Stat Comput
DOI 10.1007/s11222-012-9329-5

Regression analysis with partially labelled regressors: carbon dating of the Shroud of Turin

Marco Riani · Anthony C. Atkinson · Giulio Fanti · Fabio Crosilla

Received: 16 September 2011 / Accepted: 1 April 2012

Abstract The twelve results from the 1988 radio carbon dating of the Shroud of Turin show surprising heterogeneity. We try to explain this lack of homogeneity by regression on spatial coordinates. However, although the locations of the samples sent to the three laboratories involved are known, the locations of the 12 subsamples within these samples are not. We consider all 387,072 plausible spatial allocations and analyse the resulting distributions of statistics. Plots of robust regression residuals from the forward search indicate that some sets of allocations are implausible. We establish the existence of a trend in the results and suggest how better experimental design would have enabled stronger conclusions to have been drawn from this multi-centre experiment.

Keywords Computer-intensive method · Forward search · Robust statistics · Simulation envelope

M. Riani
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: mriani@unipr.it

A.C. Atkinson (✉)
Department of Statistics, London School of Economics, London, UK
e-mail: a.c.atkinson@lse.ac.uk

G. Fanti
Department of Industrial Engineering, University of Padua, Padua, Italy
e-mail: giulio.fanti@unipd.it

F. Crosilla
Department of Civil Engineering and Architecture, University of Udine, Udine, Italy
e-mail: fabio.crosilla@uniud.it

Published online: 27 April 2012

1 Introduction

The results of a radiocarbon dating of the Turin Shroud (TS) were published by Damon et al. (1989). Four samples of fabric were cut from a corner of the cloth and sent to three laboratories, the University of Arizona receiving two samples. Statistical analysis of the twelve resulting readings of radiocarbon age shows a surprising lack of homogeneity which was not present in control readings from three other fabrics chosen to span possible ages and sources for the TS. We use regression on spatial coordinates to model this lack of homogeneity. However, the spatial coordinates of the twelve samples are not precisely known. We consider all 387,072 plausible spatial allocations and analyse the resulting distributions of statistics, using simulation envelopes to calibrate this large number of tests on the same data. Plots of robust regression residuals from the forward search indicate that some sets of allocations of the readings from Arizona are implausible. The remaining allocations all point to the same inferential conclusion of a trend along the sample.

Undisputed historical records of the existence of the TS go back to AD 1357. The cloth shows front and back images of a thorn-crowned man. The images are much clearer in a black-and-white photographic negative than in their natural state, as was discovered in 1898 by the amateur photographer Secondo Pia. Ballabio (2006) surveys the extensive literature on scientific aspects of the date of the shroud. More recently, Fanti et al. (2010) contend that the formation mechanism of the body images has not yet been scientifically explained; only the external layer of the topmost linen fibres is coloured. However, the results of the 1988 radiocarbon dating (Damon et al. 1989) stated that the linen fabric dates from between 1260 and 1390 AD, with a confidence level of 95 %.

After publication of this result, some speculated that the sample had been contaminated due to the fire of 1532 which

seriously damaged the TS, or to the sweat of hands impregnating the linen during exhibitions, others that the date was not correct due to the presence of medieval mending and so forth. The purpose of our article is not to discuss the reliability of the various assumptions made, but to show how robust methods of statistical analysis, in particular the combination of regression analysis and the forward search (Atkinson and Riani 2000; Atkinson et al. 2010) combined with computer power and a liberal use of graphics, can help to shed light on results that are a source of scientific controversy.

The Turin Shroud (TS) is 4.4 m long and 1.1 m wide. The samples for radio carbon dating were taken from a thin strip of material cut from one corner of the TS. The strip was divided into four parts; the three larger parts were sent to laboratories in Arizona, Oxford and Zurich and the fourth, smaller, part was also sent to Arizona (see Fig. 1). These samples were divided into a total of 12 sub-samples for which datings were made, together with standard errors for each subsample based on a number of individual determinations which are not available to us.

The longitudinal locations of the four samples in the strip are known. On the assumption that the four readings from Arizona all came from the large sample (A1 in Fig. 1), Walsh (1999) showed evidence for a regression of age on the (known) centre points of the three pieces of fabric given to the three laboratories. This analysis ignored both the quoted standard deviations of the measurements and the further source of heterogeneity due to the division of the samples into subsamples, which also introduces a second spatial variable into the regression, the values of both variables depending on how the division into subsamples is assumed to have been made. Ballabio (2006) attempted an analysis based on a series of assumptions about the division of the samples, but was defeated by the number of cases to be considered.

We summarise the evidence for heterogeneity in Sect. 2. A full comparison of unweighted analyses with those weighted by the reported accuracies of the three laboratories is given in the Appendix. The possible spatial layouts of the subsamples are described in Sect. 3. We proceed in Sect. 4 by calculating the 387,072 possible bivariate regressions and looking at distributions of the resulting t statistics for the two regression variables. Only that for length along the strip is significant, but the histogram of values exhibits a surprising bimodal distribution. In Sect. 5 we use graphical methods associated with the forward search to show that the subsamples from Arizona must all have come from the single larger sample (A1 of Fig. 1); the contrary assumption leads to the generation of gross regression outliers. We conclude in Sect. 6 with a brief discussion of statistical aspects of radiocarbon dating and of how the application of statistical principles would have produced a design leading to sharper conclusions.

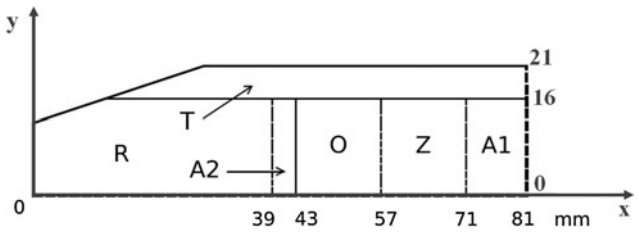

Fig. 1 Diagram showing the piece removed from the TS and how it was partitioned. *T*: trimmed strip. *R*: retained part called "Riserva" (initially including also A2). *O*, *Z*, *A1*, *A2*: subsamples given to Oxford, Zurich, and Arizona (two parts) respectively

2 Heterogeneity

Table 1, taken from Table 1 of Damon et al. (1989), gives the estimated radiocarbon age, in years BP, that is before the present which is conventionally taken as 1950, of the 12 samples of the TS. Also given in the table are the standard errors of the individual measurements. These latter are potentially misleading (at least, as we explain in the Appendix, they initially misled us).

Relative to the given standard deviations the continuous observations are all far from zero and do not have a wide range. It is then natural to consider a normal theory linear

Table 1 Estimated radiocarbon ages of the individual samples (years before 1950) with calculated standard deviations. Those for Arizona exclude one source of error (see Appendix). The standard deviations for the mean age at each laboratory come from Table 2 of Damon et al. (1989) and can be compared with those calculated from the v_{ij} using (2)

Sites		Individual observations					Weighted means	s.d. (mean) from (2)
Arizona	y	591	690	606	701		646	
	v	±30	±35	±41	±33		±31	±17
Oxford	y	795	730	745			750	
	v	±65	±45	±55			±30	±32
Zurich	y	733	722	635	639	679	676	
	v	±61	±56	±57	±45	±51	±24	±24

model. If we ignore any spatial factors, a simple general model for observation j at site i is

$$y_{ij} = \mu_i + \sigma v_{ij}\varepsilon_{ij} \quad (i = 1, 2, 3;\ j = 1, \ldots, n_i), \tag{1}$$

where the errors $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$. We briefly discuss more complicated models in Sect. 6. Our central concern is the structure of the μ_i, at this point whether they are all equal. However, to test this hypothesis we need to establish the error structure. In the Appendix we argue that an unweighted analysis is appropriate, that is that all $v_{ij} = 1$.

A major part of the argument is that, in addition to the TS, each laboratory dated three controls: linen from a Nubian tomb with Islamic patterns and Christian ink inscription, stylistically dated to the eleventh to twelfth centuries AD; an Egyptian mummy from Thebes, with a previously estimated radiocarbon age around 2010 BP and threads from a cope from Var, France, historically dated to 1290–1310 AD. None of the datings of these samples was controversial. The unweighted analysis does not reveal any inhomogeneity of either mean or variance of the three fabrics the provenance of which is not in doubt. However, as the results of Table 2 show, weighted analyses give significant differences between some means and variances for the three laboratories.

It might be expected, despite the most stringent calibrations, that there would be significant laboratory effects. However, the unweighted analysis provides no evidence of heterogeneity in the means of the three control samples and so no evidence of systematic differences between laboratories. The plots of the means for all four fabrics in Fig. 2 of Damon et al. (1989) reinforce this point; heterogeneity, or a laboratory effect, is only evident for the TS.

Christen (1994) used these data as an example of Bayesian outlier detection with a mean shift outlier model (Abraham and Box 1978) in which the null model was that the data were a homogeneous sample from a single normal population. He found that the two extreme observations, 591 and 795 were indicated as outlying. When these two observations were removed, the data appeared homogeneous, with a posterior distribution of age that agreed with the conclusion of Damon et al. (1989). We have already mentioned the regression analysis of Walsh (1999) that effectively "binned" the data and assumed that all Arizona samples came from A1. We now use a spatial analysis to try to discover the source of the egregious heterogeneity in the readings on the TS.

3 Spatial layout

We have appreciable information about the spatial layout of the samples sent to the three sites, although the detailed layout of the subsamples is uncertain. Walsh (1999) argues convincingly that the strip of TS linen fabric used for dating seems to have slightly different sizes from those reported by Damon et al. (1989). From this strip an approximately 5 mm portion was trimmed, thus removing the stitching and remnants. This process resulted in a piece of fabric that measured 81 mm × 16 mm.

This piece of TS fabric was then divided into two parts; one, called "Riserva", remained in Turin for future analyses (see Fig. 1) and the other was divided into three parts as shown. Since A1 was smaller than the other two pieces, a section of the "Riserva" was cut and these two pieces of fabric were sent to Arizona. Zurich received the sample next to A1 and Oxford, as shown in Fig. 1, the material between Zurich and A2.

We know from Damon et al. (1989) that four different pieces were dated by Arizona. The possible configurations are therefore those shown in Fig. 2. For Zurich, from photographs published on the internet (but now deleted) it is known that, after a first division, the first piece was divided into three subsamples while from the second piece two subsamples resulted; the possible configurations are those shown in Fig. 3. Oxford instead divided its piece of TS fabric into only three parts with the possible configurations as shown in Fig. 4.

Obviously Figs. 2–4 do not represent all the possible subsample configurations because, for example, triangular shapes are not considered. Those considered seem the most significant and the addition of other possibilities does not appreciably change the positions of the centres of gravity of the subsamples which we used as references for our calculations.

4 Two variable regression

To try to detect any trend in the age of the material we fit a linear regression model in x_1 (longitudinal) and x_2 (transverse) distances. Since the sample is long (in x_1) and thin (in x_2) we expect that there is more likely to be an effect, if any, in x_1 and this is what we find.

The analysis is not standard. There are 387,072 possible cases to analyse. We can permute the values of x_1 and x_2 and calculate this number of analyses. The question is how to interpret this quantity of numbers.

The left-hand panel of Fig. 5 plots, as a continuous line, the ordered significance level of the t-test for x_2 in the model with both variables. In the absence of any effect of x_2 we would expect these ordered values to fall close to a straight line. Indeed, this curve, coming from all 387,072 possible configurations of x_1 and x_2, is a relatively straight diagonal line. To calibrate it we generated 100 samples of 12 observations from a standard normal distribution and analysed each set for the 387,072 configurations. For each sample the p values were ordered. The dotted lines in the figure show the

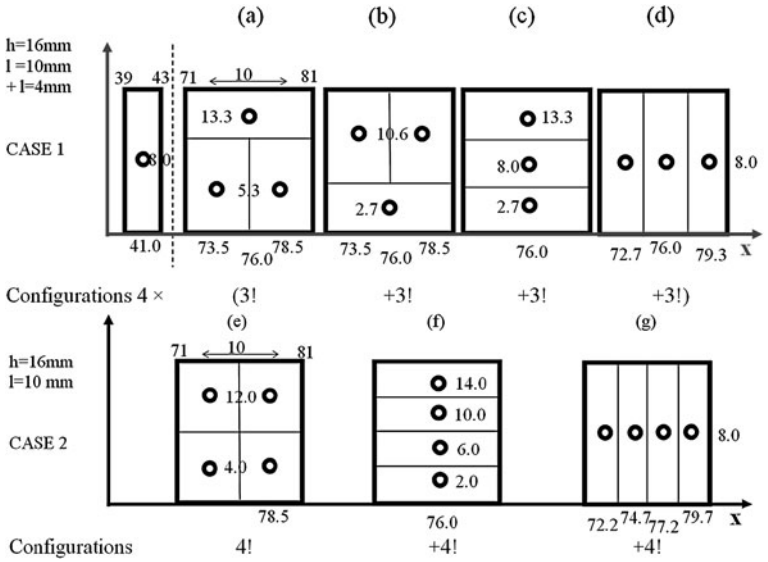

Fig. 2 Spatial arrangements investigated for the Arizona sample. The image on top assumes that Arizona dated both pieces (A1 and A2), with one reading taken on A2. The *image at the bottom* assumes that Arizona only dated piece A1. Total number of cases considered is $168 = 96 + 72$

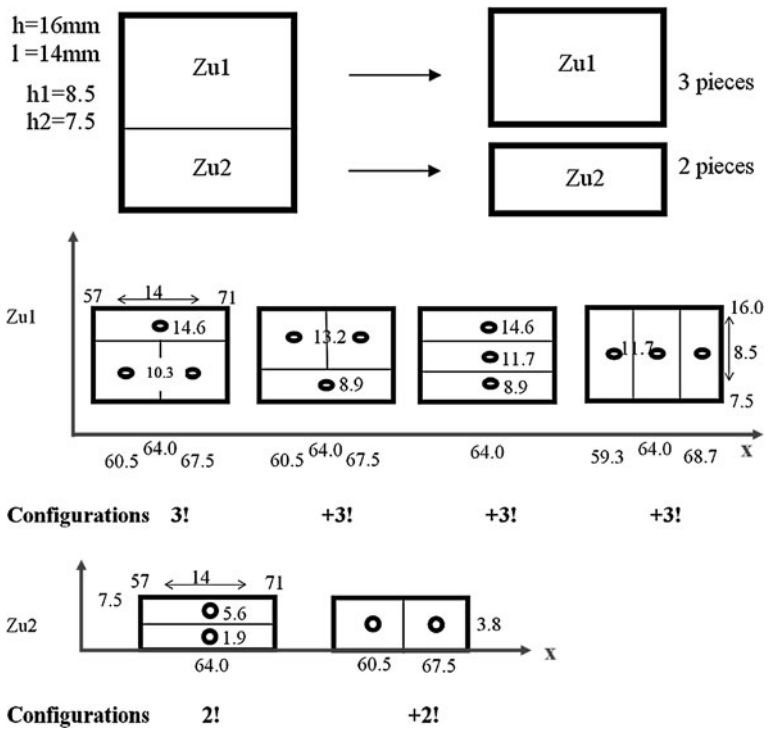

Fig. 3 Spatial arrangements investigated for the Zurich sample. It is known that this sample was divided into the two parts shown in the top panel. The *two lower panels* show possible further subdivisions. Total number of cases considered is $96 = 24 \times 4$

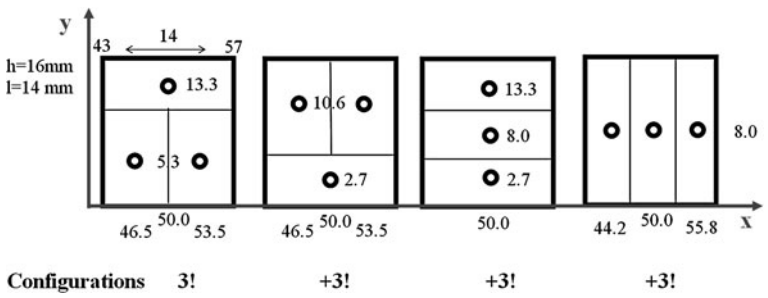

Fig. 4 Spatial arrangements investigated for the Oxford sample. The total number of cases considered is $24 = 3! \times 4$

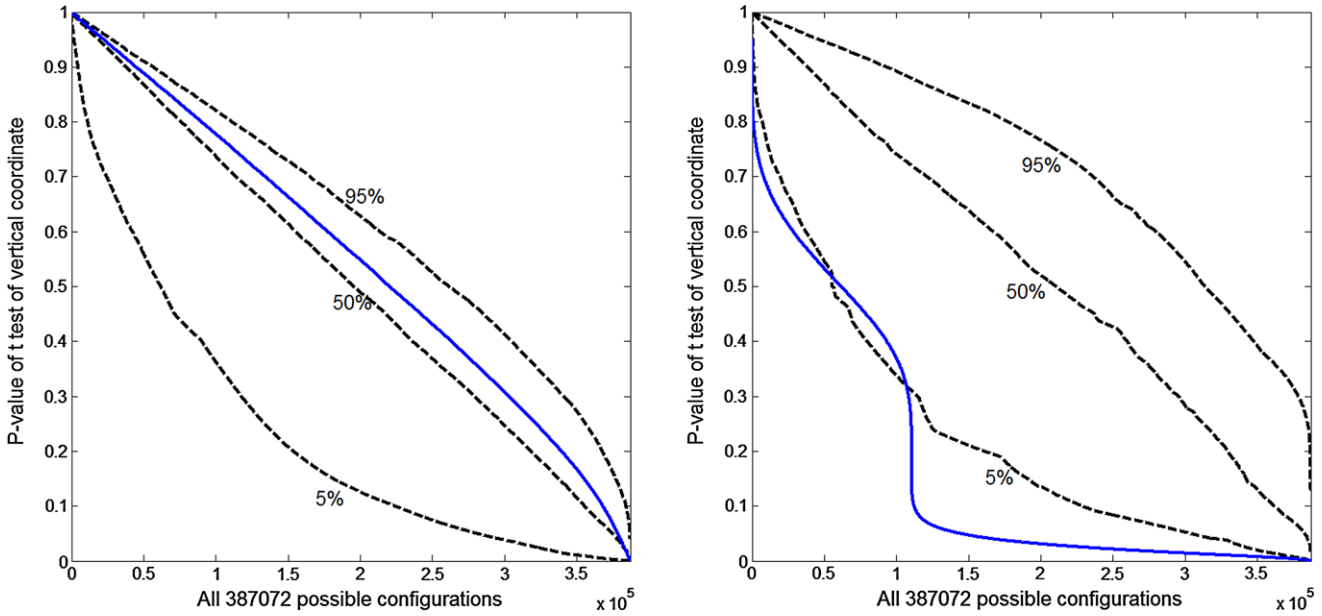

Fig. 5 Two variable regression. Significance levels of t-statistic from 387,072 possible configurations (*continuous line*) and envelopes from 100 simulations of each configuration. *Left-hand panel* x_2, *right-hand panel* x_1

5 %, 50 % and 95 % points of this empirical distribution. The observed values lie close to the 50 % point throughout. There is clearly no evidence of an effect of x_2.

The right-hand panel repeats the procedure for x_1 in the model for both variables. Now the values lie outside the lower 5 % point for virtually all configurations. The non-smooth envelopes reflect the discrete nature of the configurations, some of which include leverage points. It is clear that there is a significant effect of x_1, although the shape of the curve generated by the data merits investigation.

Histograms of the statistics help. The top panel of Fig. 6 shows the distribution of the t-statistic for x_2. This has a t like shape centred around 0.5. The bottom panel of Fig. 6, the t-statistic for x_1, is however quite different, showing two peaks. The larger peak is centred around -2.8 whereas the thinner peak is centred around -0.6. It is also interesting to notice that for each of the 387,072 configurations we obtain a negative value of the t-statistic for the longitudinal coordinate.

Although our procedure involves permutations of data, these analyses are not those associated with permutation tests. In a permutation test (for example, Box et al. 1978, Sect. 4.1) the values of x_1 and x_2 are kept fixed, the observations y being permuted over the design points and a statistic calculated for each permutation. The position of the value of the statistic corresponding to the configuration of the observations in the ordered set of statistics determines significance. In our example this procedure is the same as keeping the values of y fixed and permuting the pairs of values of x_1 and x_2. But we have some additional partial information, knowing which values of y go with each site. The permutation is of known groups of y's over sets of x configurations.

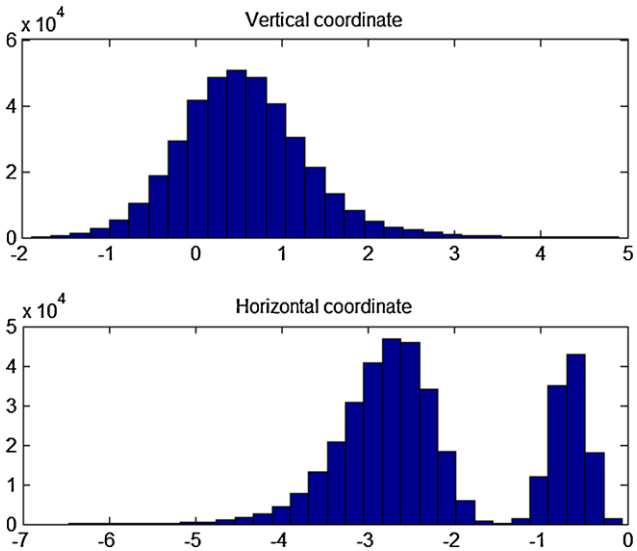

Fig. 6 Two variable regression. Histograms of values of t-statistics from 387,072 possible configurations. *Upper panel* x_2, *lower panel* x_1

5 Interesting configurations

As we have shown that x_2 is not significant, we continue our analysis with a focus on x_1. In particular, we want to discover what feature of the data leads to the bimodal distribution in Fig. 6.

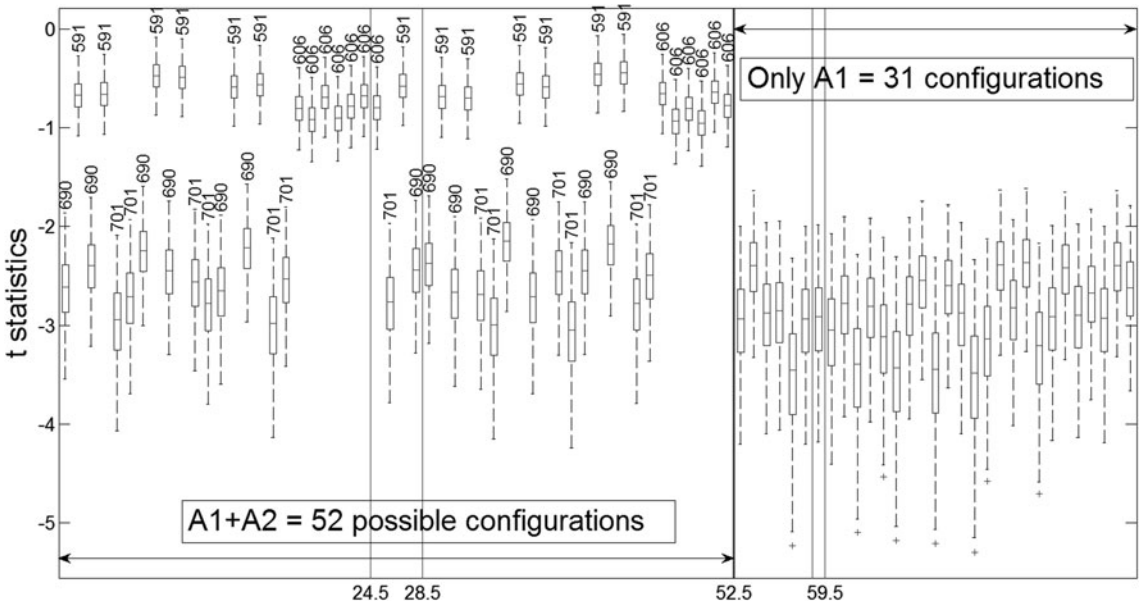

Fig. 7 Distribution of t-statistics for each longitudinal configuration for Arizona. The first 52 configurations are associated with the assumption that Arizona dated both A1 and A2. The remaining 31 configurations are associated with the assumption that Arizona only dated A1. The first 24 boxplots (to the left of the line labelled 24.5) come from the layout (a) or (b) of Fig. 2. Boxplots 25–28 and 29–52 are associated respectively with (c) and (d). Finally, boxplots 53–58, 59 and 60–83 are associated respectively with (e), (f) and (g). The labels on top of the first 52 boxplots denote the y value associated with $x_1 = 41$

If we consider the projections of the 387,072 configurations onto the longitudinal axis, we obtain 42,081 possibilities. For instance, as shown in Fig. 2, Arizona has the two most different sets of configurations. In the lower part of the figure, there are $4!/2!2! = 6$ distinct ways of allocating the four values of y to distinct values of x_1 in the left-hand arrangement, one for the central arrangement and $4! = 24$ for the right hand arrangement, making 31 in all. For the upper panel there are 52 possibilities, making 83 in total for Arizona. The other sites have 13 for Oxford, 13 for Zurich1 and 3 for Zurich2.

For each of the 83 configurations for Arizona there are 507 ($13 \times 13 \times 3$) different ways to obtain configurations for Oxford or Zurich. Figure 7 presents boxplots of the t-statistics for regression only on x_1 divided according to these 83 configurations. In Fig. 7 each boxplot is formed from the 507 values of the t-statistic for each Arizona configuration. We see two sets of values of boxplots, divided, due to the labelling, into two groups each. This structure is very clear in Fig. 8 which gives histograms of these values divided according to the value of y at $x_1 = 41$. In effect, since x_2 is not significant, we are splitting out the statistics in the bimodal bottom panel of Fig. 6. One of the sets of values in Fig. 8 centres around -0.6, the other centres around -2.8. In fact, the value -1.5 completely separates the two sets.

The ordered responses for Arizona are 591, 606, 690 and 701. Among configurations which assume that Arizona dated both A1 and A2 (see Fig. 1) the 13 which associate $y = 591$ with $x_1 = 41$ have in general, as Fig. 8 shows, the smallest absolute values of the t-statistic. The 13 configurations which associate $y = 606$ with 41, which we call 41–606, have slightly larger absolute values of the statistic, but the values are again non-significant. For all the other configurations the t-statistic is significant; there is evidence of a relationship, with a negative slope, between age and position.

It is clear that inference about the slope of the relationship depends critically on whether A2 was analysed and so on which value of y, if any, is associated with $x_1 = 41$. We now analyse the data structure, taking a typical member inside the 507 members of 41–591 and of 41–690 and look at some simple diagnostic plots.

To determine whether the proposed data configuration 41–591 is plausible we look at residuals from the fitted regression model. To overcome the potential problem of masking (when one outlier can cause another to be hidden) we use a forward search (Atkinson and Riani 2000) in which subsets of m carefully chosen observations are used to fit the regression model and see what happens as m increases from 2 to 12. The left-hand panel of Fig. 9 shows a forward plot of the residuals of all observations, scaled by the estimate of σ at the end of the search, that is when all 12 observations are used in fitting. The plot shows the pattern typical of a single outlier, here 41–591 which is distant from all the other observations until $m = n$, when it affects the fitted model. The residuals for the other 11 observations are relatively stable. The right-hand panel of the figure gives the scaled least

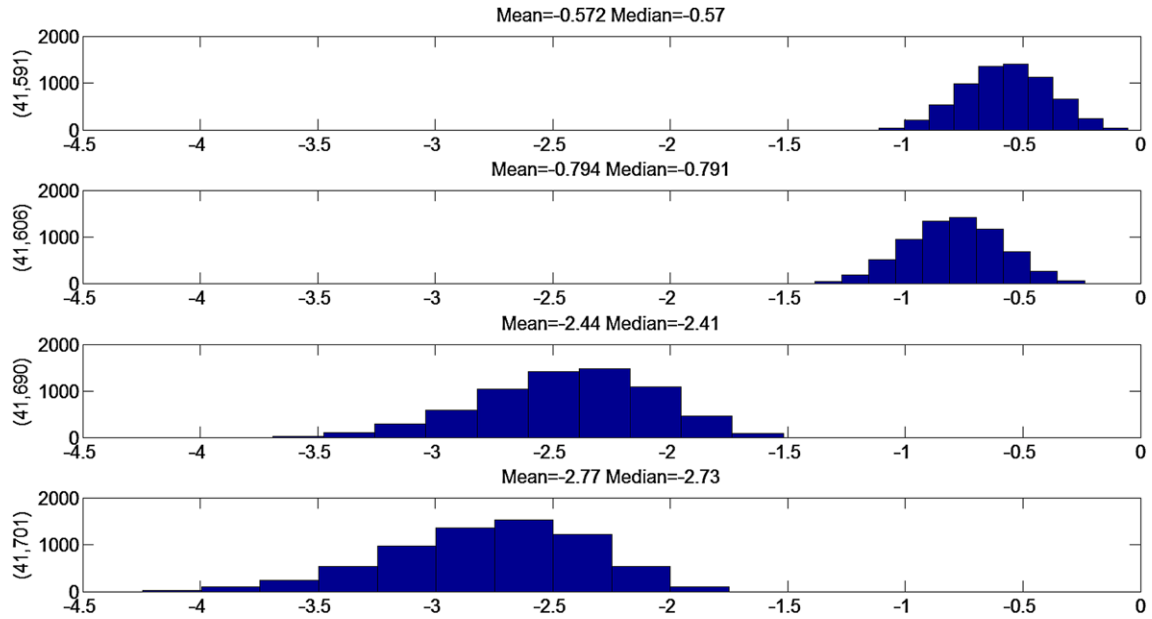

Fig. 8 Histograms of values of t-statistics from Fig. 7 divided according to the value of y from Arizona associated with $x_1 = 41$; reading down, $y = 591$, 606, 690 and 701

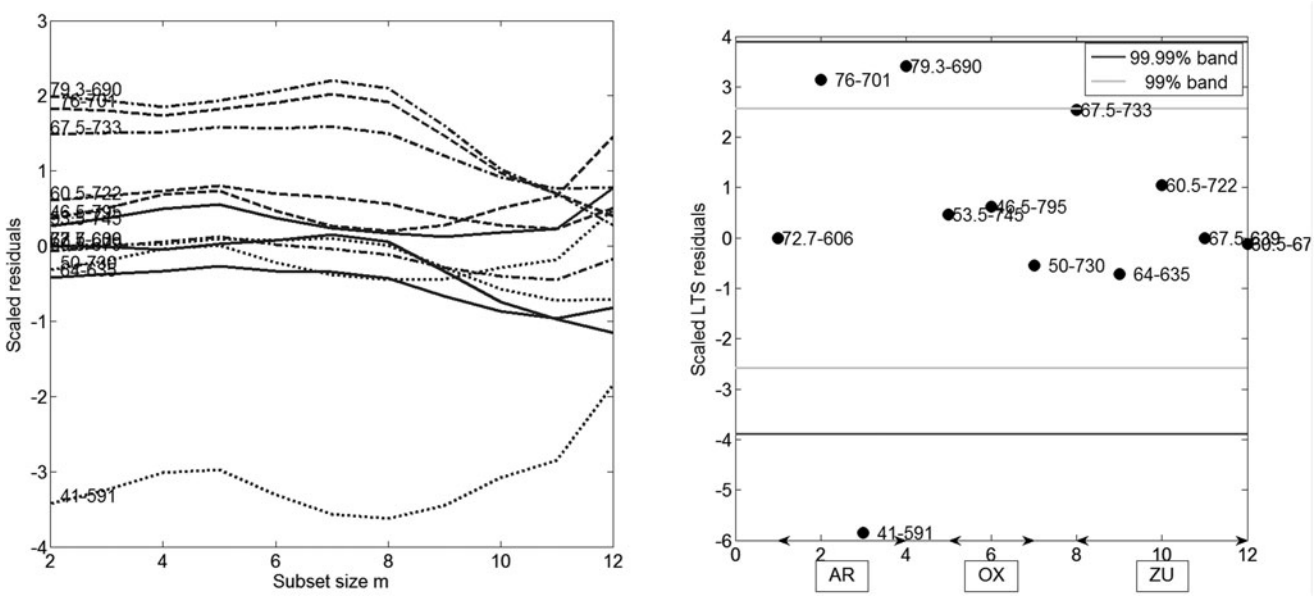

Fig. 9 Analysis of residuals for one configuration in 41–591, that is when $y_{x_1=41} = 591$. *Left-hand panel*, forward plot of scaled residuals showing that this assignment produces an outlier. *Right-hand panel*, plot of LTS residuals

trimmed squares (LTS) residuals against observation number (Rousseeuw 1984). Here again the combination 41–591 is outlying.

The configuration 41–591 led to a non-significant slope for the regression line. Figure 10 gives a similar set of plots for the configuration 41–690 which does give a significant negative slope. But, here again, there is a single outlier, the combination 41–690. This observation again lies well below all others in the forward plot of residuals, until $m = n$. The plot of LTS residuals also shows that this observation is remote from the others.

The conclusion from this analysis of the plots is that whether one of the lower y values, 591 or 606, or one of the higher y values, 690 or 701, from Arizona is assigned to $x_1 = 41$, an outlier is generated, indicating an implausible data set. The comparable plots when it is assumed that Arizona only analysed A1, for example Fig. 11, are quite different in structure. There is a stable scatter of residuals

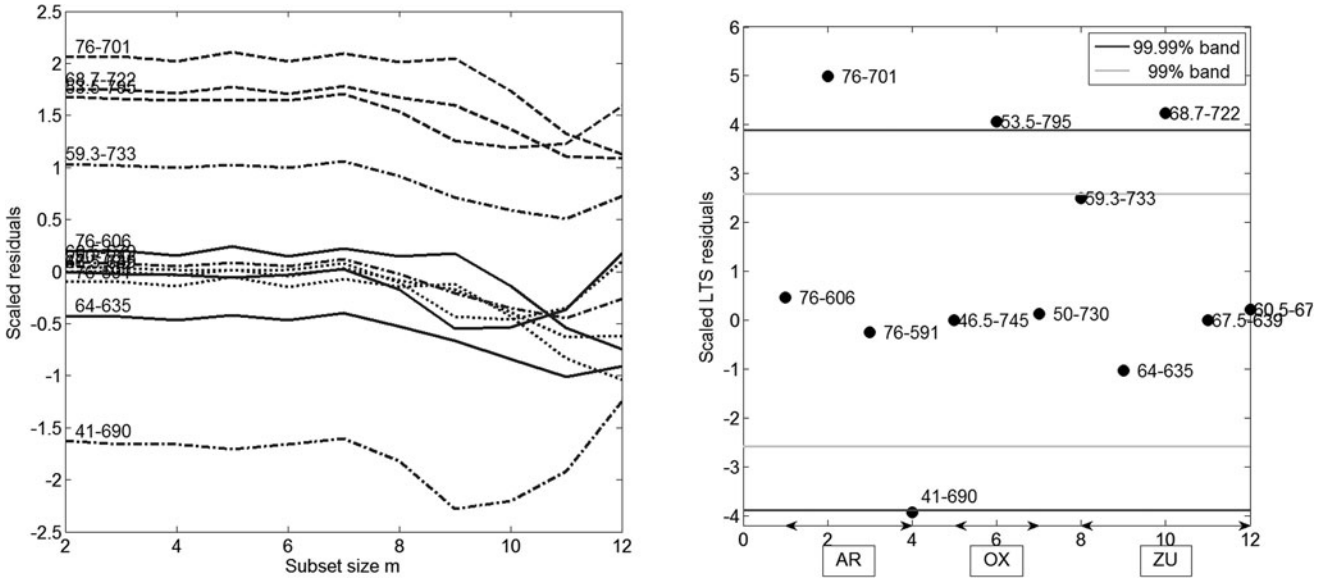

Fig. 10 Analysis of residuals for one configuration from 41–690. *Left-hand panel*, forward plot of scaled residuals showing that this assignment also produces an outlier. *Right-hand panel*, plot of LTS residuals

in the left-hand panel as the forward search progresses, with no especially remote observation. In addition, there are no large LTS residuals.

The broader conclusion of our analysis is that Arizona only analysed A1. We can therefore remove from our analysis all the combinations in which A2 was included. The distribution of the t-statistic under uncertainty about the allocations within A1 and the other sites is that in the right-hand panel of Fig. 7. The resulting histogram of values is similar to those we have already seen, such as the lower panel of Fig. 8. As a consequence there is evidence of a trend in the age of the sample with the value of x_1. The significance of this value does not depend strongly on the spatial allocation of samples within sites.

6 Discussion

Our analysis is of the radiocarbon dates which come directly from the assay of the samples. It is customary, as we have done, to treat these dates as being normally distributed (for example Buck and Blackwell 2004) so that normal theory hypothesis tests and regression analyses apply in this scale as they would in many analyses of data from physical or chemical measurements.

In our simple normal-theory model (1) there is evidence that not all means μ_i are equal. The alternative we consider is that there is a smooth trend which we approximate by a simple regression model. Justification for this physical model comes from Freer-Waters and Jull (2010) who comment that the various pretreatments used by Damon et al. (1989) ensure that the dates are not so sensitive as to fluctuate over small distances due, for example, to handling. There is also no evidence of any patching in this part of the TS which might cause a jump in dating. Debris of the kind we mentioned in Sect. 1, that had built up over the years, would have been removed. An alternative approach would be to consider that the errors were correlated. However, the errors in observation arise from the measurement process and so do not have a spatial component.

Freer-Waters and Jull report a photomicrographic investigation of the sample analysed by Arizona and conclude, partly from the structure of the fibres, that the sample studied by Arizona came from the main part of the shroud. After we had completed our analysis we received a personal communication from Prof. Jull of the University of Arizona confirming that they did indeed only analyse A1. This finding provides a nice vindication of our methodology.

The next stage in a standard analysis is conversion to calibrated years BP. Unfortunately the concentration of atmospheric carbon 14 fluctuates and the curve for conversion to calibrated years, whilst basically straight, shows series of local maxima and minima: see Fig. 3 of Damon et al. (1989) or Fig. A13 of Reimer et al. (2004). Christen and Sergio Perez (2009) provide a robust Bayesian analysis of calibrated years which can allow for different accuracies at the different laboratories. Ramsey (2009) considers a variety of errors that can occur in radiocarbon dating and describes a program for outlier detection that draws on the ideas of Christen. Importantly, these methods of outlier detection assume random samples with a fixed mean. If spatial regression is present, observations with extreme values of x will tend to have extreme values of y. These observations will

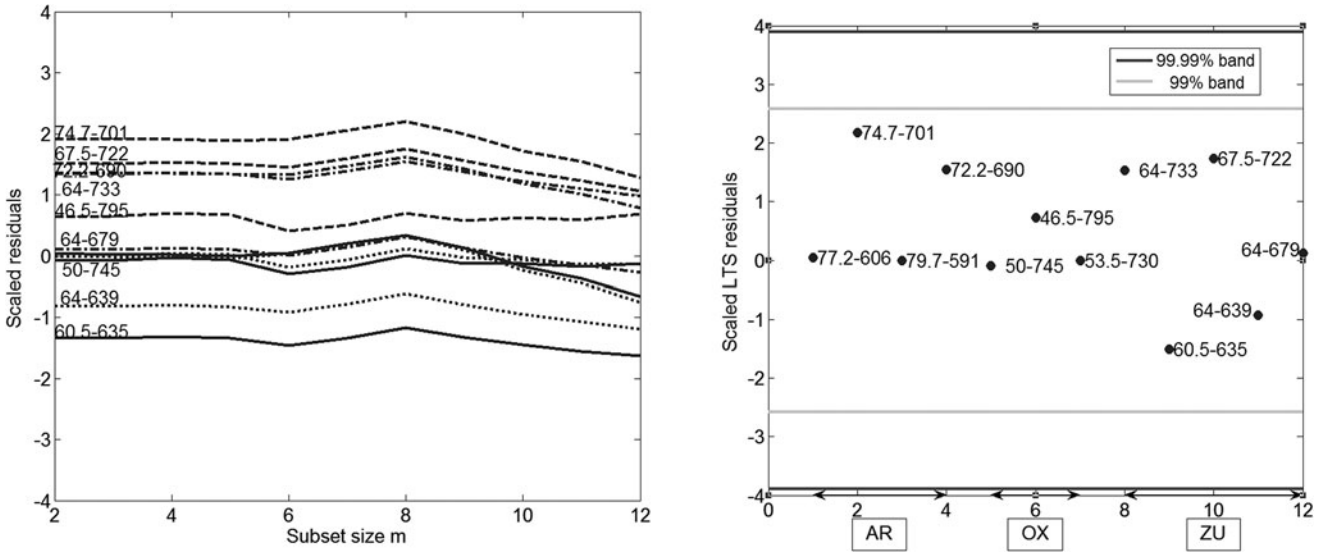

Fig. 11 Analysis of residuals for a typical configuration when all Arizona subsamples come from A1. Unlike in Figs. 9 and 10 there is now no indication of the existence of outliers

then be downweighted or identified as outliers, when the evidence for regression will be reduced or lost. However, it is not our purpose to discuss the details of dating. Our purpose is to use these data to illustrate a statistical procedure for regression analysis with partially labelled regressors which relies on simulation based inference and on graphical methods, including the forward search, for outlier detection.

One further statistical point is raised by Freer-Waters and Jull (2010) who comment on the way in which the sample was taken. It is always possible to argue that the material of the chosen corner was, in some way, different from that of the rest. This defect could have been avoided by more care in the design of the sampling experiment.

Suppose that samples could be taken anywhere in the TS. If the TS has been contaminated, the purpose of the measurements would be to establish that at least some parts of the material are old. A space-filling design is then appropriate such as are used in computer experiments (Sacks et al. 1989). In two dimensions the 'Latin-hypercube' designs would be generated by conceptually dividing the TS into n rows and n columns, creating n^2 potential experimental units. A set of n units is then chosen for experimentation and assigned to one of the laboratories. Spatial cover is achieved by choosing the units such that there is one in each row and column. The units can be chosen at random, and any seemingly unsatisfactory pattern, such as one that is spatially too regular, rejected. Alternatively, sampling can be only from a set of units which have some desirable spatial property, a procedure preferred by Bailey and Nelson (2003). The design of spatial experiments is given book-length treatment by Müller (2007).

Until less intrusive methods of age assessment are developed, samples will presumably be confined to the edges of the TS. However, in any further sampling, care should be taken to avoid confounding between location and laboratory that is an unfortunate aspect of the current data. Our permutation-based regression analysis has been able to explain the observed heterogeneity of the data without the introduction of laboratory effects. In this we are consistent with the analyses of the other fabrics, where such heterogeneity is absent. The values of the t-tests in the right-hand part of Fig. 7 show that the significance of this regression does not depend on the particular permutation chosen. Further, the residual plot in Fig. 11 shows that the residuals for the three readings from Oxford (observations 5, 6 and 7) are not in any way anomalous.

The presence of this trend explains the difference in means that was detected by Damon et al. (1989) and in our Table 1. The effect is that of a decrease in radiocarbon age BP as x_1 increases. Our results indicate that, for whatever reasons, the structure of the TS is more complicated than that of the three fabrics with which it was compared.

Appendix: Weighted and unweighted analyses

The data suggest three possibilities for the weights v_{ij} in (1):

1. Unweighted Analysis. Standard analysis of variance: all $v_{ij} = 1$.

2. Original weights. We weight all observations by $1/v_{ij}$, where the v_{ij} are given in Table 1. That is, we perform an analysis of variance using responses $z_{ij} = y_{ij}/v_{ij}$. If these v_{ij} are correct, in (1) $\sigma = 1$ and the total within groups sum of squares in the analysis of variance is distributed as χ^2 on 9 degrees of freedom, with the expected mean squared error being equal to one.

Table 2 Four fabric types: significance levels of tests of homogeneity of variances and means for unweighted and weighted analyses. The modification to the weights for Arizona uses (2) individually for each of the four fabric types

	Unweighted	Original weights	Modified weights
Shroud			
Variance Homogeneity	0.787	0.354	0.700
Difference in Means	0.0400	0.0043	0.0497
Islamic/Christian linen			
Variance Homogeneity	0.656	0.376	0.868
Difference in Means	0.8536	0.387	0.020
Egyptian mummy			
Variance Homogeneity	0.095	0.015	0.020
Difference in Means	0.712	0.126	0[a]
Cope from Var			
Variance Homogeneity	0.523	0.082	0.495
Difference in Means	0.384	0.081	0[b]

[a] 2.10×10^{-4}

[b] 2.67×10^{-4}

3. Modified weights. The v_{ij} for the TS from Arizona in Table 1 are very roughly 2/3 of those for the other sites. The text above Table 1 of Damon et al. (1989) indicates that the weights for Arizona include only two of the three additive sources of random error in the observations. Table 2 of their paper gives standard deviations for the mean observation at each site calculated to include all three sources. In terms of the v_{ij} the standard deviations of the means are

$$\text{s.d. mean}(i) = \frac{1}{n_i}\left(\sum_{j=1}^{n_i} v_{ij}^2\right)^{0.5}. \tag{2}$$

These two sets of standard deviations are also given in Table 1. Agreement with (2) is good for Oxford, and better for Zurich. However, for Arizona the ratio of the variances is 3.13. We accordingly modify the standard deviations for the individual observations for Arizona in Table 1 by multiplying by $1.77 = \sqrt{3.13}$, when the values become 53, 62, 73 and 58. The three laboratories thus appear to be of comparable accuracy, a hypothesis we now test.

We used these three forms of data to check the homogeneity of variance and the homogeneity of the means. A summary of the results for the TS is in the first two lines of Table 2.

The first line of the table gives the significance levels for the three modified likelihood ratio tests of homogeneity of variance across laboratories (Box 1953). In no case is there any evidence of non-homogeneous variance, that is whether z_{ij} is unweighted, or calculated using either set of v_{ij}, the variances across the three sites seem similar. Of course, any test for comparing three variances calculated from 12 observations is likely to have low power.

We now turn to the analysis of variance for the means of the readings. If the weights v_{ij} are correct, it follows from (1) that the error mean squares for the two weighted analyses should equal one. In fact, the values are 4.18 and 2.38. The indication is that the calculations for the three components of error leading to the standard deviations v_{ij} fail to capture all the sources of variation that are present in the measurements.

The significance levels of the F tests for differences between the means, on 2 and 9 degrees of freedom, are given in the second line of the table. All three tests are significant at the 5 % level, with that for the original weights having a significance level of 0.0043, one tenth that of the other analyses. This high value is caused by the too-small v_{ij} for Arizona making the weighted observations z_{ij} for this site relatively large. The unweighted analysis gives a significance level of 0.0400, virtually the same as the value of 0.0408 for the chi-squared test quoted by Damon et al. (1989). In calculating their test they remark "it is unlikely the errors quoted by the laboratories for sample 1 fully reflect the overall scatter", a belief strengthened by the value of 2.38 mentioned above for the mean square we calculated.

We repeated the three forms of analysis for homogeneity on the three control samples. The results are also given in Table 2. In calculating the modified weights for Arizona, we used (2) for each fabric. The unweighted analysis does not reveal any inhomogeneity of either mean or variance. However, the analysis with adjusted weights gives significant differences between the means for the three laboratories for all fabrics as well as differences in variance for the mummy sample.

One example of the effect of the weights is that of the analysis at Zurich of the mummy samples for which the values of y_{ij}/v_{ij} are $1984/50 = 39.6800$, $1886/48 = 39.2917$ and $1954/50 = 39.08$. These virtually identical values partially explain the significant values in Table 2 for the weighted analysis of this material. A footnote to the table in *Nature* comments on the physical problems (unravelling of the sample) encountered at Zurich. Since no fabric shows evidence of variance heterogeneity on the original scale, we have focused on an unweighted analysis of the TS data.

References

Abraham, B., Box, G.E.P.: Linear models and spurious observations. Appl. Stat. **27**, 131–138 (1978)

Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer, New York (2000)

Atkinson, A.C., Riani, M., Cerioli, A.: The forward search: theory and data analysis (with discussion). J. Korean Stat. Soc. **39**, 117–134 (2010). doi:10.1016/j.jkss.2010.02.007

Bailey, R.A., Nelson, P.R.: Hadamard randomization: a valid restriction of random permuted blocks. Biom. J. **45**, 554–560 (2003)

Ballabio, G.: (2006). New statistical analysis of the radiocarbon dating of the Shroud of Turin. Unpublished manuscript. See http://www.shroud.com/pdfs/doclist.pdf

Box, G.E.P.: Non-normality and tests on variances. Biometrika **40**, 318–335 (1953)

Box, G.E.P., Hunter, W.G., Hunter, J.S.: Statistics for Experimenters. Wiley, New York (1978)

Buck, C.E., Blackwell, P.G.: Formal statistical models for estimating radiocarbon calibration curves. Radiocarbon **46**, 1093–1102 (2004)

Christen, J.A.: Summarizing a set of radiocarbon determinations: a robust approach. Appl. Stat. **43**, 489–503 (1994)

Christen, J.A., Sergio Perez, E.: A new robust statistical model for radiocarbon data. Radiocarbon **51**, 1047–1059 (2009)

Damon, P.E., Donahue, D.J., Gore, B.H., et al.: Radio carbon dating of the Shroud of Turin. Nature **337**, 611–615 (1989)

Fanti, G., Botella, J.A., Di Lazzaro, P., Heimburger, T., Schneider, R., Svensson, N.: Microscopic and macroscopic characteristics of the Shroud of Turin image superficiality. J. Imaging Sci. Technol. **54**, 040201 (2010)

Freer-Waters, R.A., Jull, A.J.T.: Investigating a dated piece of the Shroud of Turin. Radiocarbon **52**, 1521–1527 (2010)

Müller, W.G.: Collecting Spatial Data, 3rd edn. Springer, Berlin (2007)

Ramsey, C.B.: Dealing with outliers and offsets in radiocarbon data. Radiocarbon **51**, 1023–1045 (2009)

Reimer, P.J., Baillie, M.G.L., Bard, E., et al.: INTCAL04 terrestrial radiocarbon age calibration. Radiocarbon **46**, 1029–1058 (2004)

Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**, 871–880 (1984)

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Stat. Sci. **4**, 409–435 (1989)

Walsh, B.: The 1988 Shroud of Turin radiocarbon tests reconsidered. In: Walsh, B. (ed.) Proceedings of the 1999 Shroud of Turin International Research Conference, Richmond, Virginia, USA, pp. 326–342. Magisterium Press, Glen Allen (1999)