An interventive study of glycerol treated freeze-dried leather

Dean Sully*
National Heritage Board
93, Stamford Road
Singapore 178897
Dean_SULLY@NHB.gov.sg

K Suenson-Taylor
Museum of London Specialist Services
150 London Wall
London EC2Y 5HN
ksuenson-taylor@museumoflondon.org.uk

Abstract

Since 1981 waterlogged leather from excavations in London has been treated by a batch process of glycerol (1,2,3 - propantriol) impregnation followed by freeze-drying. A re-evaluation of this treatment procedure has been taking place over recent years. Treatment performance was monitored and evaluated using a criterion anchored rating scale and results analysed statistically. The results of preliminary work were reported in WOAM 96. Improvements to the criterion anchored rating scale have been made and a protocol for its use determined. The additional use of a rigorous experimental design has enabled close analysis of the data using more powerful statistical techniques. The results show the usefulness of developing this technique to indicate methods of improving treatments. However further work on the development of measurement systems and the development of methodologies to model complex systems is necessary before definitive answers can be given.

Introduction

At WOAM96 in York we reported on preliminary experiments to examine and evaluate the treatment of waterlogged leather with glycerol (1,2,3 - propantriol) followed by freeze drying [1]. During that work much progress was made, however it was apparent that some refinement of experimental design was necessary in order to complete the study. It is this further work that is reported here. Our focus has become more the methodologies of how to examine leather treatments, and less the treatments themselves.

This project aimed to test whether the use of different glycerol concentrations improves the condition of the treated leather. If altering concentration has an effect, then on which basis should treatment be assigned; condition, artefact age, type etc. To do this an effective method of assessing condition before and after treatment was required. A bench-top method of object condition assessment was developed using a criterion anchored rating scale (CARS) (Fig 1). This technique focuses the conservator's judgement of object condition and standardises it into data which can be analysed statistically. It also provides a datum from which to assess the long term performance of conservation treatments.

An experimental design was adopted using methods from medical research [2]. An intervention study was chosen: this type of study tests treatment performance in order to rationalise variation of treatment process. For more details of the use of such designs and statistical data in conservation see [3].

Previous work indicated that this method was effective in examining treatment performance. Statistical tests were used that indicated that the condition of the leather after treatment was affected by changes in the concentration of the glycerol used. This suggested that material in poor condition responded better when treated with increased concentrations of glycerol. However improvements in the experimental design were needed to enable us to identify the factors which affect treatment success, and the point at which they become significant.

Refinement of methodology

- Use of the criterion anchored rating scale (CARS)

The scheme has proved to be quick and easy to use, but like any piece of analytical equipment, to achieve valid results, it must be used in a precise manner. For this reason a protocol for the use of the CARS was prepared. A full version is available from the authors.

The CARS is used in conjunction with a standardised pro forma treatment sheet. Information should be recorded in a consistent manner: assessment should be done in the order given on the sheet, adhering closely to the criteria definitions - in difficult cases the lower score should be selected. Post-treatment assessment must be done in a "blind" fashion, that is without reference to the pre-treatment condition assessment and without knowledge of the treatment process. These procedures ensure that the conservator does not introduce bias into the assessment.

- Extension of the scale

Each criteria has a range of condition representing material from worst to best described on a six point scale. The extension of the scale from four to six points provides more mid-range values and increases the subtlety of the data.

- Refinement of the scale

Each point on the scale is described discretely so that a precise relationship is maintained between different points on the scale. The range of the scale is equally divided so that the degree of difference between points is consistent. This is an attempt to derive data that conform to requirements of interval data.

There are many areas in the behavioural sciences which utilise statistical techniques that presume interval level measurements based on attitude variables (i.e. human judgements). This is achieved by collecting data with an ordinal level of measurement and then adopting certain psychological assumptions (such as Thurstone’s Law of Comparative Judgement). There are various standard methods of improving basic ordered scales which are regularly used with parametric statistical analysis in the social and biological sciences. These modes of reasoning tend to be approximate rather than exact. They are therefore effective in many real world applications where diagnosis is restricted by the arbitrary nature of quantified measurement. Elements of these approaches (for example Likert Scales, Ranked Ordered Scales, Public Evaluation, and Magnitude Scaling) have been integrated into the development of the existing CARS for leather treatments. However it should be noted that some statisticians take the view that there is no feasible method available for deriving interval data from ordinal ranked scale data [4].

• Refinement of criteria used in the scale

The criteria defined (physical integrity, cohesivity, friability and flexibility) were selected to represent the important elements of the condition of waterlogged leather and to enable an assessment of the observed effects of treatment. The criterion mineral content was also defined and recorded after discussions at WOAM 96 [5, 6]. The criteria follow a hierarchy of scale from a macro down to micro scale forcing the conservator to look at the condition of an object in a structured manner. The data were examined to ensure that each criterion was measuring an independent factor of condition. These tests also indicated that the six point scale enhanced the subtlety of recorded data and suggested a trend relating increased glycerol concentrations to improved results with leather in poor condition [7].

• Calibration of the criterion anchored rating scale.

Reference samples of leather displaying a variety of condition scores have been taken in order be used as a guide to standardised assessments. It had been hoped to carry out analytical measurements of these type samples in order to correlate the CARS measurements. This work has yet to be undertaken and will be included in future studies into the use of CARS data.

• Selection of treatments

The only systematic source of variation in the data should arise as the result of differences in the data groups. Therefore all objects in the current study have been allocated treatment in a randomised manner. Thus the first object was treated with 15% glycerol, the second with 20% and the third with 25%. Any errors in the assessment procedure should be random and therefore not influence the outcome. The random application of glycerol concentrations allows us to examine the effectiveness of each concentration on objects displaying a range of conditions. The collected data were examined (Table 1) and a Runs Test confirmed that the data are randomly distributed.

Table 1 Data frequencies

Statistical work

500 objects were treated during this study, comprising material from recent excavations and material which had been in long-term wet or frozen storage. A treatment sheet recorded administrative information, storage conditions, treatment details and condition assessments before and after treatment using the CARS.

Although valid attempts have been made to improve the rigour of the measurement process, caution is required when analysing the data. Ordinality may destroy the inferential capabilities of some of the statistical procedures used. Statistical analysis has been used which is appropriate to ordinal data. Tests normally associated with interval data are used for illustration when the results are established by appropriate ordinal techniques. Statistical analysis has been used to establish the validity of the data and to examine the effects of treatment variation. Thus for example does varying the treatment improve success? Do treatments preferentially affect artefacts in different condition?

Due to the space limitations data tables and full statistical test results are presented in summary only. All the statistically significant results listed are designated at 95% (or above) probability levels (0.05 confidence levels). Full details are available from the authors and collaboration is welcomed.

A condition score pre and post-treatment was calculated for each object from three condition criteria: physical integrity, cohesivity and friability. Flexibility and mineral content were independently assessed. Site, type and period information were included as nominal categories. The post-treatment condition score was used to evaluate the success of the treatments. Comparative tests (Wilcoxon matched pairs signed rank test) established this as the most efficient method of identifying differences in the performance of treatments.

Goodness of fit tests were carried out (Kolimogorov-Smirnov) on the condition score data. Figure 2 displays sample data distribution which conforms to normal. This enables us to be confident in the use of mean scores to represent summary information of groups within the data.

Figure 2 Normal quartile - quartile plot for pre-treatment score for 15% glycerol

Comparisons between the treatments

The data from each of the three treatment groups (15%, 20% and 25% glycerol) are compared in figure 3. Treatment with 15% glycerol seems to perform above average while 20% is below average. To test whether this apparent difference is real, univariate comparison tests (Sign, Wilcoxon matched pairs signed ranks test, Ranks, t-test) were carried out. These tests all confirmed that post-treatment scores for material treated with 15% glycerol was significantly different from that treated with 20% and 25%. No difference was established between the results for 20% and 25%. The Wilcoxon matched pairs signed ranks test suggests with 99% certainty that there is a real difference in the performance of the 15% treatment compared to 20 or 25%. Comparative tests on the pre-treatment condition scores for various glycerol groups showed no significant difference between the sample groups before treatment. This confirms that the randomisation process has been successful - any differences between the treatment groups of post condition scores is a direct effect of treatment variation.

Figure 3 Graph of mean post treatment score by pre-treatment score

Influencing factors

To assess the impact of glycerol concentration on artefacts in varying condition the data must be broken down into their constituent parts and their relative strength determined. The most powerful tool available for this is Analysis of Variance (ANOVA), although it is also associated with the most powerful set of limiting assumptions. However the Kruskal -Wallis test which performs some of the functions of ANOVA is applicable to non-parametric data.

Figure 4 Mean Gain in condition score.

Even with 500 treated artefacts there are insufficient data available in each of the 15 condition categories (3-18) once separated into 3 treatment groups. This gives rise to potential distortion of the data in these separate groups and results in a lack of coherence in the data (Figure 4). However two questions can be addressed:

- what do the data say about the relative success of the glycerol treatment with leather in different conditions.
- what do the relationships within the categories of the data tell us about the data and measurement system itself.

Figure 4 shows the amount of noise that is present in the data when condition criteria are considered. Kruskal-Wallis one way ANOVA and true one way ANOVA investigations produce an inconsistent sequence of influences from which no definitive pattern could be discerned. Tentative assertions point to the following; cohesivity and physical integrity preferentially responds to treatment with 15% (worst with 20%). friability responds to 25% (worst with 20%). Overall condition responds best to 15% (worst with 20%).

These techniques can also inform us about the relative influence of the individual condition criteria used in the study. Of the criteria used to make up condition score physical integrity is most consistent in mirroring the variations in the overall condition results. The influence of cohesivity and friability is less consistent. Interestingly mineral content also reflects the results derived from condition score, as to a lesser extent does flexibility, type and period information. It is therefore apparent that all these factors do have an impact on the success of treatments and therefore could be used to help predict the most suitable treatment process.

We have to consider the possibility that the lack of clarity in the data and the absence of obvious patterns are due to the influence of several factors:

- the relative success of the treatment cannot be fully explained by varying the effects of glycerol. Other variables in the process, such as pre-treatment storage, freeze-drying process, etc. interact with the observed effects of glycerol and influence the outcome of the treatments.
- the condition assessments used are not effectively measuring the condition variation. Elements such as flexibility and mineral content have been tentatively included, however there may be other categories of condition which are not adequately encompassed within the current condition categories and condition score (e.g. species, tannate).
- the power of the data is not sufficient to explain variations to the level of subtlety required in the study. Greater sample numbers may be required to resolve the inherent limitations of this type of data.
- patterns and trends are present within the data but are not evident with our current statistical techniques. The production of a weighted scale based on the relative influences of the individual categories could be a way of improving the use of the scale.

What we are looking for in the data are essentially trends which explain the transition between different condition states. This is more appropriate than attempting to match treatment variation to individual condition criteria which may not be within the capabilities of this type of data. An attempt was made to reduce the variation in the data by grouping into three condition groups: Poor (condition score 3-7), Medium (8-12), Good (13-18). These groups are depicted in Figure 5. This appears to show a trend in the success of the treatments, with Poor material responding best with 25% glycerol, Medium with either 15 or 25%, and Good with 15%.

Figure 5 Mean Gain in condition score for collapsed data grouping 1

This pattern was tested using Wilcoxon matched pairs signed rank test. For poor condition material 25 % glycerol yields significantly better results than 20 % and 15%, between which there is no significant difference. For the medium condition grouping no one glycerol concentration is significantly better. For the good condition grouping 15% was significantly better than both 20% and 25% glycerol.

To test this further, the grouping was amended to see if the pattern held. The composition of each of the subsets was progressively amended. Comparative tests were then carried out, which revealed that varying the nature of the groups does not contradict the assertion that there is a direct relationship between condition of the artefact, and the concentration of glycerol required to most successfully stabilise it during treatment. Thus the grouping of data into these arbitrary sets has enhanced the appearance of trends within the data.

Discussion

As a result of this experimental procedure it is clear that:

- the assessed condition factors conform to a normal distribution which enables us to use the summary data such as mean and standard deviation to define the sample groups.
- as an overall treatment, 15% glycerol proved to be the most consistent treatment for the type of material included in this study: 20% treatment tends to be the least successful of the three treatment variations.
- condition factors such as type, period, mineral content and flexibility, which are not used in the calculation of the condition score, still mirror the results observed.
- physical integrity appears to be the most meaningful of the individual condition criteria.
- there does appear to be a trend in the success of treatments with leather in poor condition responding best with higher concentrations of glycerol. This however requires verification, especially for the reactions of material in poor condition. -

Conclusion

Each phase of this study has brought a greater awareness of the limitations imposed by statistical analysis at the same time as expanding the options and possibilities of the inferences which can be drawn from collected data. Our starting point was a very simple question about the affect of treatment variation on artefact condition. This work has become a testing ground for an approach to data recording which we feel has broader application in experimental research and standardisation of conservation observations. A great deal more work is required to refine a suitable measurement scheme for use with leather treatments. Before we come back to the question of wet leather treatments, we need to look in detail at the systems available for measuring outcome. In the absence of a system for directly measuring the concept of condition in wet archaeological leather a simple convention may be required to define an agreed definition. Our future work will look at the application of psychological-physical techniques to ordered ranking systems. Other options for complex systems include Multi Dimensional Scaling and the use of linear general models which can derive predictive equations. The application of fuzzy logic techniques to model the function of whole systems rather than deconstructing elements of these systems may also be a valid solution. This has been effectively used in equivalent studies where data collection about component factors is insufficient to model the functioning of a system. Future work will focus on developing the CARS as a useful tool identify of trends in treatment success and as a datum to determine the stability of artefacts in the long-term.

Acknowledgements
The authors would like to thank the conservation staff of the Museum of London and the Museum of London Specialist Services for their help and comments. We are also grateful to colleagues and students from the Institute of Archaeology for their help with data collection and in particular Clive Orton and Bonnie Baskin for their help and advice.

References

1. Suenson-Taylor K., and Sully D., (1997) The use of condition score to determine Glycerol concentration in the treatment of waterlogged Archaeological leather. An empirical solution. Hoffman et al. Proceedings of the 6th ICOM Group on Wet Organic Archaeological Materials Conference, York 1996. Bremerhaven ICOM 1997

2. Page R., Cole G., and Timmreck T., (1995) Basic Epidemiological Methods and Biostatistics. London, Jones and Bartlett 1995

3. Suenson-Taylor K., Sully D., and Orton C., (forthcoming) Data in conservation: the missing link in the process. Studies in Conservation, forthcoming

4. Golbeck A., (1986) Evaluating Statistical Validity of Research Reports: a guide for managers, planners, and researchers. Berkeley, U.S, Dept. of Agriculture, Forest Service, General Technical Report PSW-87. 1986

5. Wallace A., (1997) Scanning electron microscopy and fibre shrinkage temperature analysis of archaeological waterlogged leather: observations on medieval leather from Swinegate, York. Hoffman et al. Proceedings of the 6th ICOM Group on Wet Organic Archaeological Materials Conference, York 1996. Bremerhaven ICOM 1997

6. Watson J., (1997) Freeze drying highly degrade waterlogged wood. Hoffman et al. Proceedings of the 6th ICOM Group on Wet Organic Archaeological Materials Conference, York 1996. Bremerhaven ICOM 1997

7. Baskin B., (1997) written communication, May 1997, unpublished