DATA
STATISTICS

ANALYSIS
MATHEMATICS

AUSTRALIA
INFORMATION

Review of Statistical Aspects of Burrup Peninsula Rock Art Monitoring

November 2016

Project: DER/1

STRATEGIC
INFORMATION
CONSULTANTS

Review of Statistical Aspects of Burrup Peninsula Rock Art Monitoring

November 2016

Client: Department of Environment Regulation

Project: DER/1

Consultants: Dr John Henstridge
Dr Kathy Haskard

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 was commissioned by the Department of Environment Regulation (DER) to review a draft paper by John Black and Simon Diffey *Reanalysis of the Colour and Mineralogy Changes From 2004 to 2014 on Burrup Peninsula Rock Art Sites*. In carrying out the review, Data Analysis Australia also reviewed the corresponding Commonwealth Scientific and Industrial Research Organisation (CSIRO) reports and the data itself.

Our findings are that the draft paper presents an analysis of a particularly poor data set. The statistical methods in the draft paper are highly appropriate (with some minor modifications) and they represent a substantial step forward in effective monitoring of the Burrup Peninsula rock art sites, for which Black and Diffey should receive appropriate recognition. However the analysis cannot overcome the lack of confidence in the current form of the data and it is not appropriate for the draft paper to be published in its current form.

We recognise that twelve years of data is a valuable resource and it should not be discarded as it is irreplaceable. However it is not appropriate for any decisions – including whether or not changes have taken place on the Burrup Peninsula – to be based on it in its current form. Hence our primary recommendation is that the problems in the data be repaired to the fullest extent possible and that its limitations then be clearly documented.

The work of Black and Diffey has highlighted the need for improved statistical methods in the monitoring of the rock art. Data Analysis Australia recognises that CSIRO has already made some steps in this direction but much more is needed.

Recommendations

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[1].
2. The CSIRO should be asked to revisit the cross calibration issues with the BYK-Gardner (BYK) portable spectrophotometer and the Konica Minolta (KM) spectrophotometer, 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.

[1] 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/.

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.

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.

In recent years questions were raised about the statistical analysis of this data, resulting in a draft paper[2] (henceforth termed the Draft Paper) suggesting that significant changes had taken place, in sharp contrast to the findings in the CSIRO reports. In this context the Department of Environment Regulation of Western Australia (DER or the Department) engaged Data Analysis Australia to review the statistical issues raised in the Draft Paper, making reference to the data itself and the CSIRO reports.[3]

1.1 Structure of This Report

The papers reviewed by Data Analysis Australia are listed in Appendix A. In addition Data Analysis Australia conducted its own analysis of various versions of the data and had discussions with Black and Diffey and members of the CSIRO team responsible for their recent reports.

Whilst the review was initially focused on the Draft Paper, many other statistical and measurement issues were highlighted. Hence this report takes a more global view starting from the design of the monitoring program, the nature of the data collected and measurement problems before it considers the respective statistical approaches of the CSIRO reports and the Draft Paper.

The review discovered a number of disturbing features in the historical data that unfortunately undermine the analyses presented in both the Draft Paper and the CSIRO Reports. Hence this review recommends further work to establish a consolidated high quality dataset and to define an appropriate analysis methodology. The review also suggests collecting additional 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.

[3] Each of the CSIRO reports covers the data over a period from 2004 and is generally published in the year following the latest data. Here we will refer to reports as "the CSIRO 2004-x Report", which is generally published in the year $x + 1$.

1.2 Acknowledgements

Data Analysis Australia was assisted in this review by the strong co-operation of the parties involves – the staff at the Department, the CSIRO and the authors of the Draft Paper – who all displayed an eagerness that a proper scientific monitoring and assessment of the rock art is achieved.

2. Design of the Monitoring Program

The monitoring program had its origins prior to 2004 and it is not clear to us just how the design was developed and approved. Hence our comments are restricted to the design as implemented in regard to the spectrophotometry.

Ideally a design would involve several elements that would enhance the ability to determine whether any change has taken place and to make inferences on the likely cause of such changes. To do this the following principles are usually applied:

- The design is based upon specified possible causes of change that may be of concern and possible nature of change. In this case the causes of concern relate to the industrial developments.
- The design aims to distinguish between such causes of concern and other causes not of direct relevance to this investigation. Causes not of direct relevance might include weathering effects that are not influenced by human activity, namely effects that would have happened regardless of the industrial developments. The common way of handling this is to include *controls* in the design, measurements that are not affected by the causes of concern, but expected to be affected (in the same way as other sites) by these other causes not of direct relevance.
    - This allows separation of the two types of causes, adjustment for the causes not of direct relevance, and estimation of the *net* effects due to the industrial developments.
- The measurement process should be able to detect changes that are of concern and should be reasonably reproducible.
- The design should have sufficient replication of measured items that it is possible to determine whether changes are “real” or just due to random fluctuations in measurement.
- Where there are multiple possible causes of change, the design should endeavour to allow these to be distinguished.

These principles typically lead to designs that include adequate controls, replication and systematic measurement methods.

2.1 Site Selection

Each “site” in the design is the location of an engraving.

The original design had seven sites of which two, Sites 1 and 2 in the North, are in some sense control sites. In 2013 two additional sites were added. We acknowledge that there were likely to have been significant constraints on the choice of sites due to cultural and physical reasons.

Our concerns relate primarily to the choice of control sites. The two sites consist of one site for each of the two dominant geological environments and hence if it is found that changes are related to the geology then the design may be reduced to only one control. This severely limits their value as controls.

A further concern is that both control sites are significantly less accessible and consequently may have had some limitations on what measurements are made at them. This was raised in discussion by the CSIRO, particularly in regard to taking bulky equipment in small boats. However it is not clear what the effect of this has been as these control sites appear to have had the same measurements each year as the other sites.

2.2 Spot Selection

At each site a number of “spots” were selected on the engraving and for each such spot a nearby position on the rock but not on the engraving was also chosen. Thus each spot is actually a pair of measurement points. The reports do not give a name to these so we will use the term “point”. Presumably the logic behind this was that the essence of an engraving is the way that it stands out on the rock face and hence the contrast between the engraving point and the background point at a spot is critical.

It is not clear how spots and their respective points were chosen or how they were physically located each year. Initially there were 3 spots per site but in 2013 this was increased to four spots per site.

2.3 Replication

Each point was measured on one occasion each year, with several measurements made on each occasion, ranging between 1 and 70 measurements, but most commonly around 22 for the BYK-Gardner (BYK) portable spectrophotometer (but very variable across years, 10 or fewer in 2004, 2006, 2008 and 2009) and predominantly 44 for Konica Minolta (KM) data. This replication, we understand, involved independently placing the instrument and taking a reading so that differences between readings would represent the measurement error itself and the placement error. Histograms of the number of observations at each point are shown in Figure 1 and Figure 2.

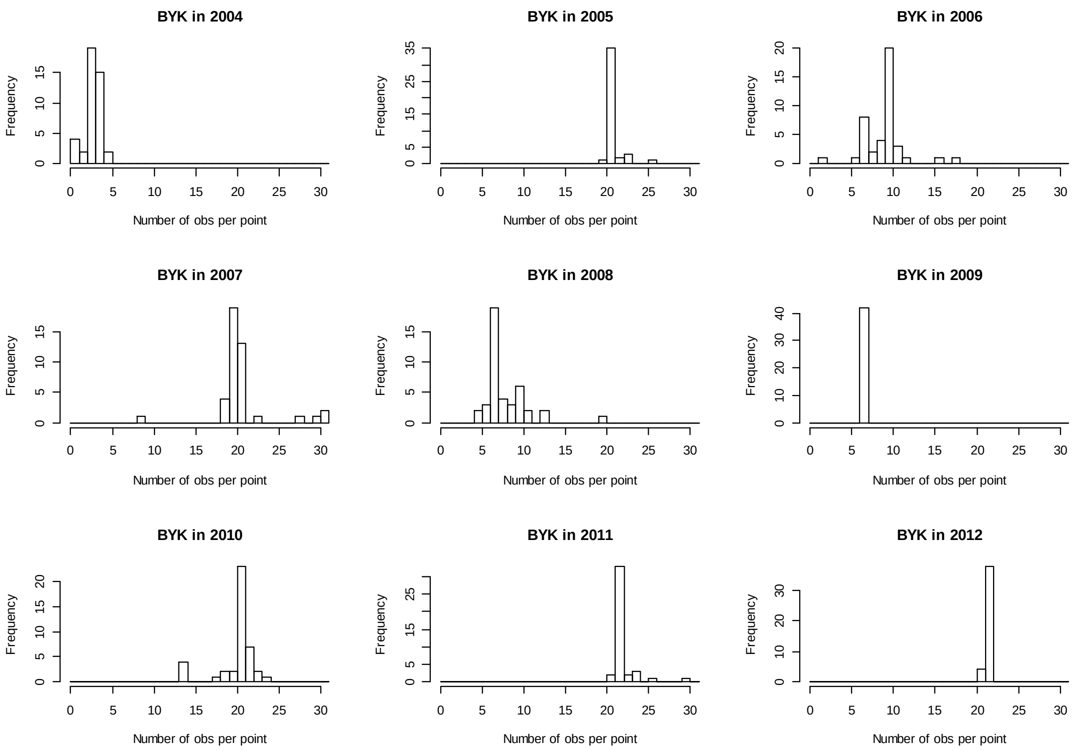


Figure 1. Numbers of observations per point in each year for BYK machine – very variable across years, ranging from 5 or fewer in 2004, to approximately 22 in some years.

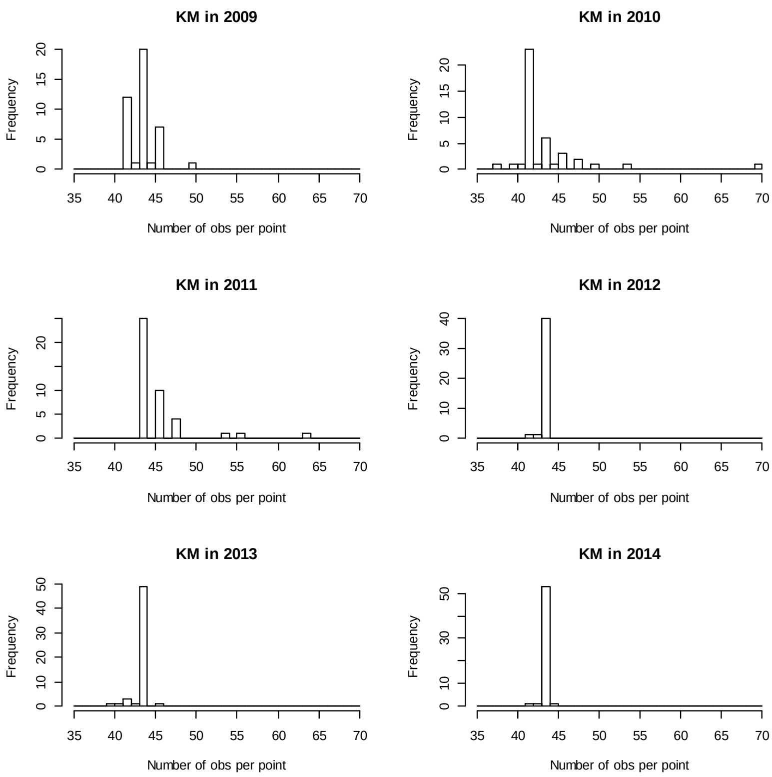


Figure 2. Numbers of observations per point in each year for KM machine – consistently around 44.

In this context it is important to note that while the KM spectrophotometer measurements had roughly twice the number of replicates, the BYK spectrophotometer was in most need of replication since it was much more variable in its measurements – in one sense, one KM measurement was equivalent in precision to the average of ten BYK measurements. We suspect that the reason for the greater replication with the KM spectrophotometer was that it had functionality that made it easier to carry out repeated measurements.

3. Measurement Processes

It appears central to the measurement process that it aimed to objectively measure colour and lightness as perceived by humans. To that end the focus in the CSIRO Reports was on the International Commission on Illumination (CIE) L^*, a^*, b^* colour measurement system that seeks to measure colour as it is perceived by the human eye. The three dimensions of this are:

- L^* – a measure of overall luminosity or lightness as perceived by the eye;
- a^* – a measure on the green/red colour axis; and
- b^* – a measure on the blue/yellow colour axis.

The scale of differences measured in this three dimensional space correspond closely to how the human eye perceives differences. The standard measure of difference is $\Delta E = \sqrt{(L_1^* - L_s^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2}$, a Euclidian measure in this space.

Colour measurement is a complex area as perceived colour of reflected light depends upon the colour of the illuminating light source, the reflecting surface and the geometry of the reflection. Laboratory measurement systems try to control these factors as far as is possible, but it is not so easy in the field. Portable spectrophotometers have their own light source but generally still require regular calibration to ensure consistency of results.

It is evident that the data collection process found it a challenge to achieve consistency in measurements over the years of monitoring. In particular, a change in the spectrophotometer was deemed necessary.

3.1 BYK and KM Spectrophotometers

The two spectrophotometers used were:

- The BYK spectrophotometer, a relatively simple instrument very much aimed at the manufacturing industry. The manufacturer's main business is in chemicals and colour additives. This instrument was used between 2004 and 2012; and
- The Konica Minolta or KM instrument is a more modern instrument. As well as giving L^*, a^*, b^* measures it is able to measure reflectance in 30 wavelength bands across the visible spectrum. It is likely but cannot be confirmed that the instrument calculates the L^*, a^*, b^* values from these spectral band measurements. This instrument was used from 2009.

Both of these instruments appear to be designed for quality management in industry where there is an ongoing need to verify consistency in colours or products. To this end, they can be used to give their results in the L^*, a^*, b^* space and to calculate differences as ΔE. Our understanding is that the more detailed spectral band information from the KM spectrophotometer was not retained.

The two instruments are most likely to differ in their illumination geometry, although the non-statistical details of this have not been investigated in this review. We understand that the compared with the KM instrument the BYK instrument has a smaller measurement area and a base less suited to use on uneven surfaces.[4]

3.2 ASD Spectrometer

The Analytical Spectral Devices, Inc. (ASD) spectrometer is conceptually a different instrument in that it outputs a reflectance spectrum from the near ultraviolet (350nm) through to the infrared (2500nm). While its standard output is given at each nanometre, its resolution varies between 3nm in the visible through to 10nm in the infrared.

We understand (from the CSIRO 2004–2013 report) that there was a change in the ASD instrument in 2012 although the effect of this change was quite small.

The ASD spectrometer gives considerably more information than the spectrophotometers:

- It covers spectral bands beyond the visual spectrum and in some of these bands it is possible to make significant inferences on mineral composition.
- It is higher resolution; and
- It is possible to calculate L^*, a^*, b^*values post hoc from the spectra.

Obviously the ASD spectrometer creates unique challenges with the amount of data collected. These may have been considered substantial in the field in 2004 but today these would be regarded as insignificant.

3.3 Cross Calibration

When trying to measure change it is important to keep the measurement process consistent so that observed changes can be attributed solely to what is being measured. Being forced to change an instrument, replacing the BYK spectrophotometer with the KM spectrophotometer, provides a significant challenge.

[4] While this was the reason given for the change in the CSIRO 2004-2015 Report, the CSIRO 2004-2014 Report indicated that "some of the automated memory retention functions of the BYK spectrophotometer started to become less reliable, requiring laborious manual data saving".

We have reviewed a CSIRO paper[5] on this, which has findings we would summarise as follows:

- It considered the years 2009 and 2010 for which there were both BYK and KM data, and investigated how the historical BYK data might be transformed to be consistent with the KM data.
- Standard linear regression methods were used to separately transform the L^*, a^*, b^* components. The regression models included allowances for the specific differences between the BYK and KM measurements at each point. The regressions were considered to be highly accurate because they had R^2 values of 0.96 or 0.97.
- Equations were given for transforming from BYK values to KM values.

We have serious concerns about this calibration process:

- Regressing the KM values upon the BYK values ignores the fact that both the BYK and KM values are *estimates* of the true reflectance; that is they are both measurements with error. Our review of the data indicates that the BYK data is substantially more variable than the KM data but the regression method used essentially assumed that the error is purely in the KM data. This is a classic "errors in variables" problem that can be expected to lead to biased estimates of the regression coefficients.
- The equations assumed that each component of the KM measure depended only upon the corresponding BYK component. That is, the L^* as measured by the KM instrument was assumed to relate to only the L^* measure from the BYK instrument and so on. Given the complexity of the colour measurement process, this does not appear to be an obviously correct assumption.
- In contrast, the equations included numerous coefficients to allow for site, spot and point specific differences. Indeed these terms explain a large proportion of the R^2 values, to the extent that the values given by the equations are affected at least as much by where the measurement is being taken as by what the BYK measure actually is.
- Whilst the inclusion of site and spot terms in the cross calibration results in a good fit, there is no pattern to these terms. The importance of these terms in the calibration regression equations effectively implies a different calibration is required at every location (spot within site).
- In the case of L^* the fitted relationship between the BYK and KM measures is very weak and negative. That is, the coefficient of L^* is -0.0088, when the *ideal* value would be 1.0. The negative coefficient implies that if the BYK and KM

[5] Alexander, D., Lau, D., Markley, T. and Ramanaidou, E., *Regression for Calibration of Burrup Peninsula Rock Art Colour Measurements*, 18 June 2013, unpublished, although some of the content appears in the CSIRO 2004-2013 Report.

spectrophotometers were used on two points at the same location they would disagree on which one was the more luminous.

- The corresponding coefficients for a^* and b^* are 0.8288 and 1.4137 respectively, which are more acceptable though it is difficult to consider them good.[6]

Collectively these raise significant concerns about the consistency and even compatibility of the BYK and KM data. The cross-calibration models used do not provide any understanding of why the two instruments give very different results.

We have not had the opportunity to consider the cross calibration of the ASD data with the KM or BYK data.

3.4 Comparisons between Instruments

A graphical examination of the data displays some of these issues. In Figure 3 the data for Site 1 is given as an example. Similar displays are given for all sites in Appendix B of this report.

It is evident that whilst the KM and ASD instruments gave similar outputs, the BYK instrument was quite different, particularly over the period 2004-2010. The 2011 and 2012 BYK data is so different that either it can barely be considered the same instrument or a dramatic change occurred in the measurements at all sites between 2010 and 2011. Since such a change was not remarked upon in the relevant CSIRO report, we are forced to consider that this is a data management problem. We suspect that the 2011 and 2012 BYK data has been transformed to match the KM data using the cross-calibration referred to above.[7]

[6] These values are taken from Appendix A of the CSIRO 2004-1013 Report.

[7] The CSIRO has been asked to comment upon this issue but at the time of writing no response has been received.

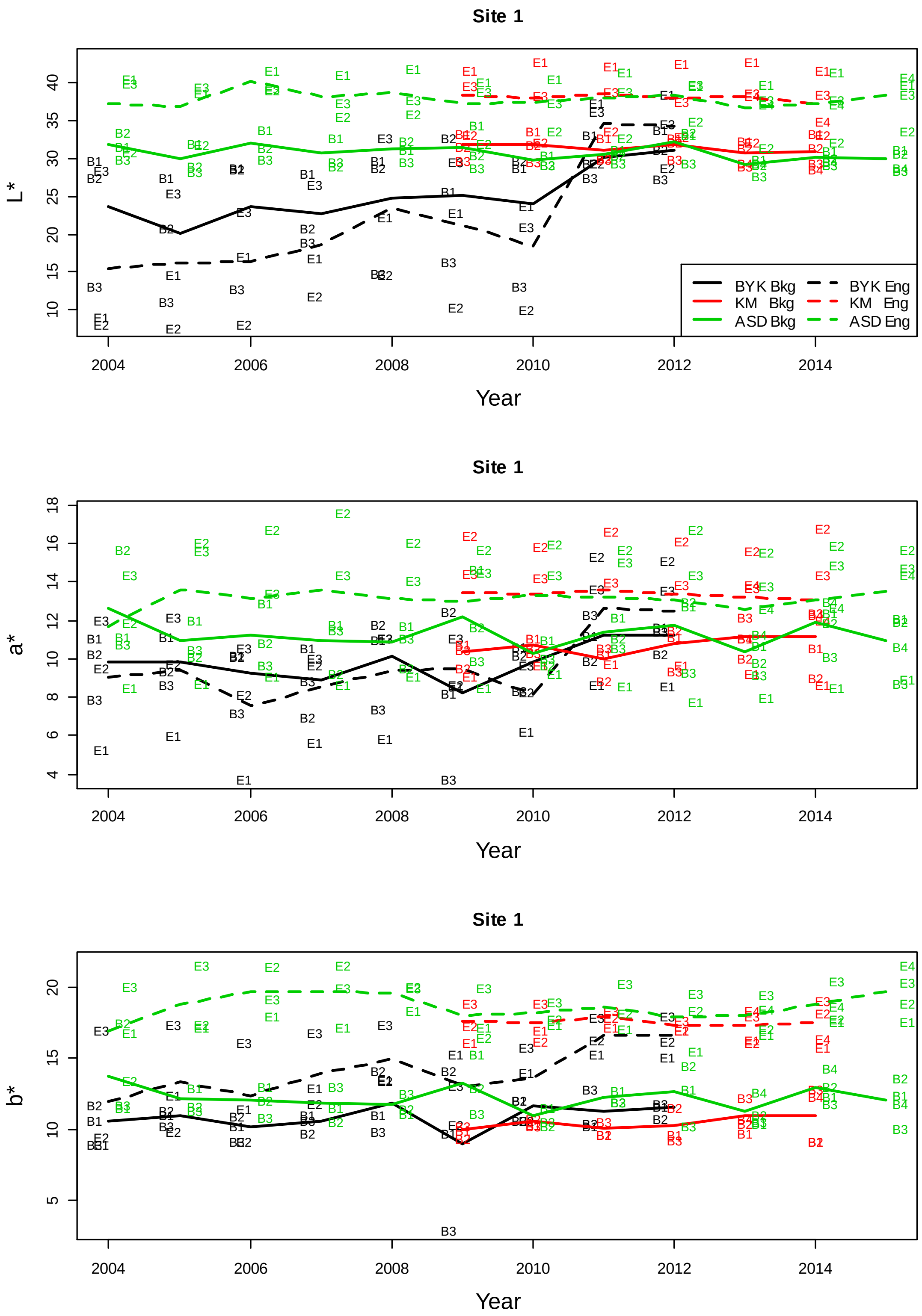


Figure 3. Empirical means of L^*, a^* and b^* values respectively for each point and means of Background and Engraving points at each time for Site 1, showing readings from all three machines, BYK, KM and ASD.

3.5 Assessment

The effective use of the spectrophotometer data back to 2004 requires a proper understanding of the cross calibration issues, problems that we do not believe to have been fully resolved by the CSIRO. We strongly recommend that the cross calibration issues be further investigated, ideally to understand what causes the differences.

We further recommend that consideration be given to considering the ASD spectrometer as the primary measurement instrument.

4. Data Management

The monitoring program has collected a considerable quantity of data, particularly in the form of the ASD spectra. Although we have not been able to explore this in detail, it is of concern to us that much of this data may be poorly managed. The reasons for our concern are:

- The ASD spectral data used a different naming convention for recording the site, the spot, and the replicate measurement each year. This suggests that the data has never been consolidated to enable proper analysis of trends.
- The spectrophotometer data that was provided to us initially (and we understand to Black and Diffey) left out some data and had some data duplicated.
- A distinct and all-pervasive change in pattern of the BYK data between 2010 and 2011 strongly suggests the dataset may have confused corrected and uncorrected (raw) BYK data.

In a project spanning many years there is a real danger that vital data is lost unless it is properly curated. Scientific enquiry is also aided by having the data readily assessable in a form that is unlikely to be misunderstood.

While we do not have sufficient information to make specific recommendations, we suggest that this should be further investigated.

5. Statistical Analysis

The data shows considerable variation, between sites, between spots and over the years. This variation is accentuated by the differences between measuring instruments. In such a context the statistical analysis becomes critical so that actual changes in the rock art can be distinguished from the "noise".

The standard statistical approach would be to set up a predictive model, whereby the observed measurements are expressed as a function of explanatory variables and a random or error term to represent the impossibility of absolute precision. That is

$$y = f(x_1, x_2, \dots x_k) + e$$

where y represents the observation, $x_1, x_2, \dots x_k$ are the explanatory variables, f the function and e is the error. The explanatory variables might include the location, spot and whether the measurement is of the engraving or the background, as well as time so that trends can be considered. The instrument used can be another variable.

Any function can be used but it is usually chosen to be linear in the variables and their combinations.

A key aspect of this modelling approach is that it is possible to incorporate all the data – all years and all sites – into a single comprehensive analysis and to determine whether the relevant terms in model are statistically significant. This is particularly important when the concern is over long term changes (and hence the need to consider all years of data) that are affecting a number of sites, some more than others. For example, a term that might be tested would be a linear trend over the years interacting with the difference between the northern and southern sites.

Whilst we do not close our minds to other statistical approaches, to experienced statisticians the modelling approach thus described is the obvious one and there would need to be good reasons for not considering it.

5.1 CSIRO Analysis

The CSIRO reports emphasised the use of the ΔE measure that combined measures of difference in the three components of the L^*, a^*, b^* system. This led to the reports focusing almost exclusively on pairwise differences – year to year changes and comparisons of engravings with backgrounds. In addition, the reports almost exclusively use average measures at each point of each spot, without considering how the repeated measurements varied. This approach has severe implications:

- The use of ΔE means that all information about the direction of change is lost. It is not even clear whether changes over time accumulate or cancel out.
- The ΔE approach does not lend itself to an analysis of the dataset as a whole – it considers many pairwise differences but does not have a framework to assemble them into a whole.
- The absence of including the replication in the analysis means that no attempt is made to distinguish between random measurement error and actual changes.

In the case of the ASD spectra, again the replication at an individual point is ignored and, until the 2015 data, formal statistical analysis is not conducted. The graphical presentation of the spectra is useful, particularly since it permits qualitative evaluation of spectral bands known to relate to certain minerals.

5.2 Draft Paper Analysis

The analysis in the Draft Paper closely follows the standard statistical approach:

- It is based upon a linear mixed model fitted with the software ASReml[8], commonly regarded as one of the most powerful statistical programs for the analysis of data of this type;
- The model takes into account a number of variables – particularly the site, the spot within the site, whether the point being measured is on the engraving or the

[8] ASReml: VSN International Ltd, Hemel Hempstead, UK. URL https://www.vsni.co.uk/software/

background and which spectrophotometer ("machine") is being used – together with interactions;

- The model is mixed (contains both fixed effects and random effects) in that some of the terms relating to replication in the model are treated as random This ensures that some variables of interest are properly considered against the background of some random effects relating to more than one observation;
- The time or year variable is treated as having three components – a linear trend over time, a smooth non-linear component, and irregular deviations. Since long term changes to the rock art are of most concern, the linear component is of most interest and focusing on this gives the greatest statistical power in detecting change.
- The principal output of the analysis is an analysis of variance that provides objective quantification of the statistical significance of various terms. The unbalanced nature of the design means that indicative upper and lower bounds rather than precise values are provided for the F ratios in this analysis.

In our opinion the general approach of Black and Diffey is one that we would strongly support. It might be described as being good quality conventional modern statistics.

Unfortunately, their analysis, as presented in the Draft Paper, has several significant drawbacks, primarily related to the data used, listed here in decreasing order of importance:

- We have serious reservations about the 2011 and 2012 BYK data – in our opinion this is so clearly different from the earlier years of BYK data that it cannot be considered the same instrument. Until the provenance of this part of the data is clarified it should, in our opinion, be removed from the analysis.
- The analysis did not take into account the substantial difference in the repeatability of the two instruments – the individual BYK spectrophotometer measurements made on a single point have about ten times the variance of the KM measurements. This is exacerbated by the fact that *means* for each point were analysed, and the numbers of observations contributing to each mean varied greatly (see Figure 1 and Figure 2), with many more observations for KM than for BYK and large variation across years for BYK.[9] This would lead the model to be fitted on the assumption that the BYK measurements were more precise than they actually were and that the KM measurements were less precise than they actually were.
- The data for the KM instrument for 2012 was actually a duplicate of the data for 2011 and the 2010 KM data was missing. We followed this up with the CSIRO and received a corrected data set, with new data for KM 2010 and 2012.

[9] Black and Diffey indicated to us that their first analysis *was* based upon the replicate measures but they them adopted the use of averages at the urging of the CSIRO.

- The model used included terms for the machine effect (the difference between the spectrophotometers) and the point that was being measured, an approach that explicitly follows the CSIRO cross-calibration method. Given the substantial reservations about CSIRO cross-calibration we are concerned that this allows an inappropriate overfitting of the data.
- The model has not explicitly used the fact that there are Control and Impact sites (although we acknowledge that the Control sites are neither ideal nor sufficient in number) and it has not explicitly considered the two rock types.
- Because the analysis used average measures at each point in each year, rather than the individual replicate measures, not only the different instrument precisions but also the different numbers of replicates were not taken into account in the analysis.

In response to our request Black and Diffey repeated their analysis allowing for different residual variances for the two spectrophotometers, using the revised dataset. The Analysis of Variance (ANOVA) output of this is displayed as Figure 4:

	Df	denDF	F.inc	F.con	Margin	Pr
(Intercept)	1	1.0	43.3100	40.8700		0.0968294375
Site	6	5.5	5.9070	5.9110	A	0.0292961223
lin(jday)	1	33.3	4.7310	4.7480	A	0.0365057560
Type	1	1.0	0.4136	0.4121	A	0.6372706563
Site:Spot	14	20.4	2.8380	2.8440	B	0.0157468931
lin(jday):Site	6	48.8	1.8130	1.8040	B	0.1179700084
Site:Type	6	5.3	6.7060	6.6990	B	0.0243856160
lin(jday):Type	1	148.5	0.2307	0.3805	B	0.5382940047
lin(jday):Site:Type	6	118.0	1.1800	1.1400	C	0.3435915281
lin(jday):Site:Spot	14	181.1	1.7720	1.7750	C	0.0453392127
Site:Spot:Type	14	20.4	3.6470	3.6470	C	0.0040824528
lin(jday):Site:Spot:Type	14	181.3	3.5780	3.5780	D	0.0000318941

Figure 4. Black and Diffey's reanalysis of the revised BYK and KM L^* data allowing for the different residual variances. Highlighted rows meet the formal 5% significance level.

Black and Diffey also produced an analysis for Luminosity using just the BYK data, shown in Figure 5.

	Df	denDF	F.inc	F.con	Margin	Pr
(Intercept)	1	11.9	20300.0000	3703.0000		0.0000000000
Site	6	27.2	172.8000	173.1000	A	0.0000000000
lin(jday)	1	10.5	10.8500	10.3400	A	0.0087435360
Type	1	85.6	30.8000	31.6400	A	0.0000002291
Site:Spot	14	90.4	12.8200	12.9600	B	0.0000000000
lin(jday):Site	6	28.8	1.8080	1.7280	B	0.1503006000
Site:Type	6	73.4	34.8300	34.7200	B	0.0000000000
lin(jday):Type	1	79.2	0.2195	0.3881	B	0.5350833000
lin(jday):Site:Type	6	73.1	1.0070	0.8779	C	0.5155299000
lin(jday):Site:Spot	14	90.5	1.3490	1.3760	C	0.1813791000
Site:Spot:Type	14	90.5	21.8300	21.8300	C	0.0000000000
lin(jday):Site:Spot:Type	14	90.7	2.0880	2.0880	D	0.0195168200

Figure 5. Black and Diffey's reanalysis of the revised BYK L^* data alone. Highlighted rows meet the formal 5% significance level.

These analyses suggest a mildly significant Time by Site by Spot by Background/Engraving interaction, although these effects were dwarfed by very strongly significant effects involving Site and/or Background/Engraving but not Time. Such situations can make interpretation difficult since any inferences become heavily sensitive to the assumption of linearity in the model. It is likely that interactions may increase or decrease (or even disappear) if a modest non-linear transformation is made.

This sensitivity means that the quality of the data is even more important than might otherwise be the case.

5.3 Alexander Analysis

The brief report of David Alexander of CSIRO's Data 61 entitled *Analysis of ASD spectra change over time in Burrup Rock Art* presents several analyses based on the ASD data between 2004 and 2015. The statistical approach is distinctly different from that in the main CSIRO Reports and is closer to that of the Draft Paper.

The first section of Alexander's report focuses on the analysis of troughs in the ASD spectra that may have a geological interpretation. While the details are not clear he apparently works with average spectra for each point each year and "standardises" them by removing the large scale features identified by fitting some form of smooth curve. This enabled the extraction of the depths of the troughs in a systematic manner and the subsequent analysis examined the trough depths. Since it is sensible to analyse each trough separately, the analyses could be reasonably simple and are presented as ANOVA tables for the troughs at 923 nm and 2255 nm. Both of these show linear trends over time while the results for the other troughs apparently did not show significant trends over time.

While this analysis suggests some trends or changes over time, little detail is provided. Furthermore, it cannot be compared with results from the BYK or KM spectrophotometers since the troughs where changes appeared to have occurred over time were outside the visible spectrum covered by those instruments. Hence this is suggestive of concern, but more detail is required and it not clear whether the changes, assuming they are real, are affecting the visible appearance of the rock art. Alexander did further analysis to investigate whether the changes related to difference between northern and southern sites, but this was not statistically significant.

The second section of Alexander's report analyses L^*, a^*, b^*values derived from the ASD spectra. We were provided with these values calculated from average spectra. We requested from the CSIRO the algorithm for the calculation but this was not provided. Hence we cannot comment upon the reliability of this calculation process.

As a first step Alexander compares the L^*, a^*, b^* values derived from the ASD instrument with the values from the BYK and KM instruments. This clearly indicates that the BYK correlates very poorly with the ASD instrument, particularly for the L^* dimension where the correlation was only 0.42. In general the correlation with the KM instrument was high. Alexander comments on the apparent unreliability of the BYK measurements.

Alexander presented the ANOVA tables for the ASD L^*, a^*, b^* data. These found no clear evidence for change. Again, little detail is provided.

Overall the Alexander report presents statistically reasonable first steps in the analysis of the ASD data. However more detailed analysis is required to understand the implications of the first findings. It does however present strong evidence that the BYK data is of doubtful utility.

5.4 Data Analysis Australia's Analyses

Data Analysis Australia's analytical efforts were initially focused primarily on the spectrophotometer CIE L^*, a^*, b^* data with the aim of understanding the data. In order to inform our understanding of the differences between machines, some analysis was later done on the L^*, a^*, b^* values derived from the spectrometer data and kindly provided to us by CSIRO. This was regarded as a third "machine", ASD.

We are unable to present definitive analyses within the scope of this review, however we present some general findings, supported by a range of preliminary analyses aimed to best elicit what is really happening in the presence of so many difficult aspects of the data, and to reveal the consequences of some of those aspects of the data.

For the spectrophotometric L^*, a^*, b^* data, we carried out a number of analyses for variables L^*, a^* and b^*, but with greater attention to L^*, many with the Year effect partitioned into linear to cubic components plus deviations (similar to Black and Diffey) and with Site partitioned into Northern Sites (1 and 2) and Southern Sites (4, 5, 6, 7, 8), plus deviations, and:

- *Both* on average data for all the observations at a point on a single occasion with a single machine, *and* on the individual observations;
- Deviations as fixed effects *or* as random effects;
- Ignoring the larger variance in the BYK data; *or* allowing for larger variance in BYK data;
- For both machines (BYK and KM) together, *and* for each machine separately.

A number of observations could be made from these analyses and the data set itself. Many of these are preliminary and highlight aspects where further work is required:

- The effect of the larger variance of the BYK data is to weaken the statistical significance for statistical tests (larger p-values), often changing significant results to non-significant or marginally significant.
- There is a complex set of interactions involving Machine, suggesting any trends or patterns are different for different machines, as well as for different sites and spots, types of points (Background or Engraving), and Northern vs Southern sites.
- Figure 6 shows fitted linear trends from a model analysing both BYK and KM data together, but with statistically significant interactions involving Machine. Separate analysis of the two machines yields a very similar looking plot. This

model fitted year effects purely as linear, and site effects purely as North (2 sites) vs South (7 sites), with deviations included in the random effects, and is therefore an overly simplistic model. However it highlights basic differences between data from the different machines.

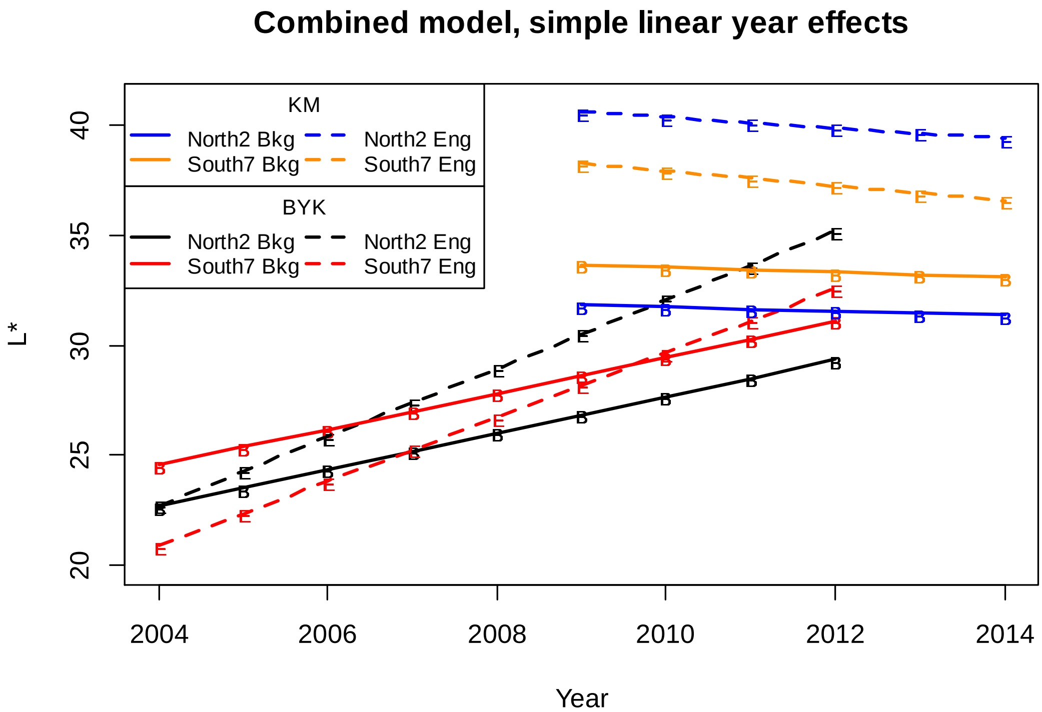


Figure 6. Combined model, including interactions with Machine – very similar outcome to fitting separate models for each Machine.

- Analysing BYK alone suggested strong increases in L^* over time, with slightly different slopes for the four types of points: Northern Sites Background, Northern Sites Engraving, Southern Sites Background and Southern Sites Engraving.
  - In fact the direction of differences reverses from 2004 to 2012 is largely due to the influence of the data for 2011 and 2012, which as we have indicated above, is possibly in error.
  - In an analysis that enables this comparison, there was no statistically significant three-factor interaction between Time, Background/Engraving and Northern/Southern sites (for the BYK data). This means the *difference* between Background and Engraving points in the North minus the *difference* between Background and Engraving points in the South did not change statistically significantly over time, even though there were clear changes in L^* for Background relative to Engraving: Background points had higher L^* than the Engraving points in the early years, and this reversed in the last two years.
- Analysing KM data alone had greater power because of smaller variance, further enhanced by the larger numbers of observations per point, although these effects are partially countered by the reduced time coverage. An analysis of this fitting

linear trends over time suggested small *negative* slopes for all four groups, beginning (in 2009) at larger L^* values than from the BYK data, and with larger differences between the four groups, Northern Sites Background, Northern Sites Engraving, Southern Sites Background and Southern Sites Engraving.

Excluding the last two years of BYK data from the analysis makes a substantial difference to the result, reducing the positive slope of the trend for the BYK data. For example a plot corresponding to Figure 6 becomes that shown in Figure 7.

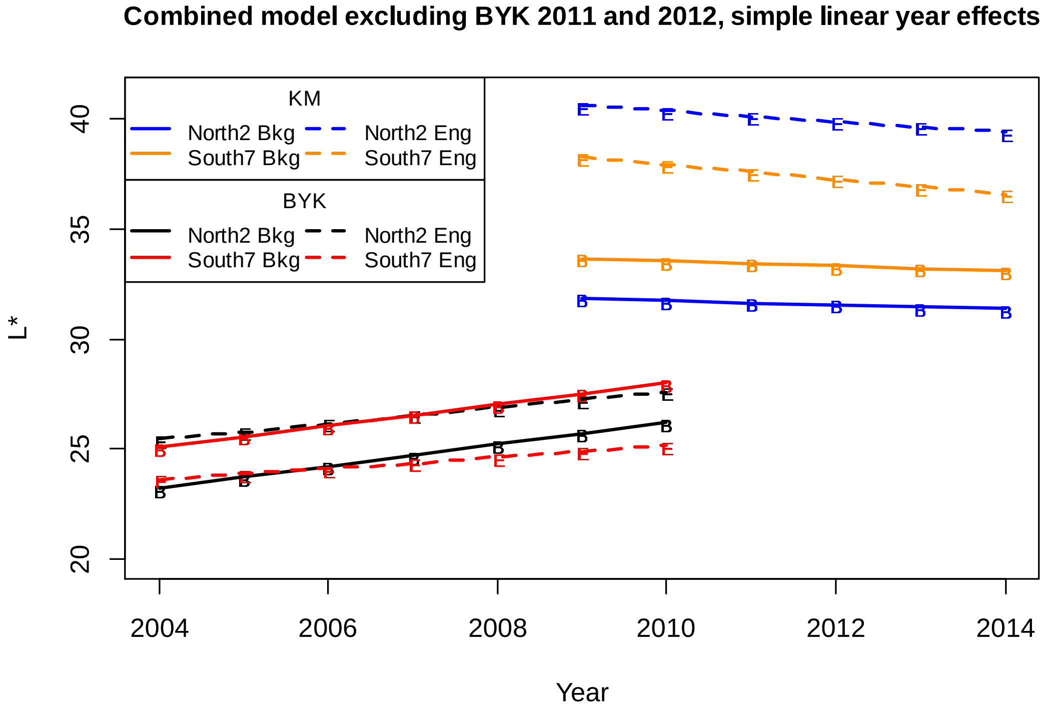


Figure 7. Analogous to Figure 6 but excluding the last two years of BYK data from the model. A non-significant Linear Year by Background/Engraving by Machine interaction has also been dropped, which made little difference.

The main aim of this monitoring is to determine whether industry on Burrup Peninsula has affected the rock art. Superficially our analyses and those of Black and Diffey suggest that some changes may have taken place, but as is explained below we have substantial doubts about the reliability of the data and hence any conclusions drawn. Figure 7 gives a hint of a difference that may be meaningful in this context. Based solely on the six years of KM data, the difference between Engraving and Background:

- Is smaller at the Southern sites, across all six years (wider gap between the blue lines than between the orange lines); and
- Appears to have reduced over time at the Southern Sites more so than at the Northern Sites – the gap between the oranges lines decreases over time, while the gap between the blue lines stays roughly the same. This is exactly the sort of effect that could suggest there has been an impact on the contrast between Background and Engraving in the sites nearer the industry.

- This is an interaction of Time by Background/Engraving by Northern/Southern Sites, and was statistically significant in the simplistic model illustrated, as well as in more sophisticated models.

Such a finding is however highly conditional:

- It assumes the Northern sites are appropriate control sites;
- It assumes that there are not further major problems with the data;
- It assumes that the linear model structure is the most appropriate;
- The discrepancies between BYK and KM data as provided cast serious doubt on the reliability of these data sets, and hence any conclusions drawn; and
- Whether the ASD data supports or conflicts with the BYK and KM data is not known.

The data issues need to be resolved, a decision made about whether the BYK data has sufficient merit to be used, and how, and an analysis such as the mixed models approach taken by Black and Diffey needs to be carried out, based on the individual observations (rather than means), and allowing for the different sized variances for data from the BYK and KM machines.

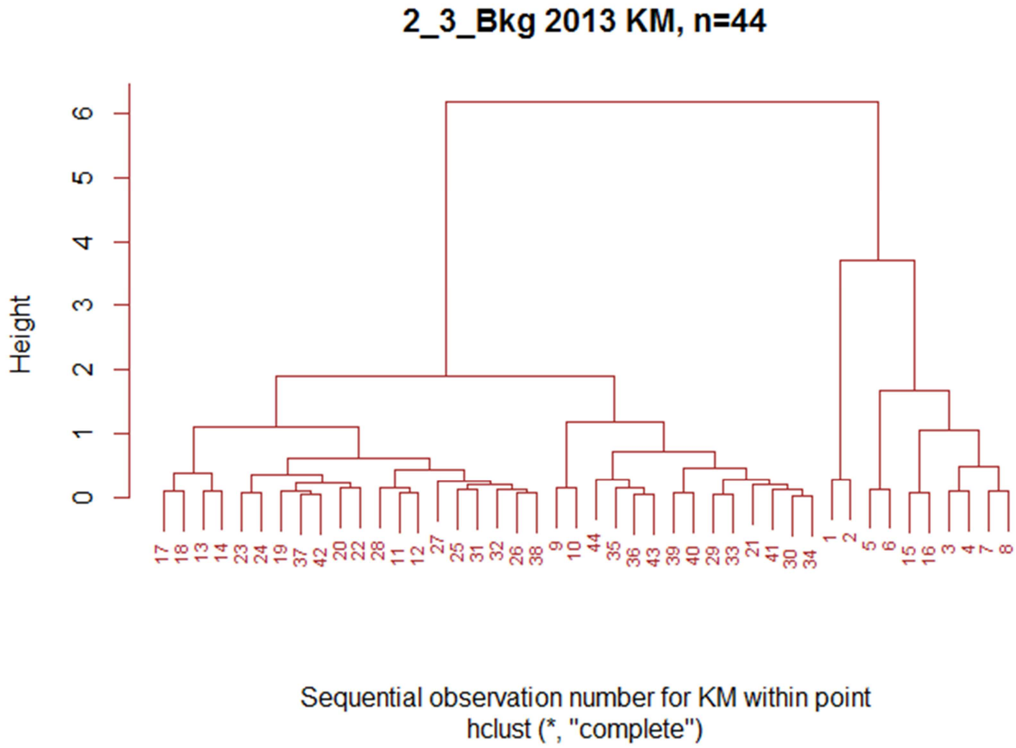
While there is evidence that some changes have taken place, it is difficult to reconcile the results from the two machines.

5.5 Finer Structure in the Data

There appears to be considerable structure in the data that has not been made explicit, nor has it been taken into account in the analyses by CSIRO and by Black and Diffey. For example, where there are multiple (e.g. 20 or 40) replicate measurements at a single point with a single machine, it is not clear whether this involved removing and replacing the machine for each individual observation.

Cluster analyses of the triple L^*, a^*, b^* (using the ΔE measure of difference between observations) from multiple readings at a single point at one time using one machine revealed that frequently pairs of adjacent readings were more similar. See for example Figure 8, with the story corroborated by the pairwise scatter plots in Figure 9 and Figure 10, in which points appear in pairs of very similar colour, meaning they were recorded close together in time.

Figure 9 also shows an apparent drift in all three colour variables across the sequence of observations, which is somewhat worrying in terms of the objectivity of the instrument. Such a drift can be seen elsewhere, though not always as clearly. One can only speculate as to the possible causes – a drift in the location being measured or a drift in the calibration of the instrument.



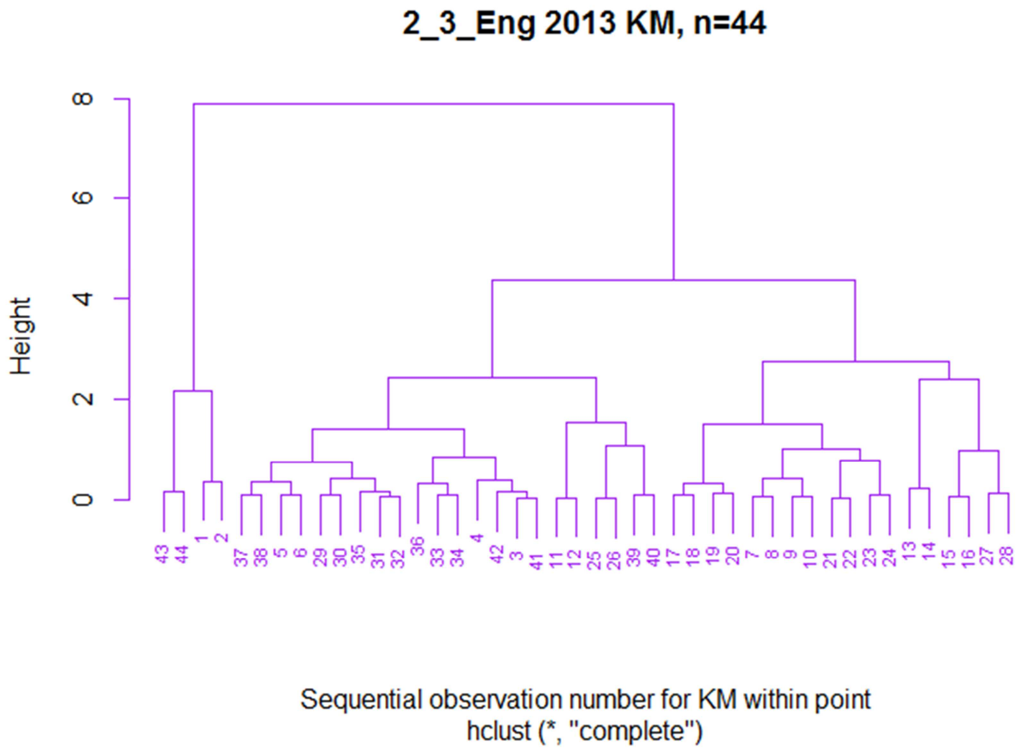


Figure 8. Clusters of observations from Site 2 Spot 3 Background and Engraving points, in 2013 with KM machine. Note how the clusters are dominated by pairs of adjacent observations.

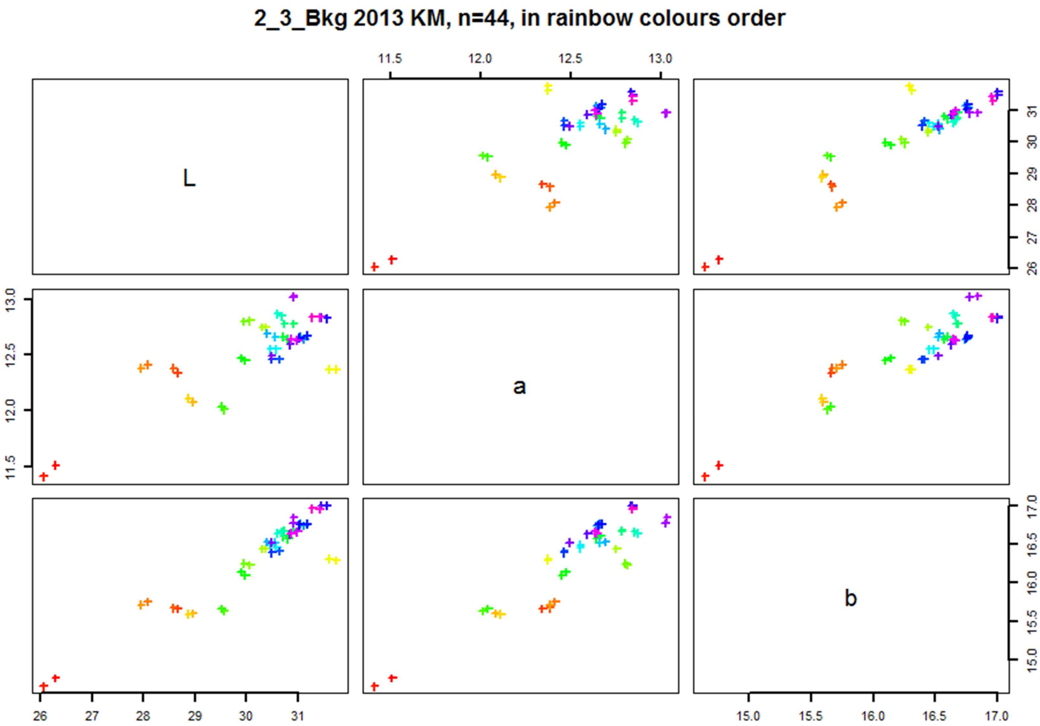


Figure 9. Pairwise plots of the three colour variables for Site 2 Spot 3 Background points in 2013 with KM machine. The rainbow colours (gradating through red, orange, yellow, green, blue, indigo, violet) track the sequence of observations. Pairs of observations close together in the sequence can be seen as pairs of points with very similar colours.

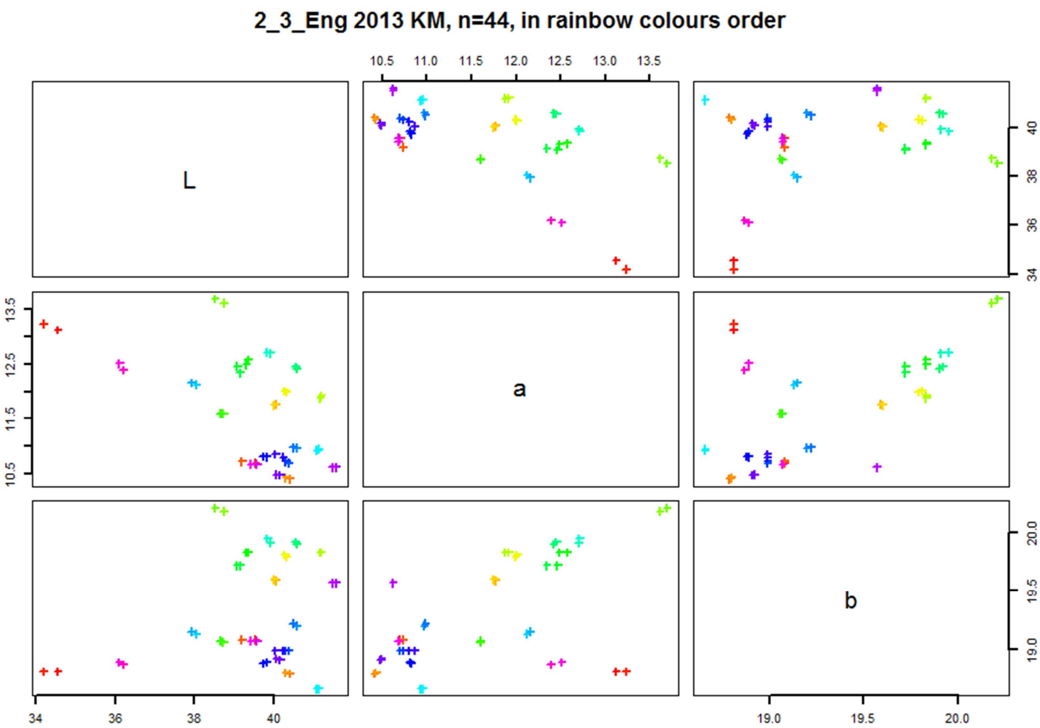


Figure 10. Pairwise plots of the three colour variables for Site 2 Spot 3 Engraving points in 2013 with KM machine.

For another example with very distinct pairing of adjacent observations, see Figure 11, Figure 12 and Figure 13.

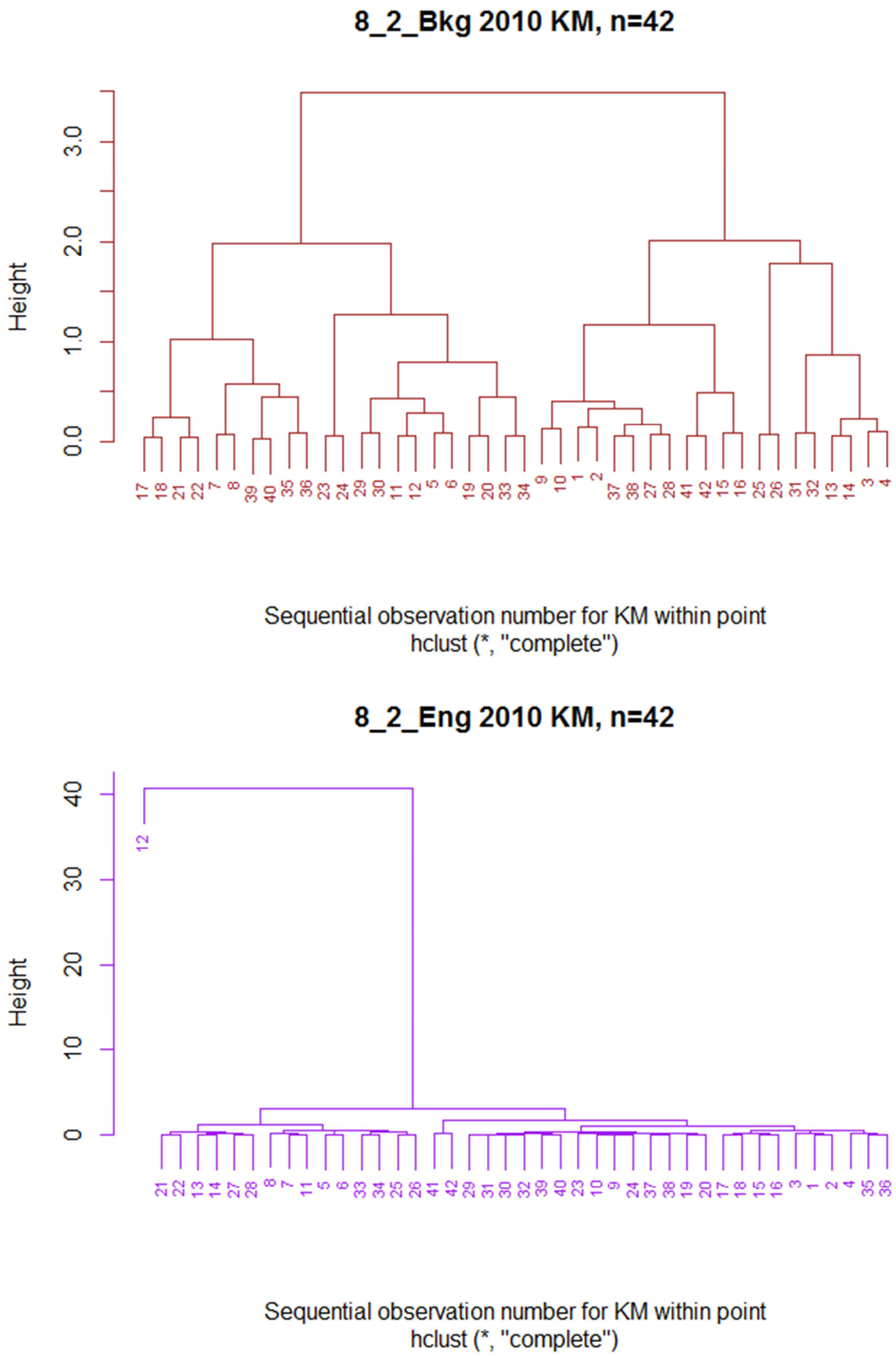


Figure 11. Clusters of observations from Site 8 Spot 2 Background and Engraving points, in 2010 with KM machine. Note how the clusters are dominated by pairs of adjacent observations. There is one serious outlier, observation 12.

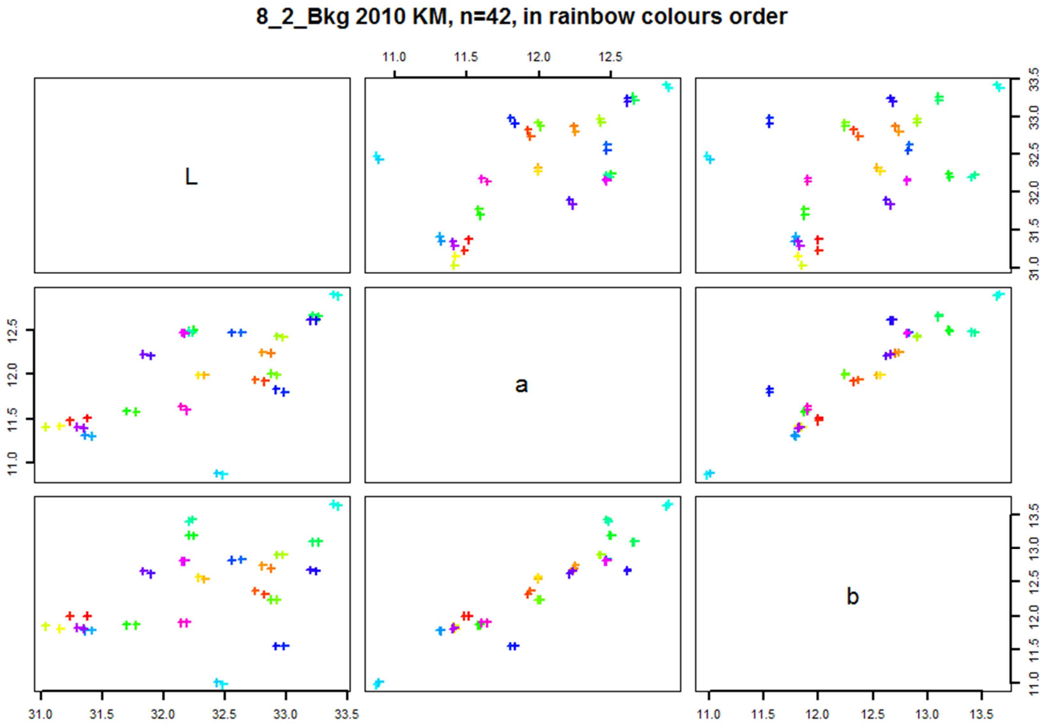


Figure 12. Pairwise plots of the three colour variables for Site 8 Spot 2 Background points in 2010 with KM machine. The pairing of adjacent values is very evident.

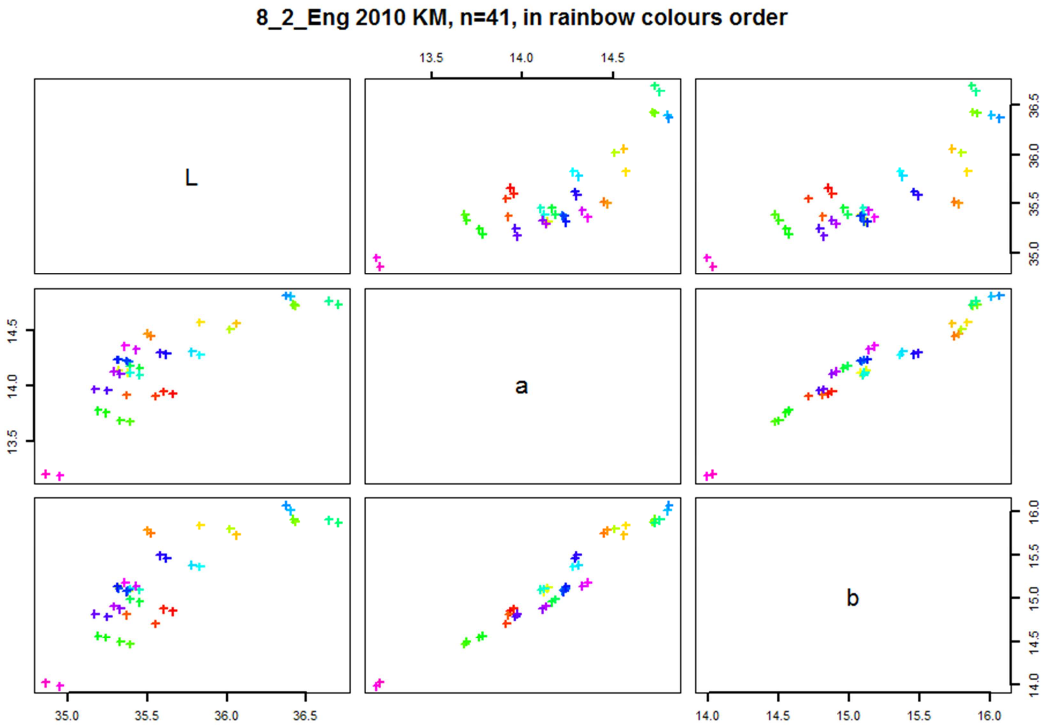


Figure 13. Pairwise plots of the three colour variables for Site 8 Spot 2 Engraving points in 2010 with KM machine, 12th observation removed (outlier).

- Pairing of the data is less obvious in the BYK data (see for example Figure 14, Figure 15 and Figure 16), possibly reflecting the greater variability of readings from that instrument, or a different protocol for taking the measurements.

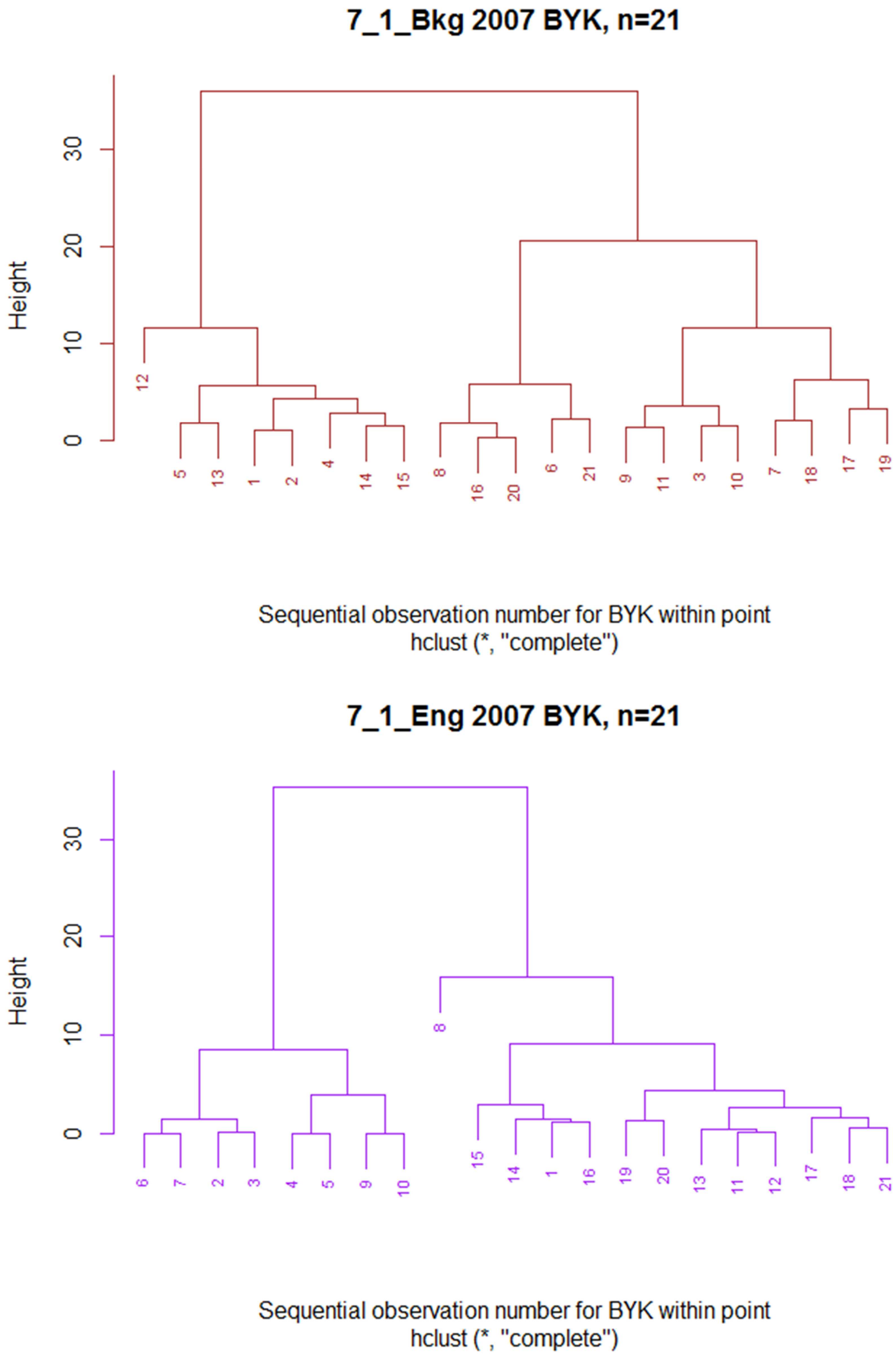


Figure 14. Clusters of observations from Site 7 Spot 1 2007 Background and Engraving points, with BYK machine. Pairing of adjacent observations is present but less obvious.

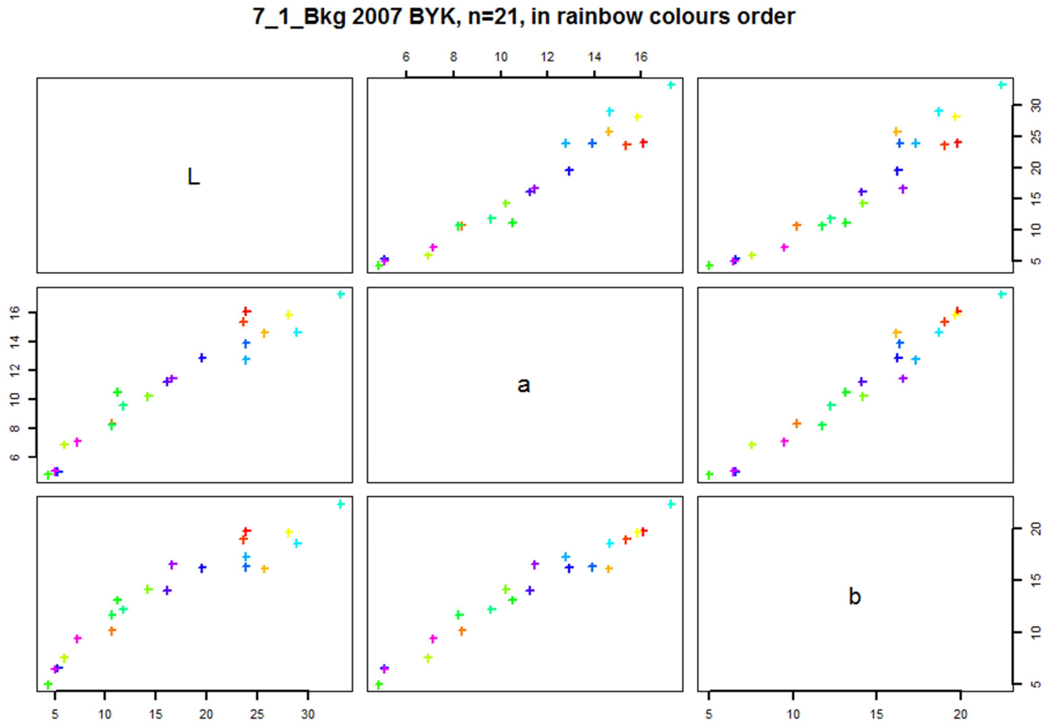


Figure 15. Pairwise plots for Background rock at Site 7, Spot 1 in 2007, BYK machine, corresponding to Figure 14. The colour variables L^*, a^*, and b^* are all highly correlated. The rainbow colour sequence indicates the order of the observations in the data.

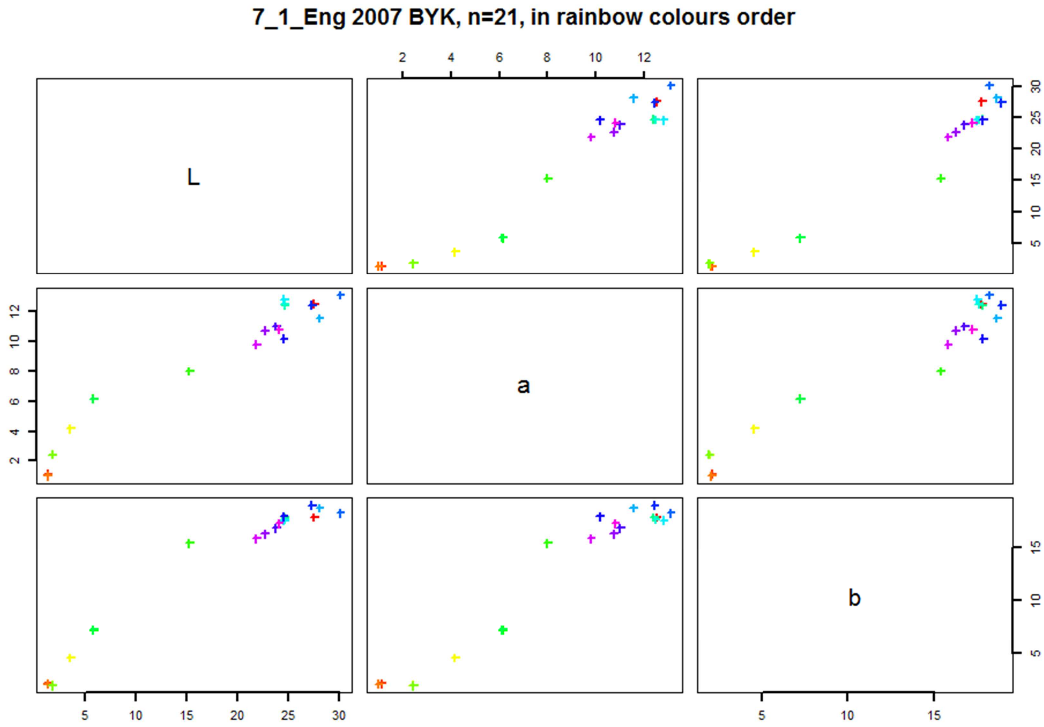


Figure 16. Pairwise plots for the Engraving at Site 7, Spot 1 in 2007, BYK machine, corresponding to Figure 14. The rainbow colour sequence indicates the order of the observations in the data.

It is important to know about such structure in the data so it can be properly allowed for in a mixed model analysis, to avoid attributing too much reliance on the data as if all the measurements were independent when in fact they were not. Mixed models are an extremely powerful tool, but must be used with skill and care.

In many respects this issue is indicative of the absence (to our knowledge) of a fully documented data collection protocol.

6. Conclusions

The Black and Diffey Draft Paper presents a very different approach to the analysis of the Burrup rock art data than that reported by CSIRO, one that in our opinion should have been used for some years. In particular it provides the opportunity to examine longer term trends, to understand whether there are issues affecting multiple sites and to potentially contrast sites close to and far from the industrial developments.

In doing this, Black and Diffey have highlighted a number of inadequacies in the CSIRO reports, particularly the absence of proper statistical analysis in the earlier reports and the over-emphasis on the ΔE measure. Our review of the CSIRO reports and the data provided has also highlighted significant problems of cross-calibration between instruments, inconsistent error-prone data management, and clear errors in the data. The more recent Alexander report using the ASD data does provide some light on the nature of the problems, particularly the unreliability of the BYK data.

It is highly unfortunate that the work of Black and Diffey was affected by the problems in the data provided to them. The data problems are such that it would not be appropriate for the Draft paper to be published in its current form – the findings are based on highly doubtful data rendering any discussion of statistical significance moot.

It is not so clear how to proceed from this position. There remain serious concerns about:

- The quality of the spectrophotometer data and whether it would be better to revert to only using the ASD spectrometer data.
- Understanding the differences between the sites, and whether it is possible to introduce the spatial aspects to the analysis and to better understand the impact of the different rock types.

6.1 Recommendations

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[10].

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.

[10] 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/.

Appendix A. List of Papers Reviewed

Alexander,D. *Analysis of ASD spectra change over time in Burrup Rock Art*, unpublished, 23 June 2016.

Black, J and Diffey, S. *Reanalysis of the Colour and Mineralogy Changes From 2004 To 2014 on Burrup Peninsula Rock Art Sites*, unpublished 2016.

Lau D., Ramanaidou E., Fonteneau L., and Markley T. *Burrup Peninsula Aboriginal Petroglyphs: Colour Change and Spectral Mineralogy (2004–2012)* July 2013.

Lau D., Ramanaidou E., and Furman S. *Burrup Peninsula Aboriginal Petroglyphs: Colour Change and Spectral Mineralogy Report (2004–2010)* March 2011.

Lau D. E. Ramanaidou, A. Hacket, M. Caccetta and S. Furman. *Burrup Peninsula Aboriginal Petroglyphs: Colour Change and Spectral Mineralogy (2004–2009)* April 2010.

Lau D., Ramanaidou E., Morin Ka S. and Furman S. *Burrup Peninsula Aboriginal Petroglyphs: Colour Change and Spectral Mineralogy Report (2004–2011)* September 2012.

Markley T., Fonteneau L., Ramanaidou E., Lau D. and Alexander D. *Burrup Peninsula Aboriginal Petroglyphs: Colour Change and Spectral Mineralogy (2004–2013)* May 2014.

Markley T., Wells M., Ramanaidou E., Lau1 D. and Alexander D. *Burrup Peninsula Aboriginal Petroglyphs: Colour Change and Spectral Mineralogy (2004–2014)* October 2015.

Appendix B. Site Summary Data

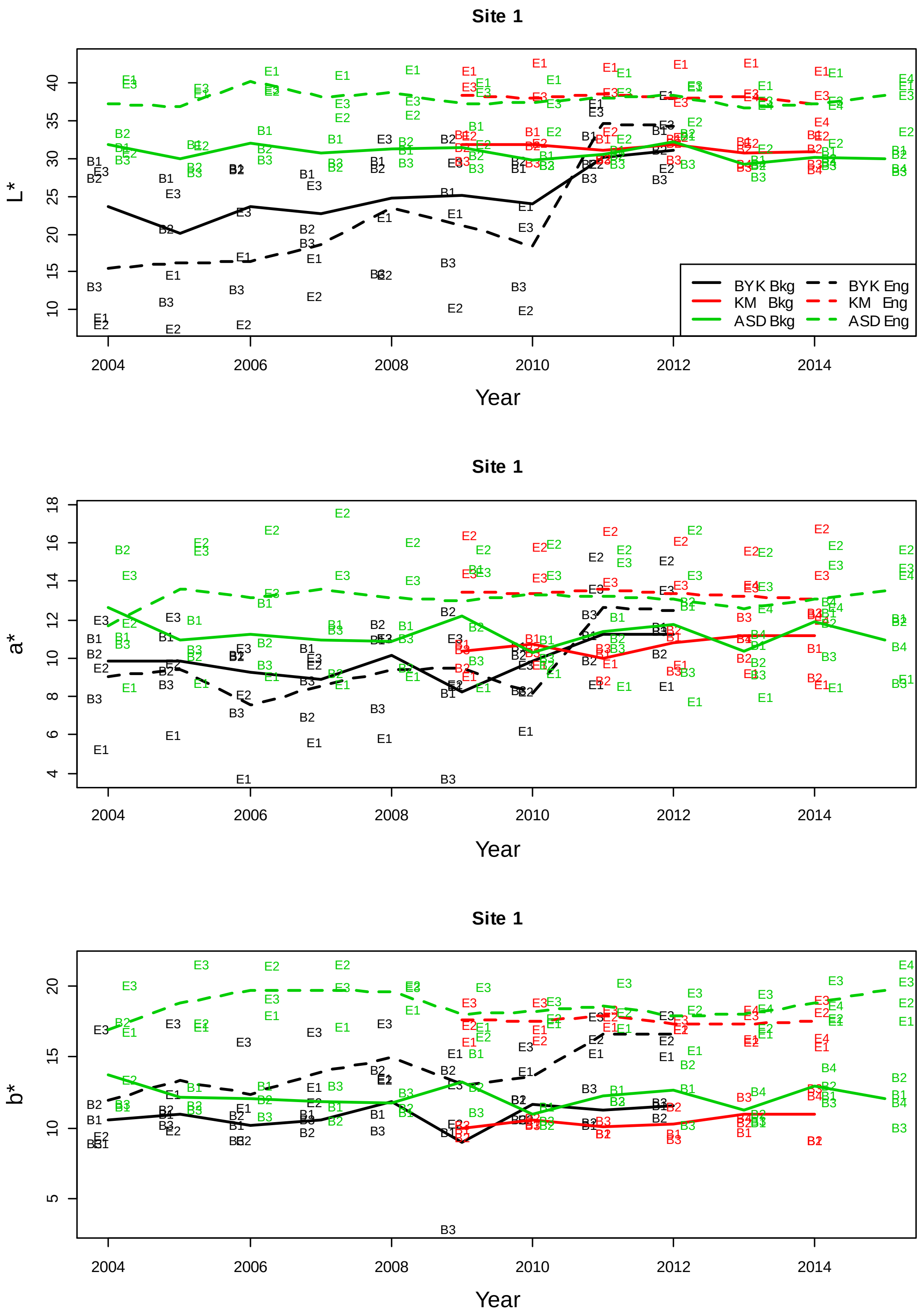


Figure 17. Average L^*, a^*, b^* values for Site 1 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

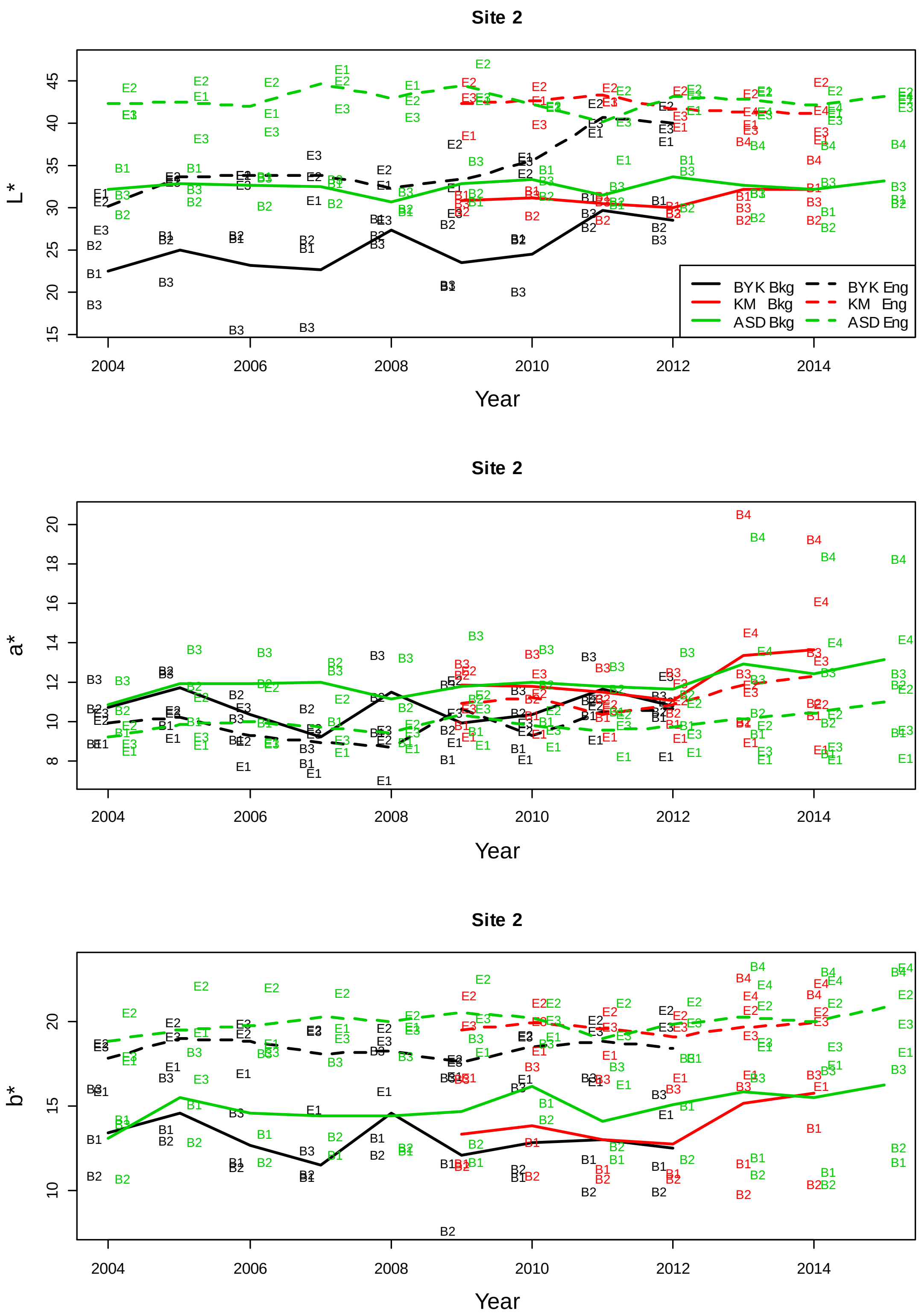


Figure 18. Average L^*, a^*, b^* values for Site 2 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

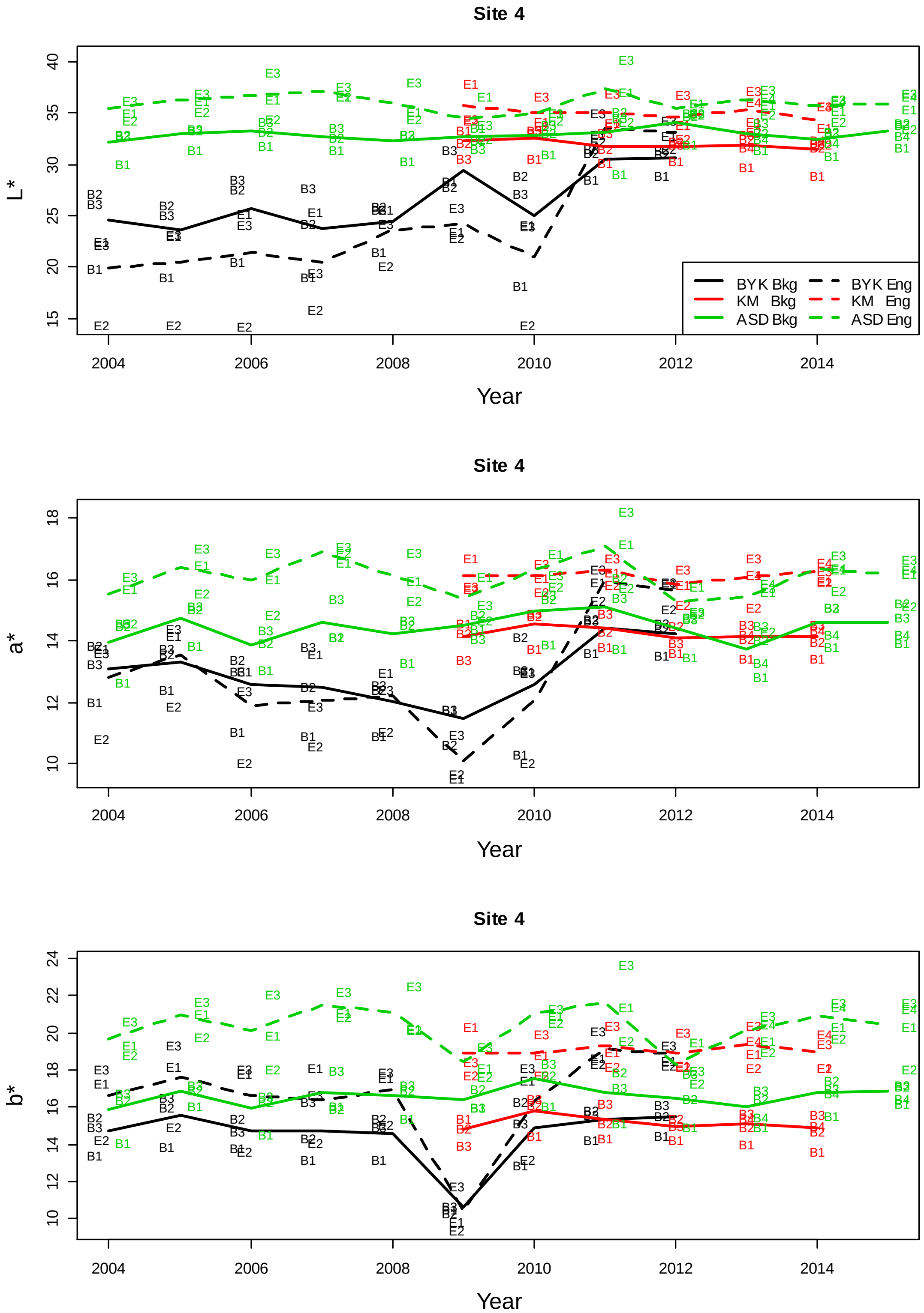


Figure 19. Average L^*, a^*, b^* values for Site 4 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

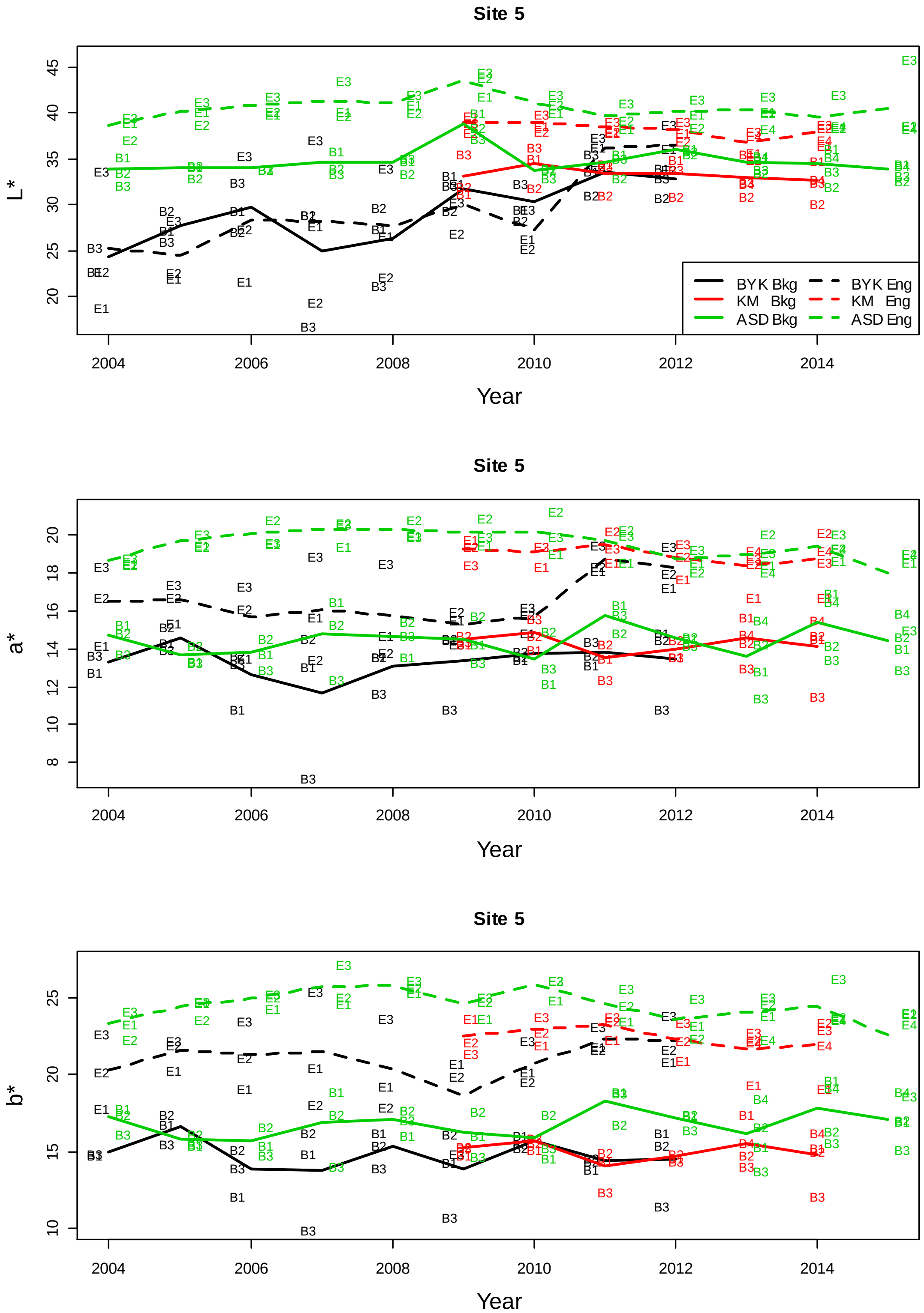


Figure 20. Average L^*, a^*, b^* values for Site 5 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

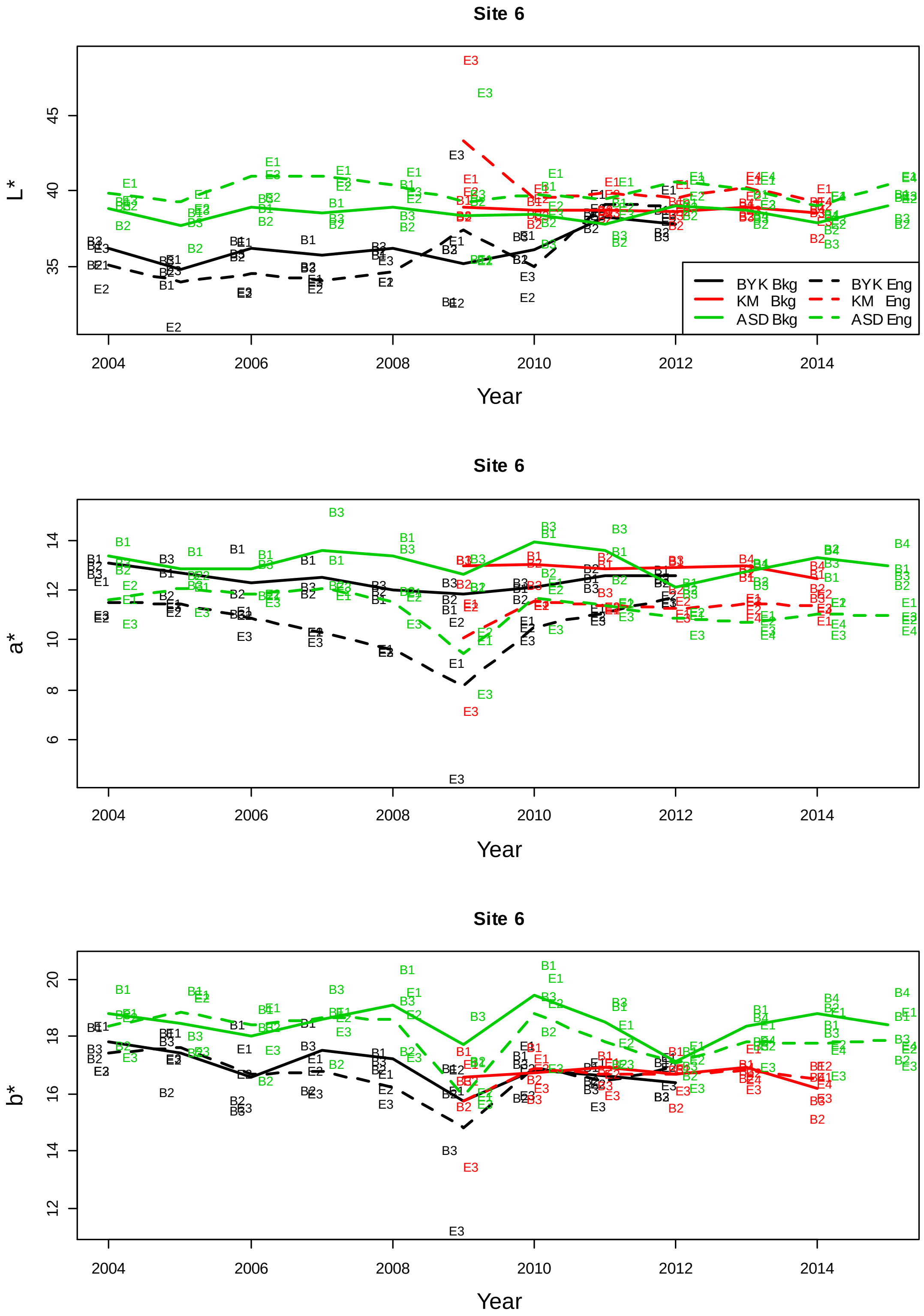


Figure 21. Average L^*, a^*, b^* values for Site 6 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

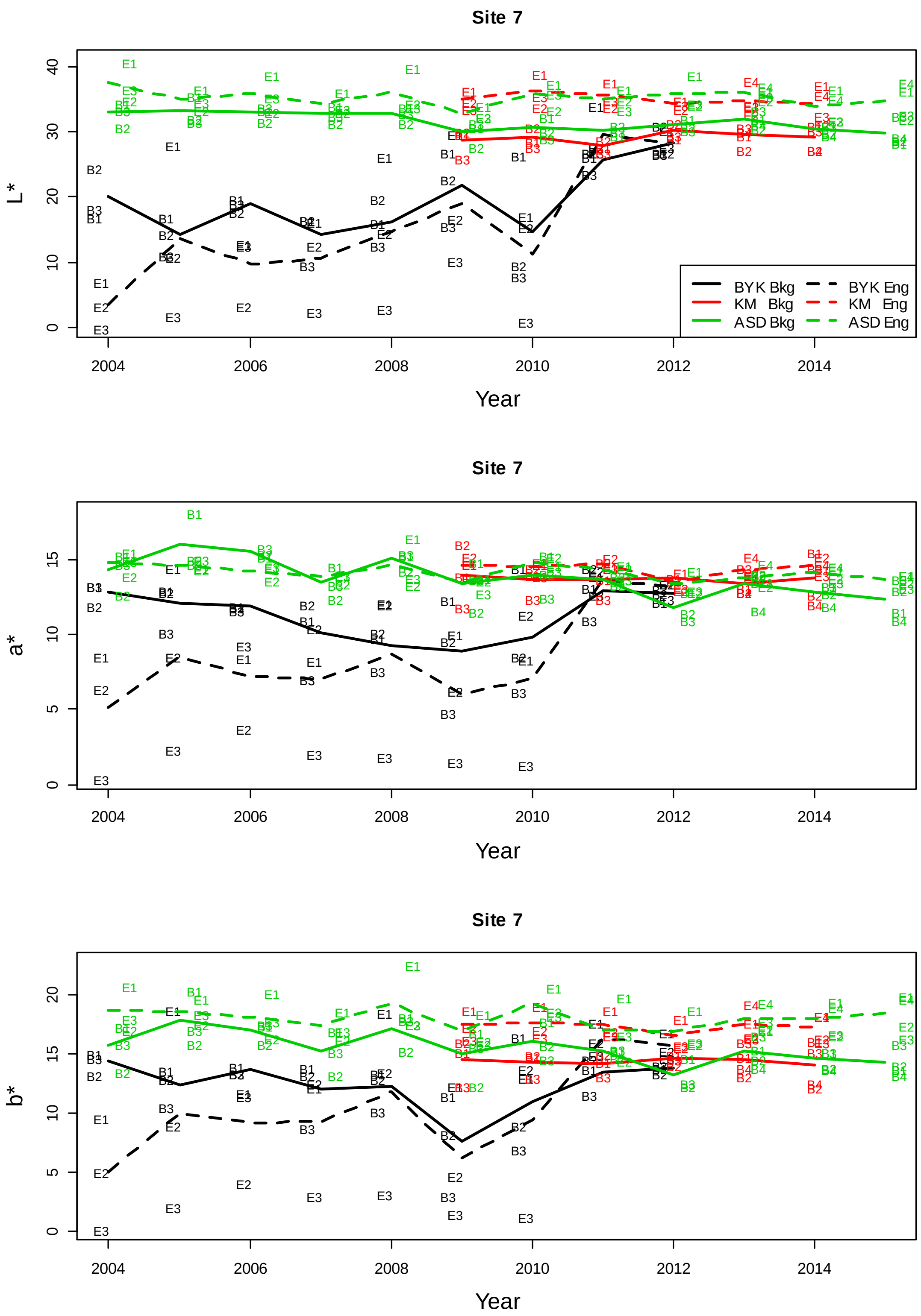


Figure 22. Average L^*, a^*, b^* values for Site 7 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

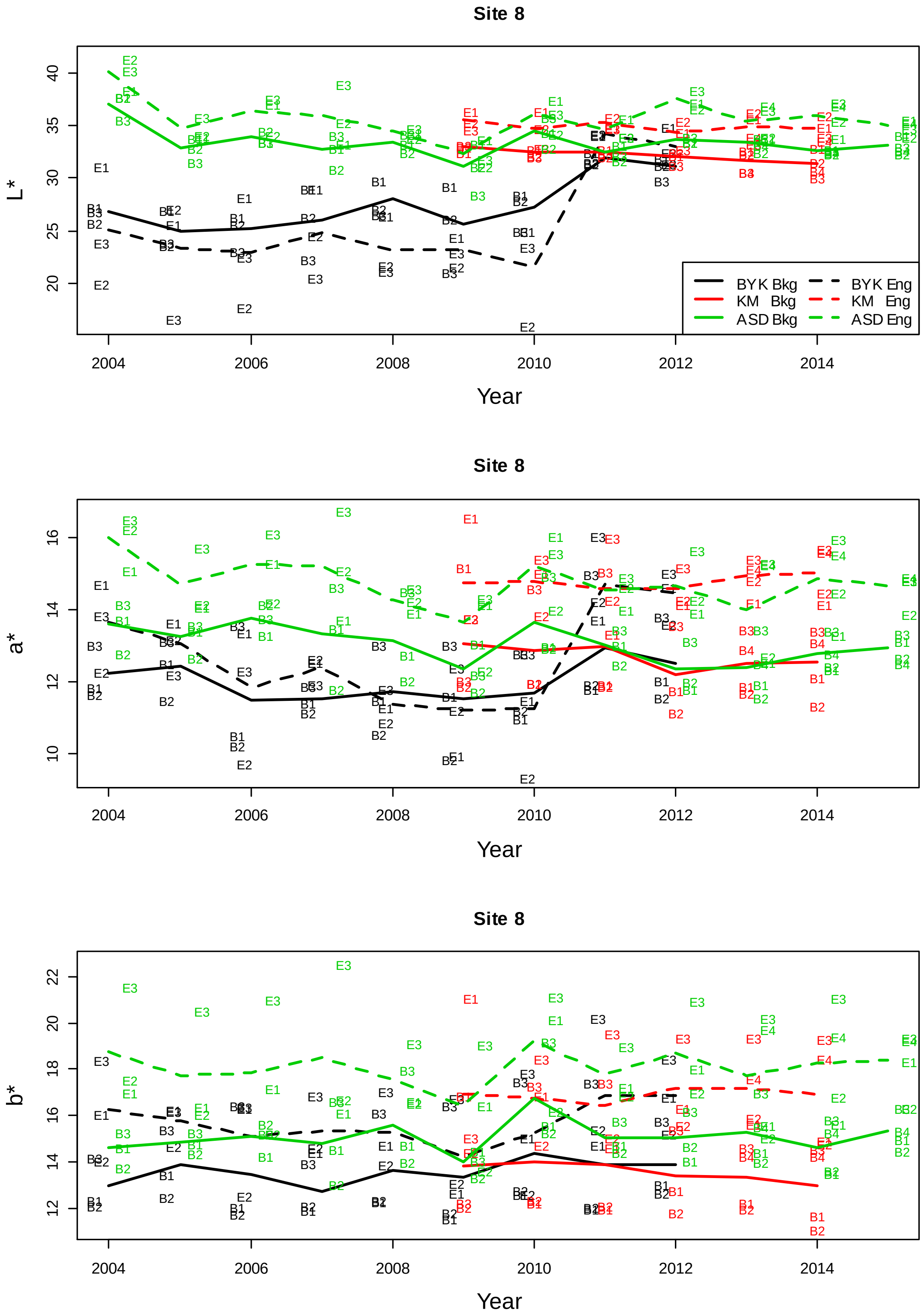


Figure 23. Average L^*, a^*, b^* values for Site 8 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

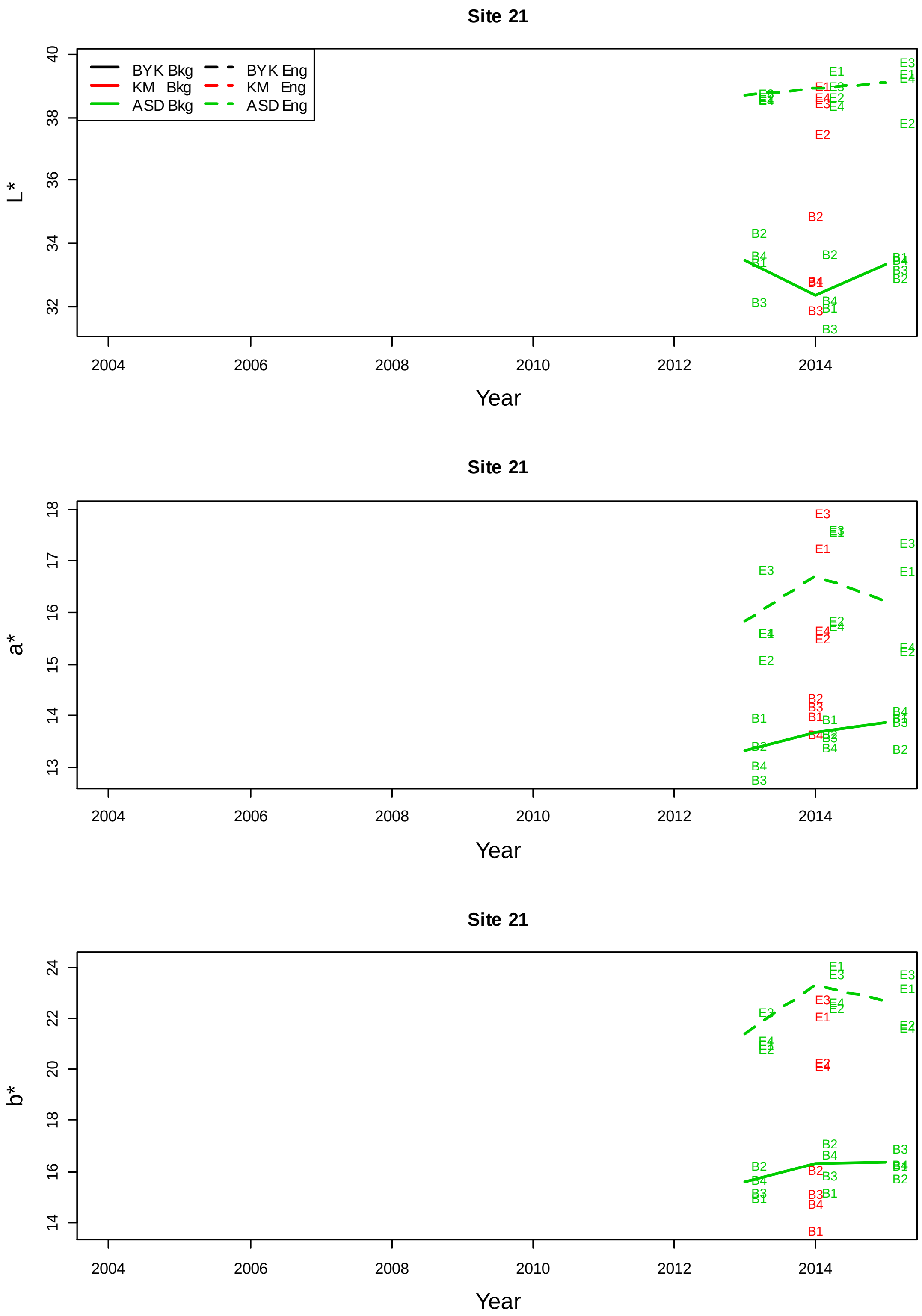


Figure 24. Average L^*, a^*, b^* values for Site 21 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

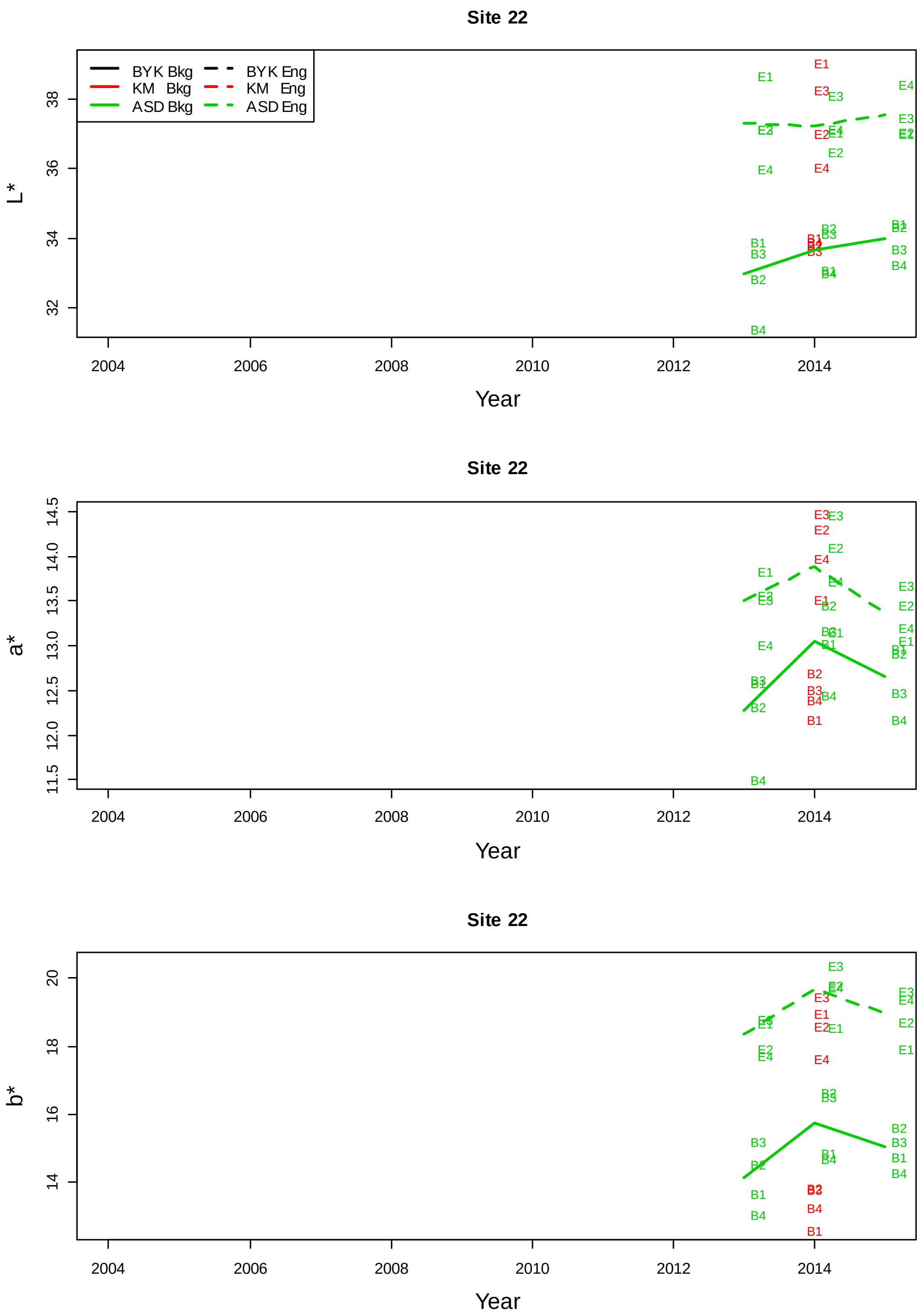


Figure 25. Average L^*, a^*, b^* values for Site 22 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

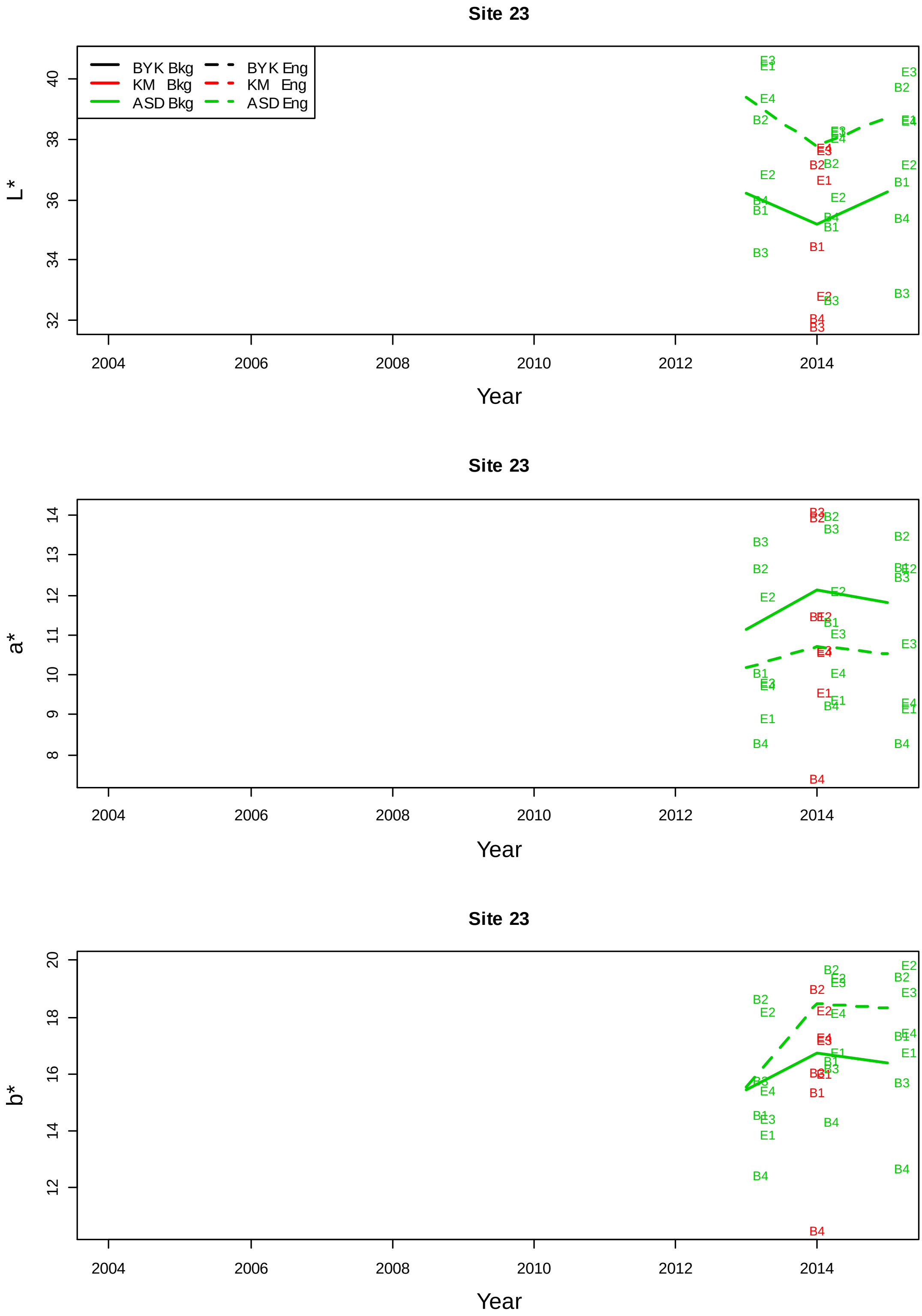


Figure 26. Average L^*, a^*, b^* values for Site 23 – red is the KM Spectrophotometer, Black the BYK spectrophotometer and green the ASD spectrometer; E and the dashed lines represent the engraving points, B and the continuous lines the background points; The point values are given the spot number and the lines are the averages across spots.

Appendix C. Abbreviations Used

Acronym	Definition
ANOVA	**An**alysis **o**f **Va**riance
ASD	**A**nalytical **S**pectral **D**evices, referring to the Analytical Spectral Devices FieldSpec Pro spectrometer
BYK	**BYK**-Gardner, referring to the BYK-Gardner *spectro-guide* spectrophotometer
CIE	**C**ommission **I**nternationale de l´**E**clairage (International Commission on Illumination)
CIELAB colour measurement variables	
L*	A measure of overall luminosity or lightness as perceived by the human eye (0 black to 100 white)
a*	A measure on the green/red colour axis (negative/positive)
b*	A measure on the blue/yellow colour axis (negative/positive)
CSIRO	**C**ommonwealth **S**cientific and **I**ndustrial **R**esearch **O**rganisation
DER	**D**epartment of **E**nvironment **R**egulation (Western Australian)
KM	**K**onica-**M**inolta, referring to the Konica Minolta CM-700d spectrophotometer