# Carbon Dating of the Shroud of Turin: Partially Labelled Regressors and the Design of Experiments

Marco Riani[*], Anthony C. Atkinson[†], Giulio Fanti[‡], Fabio Crosilla[§]

May 4, 2010

**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 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 indicate how better experimental design might have enabled stronger conclusions to have been drawn from this multi-centre experiment.

*Keywords:* computer-intensive method, D-optimality, forward search, robust statistics, simulation envelope,

## 1 Introduction

The Turin Shroud (TS) is a 4.4 m long and 1.1 m wide linen cloth showing the body front and back images of a scourged, thorn-crowned man. There are

[*] Dipartimento di Economia, Università di Parma, Italy, e-mail: mriani@unipr.it

[†] Department of Statistics, London School of Economics, London WC2A 2AE, UK, e-mail: a.c.atkinson@lse.ac.uk.

[‡] Department of Mechanical Engineering University of Padua, Italy, e-mail: giulio.fanti@unipd.it.

[§] Department of Geo-Resources and Territory, University of Udine, Italy e-mail: crosilla@dgt.uniud.it.

also many marks caused by human blood, fire, water and folding of the cloth that partially obscure the indelible double body image which first revealed itself in a negative photographic image. A tradition believes the TS to be the burial cloth Christ was enveloped in when placed in a tomb in Palestine about 2000 years ago. From a scientific point of view, many clues in favour of authenticity have been detected. For example, the formation mechanism of the body images has not yet been scientifically explained; the body image is extremely superficial in the sense that only the external layer of the topmost linen fibres are coloured (Fanti et al. 2010). Despite such indications, the results of a 1988 radiocarbon dating (Damon et al. 1989) were published as "conclusive", stating that the linen fabric dates from between 1260 and 1390 AD, with a confidence level of 95%. This scientific result agrees with the undisputed historical records of the existence of the TS, which go back to AD 1357.

However, after publication of the 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. An extensive survey of this literature is given by Ballabio (2006). 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) combined with computer power and a liberal use of graphics, can help to shed new light on results that are a source of scientific controversy.

The samples for radio carbon dating were taken from a 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 Figure 1). These samples were divided into a total of 12 sub-samples for which datings were made. A chi-squared statistic calculated by Damon et al. (1989) indicated a lack of homogeneity in the observations. It is the modelling of this inhomogeneity with which we are concerned.

The locations of the three samples in the strip are known. On the assumption that the four readings from Arizona all came from the large sample (A1 in Figure 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. However, there is a 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) at-

Table 1: TS. Estimated ages of the individual samples (years before 1950) with calculated standard deviations. Those for Arizona exclude one source of error. 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 | | | | | 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 |

tempted 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 - our analysis gives 387,072 possibilities.

We discuss the evidence for heterogeneity in §2 and compare unweighted analyses with those weighted by the reported accuracies of the three laboratories. The possible spatial layouts of the subsamples are described in §3. We proceed in §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 §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 Figure 1); the contrary assumption leads to the generation of gross regression outliers. Much of the uncertainty we have exhibited in the data comes from a poorly designed experiment. We accordingly conclude in §6 with a discussion of how the application of statistical principles would have produced a design leading to sharper conclusions.

## 2 Heterogeneity

Table 1, taken from Table 1 of Damon et al. (1989), gives the estimated age, in years BP, that is before the present which is 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 they initially misled us).

The continuous observations are all far from zero taking the standard deviations into account and do not have a wide range. It is then natural to consider a normal theory linear model. If we ignore any spatial factors, a general model for observation $j$ at site $i$ is

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

where the errors $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$. Our central concern is the structure of the $\mu_i$, at this point whether they are all equal. However, to test this hypothesis we need to establish the error structure. The data suggest three possibilities:

**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 $\chi^2$ on 9 degrees of freedom, with the expected mean squared error being equal to one.

**3. Modified weights.** The $v_{ij}$ in Table 1 for Arizona are very roughly 2/3 of those for the other sites. Part of the caption in Table 1 of Damon et al. (1989) indicates that the weights for Arizona include only two of the three sources of error. 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 deviation 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 in Table 1 by multiplying by 1.77, when the values for Arizona 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 (Box 1953). In no case

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 is individual for each fabric sample

| | 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* |
| Cope from Var | | | |
| Variance Homogeneity | 0.523 | 0.082 | 0.495 |
| Difference in Means | 0.384 | 0.081 | 0† |

Footnote: * $2.10 \times 10^{-4}$; † $2.67 \times 10^{-4}$

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. The error mean squares for the two weighted analyses are 4.18 and 2.38, far from the expected values of one. The indication is that the calculations 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 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, ten times 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.

In addition to the TS, each laboratory dated three controls: linen from a Nubian tomb with Islamic patterns and Christian ink inscription; an Egyptian mummy from Thebes and threads from a cope from Var, France. None of the datings of these samples was controversial. Accordingly, we repeated the three forms of analysis for homogeneity on the other three fabric samples. The results are also given in Table 2. In calculating the modified weights for Arizona, we used (2) for each fabric. What is noticeable about the results of the table is that 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, 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 in obtaining these values. Since there is no evidence of variance heterogeneity on the original scale, for any fabric, we focus on an unweighted analysis of the TS data. Furthermore, there is no evidence of heterogeneity in the means of the three control samples and so no evidence of systematic differences between laboratories. Accordingly, we now 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 5mm portion was trimmed, thus removing the stitching and remnants. This process resulted in a piece of fabric that measured 81mm x 16mm.

This piece of TS fabric was then divided into two parts; one, called “Riserva”, remained in Turin for future analyses (see Figure 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 Figure 1, the material between Zurich and A2.

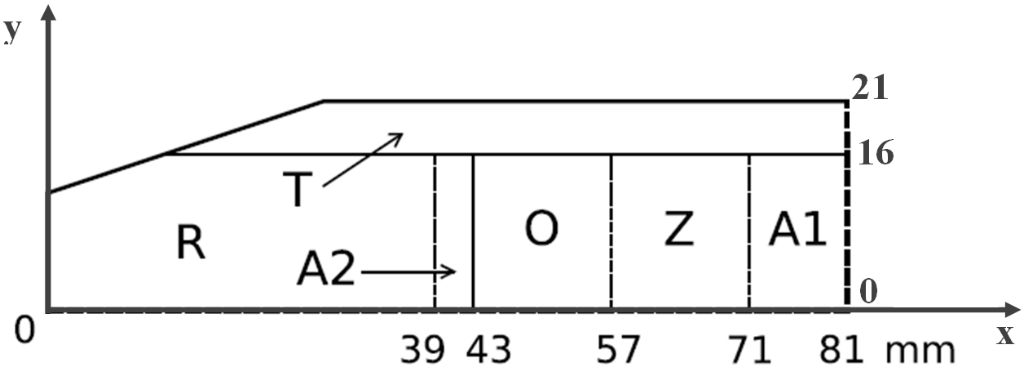

Figure 1: Diagram showing the piece removed from the TS and how it was partitioned. T: trimmed strip. R: retained part called "Riserva" (initially including also A2). O, Z, A1, A2: subsamples given to Oxford, Zurich, and Arizona (two parts) respectively.

We know from Damon et al. (1989) that four different pieces were dated by Arizona. The possible configurations are therefore those shown in Figure 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 Figure 3. Oxford instead divided its piece of TS fabric into only three parts with the possible configurations as shown in Figure 4.

Obviously Figures 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 $x_1$ (horizontal) and $x_2$ (vertical) 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 Figure 5 plots, as a continuous line, the ordered

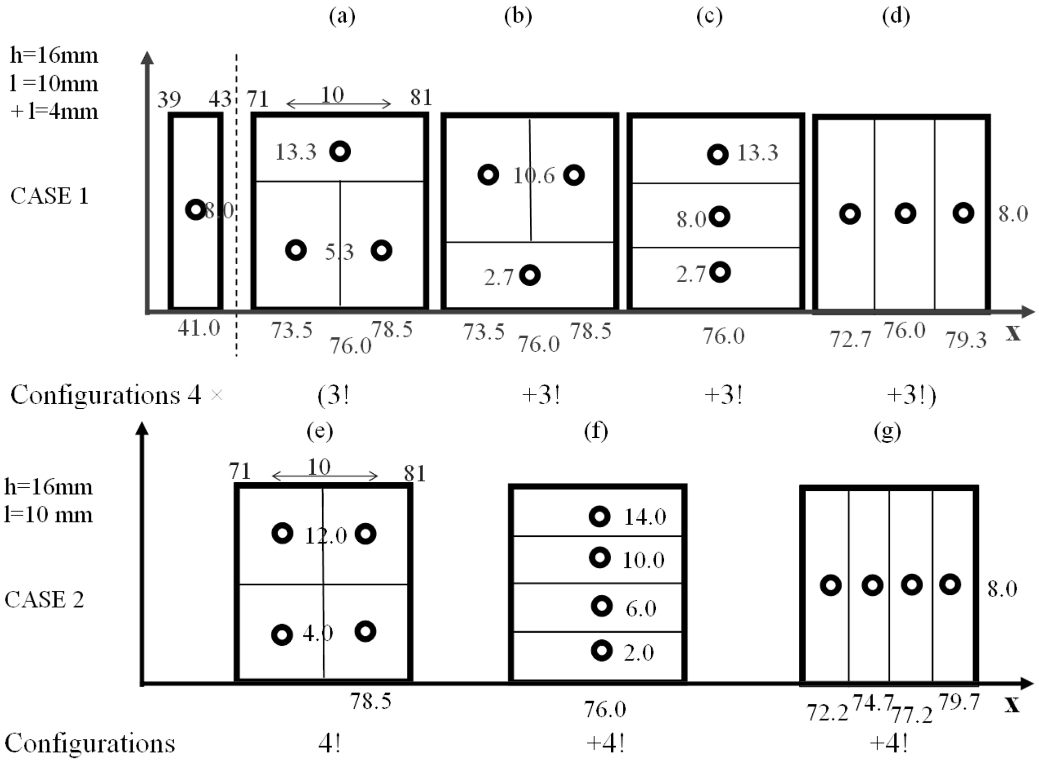

Figure 2: Arrangements investigated for the Arizona sample. The image on top assumes that Arizona dated both pieces (A1 and A2). The image at the bottom assumes that Arizona only dated piece A1. Total number of cases considered is 168 = 96+72.

significance level of the $t$-test for $x_2$ in the model with both variables. 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 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 $x_2$, although the shape of the curve generated by the data merits investigation.

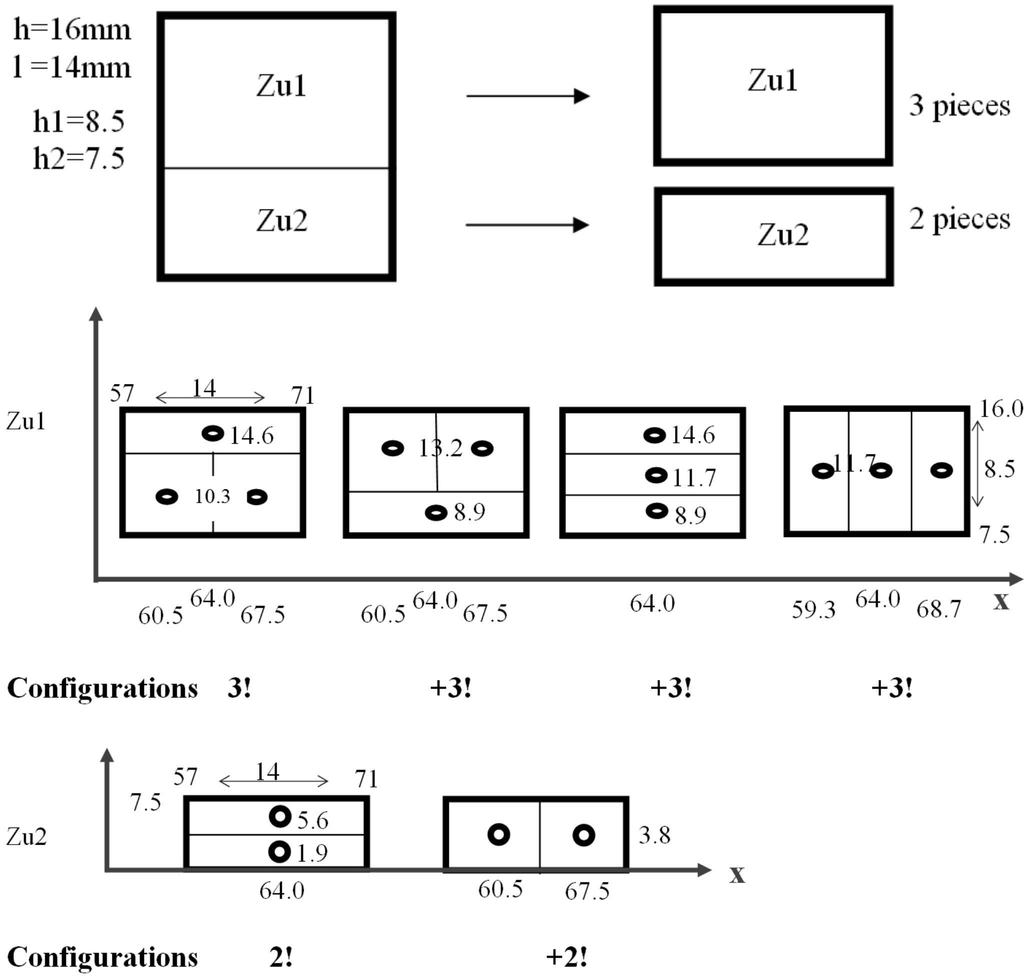

Figure 3: Arrangements investigated for the Zurich sample. Total number of cases considered is $96 = 24 \times 4$

Histograms of the statistics help. The top panel of Figure 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 Figure 6, the $t$-statistic for $x_1$, is however quite different, showing two peaks. The larger peak is centred around $-2.9$ whereas the thinner peak is centred around $-1$. 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 horizontal coordinate.

Although our procedure involves permutations of data, these analyses are not those associated with permutation tests. In a permutation test (for example Box, Hunter, and Hunter 1978, §4.1) the values of $x_1$ and $x_2$ are kept fixed, the observations $y$ being permuted over the design points and a

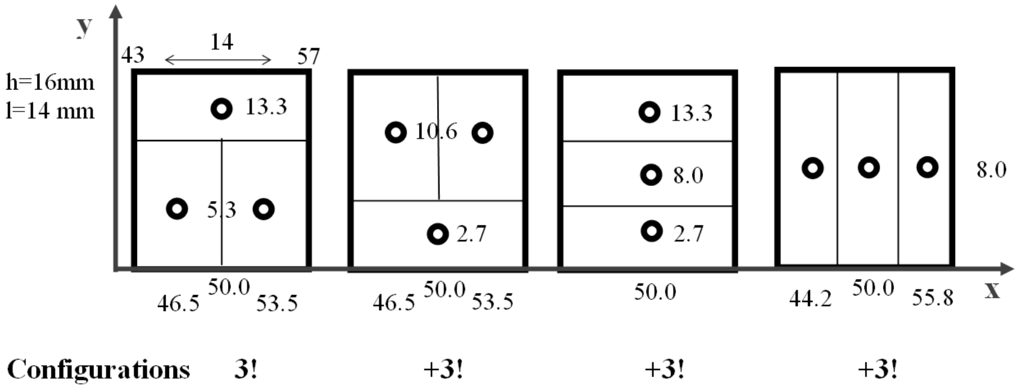

Figure 4: Arrangements investigated for the Oxford sample. Total number of cases considered is 24 = 3! × 4

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. Here 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 in our case we have some 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.

# 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 Figure 6.

If we consider the horizontal projections of the 387,072 configurations we obtain 42,081 possibilities. For instance, as shown in Figure 2, Arizona has the two most different sets of configurations. In A2, 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, A1, 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 another configuration 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 Figure 7 each boxplot is formed from the 507

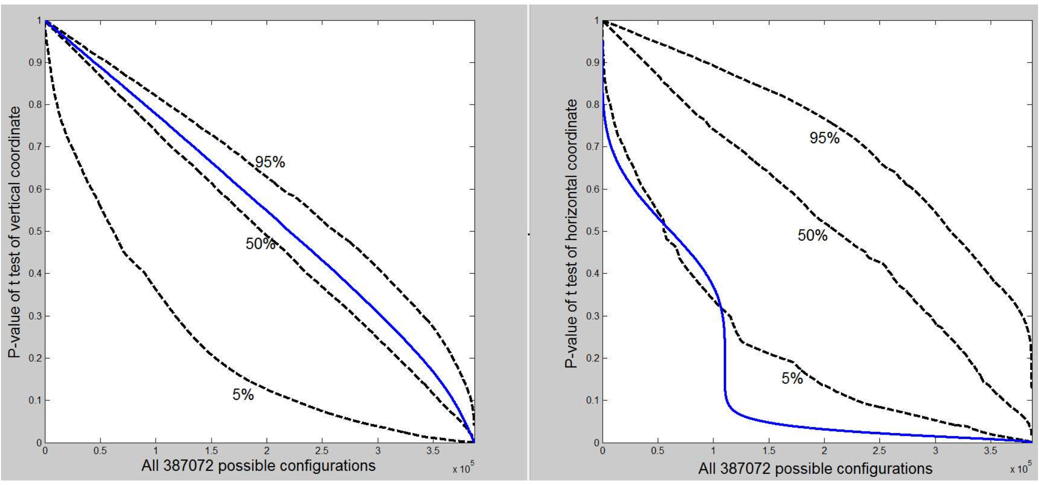

Figure 5: Two variable regression. Significance levels of $t$-statistic from 387,072 possible configurations and envelopes from 100 simulations of each configuration. Left-hand panel $x_2$, right-hand panel $x_1$

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 Figure 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 Figure 6. One of the sets of values in Figure 8 centres around $-1$, the other centres around $-2.9$. 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 Figure 1) the 13 which associate $y = 591$ with $x_1 = 41$ have in general, as Figure 8 shows, the smallest absolute values of the $t$-statistic. The 13 configurations which associate $y = 606$ with 41 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 on which value of $y$ 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 model regression model. To overcome

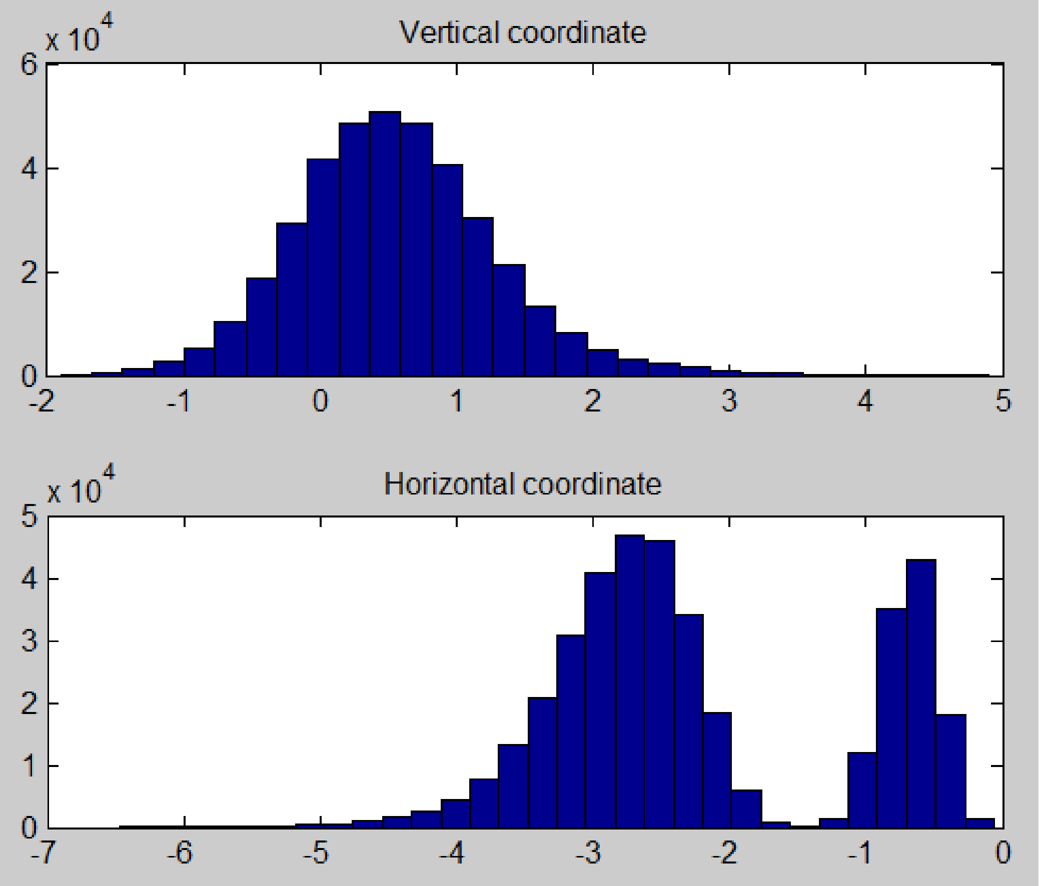

Figure 6: Two variable regression. Histograms of values of $t$-statistics from 387,072 possible configurations. Upper panel $x_2$, lower panel $x_1$

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 Figure 9 shows a forward plot of the residuals of all observations, scaled by the estimate of $\sigma$ 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 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

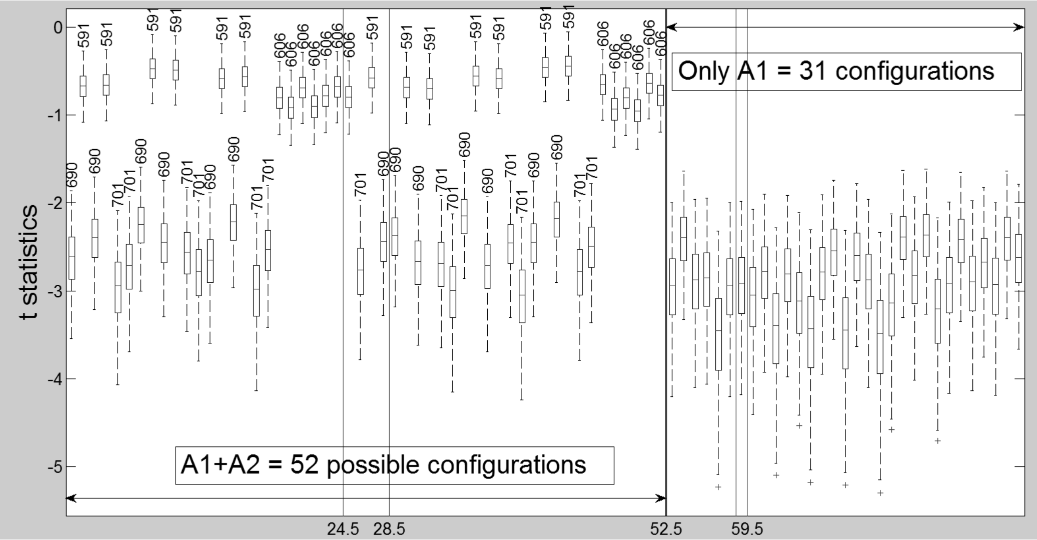

Figure 7: Distribution of $t$-statistics for each horizontal configuration for Arizona. The first 52 configurations are associated with the assumption that Arizona dated both A1 and A2. The remaining 31 configurations are associated with the assumption that Arizona only dated A1. The first 24 boxplots (to the left of the line labelled 24.5) come from the layout (a) or (b) of Figure 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 $x_1 = 41$

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 Figure 11, are quite different in structure. There is a stable scatter of 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

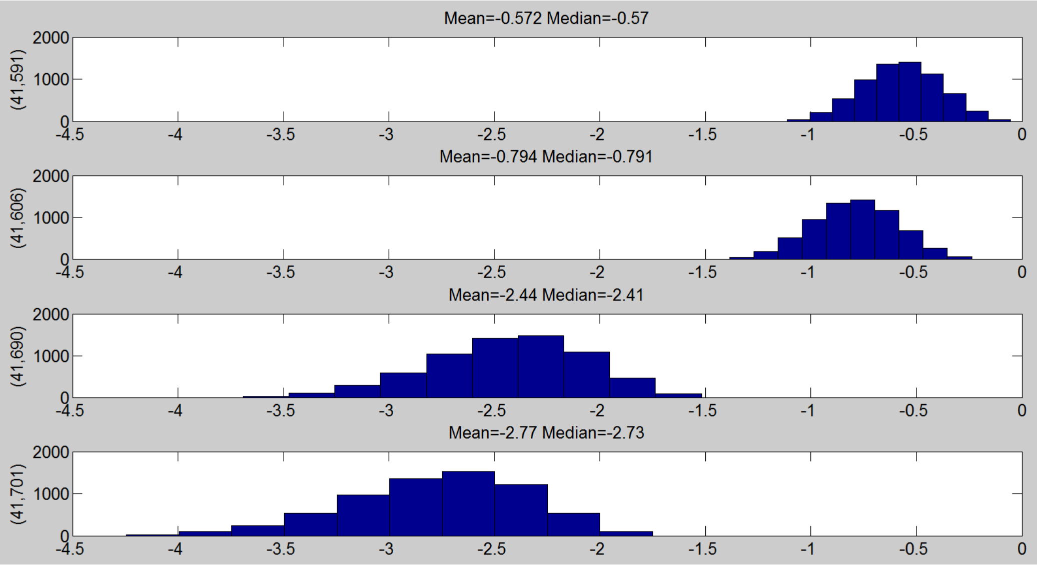

Figure 8: Histograms of values of $t$-statistics from Figure 7 divided according to the value of $y$ from Arizona associated with $x_1 = 41$, reading down, $y = 591, 606, 690$ and 701

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 Figure 7. The resulting histogram of values is similar to those we have already seen, such as the lower panel of Figure 8. As a consequence there is a 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.

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 age BP as $x_1$ increases. The effect is not large over the sampled region; between x1 = 43 and 81, our estimate of the change is about two centuries. Extrapolation of this linear trend to unsampled values of $x_1$ eventually leads to meaningless negative results. If we stay within the sampled region, the TS becomes slightly more recent as we move away from the corner. One explanation is that of greater contamination towards the centre of the cloth.

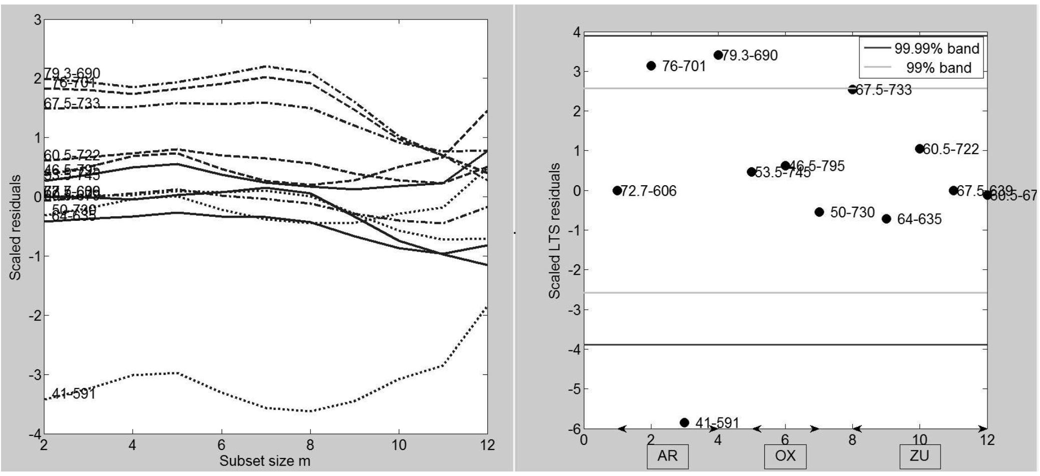

Figure 9: Analysis of residuals for one configuration when $y_{x_1=41} = 591$. Left-hand panel, forward plot of scaled residuals showing that this assignment produces an outlier. Right-hand panel, plot of LTS residuals

# 6 Design and Sampling

A major difficult in coming to conclusions about the age of (most of) the TS is the unrepresentative 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. In this section we briefly consider some better ways of sampling. First we suppose that samples could be taken anywhere in the TS. These ideas are useful when we then restrict ourselves to designs where samples are only taken from the edges.

We have fitted a first-order model in two variables. If interest is in the parameters of this model a D-optimum design would be appropriate. These designs (Fedorov and Hackl 1997; Atkinson, Donev, and Tobias 2007) minimize the volume of the confidence region for the parameters, and so allow for the covariances of the estimates, as well as making the variances small. For the first-order model the D-optimum design indicates an equal number of samples from each corner of the TS. With twelve observations in all, each laboratory would be given a subsample from each corner. An advantage of this design is that an interaction term $\beta_{12}x_1x_2$ could be included in the model.

If it is suspected that there may be curvature in the relationship between age and location a second-order model would be appropriate, that is a model also including terms in $x_1^2$ and $x_2^2$. The $3^2$ factorial, with one observation at each of the nine design points, is an efficient design, as measured by D-

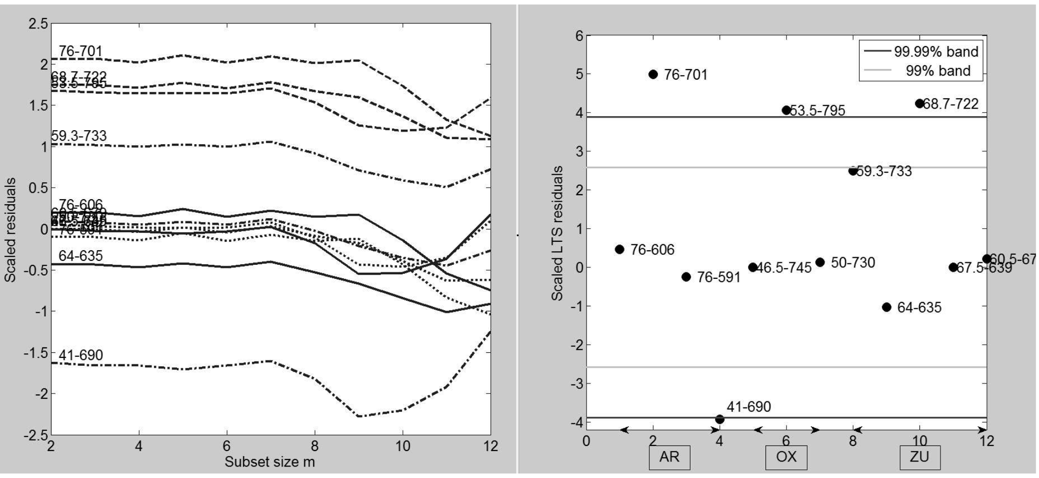

Figure 10: Analysis of residuals for one configuration when $y_{x_1=41} = 690$. Left-hand panel, forward plot of scaled residuals showing that this assignment also produces an outlier. Right-hand panel, plot of LTS residuals

optimality. Slightly more efficient designs, that is giving smaller confidence regions on a per observation basis, are found by unequal replication. A good 13 point design replicates takes two observations at each of the points of the $2^2$ factorial, that is at the corners of the region (Atkinson et al. 2007, Table 12.1). A marginally better design has 14 points, repeating the centre point (Atkinson and Tobias 2008, Table 1) as well as the points of the $2^2$ factorial.

In our context a disadvantage of these designs is that they concentrate sampling on the edges of the TS. In agricultural trials, border rows are sometimes discarded. Here, where contamination is more likely at the edges of the material, the experimental region could be shrunk away from the edges. The response-surface designs of Box and Draper (1963) achieve this automatically by focussing on mean-squared error of prediction when the model may be mis-specified.

These designs follow from the kind of analysis we have undertaken that has involved fitting a model. However, 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. Spatial cover is achieved by choosing the units such that there is one in each row and column. The

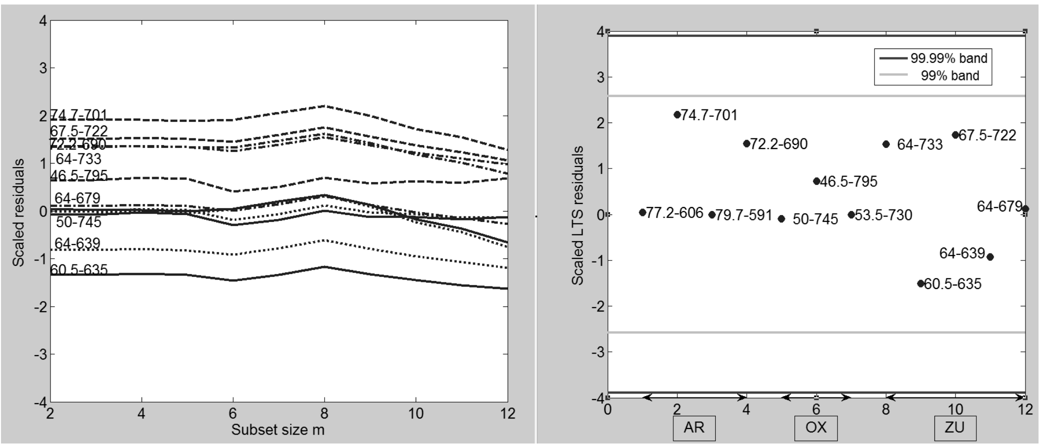

Figure 11: Analysis of residuals for a typical configuration when all Arizona subsamples come from A1. Unlike in Figures 9 and 10 there is now no indication of the existence of outliers

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. In general, Bailey and Nelson (2003) prefer the latter. Once the units for experimentation have been chosen the samples need to be assigned to the three laboratories in a suitably randomized way to, for example, avoid all samples for Arizona coming from one end or side of the material. The design of spatial experiments is given book-length treatment by Müller (2007). An interesting, non-standard, aspect of spatial design for the TS is that sampling could clearly never be from areas with images, but that areas between images are potential sites for samples.

Until less intrusive methods of age assessment are developed, samples will presumably be confined to the edges of the TS. However, the preceding discussion does provide guidance on a suitable design. If $n$ samples are to be taken, the perimeter of the material should be divided in to $n$ intervals of as equal size as possible. Locations are then selected at random within each interval, preferably subject to a restriction on the minimum distance between samples. The intervals might also be chosen to exclude corners of the material, if it is thought that contamination of these regions is more likely.

Suppose $n = 12$. Then 12 independent samples could be taken from the edges and ends. However, a good design should allow some internal estimate of inter-laboratory and random effects. A good possibility is to sample six points and to divide each sample into two subsamples. There are three possible pairs of allocations of the subsamples: AO, AZ and OZ so each can occur twice, the allocation of treatments (pair of sites) being made at random, again guarding against any obvious spatial pattern. The resulting design is a balanced incomplete block design (BIB) with two treatments per block. For a discussion of the desirable properties of such designs, see, for example, Bailey (2008, cap. 11).

The difficulties in interpretation of the results of the age of the TS show forcibly the difficulties that can be caused by inadequate experimental deign. In part, the problem arises because laboratory scientists are trained, perhaps subconsciously, to believe that they have all sources of variation under control. Statisticians, on the other hand, are aware of the possibility of unsuspected sources of heterogeneity (here the "lurking variable" of location) which have to be guarded against by suitable randomization. An illustration of this is the too small values of the $v_{ij}$ above. Further, even if all possible sources are guarded against, careful records need to be taken of any possible other variables. Obviously, the exact location of the subsamples that we have endeavoured to infer, is one example.

In experiments on variable material, such as human patients in a clinical trial, attention is paid to the need for blinded treatment allocation, randomization and adjustment for covariates (Atkinson 2002, Rosenberger and Lachin 2002). It is perhaps surprising, in hindsight, that the material of the TS has turned out to be so susceptible to claims of variability. Our analysis, and those of others, would have been more definitive had greater attention been paid to the statistical aspects of the design of the experiment for the radio carbon dating of the TS.

# 7 Conclusion

Due to the heterogeneity of the data and the evidence of a strong linear trend the twelve measurements of the age of the TS cannot be considered as repeated measurements of a single unknown quantity. The statement of Damon, Donahue, Gore, and eighteen others (1989) that "The results provide conclusive evidence that the linen of the Shroud of Turin is mediaeval" needs to be reconsidered in the light of the evidence produced by our use of robust statistical techniques.

# References

Atkinson, A. C. (2002). The comparison of designs for sequential clinical trials with covariate information. *Journal of the Royal Statistical Society, Series A 165*, 349–373.

Atkinson, A. C., A. N. Donev, and R. D. Tobias (2007). *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.

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

Atkinson, A. C. and R. D. Tobias (2008). Optimal experimental design in chromatography. *Journal of Chromatography A 1177*, 1–11.

Bailey, R. A. (2008). *Design of Comparative Experiments*. Cambridge: Cambridge University Press.

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

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. (1953). Non-normality and tests on variances. *Biometrika 40*, 318–335.

Box, G. E. P. and N. R. Draper (1963). The choice of a second order rotatable design. *Biometrika 50*, 335–352.

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

Damon, P. E., D. J. Donahue, B. H. Gore, and eighteen others (1989). Radiocarbon dating of the Shroud of Turin. *Nature 337*, 611–615.

Fanti, G., J. A. Botella, P. Di Lazzaro, T. Heimburger, R. Schneider, and N. Svensson (2010). Microscopic and macroscopic characteristics of the shroud image superficiality. *Journal of Imaging Science and Technology*. (To appear).

Fedorov, V. V. and P. Hackl (1997). *Model-Oriented Design of Experiments*. Lecture Notes in Statistics 125. New York: Springer Verlag.

Müller, W. G. (2007). *Collecting Spatial Data, 3rd edition*. Berlin: Springer-Verlag.

Rosenberger, W. F. and J. L. Lachin (2002). *Randomization in Clinical Trials: Theory and Practice*. New York: Wiley.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*, 871–880.

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

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