 (number inferred as it lacks a caption) examines the contrasts (the difference between the engraving and background points within each spot measured as ΔE^*) at each spot over time. Again the distinction between BYK and KM measurements is made clear. Simple linear regressions are applied to examine trends over time.

This Chapter addresses some of our 2016 recommendations but fails to adequately address the issues arising from the problematic BYK spectrophotometer. It reads as if there is a recognition that the BYK data is problematic but it is still reluctant to say it should not be used. Indeed, it does use the BYK data in Table 14 to provide ΔE^* comparisons with 2004, even for the more recent KM data. These comparisons rely doubly upon the questionable BYK data – for the 2004 baseline itself and for the later 2009-2010 data used in the cross calibration against the KM data.

We also have significant concerns about the structure of the cross calibration models used – these appear to be the same as we reviewed in 2016 and have not been updated despite our 2016 Recommendation that they be revisited. The models are highly inappropriate in that they are driven primarily by the KM data at each point in 2009-2010 in such a way that the resulting comparisons in the last column of Table 14 of the Draft Report are really against this period rather than against 2004.

This Chapter also fails to reference the ASD measures of L^*, a^*, b^* that are in fact the only consistent measuring process applied since 2004. This is, in a sense, the "elephant in the room", although we acknowledge the treatment of ASD colour measurements in Chapter 6.

This Chapter still uses the ΔE^* approach that is barely appropriate in this context. It is useful for detecting large changes from one year to the next but is essentially useless when considering trends and whether changes are consistent across spots or sites. For example the use of average ΔE^* in Table 14 (averaging across spots in each site) completely ignores the issue of whether spots within a site are changing in a similar fashion or not.[3] (In making our 2016 Recommendation that a modelling approach should be used, we had thought that this would lead to less use of ΔE^*.)

Finally, the analyses of Figures 16 to 25 include very simplistic fitting of trends but fail to use the replicated measurements to determine whether these trends are meaningful. Additionally there is no comment on the consistency or otherwise of the trends at different sites and spots.

Overall we believe that this Chapter needs to be substantially modified to fully recognise the changed circumstances with the BYK data being questioned. It needs to be made clear to the readers what credibility they should attach to the BYK data – with the information we have available it would seem that readers should be advised to not rely upon the BYK data.

[3] The use of ΔE^* in this manner is equivalent to describing a journey by giving the distances of each leg but not giving the directions. Adding up the distances may tell how far one has travelled, but gives little idea of how far from the origin you might finish up.

Recommendation 2. Chapter 4 of the Draft Report should more directly address the poor quality of the BYK data. We recommend that the BYK data not be used, particularly not as a reference for changes since 2004, and that it be recommended to the readers not to rely on this data, unless cogent arguments can be presented to the contrary. Introducing the ASD colour measurements is appropriate in this Chapter .

Recommendation 3. In Chapter 4 and other parts of the Draft Report less reliance should be placed on the ΔE^* measure.

2.2 Spectral Minerology (Chapter 5)

The bulk of Chapter 5 of the Draft Report presents summaries of the ASD data collected since 2004. This comprises of:

- Figure 27 (number inferred as it lacks a caption) giving the ASD average spectra at each point for each year; and
- Figure 28 (also inferred) giving plots of the "spectra parameters" over time for each point. I understand that these spectral parameters are calculated from average spectra and correspond to particular features in the near infrared spectra that have chemical or mineralogical significance.

These figures are very similar to those presented in previous years. No statistical analysis is presented and the comment (page 71) that the spectral parameters "show relatively large variations but no systematic changes or trend" is surprising in the absence of any analysis. The small scale and limited readability of Figure 28 is ill-considered and should be reviewed.

While we cannot comment upon the chemical or mineralogical significance of the spectral parameters, our preliminary statistical investigation (using data files provided by CSIRO) suggests that there are *statistically* significant trends in some of these spectral parameters. Hence the dismissal of trends without a proper statistical analysis is premature and very likely incorrect.

The average spectra presented here hide important information on the repeatability of the measurements. However to the extent that the Chapter does not include any statistical analysis, this is less of an issue. However in any proper statistical analysis it would be appropriate to consider the repeatability of the measurements.

Recommendation 4. The Draft Report should include a proper statistical analysis of the spectral parameters in Chapter 5 to determine whether or not there have been significant changes.

2.3 Statistical analysis of BYK, KM and ASD colour measurements (Chapter 6)

Chapter 6 focuses on analysis rather than colour measurement data, although in Figure 29 to Figure 68 it does graphically present the average results for each spot with the BYK, KM and ASD instruments. It is commendable that all three measures are used.

Section 6.2 considers the issue of cross calibration between BYK and KM instruments. In our opinion this is primarily of historical interest since it is not clear if the BYK data has practical value. However it is appropriate to make several comments on this cross calibration:

- The Draft Report does not contain any new work but relies upon models fitted in a note dated 18 June 2013[4]. Data Analysis Australia has previously criticised these calibration models, particularly since they really relied more on where measurements were made rather than what was actually measured. The models are not truly predictive in that they could not be used to convert from BYK values to KM values at a new site.
- The authors of Chapter 6 essentially agree with the criticism of the models, saying that the comparison with the KM and ASD instruments "tends to discredit the accuracy of the data measured using the BYK machine".
- A recommendation is made that further work be carried out to compare the ASD and KM machines.
- A final comment is made that the BYK data did not demonstrate any difference in colour change between the southern sites close to industry and the northern control sites". Given the substantial problems with the BYK data, it would be unfortunate if this was interpreted as supporting a statement of no change.

Overall the discussion in this section makes it clear that the BYK data has little place in the ongoing monitoring program.

Recommendation 5. The findings given in Section 6.2 of the Draft Report should be given greater prominence overall in the Draft Report, with the clear message that the BYK data has limited if any value.

Recommendation 6. The comments that the BYK data does not indicate change should be deleted or have strong caveats placed on it.

Section 6.3 of the Draft Report focuses on fitting models to the colour data, focusing on the KM data. This is, of necessity, somewhat technical as it involves the use of linear mixed models – simpler methods such as the presentation in Figure 72 of the Draft Report are of no use except to illustrate the scale of the variation.

Section 6.3.1 presents the main analysis and does so in a very brief manner. While it is understandable that the authors wished to avoid swamping the reader with detail, currently there is insufficient detail for the reader to be sure about what models are being discussed. We had the benefit of receiving from the author the R code for the analysis and we could hence verify that the numerical results as presented are correctly taken from certain models. The use of R is highly appropriate, as is the use

[4] David Alexander, Deborah Lau and Tracey Markley,; Erick Ramanaidou, *Regression for Calibration of Burrup Peninsula Rock Art Colour Measurements*, 18 June 2013

of mixed linear models.[5] **The bigger question is whether the most appropriate models have been fitted.**

At this point it is necessary to introduce some notation for representing these models. If we have factors that we might term A and B which we are trying to use to explain the variation in a measurement y (say the reflectance), we could have a number of possible models that by convention we write as $y \sim expression$. The possible models are:

$y \sim 1$	y does not depend upon either A or B;
$y \sim A$	y depends only on A;
$y \sim B$	y depends only on B;
$y \sim A + B$	y depends only on both A and B but they act independently of each other; and
$y \sim A * B$	y depends only on both A and B and each of A and B affects the way that y depends upon the other. This model can also be written as $y \sim A + B + A.B$ where the $A.B$ term makes the inclusion of the interaction between A and B explicit.

This notation generalises to more factors in the obvious manner and allows for complex model choices.

There are at least three factors or variables of immediate concern:

- *engraving* – the difference between an engraving and its background which is what makes the engravings visible – that is, the contrast of the engraving with its background;
- *control* – the difference between the northern control sites and the southern sites close to industry development; and
- *time* – recognising that measurements may change over time, thought of as either the year or the trend over the years.

The measurement (y) could be any of the L^*, a^*, b^* components measured with any instrument. In practice the model will also have additional terms, some of which are treated as random, to provide the most appropriate analysis.

The Draft Report does not make it clear precisely which models it thinks *should* be fitted, but in Chapter 6 it focuses on models that include a *control.time interaction* term. However it could be argued that the more relevant term is the three-factor interaction *engraving.control.time* that measures the degree to which change in the

[5] R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

contrast over time is different between the control and potentially affected areas. This term is not tested in the results described in the Draft Report.

It is clear that the models chosen for reference in the Draft Report were the result of a search amongst a number of possible models. Such a model search balances the desire for a model that fits the data well and one that is not unnecessarily complex, typically by a mixture of formal model selection using methods such as the Akaike Information Criterion (AIC) and less formal methods guided by an understanding of what the factors might be.

The Draft Report does not indicate the full range of models considered although we have been provided with the R code that was used. These include some models that do include an *engraving.control.time* term but no reference is made in the Draft Report on the significance or otherwise of such terms. We regard models with these terms as particularly important for several reasons:

- Our initial analysis of the KM and ASD using similar mixed model methods suggests that there may be significant terms overlooked in the Draft Report;
- The *engraving* factor relates to the visual perception of the rock art and is thus particularly relevant; and
- The measurements at the engraving and background points at a spot, were made by the same team on the same day under the same conditions, so the comparison can be expected to be statistically more precise than other possible comparisons. That is, the statistical power could be particularly high.

The Draft Report's conclusion of the analysis is that there is no evidence for a statistically significant effect, based in part on the number of measures analysed, models fitted and parameters estimated suggesting that the few terms reaching formal significance could have done so by chance. (These important points are not well summarised in the Executive Summary.) In this context it is important that such a null result is backed up by sufficient detail to indicate that no reasonable analysis of the data would show a statistically significant or statistically suggestive result. This requires detail on the range of models considered and significance of relevant terms. As it stands, the Draft Report neither makes it clear which models were finally reported, and which models were considered and excluded, nor why they were not preferred.

The adage that "absence of evidence is not evidence of absence" is pertinent here. It is not sufficient simply to find no problem. The question is whether the search has been sufficiently thorough and looking for the right things. It is required to demonstrate that a sufficiently diligent investigation has been carried out, together with a statement of the nature and size of any problem that would be revealed if it was present.

It might be argued that the irreplaceable significance of the rock art provides a case for applying the precautionary principle, whereby strong support is needed for a statement of "no harm".

Recommendation 7. The Draft Report needs to provide substantially more information on the mixed models considered to demonstrate reasonable support for the conclusion that there is no evidence of impact of industry on the rock art.

In investigating models with the ASD data we discovered an intriguing property in the residuals – it appears that there have been some substantial changes in the ASD data over the years. Figure 1 shows boxplots for the residuals for mixed models fitted to the L^*, a^* and b^* data. (The precise model is not important – the model effects are small compared with the scale of the residuals.)

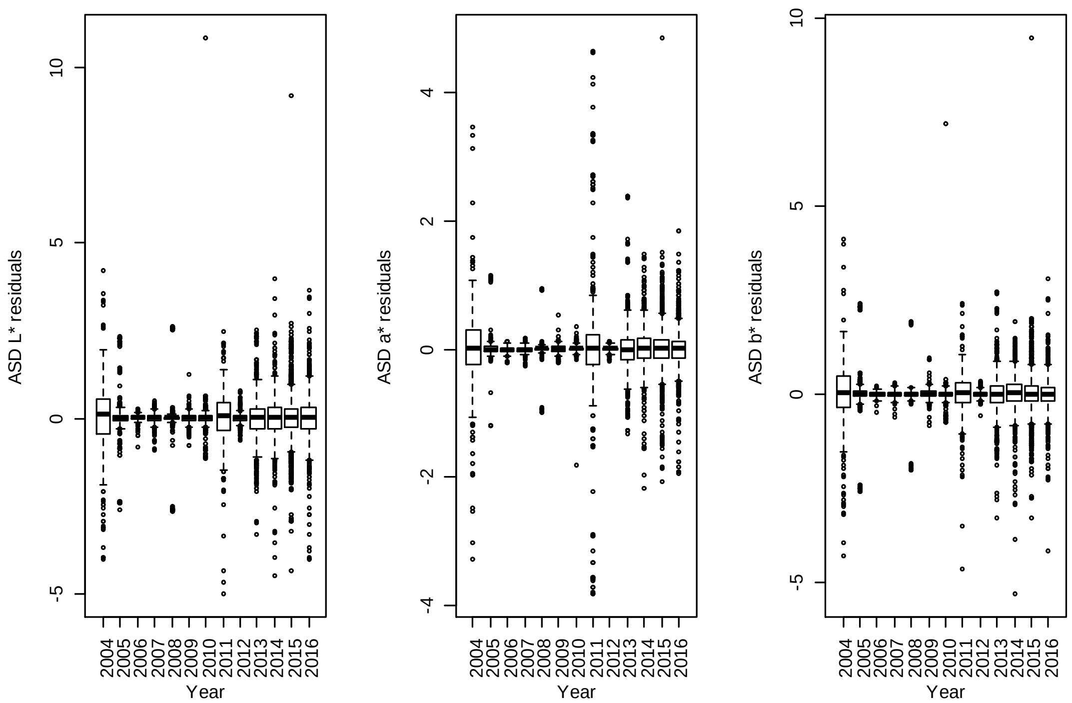

Figure 1. Box plots for the residuals of a model for the ASD L^* (left panel), a^* (centre panel) and b^* (right panel) measures for the years 2004 to 2016.

It is evident that the residuals for 2004, 2011, and 2013 to 2016 are much larger than the residuals for 2005 to 2010 and 2012. We understand that for the period 2013 to 2016 the spectrometer head was removed and replaced between readings, meaning that in our terminology they are proper replicates. This would account for the larger residuals in these years. Possibly the same practice was also in place in 2004 and 2011; however there is no statement to this effect. Within the period 2005 to 2010 the differences in the scale of residuals is also significant, suggesting further changes in practice have taken place. However such conjectures need to be clarified and documented.

If indeed there have been changes in the measurement practices, they must be taken into account in the analysis.

Improper allowance for effects such as these can lead to incorrect and misleading statistical significance tests. The fixed effect parameter estimates of interest might not change much, but an incorrect variance model can yield different standard errors and significance test results that might either imply differences where there are none, or fail to detect true effects. Appendix A illustrates an example analysis that takes the unequal residual variances into account, **and reveals a statistically significant three-factor interaction between Background vs Engraving points, Control vs Industry Sites, and Year**. This immediately raises concern and warrants investigation in relation to possible impact of industry on the rock art. The example also suggests some potential anomalies in the 2004 ASD data.

Recommendation 8. The Draft Report needs to properly document *all* changes in measurement practices and, where appropriate, incorporate these into the analyses.

3. General Comments

It is evident that the Draft Report is constrained by a less than ideal monitoring framework in place from 2004. In our opinion, this leads to a lack of clarity in purpose, inefficient design of data collection and a lack of focus in analysing the results. We cannot comment on the degree to which the 2016 project team is able to mitigate some of these shortcomings, making it unclear what a reasonable expectation of the 2016 Draft Report should be.

However, we can make some suggestions on how the design of the data collection and its reporting can be improved.

3.1 Design

The statistical theory of the design of experiments is well established and is able to guide efficient and effective monitoring. From the information available to us it does not appear to have been effectively used in the 2004 design.

Several principles of the design of experiments that should have been applied are:

- There should be a clearly stated definition of effects or changes that the monitoring should be able to detect. One can appreciate that in 2004 an understanding of the possible effects would not have been as well developed as it is today, but it should have been possible to state some hypotheses for change. The design should then have been optimised for monitoring such possible changes and the analysis method obvious from the start. Ideally most of the features of the models fitted in 2016 should have been obvious in 2004.

- The sources of error should have been understood and quantified first since the purpose of multiple measurements is to overcome the inaccuracy of single measurements. This should have considered the variation between sites, between spots on a site, between points within a spot, between times for each point, between replicate measurements (taking the recording head off the surface between readings) and repeat measurements (with the recording head kept in place). Only such an understanding can indicate the optimal number of sites, spots within sites, replicates with spots and duplicates within replicates.

- The role of the control sites should have been clearer, giving the appropriate balance between the numbers of control sites and the number of treatment sites. It is worth noting that for simple designs with a constraint on total effort it is optimal to have equal numbers of control and treatment sites.
- The size of the data collection should be based upon achieving a required level of statistical power to detect a change of a certain size. The size of changes that are *meaningful* or of practical importance should be considered in this context.

We have not been able to perform a full redesign, and it would not be appropriate to do so without better knowledge of the constraints. However we suggest the following points are considered.

- Additional control sites should be added, bringing the number close to the number of treatment sites. We recognise that a design aimed at assessing the effect of distance from the industrial developments is still likely to have fewer control sites than treatment sites.
- The small total number of sites in the current design prevents formal analysis of site specific effects, such as distance from the sea, rock types, age and orientation. If possible, additional sites should be added with thought given to balancing the selection across such factors.
- It is likely that fewer replicate measurements could be made at each spot since at that level results are more repeatable. Hence the monitoring program will almost surely be improved with more sites but fewer measurements at each site.

Recommendation 9. A formal design document should be produced before the next period of data collection, based upon established principles of the design of experiments. This document should fully explain any departures from the ideal, including the need to maintain a certain level of consistency with the existing data that, despite all its limitations, must remain part of future analysis.

We recognise that a redesign of the data collection would take time, particularly if there are cultural issues to the addressed in the choice of additional sites. We also recognise that such a redesign is not without cost. However we regard this as being sufficiently important that it would be preferable to miss a year of data collection than to delay the redesign.

3.2 Analysis

The Draft Report has a strong emphasis on presenting data, albeit in summary form. In some contexts this is an appropriate approach. For a complex data set such as in the Draft Report, this is inadequate and the meaning and significance of the data only becomes clear when it is subjected to sophisticated analysis. Hence proper analysis should become more central. This has the added benefit of clarifying what type of changes the monitoring program targets.

Recommendation 10. A formal analysis document should be produced in parallel to the design document before the next period of data collection.

3.3 Terminology

The Draft Report is not fully consistent in its terminology. For example, the term "spot" is sometimes used to represent an engraving/background pair and sometimes one of the pair. Similarly the description of the replicate measurement process is unclear.

3.4 Statistical Significance

The Draft Report implicitly uses statistical significance in the context of traditional hypothesis testing with a 5% level of significance. That is, an effect is deemed significant only if an effect of the observed size could only occur by chance alone 5% of the time.

While this is a commonly used approach, it was developed in the context of scientific research and was precautionary in that it was to "don't believe new facts until clearly demonstrated". This would mean that an inadequately sized or poorly conducted experiment would not be given much weight. It could be argued that the precautionary approach to unique art such as exists on the Burrup Peninsula should be "don't believe it is safe unless clearly demonstrated". An inadequately sized or poorly conducted monitoring program should *not* be the basis of assuming the art is safe.

Ultimately the decision on the approach to be used must be guided by statistical principles but is not simply a statistical issue. Hence we can only highlight the question, not provide the answer beyond suggesting that calculations of statistical power are likely to be relevant.

4. Summary

It is apparent that the authors of the Draft Report have made a number of changes relative to previous years' report, in part in response to Data Analysis Australia's recommendations of 2016. These changes are for the better.

However much remains to be done. In particular the conclusions for the Draft Report, namely that there is no evidence of a significant change affecting the rock art near to the industrial developments are not convincing in their current form. For that reason we recommend that the authors be asked to address the points as recommended above.

In terms of the recommendations that were made by Data Analysis Australia in 2016, the following summarises the compliance:

1. The historical data collected by the CSIRO should be systematically archived and held by DER, with consistent naming conventions, both to provide a baseline record and to facilitate comparisons with future data. The archival data format should enable ready access to the data via standard statistical software such as R.

We would describe this as **largely** met:

- The management of the data appears to have improved considerably, although it is not perfect in that the metadata – details of how the data was collected and hence what the data might therefore mean – is not systematically available.
- We remain concerned that there are undocumented features of the data collection process that should be taken into account in any analysis. For example, the substantial year-to-year variation in the ASD data is only partially explained by the differences in recording practices described in the Reports, and the descriptions that do exist are not always consistent.

2. The CSIRO should be asked to revisit the cross calibration issues with the BYK and KM spectrophotometers, both to ensure that the historical data is properly understood and to confirm whether or not the historical BYK data is capable of comparison with current and future measurement instruments.

We would describe this as being **not addressed** in the Draft Report:

- The cross validation methodology for the BYK data has not been revised and is still deficient. The Draft Report appears ambivalent about the utility of the BYK data.
- Whilst the Executive Summary does state "the BYK spectrophotometer data appears unreliable for drawing conclusions on colour change in the rock art", the data is still given undeserved prominence in the report and the ASD colour data is not discussed as a credible replacement.
- However we suggest that the first part of this recommendation from 2016 concerning the cross calibration should not be given high priority as one solution may be to largely drop reference to the BYK data.

3. An analysis similar to that of Black and Diffey should be conducted using verified ASD estimates of L^*, a^*, b^*, ideally using the original ASD spectra rather than the averaged spectra.

We would describe this being only **partially** met:

- The analysis in the Draft Report does use linear mixed models as suggested by Black and Diffey. However this is poorly reported and not convincing. Whilst the conclusion given in the Draft Report is that there is no evidence of relevant changes to the rock art in the areas close to the industrial development, it remains arguable that a more careful analysis would demonstrate changes.

5. Future work by the CSIRO should be based upon an agreed analysis plan certified by a competent statistician. Since each year the CSIRO reports have covered the full data set since 2004, it would be appropriate for the next published report to incorporate this improved analysis and in doing so, make it clear that it should replace the analyses in their previous reports.

We would describe this as **not** being met:

- No formal analysis plan appears to exist. The analysis methods in Chapters 4 and 5 of the Draft Report are essentially unchanged.

6. Consideration should be given to expanding the number of measured sites and in doing so, improving the balance of the design to include more effective controls, if feasible.

We would describe this as **not** being met:

- No change has been made to expanding the data collection or to include improved controls, although we recognise that the time of the data collection in 2016 meant it could not be affected by our 2016 Recommendations.
- The Draft Report does not discuss possible changes to the design of the data collection.

7. To maintain scientific rigour, future data collection should follow a fully documented and detailed protocol, and ensure that departures are documented.

We would describe this as **not** being met:

- Documentation of the data collection protocol does not appear to have improved.

In conclusion, we are of the opinion that while the Draft Report demonstrates substantial efforts on the part of the CSIRO to improve the reporting of the data collection and to present better analysis, more needs to be done. In particular, in its current form the Draft Report is unable to dispel what might be described as reasonable concerns about the impact of industry on the rock art.

Appendix A. Illustrative Model for the ASD L* Measure

This model fits the three-way interaction of fixed effects Background/Engraving points by Control/Industry sites by Year, and takes into account different residual variances in different years, increasing reliability of tests of the fixed effects. The result is not definitive – it may be possible to improve the model further – but the results appear meaningful based on comparison with other models, and this model illustrates what is possible in analysing the difficult features of this dataset.

A.1 Function call

```
lmer22_new7_dropcorr = lmer(L ~ CI*BE*yearF +
  (1 | BE:sitespotF) + (0+BE | sitespotF:yearF)
  + (0+as.numeric(BE %in% "Bkg") | siteF)
  + (0+as.numeric(yearF %in% 2004) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2005) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2007) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2008) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2009) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2010) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2011) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2012) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2013) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2014) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2015) | siteF:spotF:BE:yearF:repwinp_2006)
  + (0+as.numeric(yearF %in% 2016) | siteF:spotF:BE:yearF:repwinp_2006)
  , data = ww1, na.action = na.omit
   , control=lmerControl(
       check.conv.grad = .makeCC("warning", tol = 3e-3, relTol = NULL)
     )
  )
```

A.2 Model Fit Summary

```
Linear mixed model fit by REML ['lmerMod']
Formula: L ~ CI * BE * yearF + (1 | BE:sitespotF) + (0 + BE | sitespotF:yearF) +
    (0 + as.numeric(BE %in% "Bkg") | siteF) +
    (0 + as.numeric(yearF %in%      2004) | siteF:spotF:BE:yearF:repwinp_2006) +
    (0 + as.numeric(yearF %in% 2005) | siteF:spotF:BE:yearF:repwinp_2006) +
    (0 + as.numeric(yearF %in% 2007) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2008) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2009) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2010) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2011) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2012) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2013) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2014) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in% 2015) | siteF:spotF:BE:yearF:repwinp_2006) +
     (0 + as.numeric(yearF %in%      2016) | siteF:spotF:BE:yearF:repwinp_2006)
   Data: ww1
  Control: lmerControl(check.conv.grad = .makeCC("warning", tol = 0.003, relTol = NULL))

REML criterion at convergence: 17853.3
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.4036 -0.0707  0.0031  0.0741  4.4113
```

Random effects:

```
Random effects:
 Groups                                Name                         Variance Std.Dev. Corr
 siteF.spotF.BE.yearF.repwinp_2006    as.numeric(yearF %in% 2016) 0.39364  0.6274
 siteF.spotF.BE.yearF.repwinp_2006.1  as.numeric(yearF %in% 2015) 0.38497  0.6205
 siteF.spotF.BE.yearF.repwinp_2006.2  as.numeric(yearF %in% 2014) 0.45971  0.6780
 siteF.spotF.BE.yearF.repwinp_2006.3  as.numeric(yearF %in% 2013) 0.39392  0.6276
 siteF.spotF.BE.yearF.repwinp_2006.4  as.numeric(yearF %in% 2012) 0.01209  0.1100
 siteF.spotF.BE.yearF.repwinp_2006.5  as.numeric(yearF %in% 2011) 0.67330  0.8205
 siteF.spotF.BE.yearF.repwinp_2006.6  as.numeric(yearF %in% 2010) 0.32660  0.5715
 siteF.spotF.BE.yearF.repwinp_2006.7  as.numeric(yearF %in% 2009) 0.01744  0.1321
 siteF.spotF.BE.yearF.repwinp_2006.8  as.numeric(yearF %in% 2008) 0.34215  0.5849
 siteF.spotF.BE.yearF.repwinp_2006.9  as.numeric(yearF %in% 2007) 0.01408  0.1187
 siteF.spotF.BE.yearF.repwinp_2006.10 as.numeric(yearF %in% 2005) 0.34879  0.5906
 siteF.spotF.BE.yearF.repwinp_2006.11 as.numeric(yearF %in% 2004) 1.54522  1.2431
 sitespotF.yearF                      BEBkg                       1.79261  1.3389
                                      BEEng                       2.22100  1.4903   0.26
 BE.sitespotF                         (Intercept)                 3.60684  1.8992
 siteF                                as.numeric(BE %in% "Bkg")   3.57220  1.8900
 Residual                                                         0.01255  0.1120
Number of obs: 9592, groups:  siteF:spotF:BE:yearF:repwinp_2006, 9340; sitespotF:yearF, 336;
BE:sitespotF, 80; siteF, 10
```

Fixed effects:

```
                          Estimate Std. Error t value
(Intercept)                33.7836     1.6095  20.990
CIIndustry                  0.7791     1.8169   0.429
BEEng                       3.9738     1.8114   2.194
yearF2005                  -2.2490     0.8010  -2.808
yearF2006                  -1.3183     0.7975  -1.653
yearF2007                  -2.0503     0.7976  -2.571
yearF2008                  -2.6749     0.8011  -3.339
yearF2009                  -1.4953     0.7977  -1.875
yearF2010                  -2.1065     0.8006  -2.631
yearF2011                  -2.7039     0.8438  -3.204
yearF2012                  -0.7423     0.7976  -0.931
yearF2013                  -3.2188     0.7614  -4.227
yearF2014                  -3.0627     0.7636  -4.011
yearF2015                  -2.5047     0.7614  -3.289
yearF2016                  -2.7175     0.7616  -3.568
CIIndustry:BEEng           -0.3688     2.0520  -0.180
CIIndustry:yearF2005        1.5367     0.9471   1.623
CIIndustry:yearF2006        0.9439     0.9429   1.001
CIIndustry:yearF2007        1.3201     0.9430   1.400
CIIndustry:yearF2008        2.0765     0.9472   2.192
CIIndustry:yearF2009        0.7131     0.9431   0.756
CIIndustry:yearF2010        1.1045     0.9466   1.167
CIIndustry:yearF2011        1.3732     0.9845   1.395
CIIndustry:yearF2012        0.5579     0.9430   0.592
CIIndustry:yearF2013        2.5608     0.9002   2.845
CIIndustry:yearF2014        1.6498     0.8897   1.854
CIIndustry:yearF2015        1.5531     0.8874   1.750
CIIndustry:yearF2016        1.3554     0.8893   1.524
```

```
BEEng:yearF2005               3.7705     1.0357   3.640
BEEng:yearF2006               4.3145     1.0303   4.188
BEEng:yearF2007               5.3755     1.0304   5.217
BEEng:yearF2008               5.3431     1.0355   5.160
BEEng:yearF2009               4.2791     1.0305   4.153
BEEng:yearF2010               3.7927     1.0349   3.665
BEEng:yearF2011               2.5315     1.1421   2.216
BEEng:yearF2012               3.4310     1.0303   3.330
BEEng:yearF2013               4.7117     0.9845   4.786
BEEng:yearF2014               4.5030     0.9879   4.558
BEEng:yearF2015               5.1816     0.9846   5.263
BEEng:yearF2016               3.2115     0.9848   3.261
CIIndustry:BEEng:yearF2005  -4.2170     1.2311  -3.425
CIIndustry:BEEng:yearF2006  -4.0583     1.2203  -3.326
CIIndustry:BEEng:yearF2007  -4.9638     1.2204  -4.067
CIIndustry:BEEng:yearF2008  -5.3898     1.2265  -4.394
CIIndustry:BEEng:yearF2009  -5.7269     1.2249  -4.676
CIIndustry:BEEng:yearF2010  -3.5413     1.2259  -2.889
CIIndustry:BEEng:yearF2011  -2.2265     1.3192  -1.688
CIIndustry:BEEng:yearF2012  -3.5575     1.2203  -2.915
CIIndustry:BEEng:yearF2013  -4.7667     1.1662  -4.087
CIIndustry:BEEng:yearF2014  -4.5370     1.1530  -3.935
CIIndustry:BEEng:yearF2015  -5.3543     1.1495  -4.658
CIIndustry:BEEng:yearF2016  -3.2745     1.1510  -2.845

Correlation matrix not shown by default, as p = 52 > 12.
Use print(x, correlation=TRUE)  or
         vcov(x)         if you need it
```

Maximised REML Log-Likelihood at Convergence, with Degrees of Freedom

```
> summary(lmer22_new7_dropcorr)$logLik
'log Lik.' -8926.643 (df=70)
# gives df, total number of parameters estimated, fixed + random
```

A.3 ANOVA table to test fixed effects (sequentially)

```
> anova(lmer22_new7_dropcorr)
Analysis of Variance Table
            Df  Sum Sq Mean Sq F value
CI           1 0.07117 0.07117  5.6718
BE           1 0.44803 0.44803 35.7036
yearF       12 0.62786 0.05232  4.1695
CI:BE        1 0.07471 0.07471  5.9536
CI:yearF    12 0.19044 0.01587  1.2647
BE:yearF    12 0.26215 0.02185  1.7409
CI:BE:yearF 12 0.48433 0.04036  3.2163
```

Note that the three-factor interaction (the key indicator from these analyses of a potential effect of industry on the rock art) – is highly significant, F = 3.2163, numerator df=12, denominator df taken as infinity in the absence of a better estimate, approximate p-value is 0.00012.

A.4 Interaction Plots

A statistically significant three-way interaction was found, which says the *contrast between Background and Engraving* (a key indicator of visibility of the engravings) changed differently over time for Industry sites than for Control sites. This warrants further investigation to ascertain whether these differential changes could be attributable to the effect of industry.

The three-factor interaction alone does not immediately imply an impact of industry, but it does raise the possibility. The shape and patterns of the changes over time need to be investigated and interpreted. Graphical displays are particularly useful here.

Figure 2, Figure 3 and Figure 4 step through graphical displays of the statistically significant three-way interaction found in the above model for the ASD L^* variable.

In Figure 2, the year 2004 looks the most different and raises several questions. If we exclude 2004 from the analysis, would the three-way interaction still be significant? What are the implications or explanations for this difference?

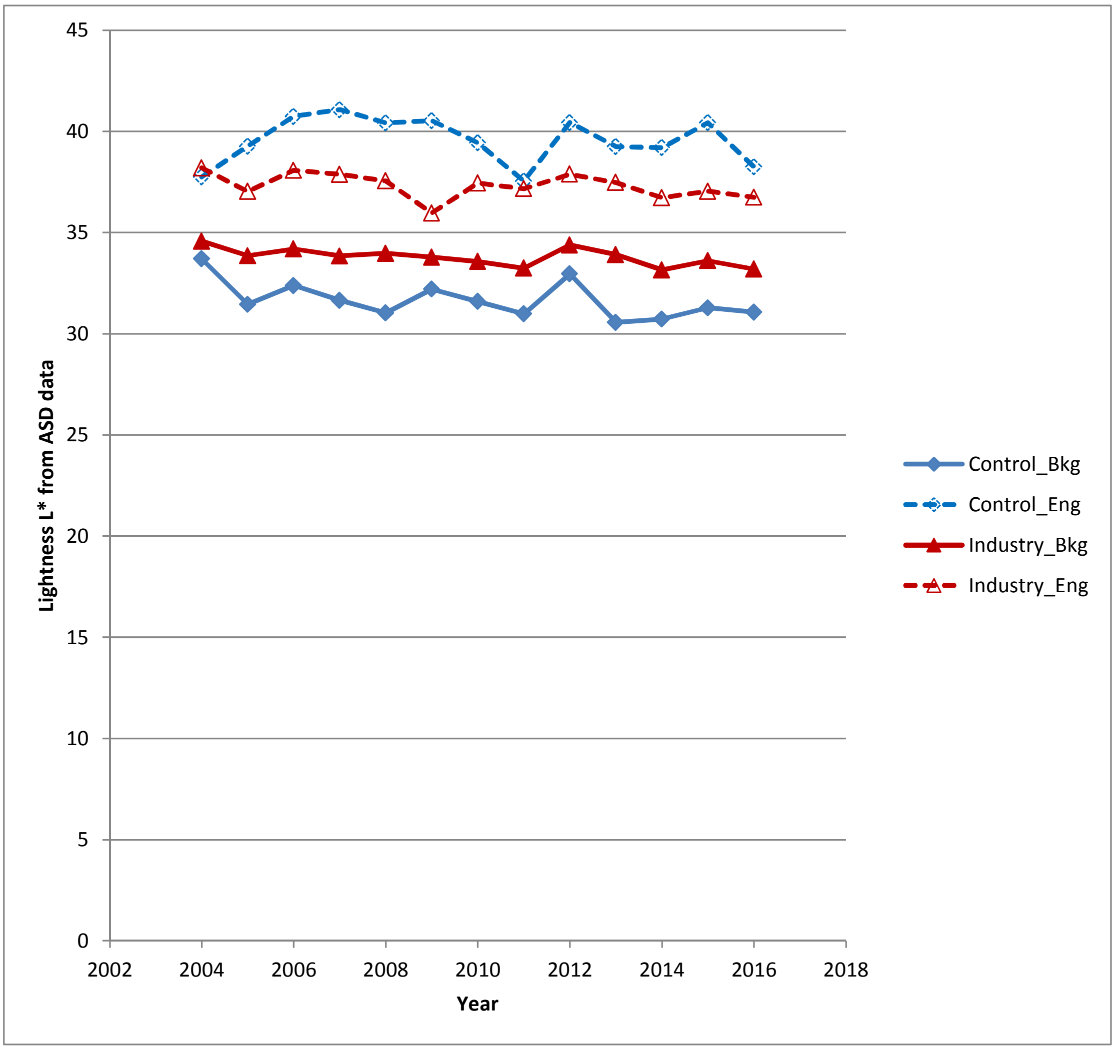

Figure 2. Three-way interaction plot.

Figure 3 reduces the display of Figure 2 to measures of the *contrast* (in terms of the "lightness" variable, *L**) between background and engraving in each year. On average, there is greater contrast at Control sites than at Industry sites, in all years except in 2004 where there is little difference. This suggests that, since 2004, the rock art has been more visible at the Control sites than at the Industry sites.

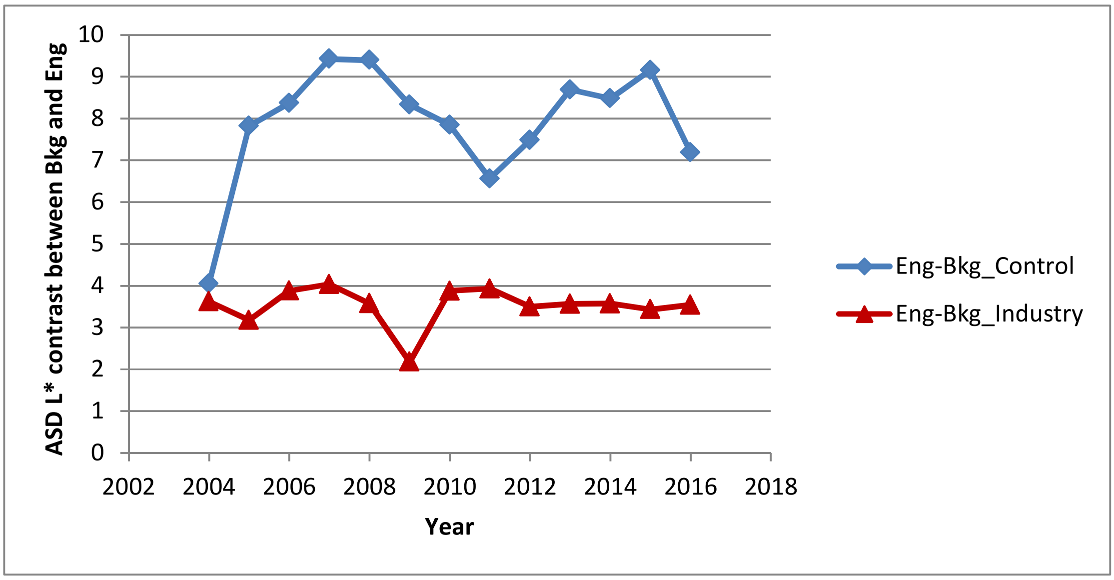

Figure 3. Contrasts in the L* variable: from Figure 2, dashed line (Engraving, lighter) minus solid line (Background, darker), at Control (blue) and Industry (red) sites.

Figure 4 directly examines the *difference* between the lightness *contrast* at Control sites and the lightness *contrast* at Industry sites. Control sites always had greater contrast on average, but there was no consistent trend over time: 2004 had very similar contrast between Background and Engraving at Control and Industry sites; other years have a larger difference, with a noticeable dip around 2011.

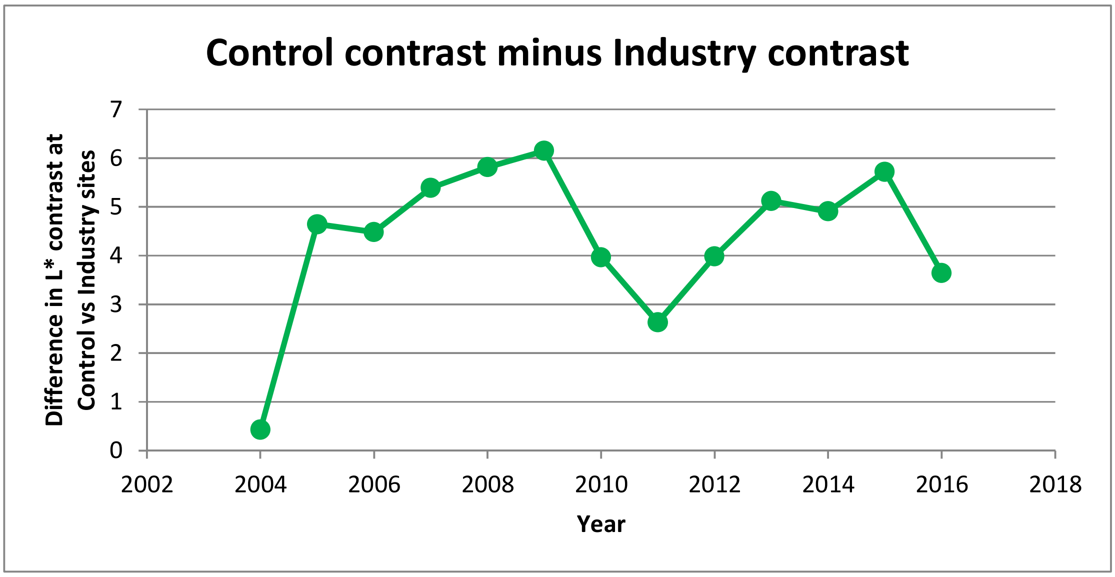

Figure 4. Three-way interaction effect: from Figure 3, blue line minus red line, plotted against Year.

If there was no three-factor interaction, the line in Figure 4 would be approximately horizontal, or at least not statistically significantly different from horizontal – the difference between the contrast at Control sites and the contrast at Industry sites would remain the same across all the years. Our analysis showed that the three-way interaction *was* statistically significant, strongly suggesting there are real differences.

If industry really was having an impact, cumulatively over the last 13 years, we might expect to see the difference between Control and Industry gradually increase over time.

These three Figures do not demonstrate such an obvious pattern – instead we see little difference (close to 0) in 2004 and a step up to larger differences (between about 3 and 6) from 2005, but no generally increasing trend – and the reasons for the statistically significant three-factor interaction should be considered further, taking advantage of logical reasoning, and any additional information, expertise or "inside knowledge" about the measurement processes.

In fact, curiously, the estimates displayed in Figure 3 suggest the contrast (Engraving minus Background) has *increased* at Control sites, but stayed roughly the same at Industry sites. This needs careful interpretation, including whether the measurements are in fact reliable for the required monitoring, i.e. able to detect/reveal the sorts of changes it is intended to monitor. If the pattern of the estimates cannot be explained, is it reasonable to think that they would reveal true impacts of the industry upon the rock art if they truly existed? And if not, then it is *not reasonable to conclude that there has been no impact*. Might there be a calibration issue between measurement at Industry sites and the more remote Control sites?

This illustrative analysis is not definitive and would benefit from further refinement, and, further, other variables should be examined by a similar analysis, actively targeting the three-factor interactions that can indicate whether the difference between Background and Engraving has changed over time in a different way for Industry sites than for Control (on average).

Other variables such as proximity to ocean, shipping and salt, proximity to industry, type of industry, rock type, the age of the exposed surface, or the age of the engravings, might also be relevant explanatory variables. The very small number of sites, especially Control sites, becomes a severe limitation when we have these extra differences between sites.

Appendix B. Glossary of Acronyms

Acronym	Definition
Measuring instruments and terms	
ASD	Analytical Spectral Devices spectroscope (ASD) – measures the full spectrum covering visible light and near infrared
BYK	BYK spectrophotometer – measures colour
KM	Konica Minolta spectrophotometer – measures colour
L*, a*, b*	Three variables that collectively describe colour mathematically in three dimensions: L* = lightness (0 = black to 100 = white), also called luminosity or luminance; a* = green (negative) to red (positive) scale; and b* = blue (negative) to yellow (positive) scale. Called the CIELAB colour space, defined by the International Commission on Illumination known as the Commission Internationale de l'Elcairage *(CIE)*
Organisations	
CSIRO	Commonwealth Scientific and Industrial Research Organisation (Australia)
DAA	Data Analysis Australia
DER	Department of Environment Regulation (Western Australian)
Statistical terms	
AIC	Akaike Information Criterion – compares adequacy of statistical models
ANOVA	Analysis of Variance
REML	Residual Maximum Likelihood – a method of fitting statistical models. Sometimes called Restricted Maximum Likelihood.