19

A technique for reducing the size of sparse contingency tables

Clive Orton

Paul Tyers

(University College London, Institute of Archaeology, 31–34 Gordon Square, London WC1H 0PY)

19.1 Introduction

The aim of this paper is to present a statistical technique developed to meet a need in the project 'Statistical Analysis of Ceramic Assemblages' (Orton & Tyers 1989), which now forms part of the PIE-SLICE package (Tyers & Orton, this volume). However, it seems to be of more general use, and is therefore presented separately here.

The project has led to the representation of ceramic assemblages as two-way tables of *pseudo-counts* (real numbers that can be treated as integers for statistical analysis). When applied to ceramic data these numbers are called PIES (pottery information equivalents). Comparison of several assemblages becomes the analysis of a three-way table, the variables usually (but not necessarily) being context, fabric and form, using established methods of quasi-log-linear analysis (Bishop *et al* 1975, pp. 177–228). Initial attempts at analysis gave suspiciously good fits to an absurdly simple model (model 1 = independence of context, fabric and form), even when interactions were visually apparent in the data. Inspection of the marginal tables showed the source of the problem: many cells with small expected values contributed very little to the χ^2 or G^2 statistic, but boosted the degrees of freedom considerably. Such tables, which are commonly associated with ceramic assemblages, are too sparse for successful analysis as they stand. The opposite potential danger was also apparent in these tables; small expected values can give erratic and misleading goodness-of-fit statistics.

This is not of course a new result. Conventional theory sets a general minimum 'expected' value of 5 per cell with an absolute minimum of 1 for the chi-squared test to be appropriate (Cochran 1954; Craddock and Flood (1970) suggest a limiting average value of around 1 in the case when the expected values are roughly equal). Two approaches seemed potentially useful:

1. examination of the significance on individual cell values,
2. rules for merging or deleting rows and columns.

19.2 Theory

19.2.1 Individual values

Since individual significant values can be diluted by a background of non-significant values, we have to ask what is the critical value for a single cell, rather than all taken together as in the chi-squared test. This is discussed by (Bishop *et al* 1975, pp. 136–155), but none of their answers seems particularly satisfactory, and they say that there is no single critical value (Bishop *et al* 1975, p. 140).

Having failed to find any theoretical criterion for the significance of individual values, we looked empirically at some 'random' two-way tables. These suggested that for tables larger than say 5 by 5, with an average cell value of 1, individual contributions to chi-squared of up to about 8 could occur frequently enough with random data for the significance of any contributions less than 8 to be in doubt.

This approach did not seem useful in practice, and was abandoned in favour of a systematic procedure for merging rows and/or columns.

19.2.2 Merging rows and/or columns

It was initially decided to adopt the criterion that, for a two-way table, all cells should have an expected value of at least 1.0, and that merging of rows and columns should be carried out with this aim in view. As experience accumulated, it became evident that this criterion was not necessary at this stage, but was best left to the final quasi-log-linear analysis. It was decided simply to merge rows and columns until no further merges seemed statistically reasonable.

19.2.3 An approach based on the chi-squared metric

We define a distance between a pair of rows or of columns, based on the chi-squared metric, i.e.

$$d^2(i_1, i_2) = \sum_j (x_{i_1,j}/x_{i_1.} - x_{i_2,j}/x_{i_2.})^2/(x_{.j}/x_{..}) \tag{19.1}$$

This is the metric used in correspondence analysis (Greenacre 1984) and just as in correspondence analysis the rows and columns can be envisaged as points in the same space. We need two sets of rules — merging rules and stopping rules.

Merging rules

1. calculate the weighted distance between each pair of rows and each pair of columns (see section 19.2.6), i.e.

$$d^2 x_{i_1.} x_{i_2.}/(x_{i_1.} + x_{i_2.}) \tag{19.2}$$

   Find the smallest weighted distance between *either* a pair of rows or a pair of columns,
2. check that the distance between them is not too great to allow them to be merged. This can be done by comparing the weighted distance with a chi-squared statistic at a chosen probability level. The number of degrees of freedom is one less than the number of non-zero columns (if the distance is between a pair

or rows) and *vice versa*. The derivation of this test is given in section 19.2.4.

3. if this condition is satisfied, merge, with the following steps:
   (a) recalculate the data matrix,
   (b) recalculate the *profiles*, i.e. the values $x_{ij}/x_{i.}$ and $x_{.j}/x_{..}$
   (c) recalculate the pairwise distances between the rows and between the columns, noting that if we have merged two rows, only the distances between the new merged row and all other rows, and between all pairs of columns, have to be recalculated. Other distances are unchanged.

Stopping rule

The merging procedure stops when the weighted distances between any pair of candidates are too great for merging.

Discussion

As well as working well in practice, this approach has some useful theoretical properties, e.g.

1. the overall chi-squared statistic is reduced by a relatively small amount at each step, subject to overriding archaeological constraints. Although the reduction at each step is the smallest possible, this seems not to be a global property.
2. the ratio of the current value of chi-squared to the initial value is a measure of the fit of the merged data to the original dataset. It is in fact the ratio of inertia accounted for/total inertia, as encountered in correspondence analysis.

This approach has general applications as a data-reduction technique. It could be used as a way of removing noise from data as a preliminary, or even an alternative, to correspondence analysis. It is here christened 'simultaneous reduction of dimension' (SRD).

19.2.4 The distance between two rows (or columns) as a chi-squared statistic

We start from the geometrical result that the total inertia of the rows, i.e. $\sum_i x_{i.}(i\text{th distance from centroid})^2$ is a chi-squared statistic whose number of degrees of freedom is one less than the number of columns (Greenacre & Hastie 1987). So for just two rows we have:

$$\chi^2 = \sum_i x_{i.} d_i^2 = x_{1.} d_1^2 + x_{2.} d_2^2 \qquad (19.3)$$

where $d_i = i$th distance to centroid and

$$d(1,2) = d_1 + d_2 \qquad (19.4)$$

and

$$x_{1.} d_1 = x_{2.} d_2, \text{ i.e. } d_2 = (x_{1.}/x_{2.}) d_1 \qquad (19.5)$$

Substitute 19.5 in 19.3:

$$\chi^2 = x_{1.} d_1^2 + x_{2.}(x_{1.}/x_{2.})^2 d_1^2 = d_1^2(x_{1.} + x_{1.}^2/x_{2.}) \qquad (19.6)$$

Substitute 19.5 in 19.4:

$$d = d(1,2) = d_1 + (x_{1.}/x_{2.})d_1 = d_1(x_{1.} + x_{2.})/x_{2.} \qquad (19.7)$$

so that $d_1 = dx_{2.}/(x_{1.} + x_{2.})$ Substitute 19.7 in 19.6:

$$\chi^2 = (dx_{2.}/(x_{1.}+x_{2.}))^2(x_{1.}+x_{2.})(x_{1.}/x_{2.}) = d^2 x_{1.} x_{2.}/(x_{1.}+x_{2.}) \qquad (19.8)$$

19.2.5 Pre-treatment of data matrices

In interpreting the outcome of an SRD, we sometimes encounter 'meaningless' merges, in which a very small unit (row or column) merges with another across a long distance simply because its low weight reduces its inertia (chi-squared) to an insignificant level. It seems likely that some units are *so* small that they *must* merge with some other unit. This leads to an inconsistency, in that the very small units survive because they must merge with something else, while larger (but still small) units may be deleted at the end of the process because they failed to merge with anything else. The merging of the very small units is often arbitrary and uninformative, and it seems better in principle to merge or delete them before carrying out SRD.

The problem is to detect such units — i.e. how 'small' is 'very small'? It seems reasonable to say that if a unit is so small that, whatever its actual profile, it could not possibly be significantly different from the overall mean profile, it should be deleted or merged in advance. In practice this means that the program should delete it; if the user wishes to reinstate it he should merge it with another unit for a second run. The critical value for the significant difference of any point of a given weight depends on the geometry of the chi-squared space, i.e. on the number of dimensions k and the mean profile $(x_{.1}, x_{.2}, \ldots, x_{.k})$. If we assume that the space is spheroidal, i.e. all dimensions have equal weight in the mean profile, we can represent this profile as

$$(1/k, 1/k, \ldots, 1/k)$$

and the worst-case profile as $(1, 0, 0, \ldots, 0)$ without loss of generality. Then using equation 19.1 we have

$$\begin{aligned} d_2 &= (1 - 1/k)^2/(1/k) + (1/k)^2/(1/k) + \ldots + (1/k)^2/(1/k), \\ &= k((k-1)/k)^2) + (k-1)(1/k), \\ &= (k-1)^2/k + (k-1)/k = k - 1 \end{aligned}$$

We note further that if the smallest weight in the mean profile is 1/rk, and the rest of the weight is distributed evenly, this result becomes

$$d^2 = rk - 1$$

From equation 19.2, the weighted distance

$$= d^2 x_{1.} x_{..}/(x_{1.} + x_{..}) = d^2/(1/x_{1.} + 1/x_{..}) \approx x_{1.} d^2$$

when $x_{..}$ is 'large'. So

$$\chi^2_{k-1} = (k-1)x_{1.}$$

and the critical value is

$$x_{1.} = \chi^2_{k-1}/(k-1)$$

at the required probability level.

This simple approach assumes that the space is spheroidal, while in practice it is often very far from being so. Since it is the smallest weight in, for example, the mean row profile that determines the critical value for the weight of a row, the critical value for rows is likely to be distorted downwards by the column weights which we would wish to eliminate had we looked at columns first, and vice versa. In other words, the very small values of column weights tend to 'protect' the very small row weights, and vice versa. It therefore seems reasonably to calculate the critical value for rows by omitting the columns which appear to be 'very small', and vice versa.

An alternative approach is to look for weighted distances that are 'significantly small' as well as those that are 'significantly large', e.g. ones which are less than the 95% probability level. In other words, we look for pairs of rows (or columns) that are significantly closer than would be expected if they were really two samples from the same parent distribution. Since the matrix of weighted distances has $k(k-1)/2$ distinct entries, we can expect such a level to be exceeded occasionally, but repeated occurrences would indicate rows (or columns) whose total weight is 'too small'. For example, a row which is 'significantly similar' to two or more other rows might be considered 'too small', and deleted before SRD. However, such rules are difficult to implement, because they can involve arbitrary decisions when, for example, two rows are significantly similar to each other, and each to one other row.

Comparison of these three approaches suggests that the first, simple, approach is roughly the union of the second and third. That is to say, any row or column which would be deleted under either the second approach or the third, is likely to be deleted under the first. We therefore recommend the use of the first rule, which has the benefit of being the simplest; the procedure is here called 'pruning'.

19.2.6 Extending the approach to three dimensions

SRD is a technique for use on two-way tables, while we are working on a three-way data structure (e.g. context, fabric, form). SRD can therefore work directly only on the marginal tables. It can be followed by examination of the following models:

1A: <2><3> : independence of fabric and form in a reduced table,

1B: <3><1> : independence of form and context in a reduced table,

1C: <1><2> : independence of context and fabric in a reduced table.

The reduced tables will not in general have the same groupings of the three variables. For example, the groupings of contexts need not be the same under models 1B and 1C — different groupings may be appropriate under different circumstances.

To extend this approach to the three-way table we introduce the idea of a 'doubly-reduced' table, which is constructed as follows:

1. suppose we have reduced the fabric-by-form marginal table, i.e. we are working with model <2><3> (exactly analogous procedures hold for the other models).
2. we construct a new three-way table in which the rows are the fabric-by-form combinations and the columns are contexts.
3. we then reduce this table by SRD, *except* that we allow only columns (i.e. context) merges. To allow row-merges would destroy the two-way nature of the marginal table.

We thus have three further models:

IIA: <23><1> : independence of fabric-by-form and context in doubly-reduced table,

IIB: <31><2> : independence of form-by-context and fabric in doubly-reduced table,

IIC: <12><3> : independence of context-by-fabric and form in doubly-reduced table.

Once again, the reduced tables do not necessarily share the same groupings of the three variables.

19.3 Examples

19.3.1 A ceramic assemblage

As an illustration of the method, we used some data from Silchester phase 1 (Fulford 1987), consisting of the pie-values of the combinations of 22 fabrics and 31 forms found in that phase. This is too large a table to examine with any ease, and it is not presented here. We shall concentrate on showing the method; archaeological aspects and implications will be discussed elsewhere (Tyers & Orton, this volume).

The initial pruning stage reduces the data matrix to 8 fabrics by 12 forms. The fabric codes E, F, G, O and S refer to fine, flint-tempered, grog-tempered, organically-tempered and sand-tempered wares respectively; the numbers refer to variant fabrics. The form codes I, II, III, VI and XI refer to jars, bowls, dishes, beakers and lids respectively; numbers refer to subdivisions of these forms and un-numbered codes refer to examples that can only be classified in general terms. Some structure can be seen but the matrix is still too large to be taken in comfortably (table 19.1).

The SRD procedure starts by merging forms which are very similar in terms of their fabrics — II and III, I4 and XI6 (a jar form and a lid, but both are only present in fabrics F1 and G1), I2 and I12, I and I6, I16 and I4/XI6 — then the majority fabric G1 absorbs the related minor fabrics GF1 and GO1 and an apparently unrelated one (S2), and I/I6 merges with I1 and with I2, leading to a matrix in which no further merges are possible (table 19.2).

The main features of the data are readily apparent from this table: the fabric E6 and the form II/III form a separable component, and beakers VI2 occur only in fabric G4. There is a marked association between form III1 (a bowl form) or rows) and *vice versa*. The derivation of this test is given in section 19.2.4.

3. if this condition is satisfied, merge, with the following steps:
   (a) recalculate the data matrix,
   (b) recalculate the *profiles*, i.e. the values $x_{ij}/x_{i.}$ and $x_{.j}/x_{..}$
   (c) recalculate the pairwise distances between the rows and between the columns, noting that if we have merged two rows, only the distances between the new merged row and all other rows, and between all pairs of columns, have to be recalculated. Other distances are unchanged.

Stopping rule

The merging procedure stops when the weighted distances between any pair of candidates are too great for merging.

Discussion

As well as working well in practice, this approach has some useful theoretical properties, e.g.

1. the overall chi-squared statistic is reduced by a relatively small amount at each step, subject to overriding archaeological constraints. Although the reduction at each step is the smallest possible, this seems not to be a global property.
2. the ratio of the current value of chi-squared to the initial value is a measure of the fit of the merged data to the original dataset. It is in fact the ratio of inertia accounted for/total inertia, as encountered in correspondence analysis.

This approach has general applications as a data-reduction technique. It could be used as a way of removing noise from data as a preliminary, or even an alternative, to correspondence analysis. It is here christened 'simultaneous reduction of dimension' (SRD).

19.2.4 The distance between two rows (or columns) as a chi-squared statistic

We start from the geometrical result that the total inertia of the rows, i.e. $\sum_i x_{i.}(i\text{th distance from centroid})^2$ is a chi-squared statistic whose number of degrees of freedom is one less than the number of columns (Greenacre & Hastie 1987). So for just two rows we have:

$$\chi^2 = \sum_i x_{i.} d_i^2 = x_{1.} d_1^2 + x_{2.} d_2^2 \qquad (19.3)$$

where $d_i = i$th distance to centroid and

$$d(1,2) = d_1 + d_2 \qquad (19.4)$$

and

$$x_{1.} d_1 = x_{2.} d_2,\ \text{i.e. } d_2 = (x_{1.}/x_{2.})d_1 \qquad (19.5)$$

Substitute 19.5 in 19.3:

$$\chi^2 = x_{1.}d_1^2 + x_{2.}(x_{1.}/x_{2.})^2 d_1^2 = d_1^2(x_{1.} + x_{1.}^2/x_{2.}) \qquad (19.6)$$

Substitute 19.5 in 19.4:

$$d = d(1,2) = d_1 + (x_{1.}/x_{2.})d_1 = d_1(x_{1.} + x_{2.})/x_{2.} \qquad (19.7)$$

so that $d_1 = dx_{2.}/(x_{1.} + x_{2.})$ Substitute 19.7 in 19.6:

$$\chi^2 = (dx_{2.}/(x_{1.}+x_{2.}))^2(x_{1.}+x_{2.})(x_{1.}/x_{2.}) = d^2 x_{1.} x_{2.}/(x_{1.}+x_{2.}) \qquad (19.8)$$

19.2.5 Pre-treatment of data matrices

In interpreting the outcome of an SRD, we sometimes encounter 'meaningless' merges, in which a very small unit (row or column) merges with another across a long distance simply because its low weight reduces its inertia (chi-squared) to an insignificant level. It seems likely that some units are *so* small that they *must* merge with some other unit. This leads to an inconsistency, in that the very small units survive because they must merge with something else, while larger (but still small) units may be deleted at the end of the process because they failed to merge with anything else. The merging of the very small units is often arbitrary and uninformative, and it seems better in principle to merge or delete them before carrying out SRD.

The problem is to detect such units — i.e. how 'small' is 'very small'? It seems reasonable to say that if a unit is so small that, whatever its actual profile, it could not possibly be significantly different from the overall mean profile, it should be deleted or merged in advance. In practice this means that the program should delete it; if the user wishes to reinstate it he should merge it with another unit for a second run. The critical value for the significant difference of any point of a given weight depends on the geometry of the chi-squared space, i.e. on the number of dimensions k and the mean profile $(x_{.1}, x_{.2}, \ldots, x_{.k})$. If we assume that the space is spheroidal, i.e. all dimensions have equal weight in the mean profile, we can represent this profile as

$$(1/k, 1/k, \ldots, 1/k)$$

and the worst-case profile as $(1, 0, 0, \ldots, 0)$ without loss of generality. Then using equation 19.1 we have

$$\begin{aligned} d_2 &= (1 - 1/k)^2/(1/k) + (1/k)^2/(1/k) + \ldots + (1/k)^2/(1/k), \\ &= k((k-1)/k)^2) + (k-1)(1/k), \\ &= (k-1)^2/k + (k-1)/k = k - 1 \end{aligned}$$

We note further that if the smallest weight in the mean profile is 1/rk, and the rest of the weight is distributed evenly, this result becomes

$$d^2 = rk - 1$$

From equation 19.2, the weighted distance

$$= d^2 x_{1.} x_{..}/(x_{1.} + x_{..}) = d^2/(1/x_{1.} + 1/x_{..}) \approx x_{1.} d^2$$

when $x_{..}$ is 'large'. So

$$\chi^2_{k-1} = (k-1)x_{1.}$$

and the critical value is

$$x_{1.} = \chi^2_{k-1}/(k-1)$$

at the required probability level.

This simple approach assumes that the space is spheroidal, while in practice it is often very far from being so. Since it is the smallest weight in, for example, the mean row profile that determines the critical value for the weight of a row, the critical value for rows is likely to be distorted downwards by the column weights which we would wish to eliminate had we looked at columns first, and vice versa. In other words, the very small values of column weights tend to 'protect' the very small row weights, and vice versa. It therefore seems reasonably to calculate the critical value for rows by omitting the columns which appear to be 'very small', and vice versa.

An alternative approach is to look for weighted distances that are 'significantly small' as well as those that are 'significantly large', e.g. ones which are less than the 95% probability level. In other words, we look for pairs of rows (or columns) that are significantly closer than would be expected if they were really two samples from the same parent distribution. Since the matrix of weighted distances has $k(k-1)/2$ distinct entries, we can expect such a level to be exceeded occasionally, but repeated occurrences would indicate rows (or columns) whose total weight is 'too small'. For example, a row which is 'significantly similar' to two or more other rows might be considered 'too small', and deleted before SRD. However, such rules are difficult to implement, because they can involve arbitrary decisions when, for example, two rows are significantly similar to each other, and each to one other row.

Comparison of these three approaches suggests that the first, simple, approach is roughly the union of the second and third. That is to say, any row or column which would be deleted under either the second approach or the third, is likely to be deleted under the first. We therefore recommend the use of the first rule, which has the benefit of being the simplest; the procedure is here called 'pruning'.

19.2.6 Extending the approach to three dimensions

SRD is a technique for use on two-way tables, while we are working on a three-way data structure (e.g. context, fabric, form). SRD can therefore work directly only on the marginal tables. It can be followed by examination of the following models:

1A: <2><3>: independence of fabric and form in a reduced table,

1B: <3><1> : independence of form and context in a reduced table,

1C: <1><2> : independence of context and fabric in a reduced table.

The reduced tables will not in general have the same groupings of the three variables. For example, the groupings of contexts need not be the same under models 1B and 1C — different groupings may be appropriate under different circumstances.

To extend this approach to the three-way table we introduce the idea of a 'doubly-reduced' table, which is constructed as follows:

1. suppose we have reduced the fabric-by-form marginal table, i.e. we are working with model <2><3> (exactly analogous procedures hold for the other models).
2. we construct a new three-way table in which the rows are the fabric-by-form combinations and the columns are contexts.
3. we then reduce this table by SRD, *except* that we allow only columns (i.e. context) merges. To allow row-merges would destroy the two-way nature of the marginal table.

We thus have three further models:

IIA: <23><1> : independence of fabric-by-form and context in doubly-reduced table,

IIB: <31><2> : independence of form-by-context and fabric in doubly-reduced table,

IIC: <12><3> : independence of context-by-fabric and form in doubly-reduced table.

Once again, the reduced tables do not necessarily share the same groupings of the three variables.

19.3 Examples

19.3.1 A ceramic assemblage

As an illustration of the method, we used some data from Silchester phase 1 (Fulford 1987), consisting of the pie-values of the combinations of 22 fabrics and 31 forms found in that phase. This is too large a table to examine with any ease, and it is not presented here. We shall concentrate on showing the method; archaeological aspects and implications will be discussed elsewhere (Tyers & Orton, this volume).

The initial pruning stage reduces the data matrix to 8 fabrics by 12 forms. The fabric codes E, F, G, O and S refer to fine, flint-tempered, grog-tempered, organically-tempered and sand-tempered wares respectively; the numbers refer to variant fabrics. The form codes I, II, III, VI and XI refer to jars, bowls, dishes, beakers and lids respectively; numbers refer to subdivisions of these forms and un-numbered codes refer to examples that can only be classified in general terms. Some structure can be seen but the matrix is still too large to be taken in comfortably (table 19.1).

The SRD procedure starts by merging forms which are very similar in terms of their fabrics — II and III, I4 and XI6 (a jar form and a lid, but both are only present in fabrics F1 and G1), I2 and I12, I and I6, I16 and I4/XI6 — then the majority fabric G1 absorbs the related minor fabrics GF1 and GO1 and an apparently unrelated one (S2), and I/I6 merges with I1 and with I2, leading to a matrix in which no further merges are possible (table 19.2).

The main features of the data are readily apparent from this table: the fabric E6 and the form II/III form a separable component, and beakers VI2 occur only in fabric G4. There is a marked association between form III1 (a bowl form)

form	fabric								total
	E6	F1	G1	G2	G4	GF1	GO1	S2	
I	0.0	0.076	1.65	0.0	0.095	0.033	0.0	0.0	3.69
I1	0.0	1.68	14.8	0.080	1.22	0.0	0.0	0.056	19.1
I2	0.0	0.0	8.20	1.42	1.26	1.00	0.094	0.043	13.3
I4	0.0	7.94	3.04	0.0	0.0	0.0	0.0	0.0	11.0
I6	0.0	0.0	10.8	0.0	2.15	1.06	0.0	1.86	15.9
I12	0.0	0.0	3.68	0.0	0.0	0.024	0.0	0.0	3.92
I16	0.0	6.09	0.0	0.0	0.030	0.0	0.0	0.0	6.39
II	1.39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.39
III	1.36	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.36
III1	0.0	0.0	1.09	2.07	0.035	0.0	0.0	0.0	3.51
VI2	0.0	0.0	0.0	0.0	1.32	0.0	0.0	0.0	1.32
XI6	0.0	1.30	1.50	0.0	0.0	0.0	0.0	0.0	2.80
total	2.75	17.8	44.8	4.29	7.54	2.64	0.094	2.85	83.6

Table 19.1: Silchester phase 1, pies by fabric and form after pruning

form	fabric					total
	E6	F1	G1	G2	G4	
I	0.0	2.44	45.6	2.22	5.58	55.9
I4	0.0	15.3	4.53	0.0	0.030	20.2
II	2.75	0.0	0.0	0.0	0.0	2.75
III1	0.0	0.0	1.09	2.07	0.035	3.51
VI2	0.0	0.0	0.0	0.0	1.32	1.32
total	2.75	17.8	51.2	4.29	7.54	83.6

Table 19.2: Silchester phase 1, pies by fabric and form after SRD. The members of the groups are:

G1 = G1 + GF1 + GO1 + S2
I = I + I1 + I2 +I6 + I12
I4 = I4 + I16 + XI6
II = II + III

and fabric G2, but the bulk of the pottery (81%) belongs of fabrics F1 and G1/GF1/GO1/S2 and forms I/I2/I2/I6/I12 and I4/I16/XI6, with a very strong association between the former fabric and the latter form, and *vice versa*.

Table 19.3 summarises the changes that come about as the original matrix is pruned and then shrunk. Pruning reduces the size of the matrix to 12% of the original, while retaining 80% of the total data and 31% of the chi-squared statistic. The SRD procedure reduces the matrix to 21% of its pruned size while retaining 89% of the chi-squared statistic. The underlying pattern is considerably 'sharpened-up' by the SRD. It should be noted that even the final data matrix is not suitable for a chi-squared goodness-of-fit treatment as it stands because of some very small expected values. However, the main points can easily be picked out be visual inspection.

19.3.2 Comparison of several assemblages across a site

Here we take yet another look at the late Roman pit groups from Portchester Castle (Fulford 1975), as re-analysed by Millett (1979). The data consist of the percentages of eight broad forms in each of 73 pit assemblages; they have been converted to absolute numbers by Hargreaves (1988), who also corrected some anomalies in the original data. It should be noted that the data are based on vessels represented, not on EVES or PIES. Because the forms have already been grouped, the procedure reduces the number of pits from 73 to 10 linked groups before merging any forms, and then makes only one further reduction in the pit groups before stopping. The merged forms — flagons and lids are unrelated archaeologically and are probably merged simply because lids are the rarest form. The outcome is shown as table 19.4.

Correspondence analysis was carried out on the data both before and after the application of srd. The original data showed a pattern in many dimensions — it took five of the seven original dimensions to account for 80% of the inertia. Despite the simplification of the reduction from 73 pits to 10 groups, and from nine to eight forms, the higher dimensions are still need to account for the inertia: only 72% is accounted for by the first three dimensions, for example.

The reason can be seen in the structure of table 19.4. Here values which appear to be significantly higher than 'expected' are underlined. It can be seen that every form has a group of pits in which it is more common than expected, and that mortaria have two such groups, while storage jars and beakers share a group as well as having their own groups. As it stands, this result is archaeologically incredible, and combined with the high dimension of the data structure suggests we may be dealing with a 'frozen' view of a random pattern. This is supported by an informal spatial

data	fabrics	forms	df	as %	pies	as %	χ^2	as %
raw	22	31	630	100	104.4	100	588.4	100
pruned	8	12	77	12	83.6	80	184.8	31
SRD	5	4	12	2	83.6	80	165.1	28

Table 19.3: Silchester phase 1 — effects on numbers of fabrics and forms, degrees of freedom, total pies and chi-squared statistic, of pruning and application of SRD

	form groups							
pit	dish	bowl	jar	s-jar	beaker	flagon	mort.	total
40	27	93	131	0	21	10	6	288
41	87	129	154	1	47	41	20	479
46	40	90	62	2	31	6	7	238
54	24	44	51	7	6	1	5	138
61	18	31	36	0	16	1	14	116
62	7	28	35	5	18	0	4	97
66	20	13	21	0	0	0	0	54
87	2	11	10	1	0	0	7	31
90	6	8	19	0	15	2	1	51
total	231	447	519	16	154	61	64	1492

Table 19.4: Number of vessels from late Roman pit groups at Portchester Castle, by broad form and pits as grouped by SRD. Each group is identified by the lowest-numbered pit that belongs to it

analysis, which revealed no apparent spatial correlates of the suggested groupings.

19.4 Discussion

19.4.1 Rejected alternatives

Unweighted distance

Earlier versions of this approach used the unweighted distance in step 1; this has now been abandoned in favour of the weighted distance. It might be thought that there would be little difference between the two approaches, since the unweighted version uses the weighted distance as a criterion for whether or not to permit a pair of rows or columns to merge. However, the order of merging differs, with small units tending to be merged sooner with the weighted distance, and to merge with other small units rather than with large ones. The effect seems to be to give a more even grouping of the units.

Archaeological intervention in the merging process

It was initially thought that some sort of archaeological intervention in the merging process would be desirable, i.e. that the archaeologists ought to have the opportunity to 'approve' (or not) any merges suggested by SRD. Second thoughts reversed this opinion; if (for example) two contexts are so similar in terms of their constituent fabrics and/or forms, is the archaeologist ever in a position to over-rule this similarity?

We now see the best place for archaeological input to be at an earlier stage — the archaeologist may if he wishes produce groupings of contexts, fabrics or forms brought together on archaeological criteria.

Use of the 'cell expectations > 1' criterion in SRD

The perceived purpose of SRD has undergone subtle changes since it was first devised. In the beginning, it was seen simply as a way of meeting the purely statistical need that no cell expectations should be less than 1. It gradually emerged that it was a valuable analytical tool in its own right, suggesting groupings of the values of each variable (usually context, fabric and form) which made sense in terms of their relationship with the values of one or both of the other variables. The values in individual cells became irrelevant at this stage of the analysis. Clearly, the criterion had to be employed somewhere for the quasi-log-linear analysis to work properly, but it seemed more logical for it to be part of that stage, and it was transferred there.

19.4.2 Statistical aspects

The SRD procedure emerges as an answer to a pressing problem, and cannot yet be regarded as a fully-developed technique. For example, being related to the simpler versions of cluster analysis, it suffers from some of the defects associated with such techniques. The most important is probably that of stability; it has been observed that, under certain circumstances, minor changes in the data can give rise to different groupings of the values of either or both variables. The answer in case of cluster analysis was the k-means approach (Doran & Hodson 1975, pp. 180–4); it seems likely that a comparable approach could be valuable here, but it has not yet been pursued.

The possibility arises in section 19.3.2 that we have created a pattern out of mainly random data, by bringing together all the pits that differ from the norm in a certain way (e.g. more bowls than 'expected'). While no one such pit may differ significantly from the overall mean, it may be that we have created an aggregate which because of its increased size does differ significantly from the overall mean. The problem is that by choosing to merge pit-assemblages on the basis of their similarity, we have distorted the significance levels on which subsequent tests are based. This aspect will be examined by simulation in the second stage of the project (1991–2).

19.5 Acknowledgements

We are very grateful to all those who allowed us to use their data for experimental purposes. Special thanks go to Andrew Scott for pointing out the problem about significance levels mentioned above.

Bibliography

BISHOP, Y. M. M., S. E. FIENBERG, & P. W. HOLLAND 1975. *Discrete Multivariate Analysis*. MIT Press, Cambridge, Massachusetts.

COCHRAN, W. G. 1954. "Some methods for strengthening the common chi-squared test", *Biometrics*, 10: 417–51.

CRADDOCK, J. M. & C. R. FLOOD 1970. "The distribution of the chi-squared statistic in small contingency tables", *Applied Statistics*, 19: 173–81.

DORAN, J. E. & F. R. HODSON 1975. *Mathematics and Computers in Archaeology*. Edinburgh University Press.

FULFORD, M. G. 1975. "The Pottery", *in* Cunliffe, B. W., (ed.), *Excavations at Portchester Castle, I: Roman*. The Society of Antiquaries, London.

FULFORD, M. G. 1987. "Calleva Atrebatum: an interim report on the excavations of the oppidum 1980–1986", *Proceedings of the Prehistoric Society*, 53: 271–8.

GREENACRE, M. J. 1984. *Theory and applications of correspondence analysis*. Academic Press, London.

GREENACRE, M. J. & T. J. HASTIE 1987. "The geometric interpretation of correspondence analysis", *Journal of the American Statistical Association*, 82: 437–47.

HARGREAVES, G. H. 1988. "Report on Project 2, Course C350". Project work held by the Prehistory Department, Institute of Archaeology, University College London.

MILLETT, M. 1979. "An approach to the functional interpretation of pottery", *in* Millett, M., (ed.), *Pottery and the Archaeologist*, pp. 35–48. Institute of Archaeology, London. Occasional Publication No. 4.

ORTON, C. R. & P. A. TYERS 1989. "Error structures of ceramic assemblages", *in* Rahtz, S. P. Q. & Richards, J. D., (eds.), *Computer Applications and Quantitative Methods in Archaeology 1989*, International Series 548, pp. 275–285. British Archaeological Reports, Oxford.

TYERS, P. A. & C. R. ORTON this volume. "Statistical Analysis of Ceramic Assemblages — a year's progress", *in* Lockyear, K. & Rahtz, S. P. Q., (eds.), *Computer Applications and Quantitative Methods in Archaeology 1990*.