ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 141, 1997

*Janusz Wywiał**

DECOMPOSITION OF TIME SERIES ON THE BASIS OF MODIFIED GROUPING METHOD OF WARD[1]

Abstract. The trend of time series can change its direction. It is assumed that the time interval is divided into subintervals where the trend is given as particular linear function. The problem is how to divide the observation of time series into disjoint and coherent groups where they have linear trend.

That is why the problem of the scatter of multivariable observation was first considered. The degree of data spread is measured by means of a coefficient called a discriminant of multivariable observation. It is equal to the sum of volumes of the parallelotops spanned on multidimensional observations. On the basis of it the modifications of the well known generalized variance were introduced. Geometrical properties of those parameters were investigated. The obtained results are used to generalize well-known clustering methods of Ward. One of the advantages of the method is that it finds clusters of high linear dependent multivariate observations.

Finally, the results are used to partition a time series into homogeneous groups where observations are close to linear trend. There is considered an example.

Key words: groupping criterion, agglomeration clustering, multidimensional variable, generalized variance, parallelotop, volume, discriminant, hyperplane, intra group spread, time series, linear trend.

1. BASIC DEFINITIONS AND NOTATION

Let $\mathbf{X} = [x_{ij}](i = 1, ..., h;\ j = 1, ..., N)$ be an $h \times N$ matrix, where x_{ij} is the j-th observation of an i-th h dimensional variable. A j-th column $(j = 1, ..., N)$ and an i-th row $(i = 1, ..., h)$ of $\mathbf{X}$ is denoted by $\mathbf{x}_j$ and $\mathbf{x}^i$, respectively. Then $\mathbf{X} = [\mathbf{x}_1 ... \mathbf{x}_N]$, $\mathbf{X}^T = [(\mathbf{x}^1)^T ... (\mathbf{x}^N)^T]$. Let $\{j_1, ..., j_k\}$ be a combination consisting of k column numbers chosen from the matrix $\mathbf{X}$. Similarly, let $\{i_1, ..., i_w\}$ be a combination consisting of w rows numbers chosen from $\mathbf{X}$, where $1 \leqslant w \leqslant h$ and $1 \leqslant k \leqslant N$. Let

* Department of Econometrics, Academy of Economics, Katowice.

[1] The research presented in the paper was support by the project KBN 1 P110 043 06.

$$\mathbf{X}(w, k) = [x_{i_s j_t}] = \begin{bmatrix} x_{i_1 j_1} & \dots & x_{i_1 j_k} \\ \dots & \dots & \dots \\ x_{i_w j_1} & \dots & x_{i_w j_k} \end{bmatrix}$$

be submatrix of $\mathbf{X}$. The $w \times k$ matrix $\mathbf{X}(w, k)$ is obtained through omitting the rows and columns in $\mathbf{X}$ except the rows and the columns of numbers $\{i_1, \dots, i_w\}$ and $\{j_1, \dots, j_k\}$ respectively. Particularly $\mathbf{X} = \mathbf{X}(h, \mathrm{N})$ and $\mathbf{X}(h, k) = [\mathbf{x}_{j_1}, \dots, \mathbf{x}_{j_k}]$. Symbol $\mathbf{P}(w, k|\mathbf{X})$ denotes a collection of all different matrices of the type $\mathbf{X}(w, k)$. The collection $\mathbf{P}(w, k|\mathbf{X})$ consists of $\binom{N}{k}\binom{h}{w}$ elements. For example if $\mathbf{X} = \begin{bmatrix} 123 \\ 456 \\ 789 \end{bmatrix}$, then $P(2,2|\mathbf{X}) = \left\{\begin{bmatrix} 12 \\ 45 \end{bmatrix}, \begin{bmatrix} 13 \\ 46 \end{bmatrix}, \begin{bmatrix} 23 \\ 56 \end{bmatrix}, \begin{bmatrix} 12 \\ 78 \end{bmatrix}, \begin{bmatrix} 13 \\ 79 \end{bmatrix}, \begin{bmatrix} 23 \\ 89 \end{bmatrix}, \begin{bmatrix} 45 \\ 78 \end{bmatrix}, \begin{bmatrix} 46 \\ 79 \end{bmatrix}, \begin{bmatrix} 56 \\ 89 \end{bmatrix}\right\}$. On the basis of the submatrix $\mathbf{X}(2, 3) = \begin{bmatrix} 123 \\ 456 \end{bmatrix}$, the following collection is generated: $P(2, 2|\mathbf{X}(2, 3)) = \left\{\begin{bmatrix} 12 \\ 45 \end{bmatrix}, \begin{bmatrix} 13 \\ 46 \end{bmatrix}, \begin{bmatrix} 23 \\ 56 \end{bmatrix}\right\}$. Moreover, $\mathbf{x}_j \in P(h, 1|\mathbf{X})$ and $\mathbf{x}^i \in P(1, N|\mathbf{X})$. The collection $P(w, k|\mathbf{X})$ can be decomposed in the following way

$$P(w, k|\mathbf{X}) = \bigcup_{\mathbf{X}(h,k) \in P(h,k|\mathbf{X})} P(w, k|\mathbf{X}(h, k)) \tag{1}$$

Let $\bar{\mathbf{x}} = [\bar{x}_1 \dots \bar{x}_h]^T = N^{-1}\mathbf{J}_N^T\mathbf{X}^T$ be the mean vector, where each element of an $N \times 1$ vector J_N is equal to one. An $h \times N$ matrix of deviations between observations of variables and their respective means is denoted by $\mathbf{B} = [b_{ij}]$ $(i = 1, \dots, h;\ j = 1, \dots, N)$, where $b_{ij} = x_{ij} - \bar{x}_i$. A submatrix $\mathbf{B}(w, k)$ is chosen from $\mathbf{B}$ in the same way as $\mathbf{X}(w, k)$ from $\mathbf{X}$. Particularly, an i-th row of $\mathbf{B}$ is $\mathbf{b}^i \in P(1, N|\mathbf{B})$ a j-th column of $\mathbf{B}$ is $\mathbf{b}_j \in P(h, 1|\mathbf{B})$. The decomposition of the $P(w, k|\mathbf{B})$ collection shows the equation:

$$P(w, k|\mathbf{B}) = \bigcup_{\mathbf{B}(w,N) \in P(w,N|\mathbf{B})} P(w, k|\mathbf{B}(w, N)) \tag{2}$$

Submatrix $\mathbf{B}(w, k)$ is the following function of $\mathbf{X}(w, k)$

$$\mathbf{B}(w, k) = \mathbf{X}(w, k) - N^{-1}X(w, N)\mathbf{J}_N\mathbf{J}_k^T \tag{3}$$

From a geometrical point of view components of a vector $\mathbf{x}_j$ are coordinates of a point in the h dimensional space. We shall denote that point as x_j, too. Components of column $\mathbf{b}_j$ are the coordinates of the vector $\vec{\mathbf{x}\mathbf{x}}_j$. The r dimensional volume of the parallelotop spanned by the vectors $\vec{\mathbf{x}_{j_{r+1}}\mathbf{x}_{j_1}}$, ..., $\vec{\mathbf{x}_{j_{r+1}}\mathbf{x}_{j_r}}$ in the h dimesional space is for $h \geqslant r$ defined by the

equation (see e.g. Jefimow and Rozendorn 1974, p. 262 or Borsuk 1969, p. 116–120):

$$m(\mathbf{x}_{j_1}, ..., \mathbf{x}_{j_{r+1}}) = \mathrm{m}(\mathbf{X}(h, r), \mathbf{x}_{j_{r+1}}) =$$
$$= \sqrt{\det(\mathbf{X}(h, r) - \mathbf{x}_{j_{r+1}}\mathbf{J}_r^T)^T(\mathbf{X}(h, r) - \mathbf{x}_{j_{r+1}}\mathbf{J}_r^T)} \tag{4}$$

The r dimensional volume of the parallelotop spanned by vectors $\vec{\mathbf{xx}}_{j_1}, ..., \vec{\mathbf{xx}}_{j_r}$ is as follows

$$m(\mathbf{x}_{j_1}, ..., \mathbf{x}_{j_{r+1}}, \bar{\mathbf{x}}) = m(\mathbf{b}_{j_1}, ..., \mathbf{b}_{j_r}, \mathbf{o}_h) = m(\mathbf{B}(h, r), \mathbf{o}_h) =$$
$$= \sqrt{\det \mathbf{B}^T(h, r)\mathbf{B}(h, r)}, \tag{5}$$

where by $\mathbf{o}_h$ is denoted the $h \times 1$ vector with all its elements equal to zero.

The r dimensional volume of the parallelotope spanned by the r vectors with their origin at the point $\mathbf{o}_N$ and the end points $\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_r}$ in the N dimensional space shows the equation:

$$m(\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_r}) = m(\mathbf{B}^T(r, N), \mathbf{o}_N) = \sqrt{\det \mathbf{B}(r, N)\mathbf{B}^T(r, N)} \tag{6}$$

Borsuk 1969, p. 64, defined the discriminant of the system of $(r+1)$ points $\{\mathbf{x}_{j_1}, ..., \mathbf{x}_{j_{r+1}}\} = \mathbf{X}(h, r+1)$ in the h dimensional space in the following way

$$q(\mathbf{X}(h, r+1) = (-1)^{r-1}2^{-r}\det\begin{bmatrix} 0 & \mathbf{J}_{r+1}^T \\ \mathbf{J}_{r+1} & \mathbf{D} \end{bmatrix} \tag{7}$$

where $\mathbf{D} = [d_{tv}]$ is the $(r+1) \times (r+1)$ matrix. Its elements are the squared distances between vectors $\mathbf{x}_{j_t}$ and $\mathbf{x}_{j_v}$, then

$$d_{tv} = (\mathbf{x}_{j_t} - \mathbf{x}_{j_v})^T(\mathbf{x}_{j_t} - \mathbf{x}_{j_v}) = \mathbf{x}_{j_t}^T\mathbf{x}_{j_t} - 2\mathbf{x}_{j_t}^T\mathbf{x}_{j_v} + \mathbf{x}_{j_v}^T\mathbf{x}_{j_v}$$

Lemma 1. (Borsuk 1969, p. 64 and 120). If $r \leqslant h$, then

$$q(\mathbf{X}(h, r+1)) = m^2(\mathbf{X}(h, r+1)) \tag{8}$$

2. MODIFIED SCATTER COEFFICIENTS

For a while let us limit our considerations to one-dimensional variable. The most simple and original way of spread measuring seems to be the way which follows from the expression:

$q = \sum_{j>i}^{N}(\mathbf{x}_j - \mathbf{x}_i)^2$ It is easy to prove that

$$q = \frac{1}{2}\sum_{j,i=1}^{N}(\mathbf{x}_j - \mathbf{x}_i)^2 = \frac{1}{2}\sum_{j,i=1}^{N}[(\mathbf{x}_j - \bar{\mathbf{x}}) - (\mathbf{x}_i - \bar{\mathbf{x}})]^2 = N\sum_{j=1}^{N}(\mathbf{x}_j - \bar{\mathbf{x}})^2 = N^2 s^2,$$

where $\bar{\mathbf{x}} = N^{-1}\sum_{j=1}^{N}\mathbf{x}_j$ is the average and s^2 is the variance of one dimensional variable. Then, the parameter q is proportionate to the variance, the most common coefficient of variability.

In order to generalize the coefficient q on multidimensional case we use the Borsuk's 1969 definition of the discriminant of a point system explained by expressions (7) and (8).

Definition 1. The discriminant of degree r of the h dimensional observation set $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ is as follows

$$q_{r/h}(\mathbf{X}) = \sum_{\mathbf{X}(h, r+1) \in P(h, r+1:\mathbf{X})} q(\mathbf{X}(h, r+1)), \tag{9}$$

where $1 \leqslant r \leqslant h \leqslant N$. Briefly, we shall call $q_{r/h}(\mathbf{X})$ the discriminant of multidimensional variable.

The equation (8) immediately causes the following:

$$q_{r/h}(\mathbf{X}) = \sum_{\mathbf{X}(h, r+1) \in P(h, r+1:\mathbf{X})} m^2(\mathbf{X}(h, r+1)) \tag{10}$$

From a geometrical point of view the defined coefficient is equal to the sum of squared volumes of the parallelotops spanned by vectors: $\vec{\mathbf{x}_{j_1}\mathbf{x}_{j_2}}$, ..., $\vec{\mathbf{x}_{j_r}\mathbf{x}_{j_{r+1}}}$ where $\mathbf{x}_j \in \mathbf{X}(h, r+1) \in P(h, r+1|\mathbf{X})$. Especially $q_{1/h}$ is equal to the sum of the squared Euclidean measure of the distances between the components of all pairs $\{\mathbf{x}_i, \mathbf{x}_j\}(i > j = 1, ..., N)$. Moreover, it is proportionate to the trace of the variance covariance matrix. The volume of the parallelotop spanned by a system consisting of $(r+1)$ points is $r!$ times greater than the volume of the simplex spaned by the same set of points (see e.g. Borsuk 1969, p. 117). Then, the coefficient $q_{2/h}$ is proportionate to the sum of the squared area of the triangles spanned by the system of points $\{\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \mathbf{x}_{j_3}\} \in \mathbf{X}$. The parameter depends on volumes of the tetrahedrons spanned by the combinations consisting of four points and so on.

Wilks 1932 introduced the generalized variance of multivariate variable as the determinant of its variance-covariance matrix. In our case the generalized variance shall be denoted by $g(\mathbf{X}) = N^{-h}\det\mathbf{B}\mathbf{B}^T$. The generalized variance of any r components of an h dimensional variable is given by the equation:

$$g(\mathbf{X}(r, N)) = g(\mathbf{B}(r, N)) = N^{-r}\det\mathbf{B}(r, N)\mathbf{B}^T(r, N) \tag{11}$$

Theorem 1. (W y w i a ł 1989, 1992).

$$q_{r/h}(\mathbf{X}) = N^{r+1} \sum_{\mathbf{X}(r, N) \in P(r, N:\mathbf{X})} q(\mathbf{X}(r, N)), \tag{12}$$

Definition 2. Modified generalized variance $q_{r/h}(\mathbf{X})$ of degree r of h dimensional variable shows the equation:

$$q_{r/h}(\mathbf{X}) = q_{r/h}(\mathbf{B}) = N^{-r-1} q_{r/h}(\mathbf{X}) \tag{13}$$

where $1 \leqslant r \leqslant h \leqslant N$ and $q_{r/h}(\mathbf{X})$ is given by Definition 1.

On the basis of (12) and (13) we infer that the coefficient $q_{r/h}$ is equal to the sum of generalized variances of all combinations consisting of r components chosen from h dimensional variable. In particular, $q_{h/h} = q(\mathbf{X})$ is the generalized variance in a simple sense of an h dimensional variable and $q_{1/h}$ is equal to the trace of a variance-covariance matrix.

A n d e r s o n 1958, p. 167 proved that $g(\mathbf{X})$ is proportionate to the squared h dimentional volume of the parallelotop spanned by vectors with the same origin point $\mathbf{o}_N$ and the end points $\mathbf{b}^1, ..., \mathbf{b}^h$. This property can be immediately generalized on the basis of the Definition 2 and expression (6) in the following way.

Theorem 2.

$$q_{r/h}(\mathbf{X}) = N^{r-1} \sum_{\mathbf{X}(r, N) \in P(r, N|\mathbf{B})} m^2(\mathbf{B}^T(r, N), \mathbf{o}_N),$$

where: $m(\mathbf{B}^T(r, N), \mathbf{o}_N)$ is the r dimensional volume of the parallelotop spanned by the vectors with the same origin at the point $\mathbf{o}_N$ and the end points $\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_r}$ in the N dimensional space.

Generalization of the second A n d e r s o n a's (1958, p. 170) theorem about the geometrical interpretation of the generalized variance is as follows:

Theorem 3. (W y w i a ł 1989, 1992).

$$q_{r/h}(\mathbf{X}) = N^{r-1} \sum_{\mathbf{X}(h, r) \in P(h, r|\mathbf{X})} m^2(\mathbf{X}(h, r), \bar{\mathbf{x}}) \tag{14}$$

$$q_{r/h}(\mathbf{X}) = N^{r-1} \sum_{\mathbf{B}(h, r) \in P(h, r|\mathbf{B})} m^2(\mathbf{B}(h, r), \mathbf{o}_h) \tag{15}$$

From the Theorem 3 we can infer that the modified generalized variance is proportionate to the sum of the squared volumes of the parallelotops spanned by the vectors $\vec{\mathbf{xx}}_{j_2}, ..., \vec{\mathbf{xx}}_{j_r}$, where $\mathbf{x}_j \in \mathbf{X}(h, r) \in P(h, r|\mathbf{X})$.

Let $\{\mathbf{ve}_j = c_j,\ j = 1, ..., h - t \geqslant 0\}$ be a system of equations. Then (see e.g. B o r s u k 1969, p. 87), solutions $\mathbf{v}$ of the system generate the t dimensional hyperplane denoted by $\mathbf{H}_{t/h}$ in the h dimensional space.

Theorem 4. (W y w i a ł 1992). $q_{r/h}(\mathbf{X}) = 0$ if and only if the points $\mathbf{x}_j$ $(j = 1, ..., N)$ are included in a not more than $(r-1)$ dimensional hyperplane.

3. COEFFICIENTS OF INTRAGROUP SPREAD

Let us simplify the notation introduced in the first chapter. Up to the end of the article we shall consider only submatrices consisting of columns of the matrix $\mathbf{X}$. So, the submatrix symbol $\mathbf{X}(h, k)$ is naturally reduced to the form: $\mathbf{X}(k)$. An $h \times k$ matrix $\mathbf{X}(k)$ consists of a k elements combination of columns chosen from the observation matrix $\mathbf{X}$. Similarly, the symbol $P(h, k|\mathbf{X})$ is reduced to $P(k|\mathbf{X})$ and it is the set consisting of all different matrixes of the type $\mathbf{X}(k)$ which can be formed on the basis of the matrix $\mathbf{X}$. Let $\mathbf{P} = \mathbf{P}(\mathbf{X})$ be the set of all submatrices (not necessarily of the same rank) made up of the columns of $\mathbf{X}$. So, it is obvious that $P(\mathbf{X}) = \sum_{k=1}^{N} P(k|\mathbf{X})$. Let $\mathbf{U} = \{\mathbf{X}(N_1), ..., \mathbf{X}(N_A)\}$ be the sequence of non-empty and disjoint sets consisting of columns chosen from $\mathbf{X}$. The columns of the submatrix $\mathbf{X}(N_a)$ represent elements of a population, which forms the a-th group.

Definition 3. The intra-set discriminant of r degree of an h dimensional variable, we call the following parameter:

$$Q_{r/h}(\mathbf{U}) = \sum_{a=1}^{a} q_{r/h}(\mathbf{X}(N_a)) \tag{16}$$

where $1 \leqslant r \leqslant h$.

The coefficient $Q_{r/h}$ is proportionate to the following linear combination of the generalized variances of r degree in the groups belonging to $\mathbf{U}$

$$G_{r/h}(\mathbf{U}) = N^{-r} Q_{r/h}(\mathbf{U}) = \sum_{a=1}^{A} w_a q_{r/h}(\mathbf{X}(N_a)) \tag{17}$$

where: $w_a = N_a^r N^{-r}$.

Definition 4. The above-written coefficient $G_{r/h}(\mathbf{U})$ will be called the weighted intra-group generalized variance of r degree of an h dimensional variable.

The parameter $G_{r/h}(\mathbf{U})$ indicates the level of the intra-group spread.

The following property of $Q_{r/h}(\mathbf{U})$ and $G_{r/h}(\mathbf{U})$ immediately results from Theorem 4 and Definition 3.

Theorem 5. If $\mathbf{U}$ consists of non-empty and mutually disjoint groups, then $Q_{r/h}(\mathbf{U}) = G_{r/h}(\mathbf{U}) = 0$ if and only if $\mathbf{X}_a = \in H_{t/h}^{(a)}$ for all $a = 1, ..., A$, and $t < r$.

4. GENERALIZING WARD'S CLUSTERING METHOD

We are going to describe an agglomeration method of clustering a fixed population into a set of disjoint groups. When the number of the algorithm stage increases, the quantity of groups decrease. At each stage of the algorithm groups are joined in such a way that the intra-group discriminant attains a minimun value. Before starting the clustering algorithm a population is treated as a collection of one element groups. The number of elements making up a group shall be called the size of that group. From Definitions 1 and 3 we infer that each created group has to consist of at least $(r+1)$ population elements because, otherwise, each distriminant $q(\mathbf{X}(k)) = 0$. In the first stage $(r+1)$ elements of a population are clustered in a group. Next, in the second stage of the algorithm there are two possibilities. A new group of size $(r+1)$ is formed or one element group is joined to a multi-element group formed in the previous stages. In the third stage there are three possible clustering options. The first two ones are the same as it was described in the second stage of the algorithm and the third is as follows: Two multi-element groups could be joined, if they were formed earlier. Generally, at the t-th stage one element groups are clustered into a group of size $(r+1)$ or two groups are joined, where one of them is at least of size $(r+1)$.

Let us suppose that the following collection of groups results from the t-th stage of the algorithm:

$$\underline{\mathbf{U}}_t = \{\mathbf{X}_\nu(N_\nu),\ \nu \leqslant t,\ \mathbf{x}_j \in \mathbf{X}_0(M_t)\} \tag{18}$$

where $\mathbf{X}_\nu(N_\nu)$ is an $h \times N_\nu$ matrix of data representing a group of size $N_\nu > r$ formed in the ν-th stage. A number of such multi-element groups is denoted by A_t. A number of one element groups remaining after t-th stage is denoted by M_t. An $h \times M_t$ matrix $\mathbf{X}_0(M_t)$ represents those one element groups.

Let $\mathbf{X}(N_{t+1})$ be a matrix representing a new group which will be formed in the $(t+1)$ stage, where $N_{t+1} > r$. Then the admissible set of groups in the $(t+1)$ stage is:

$$\mathbf{U}_{t+1} = \{\mathbf{X}_\nu(N_\nu),\ \nu \leqslant t,\ \mathbf{X}(N_{t+1}),\ \mathbf{X}_0(M_{t+1})\} \tag{19}$$

The increase of the criterion function, given by (16), is as follows

$$d_{r/h}(\mathbf{X}(N_{t+1})) = d_{r/h}(\mathbf{U}_{t+1}) = Q_{r/h}(\mathbf{U}_{t+1}) - Q_{r/h}(\underline{\mathbf{U}}_t) \tag{20}$$

As it was mentioned there are three ways of clustering:

1. If $M_t > r$, then an admissible group of size $(r+1)$ is formed on the basis of the set $\mathbf{X}_0(M_t)$, so $\mathbf{X}(r+1) \in P(r+1|\mathbf{X}_0(M_t))$. Hence, by Definition 1 and expressions (4), (8), (10), (16) we look for such a cluster $\mathbf{X}^{(1)}(U_{t+1})$ that

$$d_{r/h}(\mathbf{X}^{(1)}(N_{t+1})) = \underset{\mathbf{X}(r+1)\in P(r+1|\mathbf{X}_0(M_t))}{\text{minimum}} \{m^2(\mathbf{X}(r+1))\} \tag{21}$$

2. If $M_t > 0$, then each point $\mathbf{x}_j \in \mathbf{X}_0(M_t)$ can be joined to an earlier formed group $\mathbf{X}_v(N_v)$, $\in \underline{\mathbf{U}}_t$, so $\mathbf{X}(N_{t+1}) = \mathbf{X}_v(N_v) \cup \mathbf{x}_j$. Hence, the minimal increase of the criterion function is attained for such a set $\mathbf{X}^{(2)}(U_{t+1})$ if

$$d_{r/h}(\mathbf{X}^{(2)}(N_{t+1})) = \underset{v\leqslant t, \mathbf{x}_j\in\mathbf{X}_0(M_t)}{\text{minimum}} \{q_{r/h}(\mathbf{X}_v(N_v)\cup\mathbf{x}_j) - q_{r/h}(\mathbf{X}_v(N_v))\} \tag{22}$$

3. If set $\underline{\mathbf{U}}_t$ has at least two multi-element groups $\mathbf{X}_v(N_v)$, $\mathbf{X}_b(N_b) \in \underline{\mathbf{U}}_t$, then an admissible set is $\mathbf{X}(N_{t+1}) = \mathbf{X}_v(N_v) \cup \mathbf{X}_b(N_b)$. So we have to choose the group $\mathbf{X}^{(3)}(N_{t+1})$ which holds the expression:

$$d_{r/h}(\mathbf{X}^{(3)}(N_{t+1})) = \underset{v\neq b\leqslant t}{\text{minimum}} \{q_{r/h}(\mathbf{X}_v(N_v)\cup\mathbf{X}_b(N_b)) - q_{r/h}(\mathbf{X}_v(N_b)) - q_{r/h}(\mathbf{X}_v(N_v))\} \tag{23}$$

Finally, the optimal group is $\mathbf{X}_{t+1}(N_{t+1}) = \mathbf{X}^{(c)}(N_{t+1})$, where

$$d_{r/h}(\mathbf{X}^{(c)}(N_{t+1})) = \underset{i=1,2,3}{\text{minimum}} \{d_{r/h}(\mathbf{X}^{(i)}(N_{t+1}))\} \tag{24}$$

If $\mathbf{X}_{t+1}(N_{t+1}) = \mathbf{X}^{(1)}(N_{t+1})$ then the new $(r+1)$ element group is created. When $\mathbf{X}_{t+1}(N_{t+1}) = \mathbf{X}^{(2)}(N_{t+1})$, then the one element group is added to the appropriate multi-element one. Finally, if $\mathbf{X}_{t+1}(N_{t+1}) = \mathbf{X}^{(3)}(N_{t+1})$, then two appropriate multi-element groups are joined. Moreover, if any group obtained at the earlier steps is a subset of the optimal one formed in te current stage, then it cannot be included in the set $\underline{\mathbf{U}}_{t+1}$. Hence, if $\mathbf{X}_v(N_v) \in \underline{\mathbf{U}}_t$ and $\mathbf{X}_v(N_v) \subseteq \mathbf{X}_{t+1}(N_{t+1})$, then $\mathbf{X}_v(N_v \notin \underline{\mathbf{U}}_{t+1})$.

It is possible to prove almost immediately that if $r = 1$, then the expression (20) is reduced to the form:

$$d_{1/h}(\mathbf{X}(N_{t+1}) = N_b N_v (N_b + N_v)^{-1} (\bar{\mathbf{x}}_b - \bar{\mathbf{x}}_v)^T (\bar{\mathbf{x}}_b - \bar{\mathbf{x}}_v),$$

where $\bar{\mathbf{x}}_v$, $\bar{\mathbf{x}}_b$ are the mean vectors of variables in the groups $\mathbf{X}_v(N_v)$, $\mathbf{X}_b(N_b) \in \underline{\mathbf{U}}_t$, respectively. Hence, $d_{1/h}(\mathbf{X}(N_{t+1}))$ becames the well-known clustering criterion proposed by W a r d (1963). Therefore the Ward's rule of choice of optimal stage can be extended here. The set $\mathbf{U}_e$ obtained in the e-th stage will be optimal if it fulfils the expression:

$$d_{r/h}(\underline{\mathbf{U}}_{e+1}) = \underset{t=1,2,\ldots}{\text{maximum}} \{d_{r/h}(\underline{\mathbf{U}}_t)\} \tag{25}$$

The set $\underline{\mathbf{U}}_e$ is chosen if the increase of the criterion function is maximal at the next stage. Hence, the set $\underline{\mathbf{U}}_t$ we choose as optimal because the increase of the intra-set scatter idicated by the difference $d_{r/h}(\underline{\mathbf{U}}_{t+1}) = q_{r/h}(\underline{\mathbf{U}}_t) - q_{r/h}(\underline{\mathbf{U}}_{t+1})$ has the largest level.

Let us consider the following example of clustering population represented by the columns of the matrix: $\mathbf{X} = \begin{bmatrix} 2 & 4 & 3 & 7 & 8 & 9 & 5 \\ 8 & 8 & 9 & 2 & 3 & 4 & 1 \end{bmatrix}$. On the basis of the introduced algorithm we have: $\underline{U}_1 = \{\mathbf{x}_1, ..., \mathbf{x}_7\}$, $Q_{2/2}\underline{U}_1 = 0$;

$$\underline{U}_2 = \{[\mathbf{x}_4\mathbf{x}_5\mathbf{x}_6], \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_7\}, \quad Q_{2/2}(\underline{U}_2) = 0, \quad d_{2/2}(\underline{U}_2) = 0;$$

$$\underline{U}_3 = \{[\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3], [\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6], \mathbf{x}_7\}, \quad Q_{2/2}(\underline{U}_3) = m^2([\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3]) = \det^2 \begin{bmatrix} 2 & 8 & 1 \\ 4 & 8 & 1 \\ 3 & 9 & 1 \end{bmatrix} = 4,$$

$d_{2/2}(\underline{U}_3) = 4$;

$$\underline{U}_4 = \{[\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3], [\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7]\}, \quad Q_{2/2}(\underline{U}_4) = m^2([\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3]) + q([\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3\mathbf{x}_4]) =$$
$$= m^2([\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3]) + m^2([\mathbf{x}_4\mathbf{x}_5\mathbf{x}_6]) + m^2([\mathbf{x}_4\mathbf{x}_5\mathbf{x}_7]) + m^2([\mathbf{x}_4\mathbf{x}_6\mathbf{x}_7]) +$$
$$+ m^2([\mathbf{x}_5\mathbf{x}_6\mathbf{x}_7]) = 12, \quad d_{2/2}(\underline{U}_4) = 6;$$

$$\underline{U}_5 = \mathbf{X}, \quad Q_{2/2}(\underline{U}_5) = q_{2/2}(\mathbf{X}) = 9519, \quad d_{2/2}(\underline{U}_5) = 9507.$$

Then, the set $\mathbf{U}_4$ of two groups is optimal because the increase $d_{2/2}(\underline{U}_5)$ has the largest level.

5. CONDITIONAL METHOD OF WARD

Our main problem is how to divide the observation of time series into disjoint and coherent groups where they have a linear trend. The modified method of Ward divides the time series into such groups but under additional condition. Let $\mathbf{A} = [a_{ij}]$ be the neighbour matrix. If elements number i and j are (are not) neighbours, then $a_{ij} = 1 (a_{ij} = 0)$. The two elements of population, represented by $\mathbf{x}_i$ and $\mathbf{x}_j$, can be a cluster if and only if $a_{ij} = 1$. Similarly, two groups $\mathbf{X}_t$ and $\mathbf{X}_v$ can be joined into one cluster if and only if there exists at least one pair (x_i, x_j) such that $\mathbf{x}_i \in \mathbf{X}_v$ and $a_{ij} = 1$. For example, the neighbour matrix for the time series of five elements is as follows:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The introduced assumption leads to conditional clustering method of Ward considered in Wywiał 1994.

6. EXAMPLE OF TIME SERIES DECOMPOSITION

Let us consider the time series of electricity production in Poland from 1970 to 1991. The data are as follows (year, production in mld kWh): (1970,65), (1971,70), (1972,77), (1973,84), (1974,92), (1975,97), (1976,104), (1977,109), (1978,116), (1979,117), (1980,122), (1981,115), (1982,118), (1983,126), (1984,135), (1985,138), (1986,140), (1987,146), (1988,144), (1989,145), (1990,136), (1991,135).

Our purpose is decomposition of that time series into subintervals where the observations of electricity production are highly linear dependent. Using the conditional method of Ward we have the following decompositions of the time series[2], through minimization of the intra-set discriminant $Q_{2/2}$:

a) into four intervals: from 1970–1977 and 1978–1980 and 1981–1988 and 1989–1991, $d_{2/2} = 22900$,

b) into three intervals: from 1970–1980 and 1981–1988 and 1989–1991, $d_{2/2} = 51060$,

c) into two intervals: from 1979–1980 and 1981–1991, $d_{2/2} = 605315$.

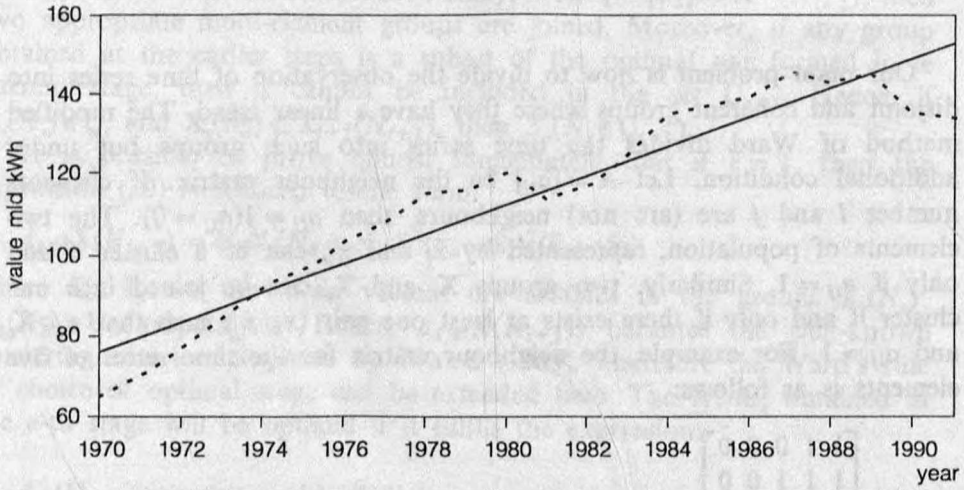

Fig. 1. The observed and the predicted production of electricity in Poland

[2] It is realized by computer program written in PASCAL.

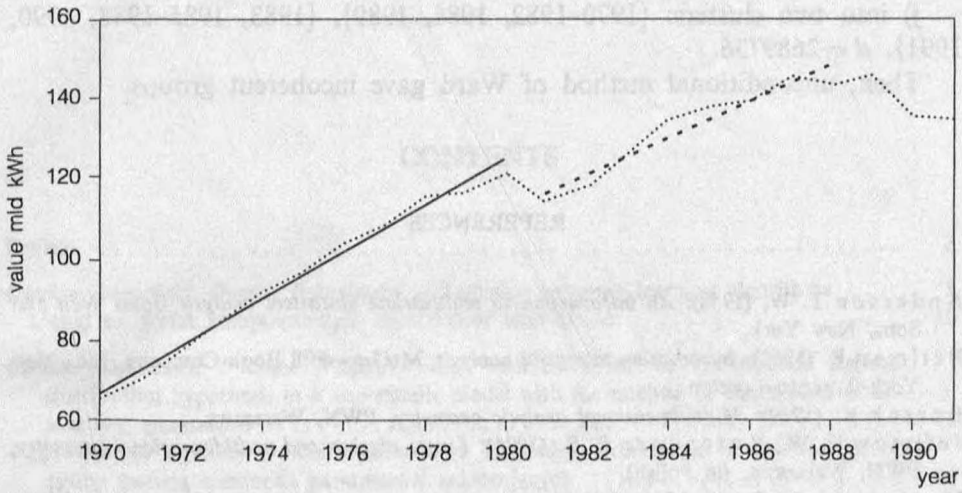

Fig. 2. The observed and the predicted production on the basis of two trends

Figure 1 represents the observations of the electricity production and the linear trend estimated on the basis of all observations. Figure 2 shows how the observations of the electricity production are approximated by two trends. The first (second) one was estimated on the basis of observations from the interval 1970–1980 (1981–1988). Those intervals were obtained through conditional minimization of $Q_{2/2}$ as it was explained above at the point b). It is obvious that the two trends (Fig. 2) are fitted better to the observations then one trend (Fig. 1). The analysis of the time series can be continued in the distinguished intervals. Especially, the significance of changing the trend parameters can be tested.

The minimization of the intra-set discriminant $Q_{1/2}$ (the order method of Ward) leads to the following system of clusters:

e) into four intervals: from 1970–1973 and 1974–1977 and 1978–1982 and 1983–1991, $d_{2/2} = 2405$,

f) into three intervals: from 1970–1973 and 1974–1982 and 1983–1991, $d_{2/2} = 7280$,

g) into two intervals: from 1979–1982 and 1983–1991, $d_{2/2} = 53708$.

The obtained intervals consist of observations which are not such linear dependent as it was in the cases a), b), c).

The unconditional method of Ward (minimizing the intra-set coefficient $Q_{2/2}$) leads to the following system of clusters.

h) into four clusters: {1970, 1971, 1984}, {1973, 1981, 1982, 1990}, {1972, 1974–1980}, {1983, 1985–1988, 1990, 1991}, $d_{2/2} = 76456$,

i) into three clusters: {1970, 1971, 1973, 1981, 1982, 1984, 1989}, {1972, 1974–1980}, {1983, 1985–1988, 1990, 1991}, $d_{2/2} = 189459$,

j) into two clusters: {1970–1982, 1984, 1989}, {1983, 1985–1988, 1990, 1991}, $d = 2689736$.

Then, unconditional method of Ward gave incoherent groups.

REFERENCES

Anderson T. W. (1958): *An introduction to multivariate statistical analysis*, John Wile and Sons, New York.

Bellman R. (1960): *Introduction to matrix analysis*, McGraw-Hill Book Company, Inc., New York–Toronto–London.

Borsuk K. (1969): *Multidimensional analytic geometry*, PWN, Warszawa.

Jefimow N. W., Rozendorn E. R. (1974): *Linear algebra and multidimensional geometry*, PWN, Warszawa, (in Polish).

Mostowski A., Stark M. (1974): *Elements of high algebra*, PWN, Warszawa, (in Polish).

Ward J. H. (1963): *Hierarchical groupping to optimize an objective function*, „Journal of the American Statistical Association".

Wilks S. S. (1932): *Certain generalization in the analysis of variance*, „Biometrika", Vol. 24, p. 471–494.

Wilks S. S. (1962): *Mathematical statistics*, John Wiley and Sons Inc., New York–London.

Wywiał J. (1989): *Decomposition of generalized variance and its application to study of population homogenity*, [in:] *Econometric analysis of structure constancy*, the grant: CPBP 10.09, edited by K. Zadora, Akademia Ekonomiczna w Katowicach, typesctript in Polish.

Wywiał J. (1992): *On some measurements of multidimensional statistical scatter and their use for grouping the finite population*, „Zeszyty Naukowe Akademii Ekonomicznej w Katowicach", Nr 120, p. 129–149 (in Polish).

Wywiał J. (1994): *On grouping method of Ward*, „Prace Naukowe Akademii Ekonomicznej we Wrocławiu", Nr 667, p. 119–122, (in Polish).

Janusz Wywiał

DEKOMPOZYCJA SZEREGÓW CZASOWYCH OPARTA NA ZMODYFIKOWANEJ METODZIE WARDA

W dłuższym czasie trend w szeregu czasowym może zmienić swój kierunek. Dlatego też proponowany jest podział obserwacji w szeregu czasowym na rozłączne i spójne podzbiory, w których trend ma postać liniową.

W pracy rozważana jest modyfikacja uogólnionej wariancji oraz przeprowadzono badanie jej geometrycznych własności. Otrzymane wyniki są wykorzystane do zaproponowania uogólnienia znanych metod Warda w tym sensie, że osłabiają założenia, przy których metody te się stosuje.