A NEW COMPUTER SERIATION ALGORITHM

Armando De Guio
Instituto di Archeologia
Universita di Padova ITALY

Giacomo Secco
Dipartimento di Geografia
Universita di Padova ITALY

ABSTRACT

The seriation problem is approached by a sequence of two algorithms, Seriat 1 and Seriat 2. The first proceeds along a logic of set theory rather blind to the spatial-topological properties of the data and offers a first ordered matrix configuration essential to the second algorithm, which in turn acts exclusively on such properties and which supplies a final ordered matrix.

INTRODUCTION

The prime objective of this paper is a practical illustration of a seriation technique including the necessary basic theoretical and epistemological references with only a brief and problem-oriented approach towards the already existing extensive literature (Marquardt 1978).

An operational definition of seriation can be given as the ranked ordering of items along a single dimension so that each item reflects its own similarity with other items. At the base of each possible operational path lies a data matrix, which, in its standard configuration, is made up of units listed horizontally as rows and variables listed vertically as columns, measured in different possible scales (numeric, ordinal, nominal). One can work with this configuration both directly and indirectly, so as to extrapolate, from this data structure, the latent vector of seriation, which from now on we will, for the sake of argument, call the time-dimenstion.

The analytical pathway now tends towards the strictly correlated definition of three types of models of incidence matrices: firstly an "ideal model" and a "realistic" one for the background data; secondly a seriation model; thirdly an iconic model for the synoptic projection of seriated data.

The last model, seen by the authors as being the most efficient one (Fig. 2a), is a matrix configuration immediately susceptible to isomorphic transformation into a two-dimensional Cartesian coordinate system, the point of origin being indifferently placed in correspondence to any of the four vertices of the matrix.

The axes measure the duration of the units and variables with two units of measurement corresponding to the intervals between rows and columns, so that chronological steps of indeterminate absolute value are unified (so as to render the system operative and functional). Units and variables are seriated on the basis of their geometric mean ("centre of gravity"). The information pertinent to beginning, end and relative duration is preserved and reflects the real diachronic pattern of the data.

Other possible matrix models appear to be less efficient: for example, the simultaneous ordering by beginning (or end) of the units and variables loses, due to its topological incompatibility, information pertaining to tne relative duration of the units and variables or, in more abstract terms, the matrix no longer allows the isomorphic transformation of the matrix within the above mentioned two-dimensional Cartesian co-ordinate system.

We should now define the necessary theoretical requisities so that the data (both units and variables) can receive an optimal seriation and become the iconic model along the lines described above.

This new "ideal" model of the background data is reducible to two simple conditions. Firstly, the unity of the latent vector (one-dimension expected) for units and variables with reference to the pertinent "parent population". Tne relative iconic matrix model must therefore show a marked diagonalisation and exclude "blanks" (o's) within the ranges of both rows and colums contemporaneously (which would otherwise indicate a sample error); each row and column can represent such blanks within their range only if they occur at the beginning or end of column and row respectively. In this case the blanks simply reflect the geometric topological results of different rates of duration of the units and variables (e.g. the rows and colums 1,5,12,16 and 1,3, respectively of Fig 2a, the background data of which can without doubt be defined as ideal). In that sense and within the constricting frame of reference outlined, a "Petrie form" of the matrix (Petrie 1899, Kendall 1963), with its rigid prescription of total contiguity of "presences" (1's) in the columns, appears to be, very simply, a "hyper-ideal" construct which would require the totally abnormal situation of each unit commencing and finishing not between two external ones.

A healthy exercise often ignored, at this point, would be to descend from the theoretical state and attempt to define, utilizing our daily professional experience, a "realistic" model based on our normal operative field of seriable data. With extreme synthesis, we can reasonably suppose that such a model can include the following distorting factors:

1. the interference by other latent vectors with respect to the desired one (for example, with respect to the temporal dimension, the spatial or "functional" ones in the widest possible sense) even after a forward-looking "clearance" of them;

2. a normal under-representative sample (with possible exceptions for specific classes and/or circumstances of find) with respect not so much to a "parent", rather to a "target" population (Doran-Hodson 1975, 75);

3. a highly differentiated sample representativity among the different types of units and variables in relation to their different frequency both pre and post depositional, spatial location, function, underlying human behaviour and last but not least, the level of taxonomic resolution of the data which we have to assume (for instance the normal marked hierarchical differentiation leads to a noticeable variation in the importance and even presence-absence of the entities);

4. variability in the relative duration among units and variables and possible temporal discontinuity-intermittence (at least for the positional units). These last factors in particular, in addition to the above mentioned ones, highlight the theoretical impractability of an approach by "abundance" matrices, whose weak supporting assumptions (specially the equal probability of sample representation, the time-span equivalence, the lack of diachronic palimpsest, the regularity of ontogenetic cycles) suggest a realistic heuristic prority for the "incidence" matrices here discussed.

Once this analytical diagnosis has been accepted two strategies are possible:

a) to return to an "ideal" model, considering the possible deviations examined as marginal or self-limiting;

b) critically to incorporate the "realistic" model and suggest a stochastic algorithm "ad hoc" capable of reproducing with close approximation the generative seriation pattern. To follow this last path a system has been devised with a rather complex, functional articulation in a sequence of two algorithms: Seriat 1 and Seriat 2. The first proceeds along a logic of set-theory rather blind to the spatial topological properties of the data and offers a first ordered matrix configuration. This is essential to the second algorithm, which in turn acts exclusively on the spatial topological properties and which supplies a final ordered matrix according to the optimal iconic model outlined above.

SERIAT 1 (ARMANDO DE GUIO)

Seriat 1 works along the following principal steps:

1. input an entry matrix of n x m dimensions (an "incidence" matrix, which one assumes has been previously "cleared" of the most visiable secondary latent vectors and presents a fairly reliable sample);

2. compute for the set rows (units) Ir (r= 1,2...n) a n x n matrix of similarity (a square symmetrical matrix) with the Jaccard coefficient (Jaccard 1908; Sneath-Sokal 1975,131; Chandon-Pinson 1981, 74);

3. percentualise row by row the similarity values (Sjp: now no longer symmetrical); the aim of this percentualisation is to introduce a factor of standardisation for the different numeric content of the sets. Each set now contains a quota, standardized in base 100, of systemic similarity (which we

would define as "bond energy") which is distributed in definite percentages to the other sets;

4. compute n vectors of tentative orientation and the relative strains with respect to the following sequential model; each element k should be situated in the ordering vector so that:

a) each element i which preceeds k (i=1,2 ....k-1) has a summation of similarity with the other j elements which follow k (ΣFi) not superior to that of k (ΣFk);

b) each element j which follows k(j = k+1, k+2 ..... n) has a summation of similarity with the i elements (Pj) not superior to that of k (Pk).

If this model is not followed, compute for each k one partial strain (sk) equal to the absolute difference between the scores ΣF and/or ΣP for the anomalous pairs. The total of such strains for each element k of the vector forms a total vector strain (sv). Proceeding from each initial element i = 1,2 ... n, n trial vectors are constructed all aligning in the same direction the elements which minimise each time the strain (sk): such a norm is avoidable only when an element k, even if it has a superior strain to others, accumulates anyway with the preceding elements of the segment of the chain all its total similarity (100). In such a case, and if the same state does not take place with other elements with a minor strain, it will link k anyway, which would otherwise contribute to the accumulation of partial strain in later steps. In the case of equality of (sk), an accessory scoring system is introduced which "weights" (with simple ranking factors) the elements proportionally more similar to the more external ones of the already linked segment;

5. choose the vector with the least (sv). In the eventuality of a hyper-ideal configuration of the data suitable to the sequential model described above:

a) there exist two vectors only with (sv)= 0;

b) the sequential order is exactly symmetrically inverted.

In the case of an un-ideal configuration, more suitable to the already-described "realistic" model:

a) the vector with least (sv) will anyway reproduce in a relatively better way the main latent vector wih possible local distortions;

b) vector with the least (sv)-scores tend anyway to have an inherent similar order, whether tne sequence is inverted or not;

6. repeat steps 2 to 5, this time for the set of columns (variables) Ic (c = 1,2......m);

7. re-order the matrix n x m according to the new reordering vectors.

SERIAT 2 (GIACOMO SECCO)

Tne second algorithm (seriat 2) proceeds along the following main steps:

1. use the matric n x m as re-ordered by Seriat 1;

2. consider the matrix as a two-dimensional Cartesian co-ordinate system with the origin corresponding to the bottom, left-hand corner, with the unit of measurement equal to the interval both of row and column (assumed to have the same width). Work out the mean for each set of colums (Ic) and of rows (Ir); compute two coefficients of strain for the columns (cs) and rows (rs) equal to the summation of the absolute differences of the values of those means, which are not aligned in monotonic increasing or decreasing order, and a third total coefficient of matrix-strain (ts) = (cs) + (rs). Compute the values of the sums of the blanks (o's) within the ranges of columns (nc), of rows (nr) and of (nt) = (nc) + (nr);

3. compute a distance-matrix between the sets of columns (Ic) on the basis of a coefficient (Ds) which takes into account both the distance between means and the dispersion of the elements according to the following formula:

   $$Ds\ (i,j) = |M_i - M_j| + (va - Vb)/N$$

   where

Ds (i,j)	=	distance between the sets i and j
M_i	=	the mean of the elements of i;
M_j	=	the mean of the elements of j;
Va	=	variance of the elements of i and j considered together;
Vb	=	variance of the elements i and j as if they were concentrated around the common mean;
N	=	the total number of elements in the sets i and j.

   The value (Va-Vb)/N works as a slight correction of the main value $|M_i - M_j|$ favouring tne coupling of sets of smaller dipersion;

4. create, on the basis of the distance matrix, a reordering vector for the sets Ic, with a clustering system of the type "single linkage-nearest neighbour" (Everitt 1981, 21), arranging each time the Ic or cluster of Ic which add themselves on the extremities of an already existing cluster, according to the highest degree of similarity;

5. re-arrange the matrix n x m in accordance to the vector of step 4 and measure the relative (cs), (rs), (ts), (nc), (nr), (nt);

6. if (ts) = 0 pass to step 7; otherwise repeat cycles 3-6, inverting, however the order between Ic and Ir up to a discretional maximum number of times, choosing the vectorial order with the least (ts) or, in the case of parity, the

least (nt);

7. memorize the vectorial order obtained at the end of steps 3-6;

8. wholly repeat steps 3-7 recommencing, however, with the configuration of the output of Seriat 1 (step 1) and inverting the order between Ic and Ir;

9. choose the vectorial order with tne least (ts) from among those steps 7 and 8, and, in the eventuality of parity, that the least (nt).

CONCLUSIONS

The sequential integration of the two algorithms Seriat 1 and Seriat 2 can not only be seen to be soundly based theoretically, but also be used experimentally with a high degree of efficiency. In a hyper-ideal situation of background data (cp. Fig. 1) Seriat 1 and Seriat 2 always give the same vectorial re-ordering; in other terms the properties of similarity based on the "set theory" and the topological ones have the one-to-one correspondence and the two seriation models find themselves with the same results. Assuming however, a body of data which is not hyper-ideal but which conforms to our realistic model, localised deviations in Seriat 1 are to be expected. the first algorithm constructs, in fact, provisional seriation vectors on the basis of a similarity coefficient (Jaccard) of a set-theory origin, without any refence to the spatial and topological properties produced by the ordering, but, in final analysis, only with reference to relationships of intersection and union between the sets. One can therefore construct a concatenation of similarity which captures the principle vector with a few possible localised deviations, derived from: a) the differentiated rates of duration of variables and units; b) the distortional factor of the possible secondary latent vectors; c) sampling limitations (our realistic model of the background data). It would now appear to be justifiable to say that such distortional factors are distributed in a tendentially randomized manner in the semi-ordered matrix produced by Seriat 1.

The aim of Seriat 2, which bases itself solely on the topological-spatial properties neglected by Seriat 1, is in fact to introduce corrections, arranging along the principal diagonal, otherwise described as the pincipal latent vector already grossly caught, the deviant sets: the seiation by "gravity point" in other words (e.g. Goldmann 1975, Wilkinson 1976) appears to be, but only at this point, the most efficient way of stochastic approximation to the presumed chronological pattern of the "target population" of units and variables: its efficiency grows with the incidence of distorting factors and with the degree of rate of diferentiation in the length of units and variables.

The criterion to optimize is that of defining the optimal equilibrium, closest to the configuration of the output of Seriat 1, in terms of minimizing strain (ts). In fact and as expected, the different cycles of Seriat 1 and Seriat 2 tend, due to their logical directive, both to diagonalise and compact the matrix ("concentration principle", (Kendall 1966, 659; Doran-Hodson 1975, 276) and to approximate the "minimum path" of the seriation chain (the "travelling salesman problem": Bellmore-Nemhouser 1968, Wilkinson 1974) in the ordering of the units and variables.

Seriat 1 and Seriat 2 were applied to both test data and real matrices for example:

1. "hyper ideal" matrix (Petrie form) (fig. 1)
2. "ideal matrix" (fig. 2)
3. "ideal matrix" with insertion of randomised blanks; (Fig 3.)
4. real matrix (from Goldmann 1975) (fig. 4)

The results of Seriat 1 and Seriat 2 in the first two instances coincide and reproduce the input model; in the third Seriat 2 betters both the values of (ts) and the approximation to the input model; in the fourth Seriat 2 beters Seriat 1 in terms of (ts) and (nt): the results are in any case very similar to those arrived at by Goldmann (Goldman 1975, Fig. 3).

REFERENCES

Bellmore, M. and Nemhauser G.L., 1968. The travelling salesman problem. A survey. Operations Research 16: 538-558

Chandon, J.L. and Pinson, S., 1981. Analyse typologique Theories et applications. Paris, Masson

Doran, J.E. and Hodson, F.R., 1975. Mathematics and computers in Archaeology. Edinburgh, Edinburgh University Press

Goldman, K., 1975. Some archaeological criteria for chronological seriation. In Hodson, F.R., Kendall, D.G., and Tautu. P. ed., Mathematics in the archaeological and historical sciences: 202;208. Edinburgh, Edinbugh Univesity Press.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. Bulle. Soc. Vaud. Sci. Nat. 44. 223-270.

Kendall, D.G., 1963. A statistical approach to Flinders Pertrie's sequence dating. International Statistical Institute, Bulletin 40:657-680.

Marquardt, W.H., 1978. Advances in archaeological seriation. In Schiffer, M.B. ed., Advances in archaeological method and theory. Vol. 1; 257-314. New York, Academic Press.

Petrie, W.M.F., 1899 Sequences in prehistoric remains. Journal of the Anthropological Institute 29:295-301.

Sneath, P.H.A. and Sokal, R.P., 1973. Numerical taxonomy. San Francisco, W.H. Freeman and Company.

Wilkinson, E.M., 1974. Techniques of data analysis-seriation theory. Archaeo-Physika 5: 1-142.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	+	+	+	+	+	+	+									1
2	+	+	+	+	+	+	+	+								2
3		+	+	+	+	+	+	+	+	+	+					3
4			+	+	+	+	+	+	+	+	+					4
5				+	+	+	+	+	+	+	+					5
6					+	+	+	+	+	+	+	+				6
7						+	+	+	+	+	+	+				7
8						+	+	+	+	+	+	+	+			8
9							+	+	+	+	+	+	+	+		9
10									+	+	+	+	+	+		10
11										+	+	+	+	+	+	11
12											+	+	+	+	+	12
13												+	+	+	+	13
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

a

CSTRAIN .000E+00 NC 0
RSTRAIN .000E+00 NR 0
TSTRAIN .000E+00 NT 0

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	+		+		+		+		+		+		+			1
2		+		+		+		+		+		+	+		+	7
3		+		+	+	+	+		+		+		+		+	4
4						+		+		+		+		+		12
5		+		+		+		+			+		+		+	7
6	+		+		+		+		+		+		+		+	2
7		+		+		+		+		+		+				10
8		+		+		+	+		+		+		+		+	5
9								+		+		+		+		13
10		+		+		+		+		+	+		+		+	8
11		+	+	+	+	+	+		+		+		+		+	3
12				+		+		+		+		+		+		11
13		+		+		+		+	+		+		+		+	6
	1	9	2	10	3	11	4	12	5	13	6	14	7	15	8	

b

CSTRAIN 32.3 NC 75
RSTRAIN 39.0 NR 70
TSTRAIN 71.3 NT 145

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	+	+	+	+	+	+	+									1
2	+	+	+	+	+	+	+	+								2
3		+	+	+	+	+	+	+	+	+	+					3
4			+	+	+	+	+	+	+	+	+					4
5				+	+	+	+	+	+	+	+					5
6					+	+	+	+	+	+	+	+				6
7						+	+	+	+	+	+	+				7
8						+	+	+	+	+	+	+	+			8
9							+	+	+	+	+	+	+	+		9
10									+	+	+	+	+	+		10
11										+	+	+	+	+	+	11
12											+	+	+	+	+	12
13												+	+	+	+	13
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

c

CSTRAIN .000E+00 NC 0
RSTRAIN .000E+00 NR 0
TSTRAIN .000E+00 NT 0

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	+	+	+	+	+	+	+									1
2	+	+	+	+	+	+	+	+								2
3		+	+	+	+	+	+	+	+	+	+					3
4			+	+	+	+	+	+	+	+	+					4
5				+	+	+	+	+	+	+	+					5
6					+	+	+	+	+	+	+	+				6
7						+	+	+	+	+	+	+				7
8						+	+	+	+	+	+	+	+			8
9							+	+	+	+	+	+	+	+		9
10									+	+	+	+	+	+		10
11										+	+	+	+	+	+	11
12											+	+	+	+	+	12
13												+	+	+	+	13
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

d

CSTRAIN .000E+00 NC 0
RSTRAIN .000E+00 NR 0
TSTRAIN .000E+00 NT 0

Figure 1. a) hyper-ideal matrix model (Petrie form); b) permutations of rows and columns (with the indication at the end of each row and column of the previous order-number); c) output of Seriat 1; d)output of Seriat 2;(all which (cs),(rs),(ts),(nc),(nr),(nt) values).

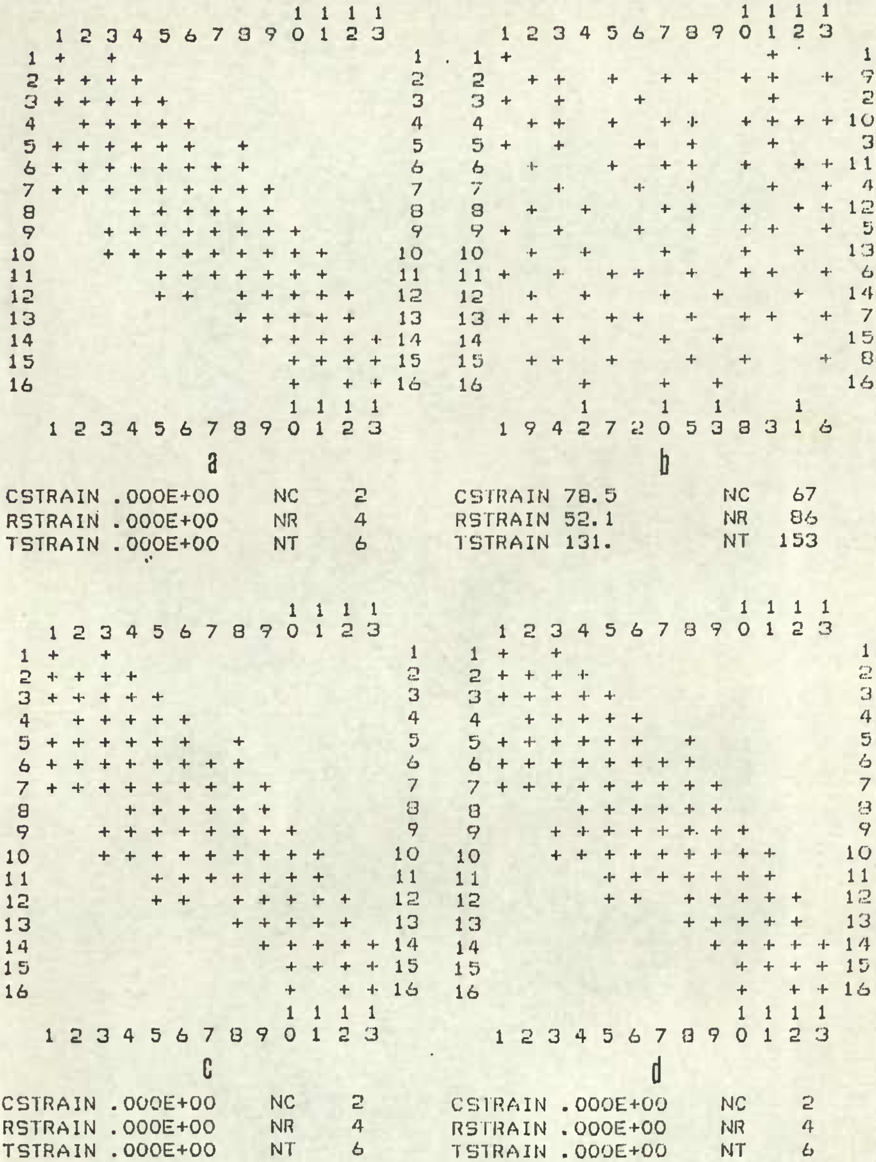

Figure 2. a) ideal matrix model; b) permutations of rows and columns; c) output of Seriat 1; d) output of Seriat 2.

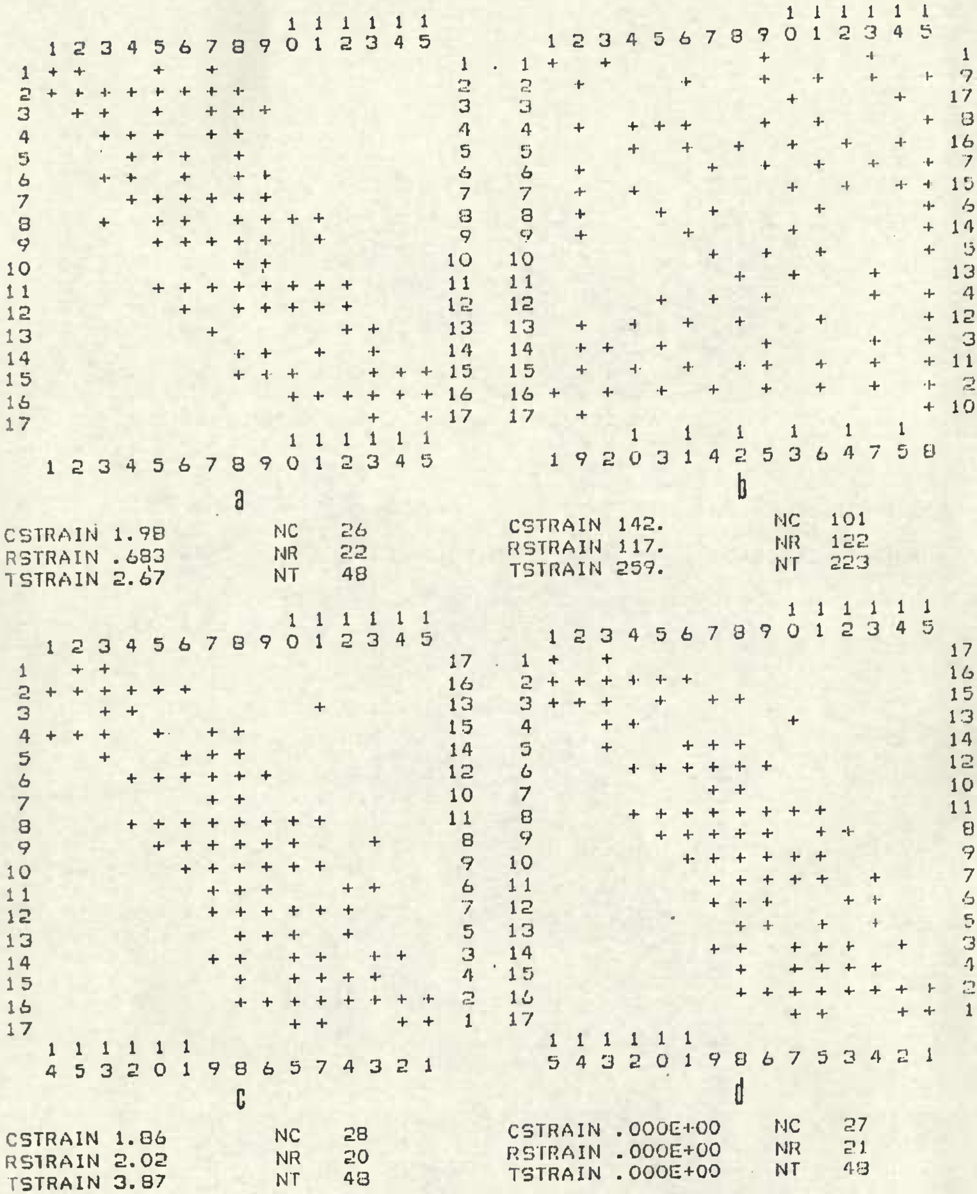

Figure 3. a) matrix derived from an ideal one with the insertion of randomized blanks; b) permutations of rows and columns; c) output of Seriat 1; d) output of Seriat 2.

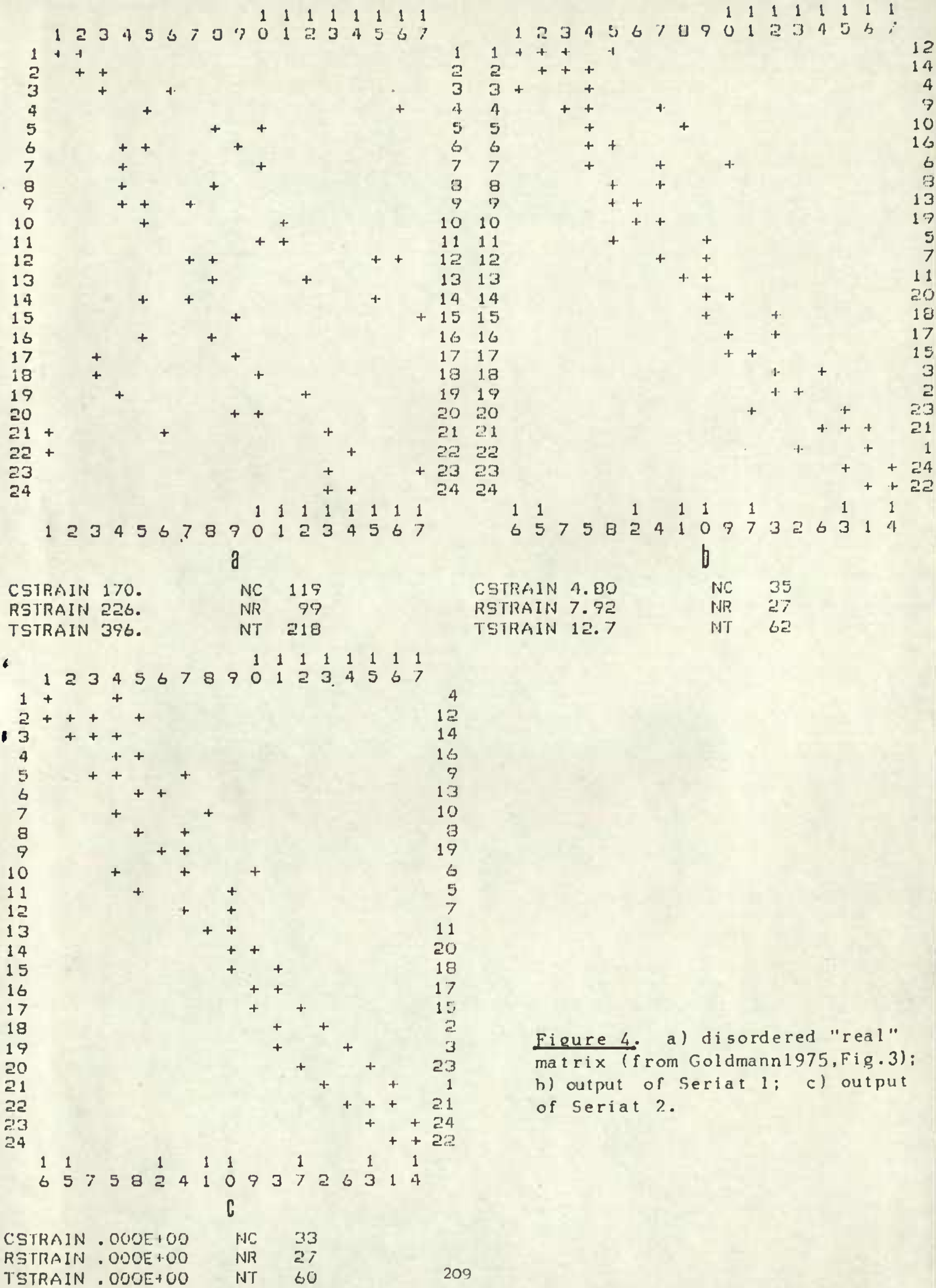

Figure 4. a) disordered "real" matrix (from Goldmann1975,Fig.3); b) output of Seriat 1; c) output of Seriat 2.