Control and Cybernetics

vol. **29** (2000) No. 4

On decomposition of $m \times m$ statistical tables based on grade dependence measures

by

Teresa Kowalczyk

Institute of Computer Science, Polish Academy of Sciences
Ordona 21, Warsaw, Poland

Abstract: The aim of the paper is two-fold. The first is to introduce, for any given partition of $\mathcal{R}^2$, the decomposition of Spearman's *rho* into three terms: *between*, *within* and *remainder*. This decomposition, presented in Section 4, is strictly connected with that of the concentration index *ar* as introduced in Kowalczyk (1998), and with the decomposition of Kendall's *tau* as introduced in Kowalczyk and Niewiadomska-Bugaj (2000). Those earlier results are reminded in Sections 2 and 3.

The second aim of the paper is to show and exemplify how one can use the decompositions of ρ^* (Spearman's *rho*) and τ (Kendall's *tau*) to analyse, decompose and compare $m \times m$ contingency tables with the same categories for the row variable and the column variable. The examples given in Section 5 concern social mobility tables with data from Britain and Poland. An important observation from the analysis of these data is that ρ^* and τ lead there to practically equivalent decompositions.

Keywords: aggregation, clustering, decomposition, dependence, grade correspondence analysis, Kendall's *tau*, regularity of dependence, Spearman's *rho*.

1. Introduction

Decomposition of $m \times k$ contingency tables is often considered in statistical literature. In the present paper we deal with decomposition based on the grade correspondence analysis (GCA). In Ciok et al. (1996), GCA is presented as a modification of the bivariate correspondence analysis maximizing the grade correlation coefficient ρ^*, called Spearman's *rho*. In the same paper there a *non-overlapping* aggregation of rows and of columns of an $m \times k$ table transformed by GCA is considered, and this operation is called grade correspondence-cluster analysis (GCCA). Various numerical aspects of such aggregation were then considered by Ciok (1998). There also exists a modification of GCA which maximizes Kendall's *tau*; this procedure and the related aggregation of rows and of columns considered in Kowalczyk and Niewiadomska-Bugaj (1998, 2000) is

Aggregation of rows and columns of a contingency table is linked to decomposition of related monotone dependence measures, in particular — to decompositions of ρ^* and τ. Especially interesting is the decomposition of *maximal values* of ρ^* and τ (obtained on the set of all pairs of permutations of rows and columns), denoted $\rho^*_{\max}$ and $\tau_{\max}$. A full account of the decomposition of τ and its implications for the decomposition of $\tau_{\max}$ is given in Kowalczyk and Niewiadomska-Bugaj (2000). A similar decomposition for ρ^* is considered in Kowalczyk (2000), mainly for the case when only one variable is aggregated. This is based on the decomposition of the concentration index ar and the corresponding decomposition of its maximal value $ar_{\max}$ which is obtained due to a suitable permutation of categories (Kowalczyk, 1998, Kowalczyk and Pleszczyńska, 1998).

The decompositions of ar and $ar_{\max}$ are reminded in Section 2, decomposition of τ in Section 3. Decomposition of ρ^* for non-overlapping aggregation of both variables is given in Section 4. This decomposition is especially useful in the case of $m \times m$ tables with the same categories of rows and columns and the same order imposed on the categories. Occupational mobility data which link occupations of fathers and sons may serve as an example. Usually one is interested in aggregation of such tables which provides identical and non-overlapping clusters of rows (fathers) and columns (sons).

There is a long tradition of analysing the father/son occupational status. The occupational mobility data form tables $\{n_{ij},\, i,j = 1,\ldots,m\}$, where n_{ij} is the number of pairs (father, son) with the first category i and the second category j. As the number of categories and their definitions vary from one study to another, it is difficult to compare the results. In Sections 5 and 6 we analyse two mobility data sets on father/son occupational status: a table with 7 categories given in Gifi (1990) and a table with 12 categories given in Pohoski (1983). These tables are first transformed into 6×6 tables with the same labels of six categories. Then, grade correspondence analysis GCA is performed for both 6×6 tables, and graphical displays in the form of over-representation maps (described in several papers, e.g. Ciok et al. (1994)) are presented and commented for each of them. A deeper insight into the common structure of these tables is obtained when the set of six ordered categories is non-overlappingly aggregated into three subgroups. This is done separately for each table. The clusters are chosen so that the resulting 3×3 tables have maximal strength of dependence under the condition that clustering is the same for fathers and for sons.

Decompositions of the 6×6 tables into 3×3 tables, based on ρ^* and τ separately, are described in Section 6 by partitioning both measures of dependence into the sum $B + W + R$.

2. Indices ar and $ar_{\max}$ and its decompositions

Let P, Q be probability measures defined on $(\Omega, \mathcal{B}(\Omega))$ and let $\varphi : \Omega \rightarrow \mathcal{R}$ be a well described by the concentration curve $C(Q\varphi^{-1} : P\varphi^{-1})$ (see e.g. Bamber, 1975, for real measures P, Q, and Kowalczyk, 1994, for the general case) defined on the square $[0,1]^2$ as the set

$$\begin{aligned} &C(Q\varphi^{-1} : P\varphi^{-1}) \\ &= \{(P(\{\omega : \varphi(\omega) \le z\}), Q(\{\omega : \varphi(\omega) \le z\}));\ z \in [-\infty, \infty]\} \end{aligned}$$

complemented, if necessary, by the points (0,0), (1,1) and by linear interpolation. The graph of $C(Q\varphi^{-1} : P\varphi^{-1})$ is a nondecreasing relation on the square $[0,1]^2$. As a special case we can take $\varphi = h = \frac{dQ}{dP}$ and we obtain convex curve $C(Qh^{-1} : Ph^{-1})$ which is equal to the Lorenz curve $L(Q : P)$, defined by Cifarelli and Regazzini (1987). For any φ we have

$$L(Q : P) = C(Qh^{-1} : Ph^{-1}) \le C(Q\varphi^{-1} : P\varphi^{-1}) \le C(Q\tilde{h}^{-1} : P\tilde{h}^{-1})$$

where $\tilde{h} = \frac{dP}{dQ} = \frac{1}{h}$.

If $\Omega = \mathcal{R}$ and $\varphi(x) = x$, the symbol φ will be omitted.

The curves $C(Q\varphi^{-1} : P\varphi^{-1})$ and $L(Q : P)$ lead to various *numerical* measures of monotone and absolute separation. Two of them are of particular importance: *the monotone separation index ar*

$$ar(Q\varphi^{-1} : P\varphi^{-1}) = 1 - 2\int_0^1 C(Q\varphi^{-1} : P\varphi^{-1})(u)\, du$$

and *the maximal separation index* $ar_{\max}$:

$$ar_{\max}(Q : P) = 1 - 2\int_0^1 L(Q : P)(u)\, du.$$

The decomposition of indices ar and $ar_{\max}$ into three terms: *between*, *within* and *remainder* is as follows:

THEOREM 1 (Kowalczyk, 1998) *Let P, Q be probability measures defined on a measurable space $(\Omega, \mathcal{A})$ and let*

$$h(\omega) = \frac{dQ}{dP}(\omega).$$

Let $\Omega_1, \ldots, \Omega_k$ be a partition of Ω, let $P_i = P|_{\Omega_i}$, $Q_i = Q|_{\Omega_i}$ be the conditional distributions on Ω_i, and let

$$K_i(x) = P_i(\{\omega : h(\omega) \le x\}),\ i = 1, \cdots, k.$$

Further, let $\alpha = (\alpha_1, ..., \alpha_k)$ and $\beta = (\beta_1, ..., \beta_k)$ be the distributions of the

(i) $ar_{\max} = ar^B_{\max}(P,Q) + ar^W_{\max}(P,Q) + ar^R_{\max}(P,Q)$ *where*

$$ar^B_{\max}(P,Q) = ar_{\max}(\alpha,\beta) = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}|\alpha_i\beta_j - \alpha_j\beta_i|,$$

$$ar^W_{\max}(P,Q) = \sum_{i=1}^{k}\alpha_i\beta_i ar_{\max}(P_i,Q_i).$$

$$ar^R_{\max}(P,Q) = \sum_{i=1}^{k}\sum_{j\neq i}\left(\alpha_i\beta_j ar(Q_j h^{-1} : P_i h^{-1}) - \frac{1}{2}|\alpha_i\beta_j - \alpha_j\beta_i|\right)$$

$$= 2\sum_{i=1}^{k}\sum_{j=1}^{i-1}\alpha_i\alpha_j\int_0^\infty K_i(x)(1-K_j(x))\,dx.$$

(ii) *for any measurable function* $\varphi : \Omega \to \mathcal{R}$,

$$ar(Q\varphi^{-1} : P\varphi^{-1})$$
$$= ar^B(Q\varphi^{-1} : P\varphi^{-1}) + ar^W(Q\varphi^{-1} : P\varphi^{-1}) + ar^R(Q\varphi^{-1} : P\varphi^{-1})$$

where

$$ar^B(Q\varphi^{-1} : P\varphi^{-1}) = ar(\tilde{Q}\varphi^{-1} : P\varphi^{-1}) \text{ for } \tilde{Q} = \sum_{i=1}^{k}\beta_i P_i,$$

$$ar^W(Q\varphi^{-1} : P\varphi^{-1}) = \sum_{i=1}^{k}\alpha_i\beta_i ar(Q_i\varphi^{-1} : P_i\varphi^{-1}),$$

$$ar^R(Q\varphi^{-1} : P\varphi^{-1})$$
$$= \sum_{i=1}^{k}\sum_{j\neq i}\alpha_i\beta_j(ar(Q_j\varphi^{-1} : P_i\varphi^{-1}) - ar(P_j\varphi^{-1} : P_i\varphi^{-1})).$$

These decompositions have the property that the reminder term is equal to zero if the sets $\{h(\Omega_j)\}$ are non-overlapping. This occurs for $ar_{\max}$ if there exists a permutation $(i_1,\ldots,i_k)$ of $(1,\ldots,k)$ such that

$$h(\Omega_{i_1}) \prec \cdots \prec h(\Omega_{i_k}),$$

and for ar under an analogous condition with h replaced by φ.

The between term in the decomposition of $ar_{\max}$ is equal to $ar_{\max}$ applied to the aggregated table. In case of ar, the between term is equal to ar of the aggregated table if this aggregation is non-overlapping.

To illustrate, let P and Q be defined on $\Omega = \{1,2,\ldots,6\}$ by and

$$(q_1, \dots, q_6) = (0.1025, 0.1311, 0.04, 0.4338, 0.1183, 0.1746).$$

Then

$$ar(Q : P) = 0.2206,\ ar_{\max}(Q : P) = 0.3073.$$

When the categories are aggregated non-overlappingly (with respect to the initial order) onto $(1, 2, 3)$, (4), $(5, 6)$, we obtain:

$$\begin{aligned}
&\alpha = (\alpha_1, \alpha_2, \alpha_3) = (.409, .4256, .1657),\\
&\beta = (\beta_1, \beta_2, \beta_3) = (.2736, .4338, .2929),\\
&ar^B(Q : P) = ar^B_{\max}(Q : P) = ar_{\max}(\beta : \alpha) = ar(\beta : \alpha) = 0.1882,\\
&ar^W(Q : P) = 0.0325,\ ar^W_{\max}(Q : P) = 0.0496,
\end{aligned}$$

the remainder term in the decomposition of $ar(Q : P)$ is zero $(ar^R(Q : P) = 0)$, while this term in the decomposition of $ar_{\max}(Q : P)$ is $ar^R_{\max} = 0.0694$ since the sets $\{h(\Omega_i)\}$ overlap.

3. Kendall's τ and its decomposition

Let (X, Y) be any pair of random variables on $(\Omega, \mathcal{B}, P)$ with joint distribution H and marginal cdf's (right continuous) F, G, respectively. Let

$$\widetilde{H}(x, y) = \frac{1}{4}(H(x-, y-) + H(x-, y) + H(x, y-) + H(x, y))$$

and consequently

$$\widetilde{F}(x) = \frac{1}{2}(F(x-) + F(x)),\ \widetilde{G}(y) = \frac{1}{2}(G(y-) + G(y)).$$

Kendall's *tau* is defined by

$$\tau(X, Y) = 4E(\widetilde{H}(X, Y)) - 1$$

and can be expressed as functions of index ar applied to all pairs of conditional distributions (see Kowalczyk, 2000):

$$\begin{aligned}
\tau(X, Y) &= 2\int_{x<t}\int ar(P_{Y|X=t} : P_{Y|X=x})\, dF(x)\, dF(t)\\
&= 2\int_{y<z}\int ar(P_{X|Y=z} : P_{X|Y=y})\, dG(y)\, dG(z).
\end{aligned}$$

By replacing indices ar by $ar_{\max}$, we obtain index τ_{abs} which was considered in Kowalczyk (2000) and used there to construct a measure of departure from

Lehmann (1966) has formulated a few increasingly strong conditions expressing intuitions of monotone (positive or negative) dependence of bivariate distributions. In effect these are conditions expressing increasingly strong regularity of distributions. The strongest condition given by Lehmann is called "total positivity of order two" (TP_2). When expressed by means of the density function, it has the form

$$f(x,y)f(x',y') \geq f(x',y)f(x,y') \text{ for any } x < x',\ y < y'.$$

As noticed in Kowalczyk (2000), the distribution is TP_2 if and only if all indices ar for ordered pairs of conditional distributions are equal to $ar_{\max}$. Basing on this fact, a measure of departure from TP_2 of a distribution transformed by GCA was introduced in the above mentioned paper; this measure is defined as the suitably normalized difference of indices $\tau_{abs}(X,Y)$ and $\tau_{\max}(X,Y)$. In the general case, it is sufficient to replace index $\tau_{\max}$ in this expression by $|\tau(X,Y)|$ to measure how distant from TP_2 is the distribution of (X,Y) or of $(X,-Y)$ (it depends on in which of them the variables are positively dependent).

Decomposition of Kendall's *tau* was presented in Kowalczyk and Niewiadomska-Bugaj (1999, 2000). Here we restrict ourselves to reminding the case of non-overlapping aggregation of both variables.

THEOREM 2 (Kowalczyk and Niewiadomska-Bugaj, 2000) *Let Ω_{ij} be a partition of Ω where $\Omega_{ij} = \Omega_i \cap \Omega'_j = \mathcal{X}_i \times \mathcal{Y}_j$, $\Omega_i = \mathcal{X}_i \times \mathcal{Y}$, $i = 1, \ldots, M$, $\Omega'_j = \mathcal{X} \times \mathcal{Y}_j$, $j = 1, \ldots, K$, $\mathcal{X}_1 \prec \cdots \prec \mathcal{X}_M$, $\mathcal{Y}_1 \prec \cdots \prec \mathcal{Y}_K$. Let $(X_i, Y_i) = (X,Y)|_{\Omega_i}$, $(X'_j, Y'_j) = (X,Y)|_{\Omega'_j}$, $(X_{ij}, Y_{ij}) = (X,Y)|_{\Omega_{ij}}$, $\alpha_i = P(\mathcal{X}_i)$, $\beta_j = P(\mathcal{Y}_j)$, $\gamma_{ij} = P(\Omega_{ij}) > 0$, F_i, G_i, F'_j, G'_j, F_{ij}, G_{ij} be the marginal distribution functions of X_i, Y_i, X'_j, Y'_j, and X_{ij}, Y_{ij} for $i = 1, \ldots, M$, $j = 1, \ldots, K$, and let $T_{M\times K}$ be the aggregated table $[\gamma_{ij}]$. Let (X^o, Y^o), (X'^o, Y'^o) and (X^{oo}, Y^{oo}) be pairs of random variables with the distribution functions $H^o(x,y) = \sum_{i=1}^M \alpha_i F_i(x) G_i(y)$, $H'^o(x,y) = \sum_{j=1}^K \beta_j F'_j(x) G'_j(y)$, $H^{oo}(x,y) = \sum_{i=1}^M \sum_{j=1}^K \gamma_{ij} F_{ij}(x) G_{ij}(y)$, respectively, and let $\tau^B(X,Y) = \tau(X^{oo}, Y^{oo})$. Then*

$$\tau(X,Y) = \tau^B(X,Y;\{\Omega_{ij}\}) + \tau^W(X,Y;\{\Omega_{ij}\}),$$

where

$$\begin{aligned}
&\tau^B(X,Y;\{\Omega_{ij}\}) = \tau^{BB}(X,Y;\{\Omega_{ij}\})\\
&+\sum_{i=1}^{M} \alpha_i^2 \tau^B(X_i,Y_i;\{\Omega_{ij}\}_{j=1}^K) + \sum_{j=1}^{K} \beta_j^2 \tau^B(X'_j,Y'_j;\{\Omega_{ij}\}_{i=1}^M),\\
&\tau^B(X_i,Y_i;\{\Omega_{ij}\}_{j=1}^K) = 2\sum_{j<s} \frac{\gamma_{ij}\gamma_{is}}{\alpha_i^2} a_C(F_{ij},F_{is}),\\
&\tau^B(X'_j,Y'_j;\{\Omega_{ij}\}_{i=1}^M) = 2\sum \frac{\gamma_{ij}\gamma_{sj}}{\beta_j^2} a_C(G_{ij},G_{sj}),
\end{aligned}$$

$$\tau^{BB}(X,Y;\{\Omega_{ij}\}) = 2\sum_{j<s}\Big(\sum_{i<t}\gamma_{ij}\gamma_{ts} - \sum_{i>t}\gamma_{ij}\gamma_{ts}\Big) = \tau(T_{M\times K})$$
$$= \tau^{B}(X^{o},Y^{o};\{\Omega'_j\}) = \tau^{B}(X'^{o},Y'^{o};\{\Omega_i\}).$$
$$\tau^{W}(X,Y;\{\Omega_{ij}\}) = \sum_{i=1}^{M}\sum_{j=1}^{K}\gamma_{ij}^2\tau(X_{ij},Y_{ij}) = \sum_{i=1}^{M}\alpha_i^2\tau^{W}(X_i,Y_i;\{\Omega_{ij}\}_{j=1}^{K})$$
$$= \sum_{j=1}^{K}\beta_j^2\tau^{W}(X'_j,Y'_j;\{\Omega_{ij}\}_{i=1}^{M}).$$

We see that in this decomposition the "between" term deals only with the marginal distributions of X and Y, with the marginal distributions of the suitably chosen subtables and with the aggregated table $T_{M\times K}$. The "within" term takes into account joint distributions in the suitably chosen subtables. Moreover, this decomposition has a property, presented in the sequel, which is analogous as in the decomposition with aggregated one variable, X or Y. Let B and W denote the between and within terms in decompositions of $\tau(X,Y)$ with two variables X and Y which are aggregated, and let B_X, W_X, B_Y, W_Y denote the between and within terms in decomposition with respect to X or with respect to Y. The terms B_X, W_X can be decomposed with respect variable Y onto $B_XB_Y + B_XW_Y$, and $W_XB_Y + W_XW_Y$, respectively. Similarly, we introduce symbols $B_YB_X + B_YW_X$ and $W_YB_X + W_YW_X$.

By Theorem 2, these terms fulfil the equalities:

$$B_XB_Y = B_YB_X = \tau^{BB},\ W_XW_Y = W_YW_X = \tau^{W},$$
$$W_XB_Y = B_YW_X,\ W_YB_X = B_XW_Y.$$

Thus, we have $B = B_X + B_Y - B_XB_Y$, i.e.

$$\tau^{B}(X,Y;\{\Omega_{ij}\}) = \tau^{B}(X,Y;\{\Omega_i\}) + \tau^{B}(X,Y;\{\Omega'_j\}) - \tau^{BB}(X,Y).$$

This will be exemplified on two tables considered in Section 6.

4. Spearman's *rho* and its decomposition

We start with the notions of the regression and correlation functions, needed in the decompositions of ρ^* and $\rho^*_{\max}$ when only X is aggregated.

Let $r_{\widetilde{G}(Y)|X}$ and $C_{cor}[\widetilde{G}(Y)|X]$ be the *regression* and *correlation functions* of $\widetilde{G}(Y)$ on X (called also the grade regression and grade correlation of Y on X), defined by

$$r_{\widetilde{G}(Y)|X}(x) = E(\widetilde{G}(Y)|X = x)$$
$$C_{cor}[\widetilde{G}(Y)|X](x) = \frac{E(\widetilde{G}(Y); X \le x)}{} = 2\int^{x} r_{\widetilde{G}(Y)|X}(z)\,dF(z).$$

Since

$$r_{\widetilde{G}(Y)|X}(x) \geq 0, \ \lim_{x\to-\infty} C_{cor}[\widetilde{G}(Y)|X](x) = 0, \ \lim_{x\to\infty} C_{cor}[\widetilde{G}(Y)|X](x) = 1,$$

the correlation curve can be treated as a distribution function on $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$. Let P_X, Q_X be probability measures corresponding to F and $C_{cor}[\widetilde{G}(Y)|X]$, respectively. Let $x_u = \inf\{x : F(x) \geq u\}$ for $u \in (0,1)$, and $\rho^*_{1\,\max}(X,Y)$ denote the maximal value of rho^* obtained under all measurable transformations of variable X. Let

$$C^*_{cor}[\widetilde{G}(Y)|X](u) = 2\int_0^u r_{\widetilde{G}(Y)|X}(x_t)\,dt.$$

The Spearman's coefficients $\rho^*(X,Y)$ and $\rho^*_{1\,\max}(X,Y)$ fulfil the equalities

$$\begin{aligned}\rho^*(X,Y) &= 3\Big(1 - 2\int_0^1 C^*_{cor}[\widetilde{G}(Y)|X](u)\,du\Big)\\ &= 3\Big(1 - 2\int_0^1 C(Q_X : P_X)(u)\,du\Big)\end{aligned}$$

and consequently

$$\begin{aligned}\rho^*(X,Y) &= 3ar(Q_X : P_X)\\ &= 6\int_{t<x}\int (r_{\widetilde{G}(Y)|X}(x) - r_{\widetilde{G}(Y)|X}(t))\,dF(t)\,dF(x)\\ &= 3\Big(2\int_{-\infty}^{\infty} \widetilde{F}(x) r_{\widetilde{G}(Y)|X}(x)\,dF(x) - 1\Big);\\ \rho^*_{1\,\max}(X,Y) &= ar_{\max}(Q_X : P_X) = \Big(1 - 2\int_0^1 L(Q_X : P_X)(u)\,du\Big)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |r_{\widetilde{G}(Y)|X}(x) - r_{\widetilde{G}(Y)|X}(t)|\,dF(t)\,dF(x)\end{aligned}$$

i.e. $ar_{\max}(Q_X : P_X)$ is the Gini index of random variable $r_{\widetilde{G}(Y)|X}(X)$.

Since

$$\frac{dQ_X}{dP_X}(x) = 2r(x),$$

then $3ar_{\max}(Q_X : P_X) = \rho^*(X,Y)$ if and only if $r_{\widetilde{G}(Y)|X}(x)$ is non-decreasing.

We remind also the formula expressing ρ^* as a function of concentration indices for pairs of conditional distributions and marginal distributions:

$$\begin{aligned}\rho^*(X,Y) &= 3\int_{-\infty}^{\infty} \widetilde{F}(x) ar(P_{Y|X=x} : P_Y)\,dF(x)\\ &= 3\int^{\infty} \widetilde{G}(y) ar(P_{X|Y=y} : P_X)\,dG(y).\end{aligned}$$

Now we turn to the decomposition of ρ^*. When we deal with aggregation of only one variable, say X, we can use the decomposition of the concentration index ar (reminded in Section 2), applied to the distribution of X and of the distribution of random variable with (generalized) density equal to the grade regression function $2E(\widetilde{G}(Y)|X = x)$.

PROPOSITION 1 *Let us denote $r_{\widetilde{G}(Y)|X}(x) = r(x)$. Let the support $\mathcal{X}$ of random variable X be partitioned onto M disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_M$, $P_X(\mathcal{X}_i) = \alpha_i$, $(X_i, Y_i) = (X, Y)|_{\mathcal{X}_i \times \mathcal{Y}}$ and let F_i be the cdf of X_i, $F_i^r = F_i \circ r^{-1}$, and $R_i = E(r(X)|X \in \mathcal{X}_i) = E(r(X_i)) = E(\widetilde{G}(Y_i))$ for $i = 1, \ldots, M$. Then*

(i)

$$\rho^*_{1\,\max}(X,Y) = 3ar^B_{\max}(Q_X : P_X) + 3ar^W_{\max}(Q_X : P_X) + 3ar^R_{\max}(Q_X : P_X)$$

where

$$ar^B_{\max}(Q_X : P_X) = \sum_{i=1}^{M}\sum_{j=1}^{M} \alpha_i \alpha_j |R_i - R_j|,$$

$$ar^W_{\max}(Q_X : P_X) = \sum_{i=1}^{M} \alpha_i^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |r(x) - r(t)|\, dF_i(t)\, dF_i(x),$$

$$ar^R_{\max}(Q_X : P_X)$$
$$= \sum_{s=1}^{M}\sum_{l \neq s} \alpha_l \alpha_s \left(\int\int |r(t) - r(x)|\, dF_s(t)\, dF_l(x) - |R_s - R_l| \right)$$

(ii)

$$\rho^*(X,Y) = 3ar^B(Q_X : P_X) + 3ar^W(Q_X : P_X) + 3ar^R(Q_X : P_X)$$

where

$$ar^B(Q_X : P_X) = \sum_{i=1}^{M} \alpha_i ar(F_i : F) ar(G_i : G),$$

$$ar^W(Q_X : P_X) = 2\sum_{i=1}^{M} \alpha_i^2 \int_{t<x}\int (r(x) - r(t))\, dF_i(t)\, dF_i(x)$$

$$= 4\sum_{i=1}^{M} \alpha_i^2 \int_{-\infty}^{\infty} \widetilde{F}_i(x)(r(x) - R_i)\, dF_i(x),$$

$$ar^R(Q_X : P_X) = 4\sum^{M}\sum \alpha_i \alpha_j \int \tilde{F}_i(x)(r(x) - R_j)\, dF_j(x).$$

Let us note that $R_1, \ldots, R_M$ are the values of regression of $\widetilde{G}(Y)$ on X_a, where X_a is equal to X after aggregation, taking values $1, \ldots, M$ and such that the distribution function $H_a(i,y)$ of (X_a, Y) is equal to

$$dH_a(i,y) = \int_{\mathcal{X}_i} dH(x,y),\ i = 1, \ldots, M,\ y \in \mathcal{R};$$

on the other hand, $R_1, \ldots, R_M$ are the values of regression of random variable $\widetilde{G}(Y^o)$ on X^o, where (X^o, Y^o) is a pair of random variables such that $(X_i^o, Y_i^o) \overset{df}{=} (X^o, Y^o)|_{\mathcal{X}_i}$ are independent and $X_i^o \sim X_i, Y_i^o \sim Y_i$ for $i = 1, \ldots, M$ (i.e. the distribution function H^o of (X^o, Y^o) is of the form $H^o(x,y) = \sum_i \alpha_i F_i(x) G_i(y)$). Hence, we have

$$3ar^B_{\max}(Q_X : P_X) = \rho^*_{1\max}(X^o, Y^o) = \rho^*_{1\max}(X_a, Y),$$
$$3ar^B(Q_X : P_X) = \rho^*(X^o, Y^o).$$

COROLLARY 1 (i) $ar^R_{\max}(Q_X : P_X) \geq 0$; *for any pair of random variables* (X,Y), *with equality holding if* (X_i, Y_i) *are independent or if there exists a permutation* $(i_1, \ldots, i_M)$ *of* $(1, \ldots, M)$ *such that*

$$r(\{\mathcal{X}_{i_1}\}) \prec \cdots \prec r(\{\mathcal{X}_{i_M}\}), \tag{1}$$

if $\rho^*(X,Y)$ *is equal to the maximal value* $\rho^*_{1\max}$, *condition* (1) *is equivalent to* $\mathcal{X}_{i_1} \prec \cdots \prec \mathcal{X}_{i_k}$;

(ii) *If* $r(X_i) = const = R_i$ *for* $i = 1, \ldots, M$ *then* $ar^B_{\max}(Q_X : P_X) = \rho^*_{1\max}(X,Y)$, $ar^W_{\max}(Q_X : P_X) = ar^R_{\max}(Q_X : P_X) = 0$;

(iii) *Let* $S^*_i = \sum_{s=1}^i \alpha_s$; *if* $\rho^*(X,Y) = \rho^*_{1\max}(X,Y)$ *and* $\mathcal{X}_1 \prec \cdots \prec \mathcal{X}_M$, *then decompositions* (i) *and* (ii) *in Proposition* 1 *are identical, and*

$$3ar^B_{\max}(Q_X : P_X) = 3\sum_{i=1}^M \alpha_i (S^*_i + S^*_{i-1}) ar(G_i : G) = \rho^*(X_a, Y).$$

Note that Proposition 1 and Corollary 1 can be analogously rewritten when Y is aggregated instead of X, with obvious changes in notation, and with $T^*_j = \sum_{s=1}^j \beta_j$, $j = 1, \ldots, K$, replacing S^*_i, $i = 1, \ldots, M$.

We note that the decomposition of $\rho^*(X,Y)$ and $\rho^*_{1\max}$ given in Proposition 1 in case of non-overlapping sets is equivalent to the decomposition of $cov(\widetilde{F}(X), \widetilde{G}(Y))$ into $cov^B(\widetilde{F}(X), \widetilde{G}(Y)) + cov^W(\widetilde{F}(X), \widetilde{G}(Y))$.

Suppose now that both variables (X,Y) are non-overlappingly aggregated (the general case for any partition of $\mathcal{R}^2$ is considered in Kowalczyk, 2000). In the decomposition of ρ^* related to this case we will use the notation introduced in Theorem 2 and also the expressions S^*_i, T^*_j related to Corollary 1.

The general decomposition of ρ^* for non-overlapping aggregation of X and Y is of the form $B + W$ (and similar to decomposition with respect to one variable, The between term B is equal to $\rho^*(X^{oo}, Y^{oo})$ so that

$$B = 3\sum_{i=1}^{M}\sum_{j=1}^{K}\gamma_{ij}ar(P_{X_{ij}} : P_X)ar(P_{Y_{ij}} : P_Y) = \rho^{*B}(X,Y)$$

and it is further decomposed. We can express the elements of B as:

$$B = B_X + B_Y - BB + c$$

where

$$B_X = 3\sum_i \alpha_i(S_i^* + S_{i-1}^* - 1)ar(P_{Y_i} : P_Y) = \rho^*(X_a, Y)$$

$$B_Y = 3\sum_j \beta_j(T_j^* + T_{j-1}^* - 1)ar(P_{X_j'} : P_X) = \rho^*(X, Y_a)$$

$$BB = 3\sum_i\sum_j \gamma_{ij}(S_i^* + S_{i-1}^* - 1)(T_j^* + T_{j-1}^* - 1) = \rho^*(X_a, Y_a) = \rho^{*BB}$$

$$c = 3\sum_i\sum_j \gamma_{ij}\alpha_i\beta_j ar(P_{X_{ij}} : P_{X_i})ar(P_{Y_{ij}} : P_{Y_j'}).$$

The within term W is equal to

$$W = \sum_{i=1}^{M}\sum_{j=1}^{K}\gamma_{ij}cov(\widetilde{F}(X_{ij}), \widetilde{G}(Y_{ij}) = cov^W(\widetilde{F}(X), \widetilde{G}(Y); \{\Omega_{ij}\})$$

and can be presented as

$$W = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{K}\gamma_{ij}\int ar(P_{Y_{ij}|X_{ij}=x} : P_Y)\widetilde{F}(x)\,dF_{ij}(x)$$

$$-\frac{1}{4}\sum_{i=1}^{M}\sum_{j=1}^{K}\gamma_{ij}ar(P_{X_{ij}} : P_X)ar(P_{Y_{ij}} : P_Y).$$

Analogously as in Section 3, W_X, W_Y will be defined by $\rho^*(X,Y) = B_X + W_X = B_Y + W_Y$; let $B_X = B_XB_Y + B_XW_Y$, and $W_X = W_XB_Y + W_XW_Y$, and similarly for definitions of B_Y and W_Y.

We have

$$B_XB_Y = B_YB_X = BB = \rho^{*BB}(X,Y),\ W_XW_Y = W_YW_X = W,$$
$$B_XW_Y = B_X - BB,\ W_YB_X = B_X - BB + c,\ B_YW_X = B_Y - BB$$

and $W_XB_Y = B_Y - BB + c$.

These formulas for the decomposition of ρ^* in case of non-overlapping aggregation of X and of Y are *not* quite analogous to such formulas given in Section 3 for τ unless the distribution of (X,Y) is concentrated on $\bigcup_{j=1}^{k}(\mathcal{X}_j \times \mathcal{Y}_j)$, when

5. Decomposition and comparison of two data sets on the father/son occupational status

The first data set is based on a sample of 3497 families in Britain. The study resulted in a 7 × 7 table $BRIT_{7\times 7}$ on the relationship between father's occupational status and son's status. These categories are the following: PROF (professional and high administrative), EXEC (managerial and executive), HSUP (higher supervisory), LSUP (lower supervisory), SKIL (skilled manual and routine nonmanual), SEMI (semi-skilled manual), and UNSK (unskilled manual). The seven categories are ordered according to the *social prestige* scale, from high to low. It happened that this ordering is strictly preserved by the grade correspondence analysis (both for rows and for columns) which means that the strongest monotone trend in the $BRIT_{7\times 7}$ is concordant with the prestige scale: sons of a father in category i at this scale *tend* to be in categories i, $i+1$ or $i-1$, i.e. preserve fathers occupation or choose one close to it on the prestige scale. In terms of GCA, this means that the initial table with rows and columns ordered according to the prestige scale is that one which maximizes the value of ρ^* as well as of τ in the whole set of tables with arbitrary permutations of rows and columns. We obtain $\rho^*(BRIT_{7\times 7}) = 0.3720$ and $\tau(BRIT_{7\times 7}) = 0.2566$.

The second set of data taken from Pohoski (1983), is a father/son occupational table $POH_{12\times 12}$ for 8767 families in Poland, analogous to $BRIT_{7\times 7}$, but possessing 12 categories which strongly differ from 7 categories considered in $BRIT_{7\times 7}$. The initial ordering introduced by Pohoski is preserved by GCA *neither for ρ^* nor for τ*; the optimal GCA permutations of father's and son's categories are *different for rows and for columns* and, moreover, differ for GCA's based on ρ^* and on τ. Thus, the monotone trend in $POH_{12\times 12}$ is less regular than in $BRIT_{7\times 7}$. It is not concordant with the prestige scale and also weaker: $\rho^*_{\max}(POH_{12\times 12}) = 0.324$, $\tau_{\max}(POH_{12\times 12}) = 0.223$, where $\rho^*_{\max}$ and $\tau_{\max}$ mean ρ^* and τ for $POH_{12\times 12}$ transformed by GCA.

To compare the $BRIT$ and POH data, an attempt was made to bring into agreement the labels of categories by suitable aggregation in both sets and rejection in POH. Two adjacent categories were aggregated in $BRIT_{7\times 7}$, namely HSUP+LSUP called SUP. In the set of categories specified by Pohoski, two categories referring to farms (farm owners and farm workers) were rejected, and the remaining ones were aggregated into six categories which are hoped to correspond to PROF, EXEC, SUP, SKIL, UNSK, and SEMI in the $BRIT$ data. Consequently, two 6 × 6 tables with rows and columns identically labelled were formed and each of them was transformed twice by the GCA based on ρ^* and on τ. As before, the table $BRIT_{6\times 6}$ remained unchanged under both GCA's and thus still concordant with the prestige scale: PROF, EXEC, SUP, SKIL, SEMI, UNSK. Dependence strength decreased slightly due to the aggregation of HSUP and SUP: 0.3596 instead of 0.3720 for ρ^* and 0.2469 instead of 0.2566 for τ.

Turning to the POH data, we find that the permutations due to GCA based sons is (PROF, EXEC, SUP, SKIL, UNSK, SEMI), (PROF, EXEC, SKIL, SUP, UNSK, SEMI), i.e. the two orderings are slightly different. The value of ρ^* for this table is 0.2061, the value of τ is 0.1381. An interchange of categories SKIL and SUP for sons results in a slight decrease of ρ^* and τ : 0.1972 and 0.1322,

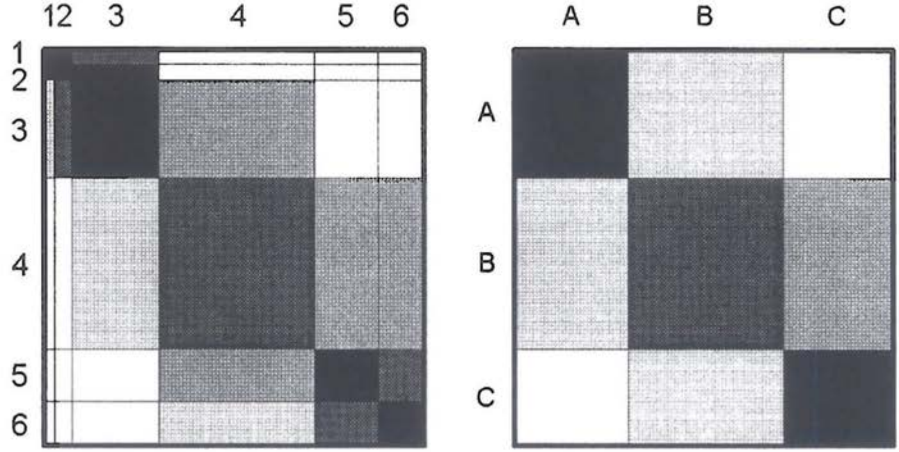

Figure 1. Visualization of $BRIT_{6\times6}$ and $BRIT_{3\times3}$. 1 — PROF, 2 — EXEC, 3 — SUP, 4 — SKIL, 5 — SEMI, 6 — UNSK, A = PROF + EXEC + SUP, B = SKIL, C = SEMI + UNSK.

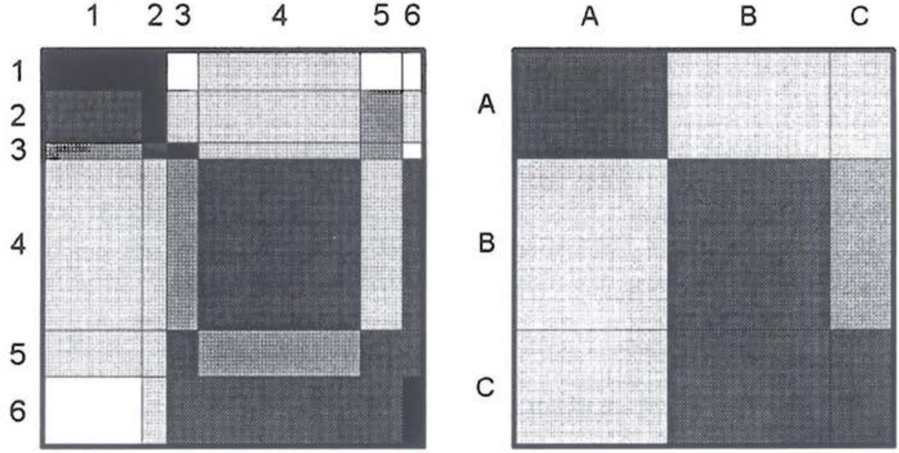

Figure 2. Visualization of $POH_{6\times6}$ and $POH_{3\times3}$. 1 — PROF, 2 — EXEC, 3 — SUP, 4 — SKIL, 5 — SEMI, 6 — UNSK, A = PROF + EXEC + SUP,

respectively, and this permutation maximizes both ρ^* and τ in the set of tables with rows and columns identically permuted. The corresponding table will be denoted here $POH_{6\times6}$.

The tables $BRIT_{6\times6}$ and $POH_{6\times6}$ are graphically displayed in Figs. 1 and 2 in a way described in detail e.g. in Ciok et al. (1994): black and dark grey rectangles correspond to combinations of father/son categories which are strongly and weakly overrepresented, white and light grey rectangles to combinations of categories which are strongly and weakly underrepresented, white rectangles with black vertical lines are close to *fair representation.*

It is seen that the monotone trend in $BRIT_{6\times6}$ is very regular and the table is almost symmetric. In particular, the marginal distributions for fathers and for sons are very similar; the respective concentration index ar measuring differentiation of marginals is merely 0.06. Table $POH_{6\times6}$ is not quite monotone and less symmetric. Category 1 (PROF) is much more frequent for sons than for fathers, categories 2 and 5 (EXEC and SEMI) more frequent for fathers than for sons; the respective concentration index for marginals is 0.31. The marginals for POH and $BRIT$ are not similar (cf. the example in Section 2).

6. Decompositions of the occupational mobility tables

As announced in the Introduction, we optimally aggregate tables $BRIT_{6\times6}$ and $POH_{6\times6}$ into 3×3 tables. For $BRIT_{6\times6}$, which has rows and columns optimally permuted (w.r.t. ρ^* and w.r.t. τ), we get $BRIT_{3\times3}$ as presented in Fig. 1, with the following decompositions in which $\rho^*(T_{3\times3})$ and $\tau(T_{3\times3})$ for $BRIT$ and POH correspond to the term BB :

$\rho^*(BRIT_{3\times3}) = 0.3158 = 87.8\%$	$\tau(BRIT_{3\times3}) = 0.2132 = 86.4\%$
$B = 0.3531 = 98.2\%$	$B = 0.2395 = 97\%$
$W = 0.0065 = 1.8\%$	$W = 0.0074 = 3\%$
$R = 0$	$R = 0$
$\rho^*(BRIT_{6\times6}) = 0.3596 = 100\%$	$\tau(BRIT_{6\times6}) = 0.2469 = 100\%$

In particular, in decomposition of $\rho^*(BRIT_{6\times6})$ we have: $B_X = 0.3334$, $B_Y = 0.3352$, $c = 0.00046$. In decomposition of $\tau(BRIT_{6\times6})$ we have: $B_X = 0.2254$, $B_Y = 0.2274$. The measure of departure $BRIT_{6\times6}$ from TP_2 takes the value 0.0214.

We see that the quotient $\rho^*(BRIT_{3\times3})/\rho^*(BRIT_{6\times6})$ is very high, and it is very close to the analogous quotient for τ; the same concerns the terms B. This is another proof of regularity of $BRIT_{6\times6}$. The interpretation of the three optimal clusters is obvious and intuitively convincing. We also see that $\rho^*(BRIT_{6\times6})/\tau(BRIT_{6\times6}) = 1.4565$, $\rho^*(BRIT_{3\times3})/\tau(BRIT_{3\times3}) = 1.4812$.

Turning to $POH_{3\times3}$, we have three optimal clusters slightly different than in $BRIT_{3\times3}$ (see Fig. 2). The decomposition of $POH_{6\times6}$ is slightly less transparfor ρ^* and for τ are almost identical as in the $BRIT$ data, and the quotients of the respective values for ρ^* and τ are very close to 3/2. We have:

$\rho^*(POH_{3\times3}) = 0.1837 = 93\%$	$\tau(POH_{3\times3}) = 0.1230 = 93\%$
$B = 0.1950 = 98.9\%$	$B = 0.1303 = 98.6\%$
$W = 0.0022 = 1.1\%$	$W = 0.0019 = 1.4\%$
$R = 0$	$R = 0$
$\rho^*(POH_{6\times6}) = 0.1972 = 100\%$	$\tau(POH_{6\times6}) = 0.1322 = 100\%$

In particular, in decomposition of $\rho^*(POH_{6\times6})$ we have: $B_X = 0.1792$, $B_Y = 0.1717$, $c = 0.0015$. In decomposition of $\tau(POH_{6\times6})$ we obtain: $B_X = 0.1196$, $B_Y = 0.1149$. The measure of departure $POH_{6\times6}$ from TP_2 takes the value 0.1271.

In this case $\rho^*(POH_{6\times6})/\tau(POH_{6\times6}) = 1.4917$, $\rho^*(POH_{3\times3})/\tau(POH_{3\times3}) = 1.4935$.

References

BAMBER, D. (1975) The area above the ordinal dominance graph and area below the receiver operating characteristic graph. *J. Math. Psych.*, **12**, 387–415.

CIFARELLI, D.M. and REGAZZINI, E. (1987) On a general definition of concentration function. *Sankhya B*, **49**, 307–319.

CIOK, A. (1998) Discretization as a tool in cluster analysis. In: *Advances in Data Science and Classification, Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Rome, July 21–24, 1998*, Rizzi, A., Vichi, M. and Bock, H., eds., 349–354.

CIOK, A., KOWALCZYK, T. and PLESZCZYŃSKA, E. (1998) How a New Statistical Infrastructure Induced a New Computing Trend in Data Analysis. In: *Rough Sets and Current Trends in Computing*, Polkowski, L. and Skowron, A., eds., *Lecture Notes in Artifical Intelligence*, 1424, *Proceedings of the First International Conference, RSCTC'98, Warsaw, Poland, June 22–26, 1998*, Springer, 75–82.

CIOK, A., KOWALCZYK, T., PLESZCZYŃSKA, E. and SZCZĘSNY, W. (1994) Visualization of a two-way table as a Data Mining Tool. *Intelligent Information Systems III, Proceedings of the Workshop held in Wigry, Poland, 6–10 June, 1994.*

CIOK, A., KOWALCZYK, T., PLESZCZYŃSKA, E. and SZCZĘSNY, W. (1995) Algorithms of grade correspondence-cluster analysis. *The Collected Papers on Theoretical and Applied Computer Science*, **7**, 5–22.

GIFI, A. (1990) *Nonlinear multivariate analysis.* J. Wiley & Sons, New York.

KOWALCZYK, T. (1994) A unified Lorenz-type approach to divergence and dependence. *Dissertationes Math.*, 335.

KOWALCZYK, T. (1998) Decomposition of the Concentration Index and its implications for the Gini inequality and dependence measures. *Statistics in*

KOWALCZYK, T. (2000) Link between grade measures of dependence and of separability in pairs of conditional distributions. *Statistics and Probability Letters*, **46** (4), 371–379.

KOWALCZYK, T. (2000) Decomposition of Spearman's *rho* with implication to clustering. Manuscript.

KOWALCZYK, T. and NIEWIADOMSKA-BUGAJ, M. (1998) Grade correspondence analysis based on Kendall's tau. *Advances in Data Science and Classification, 6th Conference of the International Federation of Classification Societies (IFCS-98), Short Papers, Rome, 21–24 July*, Instituto Nazionale di Statistica, Rome, 182–185.

KOWALCZYK, T. and NIEWIADOMSKA-BUGAJ, M. (2000) Decomposition of Kendall's *tau*: implications for clustering. *Statistics and Probability Letters*, **48**, 375–383.

KOWALCZYK, T. and NIEWIADOMSKA-BUGAJ, M. (2001) An algorithm for maximizing Kendall's *tau*. (To appear in *Computational Statistics and Data Analysis.*)

LEHMANN, E.L. (1966) Some concepts of dependence. *Ann. Math. Statist.*, **37**, 1137–1153.

POHOSKI, M. (1983) Social mobility and social inequalities (in Polish). *Kultura i Społeczeństwo*, **27**, 135–164.