

Three Approaches to Aggregation of Interaction Tables¹

Boris Mirkin
DIMACS, Rutgers University
Piscataway, NJ 08854-8018 USA

Abstract

An interaction table is a summable square matrix emerging in analysis of inter-citation, international trade, brand-switching, mobility, or input-output industrial data. Three approaches to aggregation of interaction data are theoretically compared: (i) loglinear modeling, (ii) aggregation of Markov chains, and (iii) principle of equivalence in the correspondence analysis. This way an empirical clustering algorithm, developed in the framework (iii), is justified and amended by substantively modeling the interaction processes.

1. Introduction

An interaction table is a square matrix $P = (p_{ij})$ whose entries show volumes (flow) of interaction between entities being elements of the same set I . Examples: (a) an intergenerational mobility table, with I the set of occupations and p_{ij} the number of families whose head has occupation i while his son occupation j ; (b) a brand switching table, with I the set of car brands and p_{ij} the number of consumers who purchased a model of brand j after they had purchased brand i ; (c) a confusion table, with I the set of stimuli and p_{ij} the number of respondents who perceived the stimulus i as j . The aggregation problem emerges when the existence of wider categories, or classes, of entities is hypothetically assumed, whose pattern of interaction yields the observed interaction table.

For instance, based on their analysis of intergenerational mobility in the USA, Featherman and Hauser (1978) proposed a five-class aggregation of occupations that has become classic in occupation research. The classes are listed in Table 1 as aggregates of seventeen finer occupation categories for which a corrected mobility table is presented in Breiger (1981).

There are three approaches to reveal the aggregate classes and their interaction patterns: (i) loglinear modeling, (ii) aggregation of Markov chains, and (iii) principle of equivalence in the correspondence analysis that are presented in the subsequent sections. Our major concern is compatibility of the correspondence-analysis-based approach with the two others.

2. Correspondence Analysis Based Approach

In analysis of contingency data $P(I, J) = (p_{ij})$, where I and J are sets of categories, $p_{ij} \geq 0$, and, without any loss of generality, $\sum_{i,j} p_{ij} = 1$, a method called Correspondence Analysis (CA) has proven to be useful (see, for instance, Benzecri (1973), Greenacre (1993), and Lebart, Morineau & Piron (1995)). CA is based on a specific form of the singular-value decomposition and takes into account what has been called the Equivalence Principle. The principle states that two rows, $i', i'' \in I$, must be treated as the same if corresponding conditional probability profiles, $p_{i'} = (p_{i'j}/p_{i'+})$

¹Published in: H. Bacelar-Nicolau, F. Costa Nicolau and J. Janssen (Eds.) *Applied Stochastic Models and Data Analysis*, Lisbon: Nat. Institute of Statistics, 30-35.

| No | Occupation | Aggregate |
|----|------------------------------|--------------------|
| 1 | Professionals, self-employed | Upper nonmanual |
| 2 | Professionals, salaried | |
| 3 | Managers | |
| 4 | Sales, other | |
| 5 | Proprietors | Lower nonmanual |
| 6 | Clerks | |
| 7 | Sales, retail | |
| 8 | Crafts, manufacturing | Upper manual |
| 9 | Crafts, other | |
| 10 | Crafts, construction | |
| 11 | Service | Lower manual |
| 12 | Operatives, other | |
| 13 | Operatives, manufacturing | |
| 14 | Laborers, manufacturing | |
| 15 | Laborers, other | |
| 16 | Farmers | Farm |
| 17 | Farm laborers | |

Table 1: Seventeen occupations and their aggregation according to Featherman and Hauser (1978).

and $p_{i''} = (p_{i''j}/p_{i''+})$, are equal to each other, where, for any $i \in I$, its marginal distribution (p_{i+}) is defined as usual by $p_{i+} = \sum_{j \in J} p_{ij}$. The principle is defined dually for columns $j \in J$. It is not difficult to prove that the principle can be reformulated with the so-called Quetelet coefficients as follows.

Let $S = \{S_1, \dots, S_m\}$ be a partition of I and $T = \{T_1, \dots, T_n\}$ of J . For a pair of elements, $i \in S_u$, $u = 1, \dots, m$, and $j \in T_v$, $v = 1, \dots, n$, let us consider $q_{ij} = (Prob(j/i) - Prob(j))/Prob(j) = p_{ij}/(p_{i+}p_{+j}) - 1$ that was proposed by A. Quetelet (1832) as a measure of association between i and j . Similarly, the aggregate Quetelet coefficient is defined as $q_{uv} = p_{uv}/(p_{u+}p_{+v}) - 1$ based on the aggregate table $P(S, T)$ with entries $p_{uv} = \sum_{i \in S_u} \sum_{j \in T_v} p_{ij}$. Then, the Equivalence Principle can be stated as follows: All the rows within classes of S and all the columns within classes of T are equivalent if and only if $q_{ij} = q_{uv}$ for all $i \in S_u$ and for all $j \in T_v$, for any u and v (Lebart & Mirkin, 1993; see also Govaert, 1989).

To approximately aggregate a contingency matrix $P(I, J)$ into an aggregate table $P(S, T)$ according to the Equivalence Principle, one should consider a model like

$$q_{ij} = \sum_u \sum_v \mu_{uv} s_{iu} t_{jv} + e_{ij} \quad (1)$$

where s_{iu} and t_{jv} are matrices of partitions S and T , respectively: $s_{iu} = 1$ when $j \in S_u$ and $t_{jv} = 1$ when $j \in T_v$ and $s_{iu} = t_{jv} = 0$, otherwise. The residuals in (1) are to be minimized with regard to unknown S, T , and reals μ_{uv} according to the least-squares criterion weighted by row/column weights:

$$E^2(S, T) = \sum_{i \in I} \sum_{j \in J} p_{i+} p_{+j} (q_{ij} - \sum_u \sum_v \mu_{uv} s_{iu} t_{jv})^2 \quad (2)$$

Given S, T , the optimal μ_{uv} is q_{uv} , which justifies selection of the weights in (2). Another point is that actually, up to minor changes, the same criterion is exploited in CA itself (with regard to arbitrary, not just binary, entries s_{iu} and t_{jv}).

With the optimal $\mu_{uv} = q_{uv}$ substituted in (2), the following data scatter decomposition can be proven:

$$X^2(I, J) = X^2(S, T) + E^2(S, T) \quad (3)$$

where $X^2(I, J)$ is the data scatter equal to the well-known Pearson chi-squared coefficient, $E^2(S, T)$ is the criterion minimized, and $X^2(S, T)$, the Pearson chi-squared for $P(S, T)$, is the part of the data scatter explained by bipartition S, T .

Applied to the interaction data, this construction involves $I = J$ as a given category set and $S = T$ as a sought aggregate category set, so that the decomposition (3) becomes

$$X^2(I, I) = X^2(S, S) + E^2(S, S) \quad (4)$$

where minimized criterion is:

$$E^2(S, S) = \sum_{u,v} \sum_{i \in S_u} \sum_{j \in S_v} p_{i+p+j} (q_{ij} - q_{uv})^2 \quad (5)$$

According to (4), the aggregation problem is of maximizing $X^2(S, S)$, which can be attacked with local search algorithms such as sequential pair-wise merging (agglomeration).

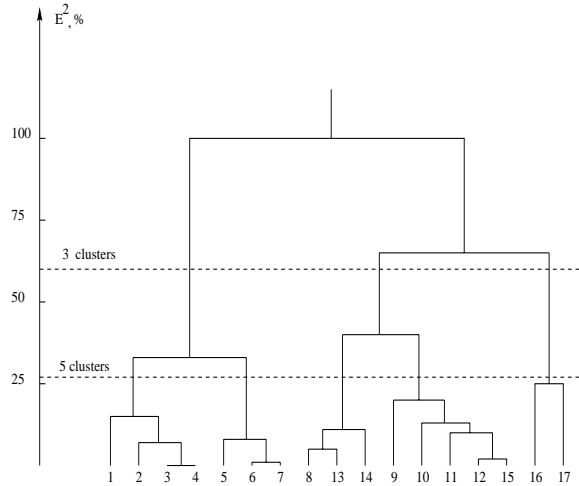


Figure 1: Results of agglomerative chi-square based aggregation for the 17×17 mobility table data in Breiger (1981).

The dendrogram in Fig.1 represents results of the chi-square based agglomeration algorithm applied to the 17×17 intergenerational mobility data in Breiger (1981). The tree differentiates three major divisions well: Nonmanual (1 to 7), Manual (8 to 15) and Farm (16-17). Five classes produced by the algorithm, basically, coincide with the Featherman-Hauser aggregation in Table 1 above. However, there is a difference in partitioning of manual workers (occupations 8 to 15): the algorithm separates manufacturing workers, 8-13-14, rather than maintaining Featherman-Hauser's Upper-Lower division. Separation of the manufacturing workers as a class has been suggested earlier

by Breiger (1981). Our 5-class partition is somewhat better than that of Featherman-Hauser: it takes 74.8% of the original X^2 value while the latter partition accounts for 72.5%.

3. Homogeneity Clustering with Loglinear Models

The problem of aggregation of interaction data, in terms of intergenerational mobility, was analyzed in mathematical sociology in terms, basically, related to multiplicative/loglinear modeling (see, for instance, Breiger (1981) and Goodman (1981)). In particular, L. Goodman (1981) presents the following framework.

For a partition $S = \{S_1, \dots, S_m\}$ on I , let us denote by $F_{ij,uv}$ the (i, j) entry of an interaction matrix $P(I, I)$, where $i \in S_u$ and $j \in S_v$. Let us assume that any $F_{ij,uv}$ can be expressed as the product of factors pertaining to row i , column j , and the corresponding pair of classes (u, v) , that is, as $\alpha_i \beta_j \gamma_{uv}$. Then, for any $i, j \in I$ (with $i \in S_u, j \in S_v$), (i, j) -th entry $F_{ij,uv}$ can be estimated due to partition S as “a product of three factors: (1) a factor pertaining to the i th row category (namely, p_{i+}), (2) a factor pertaining to the j th column category (namely, p_{+j}), (3) a factor pertaining to u th row class and v th column class (namely, $p_{uv}/[p_{u+}p_{+v}]$)” (see Goodman (1981), p.648); that is,

$$F_{ij,uv} = p_{i+}p_{+j} \frac{p_{uv}}{p_{u+}p_{+v}}. \quad (6)$$

Equation (6), actually, has a meaning on its own. It can be considered a model of the interaction process on I as governed by the aggregate process on S in such a manner that the parent's (row) and son's (column) occupation distributions within S classes are proportional to the proportions observed.

Obviously, the equation in (6) is equivalent to equation $q_{ij} = q_{uv}$ in terms of the preceding section. However, to find an appropriate S , the likelihood-ratio is employed in Goodman (1981) rather than the goodness-of-fit X^2 of the preceding section. This is based on equation:

$$F_{ij} = F_{ij,uv}/(p_{uv}/F_{uv}) \quad (7)$$

where $F_{ij} = p_{i+}p_{+j}$ and $F_{uv} = p_{u+}p_{+v}$ are expected entries for the processes on I and on S , respectively, under the hypothesis of “perfect” behavior, that is, of null association. Applying to (7) the likelihood-ratio χ^2 statistic for testing the null association hypothesis,

$$L^2(p_{ij}, F_{ij}) = \sum_{i,j \in I} p_{ij} \log \frac{p_{ij}}{F_{ij}}$$

the following decomposition holds:

$$L^2(p_{ij}, F_{ij}) = L^2(p_{ij}, F_{ij,uv}) + L^2(p_{uv}, F_{uv}) \quad (8)$$

This is an analogue to decomposition (4) implying that, in this setting, although the model equation is essentially the same as in CA-based approach, the aggregation criterion is maximizing $L^2(p_{uv}, F_{uv})$ rather than $X^2(p_{uv}, F_{uv})$. This criterion has been exploited by Krymkowski, Sawinski and Domanski (1996).

However, an alternative “multiplicative” model for governing interactions on I by interactions on S can be formulated in such a way that the equivalence to the CA-based approach becomes complete; that is, both the model equations and aggregation criterion remain the same (Mirkin, 1996). The alternative model is based upon the

assumption that the observed mobility (interaction) is a result of two distinct processes, perfect mobility and imperfect mobility, where perfect mobility runs in terms of I categories while imperfect mobility is governed by the “imperfectness” of the aggregate process. More explicitly, let the value $p_{uv} - p_{u+}p_{+v}$ in the theoretical aggregate matrix $P(S, S)$ reflect the difference between the “real” and “perfect”, at the aggregate level, mobility. In terms of the original categories, this gives $p(i/u)(p_{uv} - p_{u+}p_{+v})p(j/v)$ as the imperfect part of the overall mobility, where $p(i/u) = p_{i+}/p_{u+}$ and $p(j/v)$ is defined dually. This leads us to the following alternative model:

$$p_{ij} = e_{ij} + p(i/u)(p_{uv} - p_{u+}p_{+v})p(j/v) \quad (9)$$

where e_{ij} is that part of the observed mobility which counts for moves that are subject to the hypothesis of perfect mobility (null association) and, thus, are to be tested with the chi-squared coefficient. The model implies: (i) in a “natural” process, e_{ij} must be positive, and (ii) decomposition (4) of the CA-based approach remains with $X^2(S, S)$ as the aggregation criterion. Corollary (i) can be regarded a model-based stop-criterion for chi-squared agglomeration clustering.

With Featherman-Hauser five-class aggregation, there are three negative values of e_{ij} present. For the five-class aggregation produced by the algorithm (Fig.1), there is only one negative e_{ij} , which is quite close to zero, some -0.00001 . This can be considered yet another argument in favor of the aggregation found.

4. Aggregation in Terms of Markov Chains

Matrix $\Pi(I) = (\pi_{ij})$ of conditional probabilities $\pi_{ij} = p_{ij}/p_{i+}$ is a Markov chain on I . A Markov chain is referred to as aggregable due to partition $S = \{S_1, \dots, S_m\}$, or S -aggregable, if, for any initial distribution $p(I) = (p_i)$, the probabilities to reach any S_v from any S_u depend on $p(S)$ only. Here $p(S) = (p_u)$ where $p_u = \sum_{i \in S_u} p_i$. A Markov chain is S -aggregable if and only if, for any $u, v = 1, \dots, m$, the totals $\pi_{iv} = \sum_{j \in S_v} \pi_{ij}$ are equal to each other for any $i \in R_u$ (Kemeny and Snell, 1976). This statement can be equivalently reformulated as follows: A Markov chain $\Pi(I)$ is S -aggregable if and only if there exists an aggregate Markov chain $\Pi(S)$ such that $\sum_{j \in S_v} \pi_{ij} = \pi_{uv}$ for all $u, v = 1, \dots, m$ and $i \in S_u$.

The contents of the preceding sections suggests one more concept of aggregability. Let us refer to an interaction matrix, $P(I) = (p_{ij})$, as CA-aggregable via partition $S = \{S_1, \dots, S_m\}$ iff $q_{ij} = q_{uv}$ for all $i \in S_u, j \in S_v$ and all $u, v = 1, \dots, m$. Obviously, the corresponding Markov chain $\Pi(I)$ is S -aggregable if P is CA-aggregable via S , and so is the dual chain $\Phi(I) = (\phi_{ij})$ with ϕ_{ij} defined as $\phi_{ij} = p_{ij}/p_{+j}$.

An independent characteristic of CA-aggregable interaction matrices in terms of related Markov chains can be stated as this. An interaction matrix $P(I)$ is S -CA-aggregable if and only if corresponding Markov chains $\Pi(I)$ and $\Phi(I)$ are T -aggregable for any partition T which is finer than S , that is, $T \subseteq S$.

Also, the Kemeny-Snell characteristic of aggregable Markov chains suggests a weaker form of the CA-based aggregation model (1), involving conditional probabilities rather than Quetelet coefficients. According to this model, for any pair of classes, S_u, S_v of S , sums $\pi_{iv} = \sum_{j \in S_v} p_{ij}/p_{+j}$ ($i \in S_u$) must be approximated by unknown μ_{uv} according to criterion

$$\sum_{v=1}^m \sum_{u=1}^m \sum_{i \in S_u} p_{i+} (\pi_{iv} - \mu_{uv})^2 \quad (10)$$

The optimal μ_{uv} can be proven to form the aggregate Markov chain, $\Pi(S)$, thus leading to an aggregation S approximating properties of $\Pi(I)$ with regard to aggregability.

Theoretical and computational exploring of this criterion and its relation to the CA-induced criterion $X^2(S, S)$ yet remains to be done.

5. Conclusion

The paper provides an empirical aggregation method that is supported by two kinds of theoretical models of interactions governed by an aggregate interaction matrix. Each of the theoretical models suggests supplementary insights to the nature of the aggregation sought.

References

- [1] Benzécri, J.-P. (1973) *L'Analyse des Données*, Paris: Dunod.
- [2] Breiger, R.L. (1981) The social class structure of occupational mobility, *American Journal of Sociology*, 87, 578-611.
- [3] Featherman, D.L., and Hauser, R.M. (1978) *Opportunity and Change*, New York: Academic Press.
- [4] Greenacre, M.J. (1993) *Correspondence Analysis in Practice*, San Diego, Ca: Academic Press.
- [5] Goodman, L.A. (1981) Criteria for determining whether certain categories in a cross-classification table should be combined, with special reference to occupational categories in an occupational mobility table, *American Journal of Sociology*, 87, 612-650.
- [6] Govaert, G. (1989) La classification croisée, *La Revue de MODULAD*, 4, 9-36.
- [7] Kemeny, J.G., and Snell, J.L. (1976) *Finite Markov Chains*. New-York: Springer-Verlag.
- [8] Krymkowski, D., Sawinski, Z., and Domanski, H. (1996) Classification Schemes and the Study of Social Mobility, *Quality and Quantity*, 8.
- [9] Lebart, L., and Mirkin, B. (1993) Correspondence analysis and classification. *Multivariate Analysis: Future Directions 2*, C.M. Quadras and C.R. Rao (Eds.) Amsterdam: North-Holland, 341-357.
- [10] Lebart, L., Morineau, A., and Piron, M. (1995) *Statistique Exploratoire Multidimensionnelle*, Paris: Dunod.
- [11] Mirkin, B. (1996) *Mathematical Classification and Clustering*, Dordrecht: Kluwer Academic Press.
- [12] Quetelet, A. (1832) Sur la possibilité de mesurer l'influence des causes qui modifient les éléments sociaux, *Lettre à M. Willermé de l'Institut de France*, Bruxelles.