The Optimal Valid Partitioning Procedures

Senko Oleg V. [1] *Kuznetsova A.V.* [2]

Computer Center of RAS [1]

Institute of Biochemical Physics of RAS [2]

Moscow,Russia

The purpose of discussed optimal valid partitioning (OVP) methods is uncovering of ordinal or continuous explanatory variables effect on outcome variables of different types. The OVP approach is based on searching partitions of explanatory variables space that in the best way separate observations with different levels of outcomes. Partitions of single variables ranges or two-dimensional admissible areas for pairs of variables are searched inside corresponding families. Statistical validity associated with revealed regularities is estimated with the help of permutation test repeating search of optimal partition for each permuted dataset. Monte Carlo simulation was used to test performance of OVP procedures both on ability to uncover regularities specified by experiments scenario and probability of false regularities that partially or completely do not agree with scenario. At the first stage OVP method was examined with the same technique for estimating statistical validities associated with simplest and more complicated partitions. However probability of partially false regularities appeared to be too high for this procedure. So alternative technique was suggested where statistical validity associated with more complicated partitions is calculated using statistically valid simplest partitions previously found for the same explanatory variables.

Keywords: Optimal partitioning, statistical validity, permutation test, regularities, explanatory variables effect, complexity, Monte Carlo simulation

Introduction.

Within the scope of techniques for evaluating explanatory variables effect on dependent one we can mark out methods that are based on searching of change points within explanatory variables range. Such change points allow to separate different levels of outcome. One of the most common techniques of this type is Kolmogorov-Smirnov test. We can remind that this test is based on searching of such value of explanatory variable where deviation between empirical distribution function and a hypothetical cumulative distribution function is maximal (Borovkov).

Optimal boundaries searching is widely used in recognition or regression methods. The most known approaches of such type are classification and

regression trees, (see Breiman et all). However the goal of regression modeling is not evaluating effects of all explanatory variables but exact forecasting of dependent by some optimal subset of explanatories.

Based on optimal partitioning methods for evaluating effects of explanatory variable were considered by different groups of researchers. So Chitchian and Safaryan (2001) proposed method for testing independence between two ordinal or continuous variables that includes detection of changepoints in sequences of induced order statistics. Abdolell et all.(2002) used permutation test to asses the statistical validity of regularities associated with optimal dichotomic partitions of continuous prognostic variable.

In present paper the optimal valid partitioning (OVP) approach to data analysis is discussed. The OVP procedures calculate the sets of optimal partitions of one-dimensional admissible intervals of single variables or two-dimensional admissible areas of pairs of variables and evaluate statistical validity of regularities associated with these partitions. It must be noted that applying standard techniques (F-test, Chi-square and others) for assessing validity by the same datasets which previously has been used for boundaries calculating come across problem of multiple testing (`see Mazumdar and Glassman). So validity estimates appeared to be too optimistic. One of the ways to calculate adequate estimate is randomized splitting of initial data on two subsets. The first one is used for the boundaries calculating and the second one is used for evaluating of statistical validity. But such approach leads to significant loss of both boundaries exactness and validity levels due to decrease of observations numbers in two datasets. The another way to verify nonrandom character of differences between dependent variable levels in groups of observations formed by partitions is using permutation tests. Discussed below technique that is based on random permutations allows using the same dataset for both purposes: boundaries search and evaluating statistical significance. One more advantage of permutation tests is absence of necessity for any suppositions about variables distribution or any restrictions on groups sizes. Today rather many examples of successful use of permutation technique in

different types of tasks (O'Gorman (2001), Abdolell et all.(2002)) Variants of OVP methods using search of optimal partitions inside families of different complexity levels was previously considered by Senko and Kuznetsova (1998), Kuznetsova et all(2000), Senko et. all(2003).

2 The Optimal Valid Partitioning.

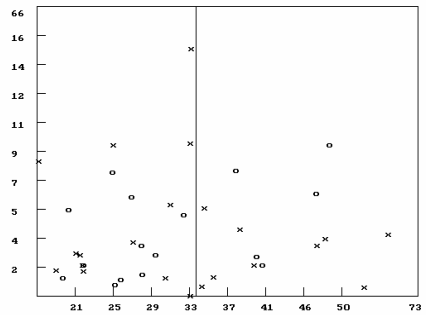
Suppose that we study dependence of variable Y on explanatory variables $X_1,\ldots,X_n$ by some empirical dataset $\tilde{S}_0$. Various types of dependent variable are admissible: Y may be continuous variables that are directly observed, vectors of probabilities of several types of events at points in X space, survival curves and so on. The observations from data set $\tilde{S}_0$ must include the vectors of independent variables **x** and information $\mathcal{Y}$ related to dependent variable Y that allows to evaluate mean values of Y by sets of observations with the help of some common procedure. In case Y is directly observed continuous variable $\mathcal{Y}$ is simply value of Y and abovementioned evaluating procedure is calculating of normal means, evaluating procedure is also reduced to calculating of normal means (fractions of events types) when Y is probabilities vector and $\mathcal{Y}$ is binary vector indicating type of events, in case Y is survival curve $\mathcal{Y}$ is pair including time of last observation and binary indicating if patient is alive. In the last case the Kaplan-Mayer technique is the example of evaluating procedure. Let Y belongs to some set M_y. It is supposed that distance function ρ defined on Cartesian product $M_y \times M_y$ satisfies following conditions:

a) $\rho(y',y'') \geq 0$, b) $\rho(y',y'') = \rho(y'',y')$, c) $\rho(y',y') = 0 \ \ \forall y', y'' \in M_y$.

The OVP methods are based on optimal partitioning of independent variables admissible regions. The partitions that provide for best separation of observations from dataset $\tilde{S}_0$ with different levels of dependent variable are searched inside apriori defined families by optimizing of quality functional.

2.1 Partitions families.

The partition family is defined as the set of partitions with limited number of elements that are constructed by the same procedure. The unidimensional and two-dimensional families are considered. The unidimensional families includes partitions of admissible intervals of single variables. The simplest Family I (Fig 1) includes all partitions with two elements that are divided by one boundary point. The more complex Family II (Fig 2) includes all partitions with no more than three elements that are divided by two boundary points. The two-dimebsional Family III (Fig. 3) includes all partitions of two-dimensional admissible areas with no more than four elements that are separated by two boundary lines parallel to coordinate axes. The examples of partitions from each of three abovementioned families are given at figures 1-3.



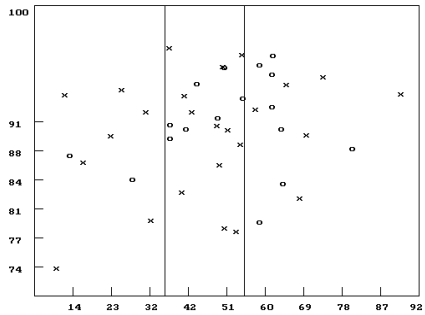


Fig . 1 . Partition from Family I **Fig. 2. Partition from Family II**

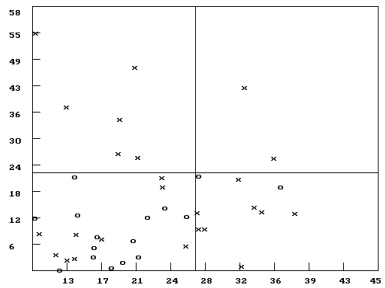


Fig . 3 . Partition from Family III

2.2 Quality functional.

Let $\tilde{Q}$ is partition of admissible region of independent variables with elements $q_1,\ldots,q_r$. The partition $\tilde{Q}$ produces partition of dataset $\tilde{S}_0$ on subsets $\tilde{S}_1,\ldots,\tilde{S}_r$, where $\tilde{S}_j$ ($j=1,\ldots,r$) is subset of observations with independent variables vectors belonging to q_j. The evaluated Y mean value of subsets $\tilde{S}_j$ is denoted as $\hat{y}(\tilde{S}_j)$. The integral quality functional $F_I(\tilde{Q},\tilde{S}_0)$ is defined as the sum:

$F_I(\tilde{Q},\tilde{S}_0)=\sum_{j=1}^{r}\rho[\hat{y}(\tilde{S}_0),\hat{y}(\tilde{S}_j)]m_j$, where m_j- is number of observations in subset $\tilde{S}_j$.

Besides integral functional $F_I(\tilde{Q},\tilde{S}_0)$ local functional $F_L(\tilde{Q},\tilde{S}_0)$ is possible that is defined as $F_L(\tilde{Q},\tilde{S}_0)=\max_{j=1,\ldots,r}\{\rho[\hat{y}(\tilde{S}_0),\hat{y}(\tilde{S}_j)]m_j\}$. Unlike integral functional $F_I(\tilde{Q},\tilde{S}_0)$ local functional $F_L(\tilde{Q},\tilde{S}_0)$ allows to pick out the most distant from remaining part of $\tilde{S}_0$ subregion of partition.

The optimal value of quality functional in dataset $\tilde{S}$ will be further referred to as $F_I^o(\tilde{S})$ or $F_L^o(\tilde{S})$.

2.3 Optimal Partitioning Procedure

The represented in this section optimal partitioning technique may be used in cases when mean values of Y in some $\tilde{S}\subseteq\tilde{S}_o$ is calculated as simple normal mean of corresponding $\mathcal{Y}$ value. It is based on use for evaluating of a new partition results of evaluating of previous one. So it is more economical than direct recalculating.

2.3.1 Univariate Partitions families.

Suppose that initial information is set of pairs $\tilde{S}_0=\{(\mathcal{Y}_1,x_1),\ldots,(\mathcal{Y}_m,x_m)\}$, where $x_1,\ldots,x_m$ are values of explanatory variable X.

Assume that explanatory $X\in\{a_1,\ldots,a_r\}$, where $a_1<\ldots<a_r$. Boundary points are calculated as $b_1=\frac{a_1+a_2}{2},\ldots,b_{r-1}=\frac{a_{r-1}+a_r}{2}$. Then we calculate sets $\{y_1,\ldots,y_r\}$ and

$\{m_1,\ldots,m_r\}$ corresponding to numbers from $\{a_1,\ldots,a_r\}$, where m_i is number of objects in $\tilde{S}_o$ with explanatory X equal a_i, $y_i = \sum_{j=1}^{m} I(x_j \in a_i)\mathcal{Y}_j$, $I(x_j,a_j)=1$ if $x_j = a_i$ and $I(x_j,a_i)=0$ otherwise. Suppose that optimal partition is searched inside simplest family 1.

Optimal partitions searching inside family I.

a) At the initial step boundary quality functional is calculate for boundary point b_1. It forms following subsets of $\tilde{S}_o$: left $\tilde{S}_l$ with explanatory $X < b_1$, right $\tilde{S}_r$ with explanatory $X > b_1$. Let $\hat{m}_l$ is the number of observations in $\tilde{S}_l$, $\hat{m}_r$ is the number of observations in $\tilde{S}_r$, $\hat{y}_l$ is the sum of all $\mathcal{Y}$- components in $\tilde{S}_l$, $\hat{y}_r$ is the sum of all $\mathcal{Y}$- components in $\tilde{S}_r$. Then values $\hat{y}(\tilde{S}_o) = \frac{\sum_{i=1}^{r} y_i}{m}$, $\hat{y}(\tilde{S}_l) = \frac{\hat{y}_l}{\hat{m}_l}$ and

$\hat{y}(\tilde{S}_r) = \frac{\hat{y}_r}{\hat{m}_r}$ are calculated, where $\hat{y}_l = y_1$, $\hat{y}_r = \sum_{i=2}^{r} y_i$, $\hat{m}_l = m_1$, $\hat{m}_r = \sum_{i=2}^{r} m_i$. At last optimized quality functional is calculated:

$F_I(\tilde{Q},\tilde{S}_0) = \rho[\hat{y}(\tilde{S}_0),\hat{y}(\tilde{S}_l)]\hat{m}_l + \rho[\hat{y}(\tilde{S}_0),\hat{y}(\tilde{S}_r)]\hat{m}_r$ or

$F_L(\tilde{Q},\tilde{S}_0) = \max\{\rho[\hat{y}(\tilde{S}_0),\hat{y}(\tilde{S}_l)]\hat{m}_l, \rho[\hat{y}(\tilde{S}_0),\hat{y}(\tilde{S}_r)]\hat{m}_r\}$.

b) Assume that values $\hat{m}_l$, $\hat{m}_r$, $\hat{y}_l$, $\hat{y}_r$ were used when quality functional for boundary point b_{i-1} was calculated. To receive the quality functional for boundary point b_i it is sufficient to take $\hat{y}_l = \hat{y}_l + y_i$, $\hat{y}_r = \hat{y}_r - y_i$, $\hat{m}_l = \hat{m}_l + m_2$, $\hat{m}_r = \hat{m}_r - m_2$ and to recalculate quality functional with these new values.

c) procedure is repeated until all boundary points are tested.

Optimal partitions searching inside family II.

a) At the initial step boundary quality functional is calculate for pair of boundary points b_1 and b_2. It forms to subsets of $\tilde{S}_o$: left $\tilde{S}_l$ with explanatory $X < b_1$, middle $\tilde{S}_m$ with explanatory $b_1 < X < b_2$ and right $\tilde{S}_r$ with explanatory $X > b_1$. As in previous section $\hat{m}_*$ is the number of observations in $\tilde{S}_*$, $\hat{y}_*$ is the sum of all $\mathcal{Y}$- components in $\tilde{S}_*$. Values $\hat{y}(\tilde{S}_o) = \frac{\sum_{i=1}^{r} y_i}{m}$, $\hat{y}(\tilde{S}_l) = \frac{\hat{y}_l}{\hat{m}_l}$ and $\hat{y}(\tilde{S}_r) = \frac{\hat{y}_r}{\hat{m}_r}$ are calculated, where $\hat{y}_l = y_1$, $\hat{y}_m = y_2$, $\hat{y}_r = \sum_{i=3}^{r} y_i$, $\hat{m}_l = m_1$, $\hat{m}_l = m_2$, $\hat{m}_r = \sum_{i=3}^{r} m_i$. Then quality functional is calculated as

$$F_I(\tilde{Q}, \tilde{S}_0) = \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_l)]m_l + \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_m)]m_m + \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_r)]m_r \quad \text{or}$$

$$F_L(\tilde{Q}, \tilde{S}_0) = \max\{\rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_l)]m_l, \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_m)]m_m, \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_r)]m_r\}.$$

b) Assume that values $\hat{m}_l$, $\hat{m}_m$, $\hat{m}_r$, $\hat{y}_l$, $\hat{y}_m$, $\hat{y}_r$ were used when quality functional for pair of boundary points $b_{i'}$ and $b_{i''-1}$ was calculated. To receive value of quality functional for boundary point $b_{i'}$ and $b_{i''}$ it is sufficiently to take $\hat{y}_l = \hat{y}_l$, $\hat{y}_r = \hat{y}_r - y_{i''}$, $\hat{m}_l = \hat{m}_l$, $\hat{m}_r = \hat{m}_r - m_{i''}$ and to recalculate quality functional with these new values.

c) Assume that values $\hat{m}_l$, $\hat{m}_m$, $\hat{m}_r$, $\hat{y}_l$, $\hat{y}_m$, $\hat{y}_r$ were used when quality functional for pair of boundary points $b_{i'-1}$ and $b_{i''}$ was calculated. To receive the value of quality functional for boundary point $b_{i'}$ and $b_{i''}$ it is sufficiently to take $\hat{y}_l = \hat{y}_l + y_i$, $\hat{y}_r = \hat{y}_r$, $\hat{m}_l = \hat{m}_l + m_{i'}$, $\hat{m}_r = \hat{m}_r$ and to recalculate quality functional with these new values.

d) procedure is repeated until all boundary points are evaluated.

2.3.2 Two-variate partitions family II

Suppose that initial information is set $\tilde{S}_0 = \{(y_1, x_1^1, x_1^2), \ldots, (y_m, x_m^1, x_m^2)\}$, where $x_1^i, \ldots, x_m^i$ are values of explanatory variable X_i, $i = 1,2$. Assume that we consider pair of explanatory $X_1 \in \{a_1^1, \ldots, a_{r_1}^1\}$ and $X_2 \in \{a_1^2, \ldots, a_{r_2}^2\}$, where $a_1^1 < \ldots < a_{r_1}^1$, $a_1^2 < \ldots < a_{r_2}^2$. At initial stage boundary points are calculated as $b_1^1 = \frac{a_1^1 + a_2^1}{2}, \ldots, b_{r-1}^1 = \frac{a_{r-1}^1 + a_r^1}{2}$, $b_1^2 = \frac{a_1^2 + a_2^2}{2}, \ldots, b_{r-1}^2 = \frac{a_{r-1}^2 + a_r^2}{2}$. Then we calculate sets $\{y_{i'i''} \mid i' = 1, \ldots, r; i'' = 1, \ldots, r\}$ and $\{m_{i'i''} \mid i = 1, \ldots, r;\ j = 1, \ldots, r\}$ corresponding to pairs from $\{(a_{i'}^1, a_{i''}^2) \mid i' = 1, \ldots, r; i'' = 1, \ldots, r\}$, where $m_{i'i''}$ is number of observations in $\tilde{S}_o$ with $X_1 = a_{i'}^1$, $X_2 = a_{i''}^2$, $y_{i'i''} == \sum_{j=1}^{m} I(x_i^1, a_{i'}^1) I(x_j^2, a_{i''}^2) y_j$.

At first optimal partition with boundary point b_1^1 for explanatory X_1 is searched. This task is very close to univariate optimization inside partitions family I, boundary points from set $b_1^2, \ldots, b_{r-1}^2$ are evaluated. The main difference from univariate optimization inside family I that was discussed in previous section is the type of quality functional. In previous section it incorporates 2 subregions and now it incorporates 4 subregions.

a) At the first step value of quality functional is calculated for partition with boundary points b_1^1 for x_1 and b_1^2 for x_2:

$$F_I(\tilde{Q}, \tilde{S}_0) = \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{ul})]\hat{m}_{ul} + \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{ur})]\hat{m}_{ur} + \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{lr})]\hat{m}_{lr} + \\ + \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{ll})]\hat{m}_{ll}$$

or

$$F_L(\tilde{Q}, \tilde{S}_0) = \max\{\rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{ul})]\hat{m}_{ul}, \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{ur})]\hat{m}_{ur}, \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{lr})]\hat{m}_{lr}, \\ \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_{ll})]\hat{m}_{ll}\}$$

Here following notations are used:

$\tilde{S}_{ul}$ is upper left subset with explanatory $X_1 < b_1^1$, $X_2 > b_1^2$;

$\tilde{S}_{ur}$ is upper left subset with explanatory $X_1 > b_1^1$, $X_2 > b_1^2$;

$\tilde{S}_{lr}$ is upper left subset with explanatory $X_1 < b_1^1$, $X_2 > b_1^2$;

$\tilde{S}_{ll}$ is upper left subset with explanatory $X_1 < b_1^1$, $X_2 < b_1^2$;

$\hat{\mathrm{m}}_{**}$ is the number of observations in $\tilde{S}_{**}$;

$\hat{\mathrm{y}}_{**}$ is the sum of y-components from $\tilde{S}_{**}$;

$\hat{y}(\tilde{S}_{**}) = \frac{\hat{\mathrm{y}}_{**}}{\hat{\mathrm{m}}_{**}}$; $\hat{\mathrm{y}}_{ul} = \sum_{i=2}^{r_2} \bar{\mathrm{y}}_i^l$, $\hat{\mathrm{y}}_{ll} = \bar{\mathrm{y}}_1^l$; $\hat{\mathrm{y}}_{lr} = \bar{\mathrm{y}}_1^r$; $\hat{\mathrm{y}}_{ur} = \sum_{i=2}^{r_2} \bar{\mathrm{y}}_i^r$;

$\bar{\mathrm{y}}_i^l = \mathrm{y}_{1i}$; $\bar{\mathrm{y}}_i^r = \sum_{k=2}^{r_1} \mathrm{y}_{ki}$;

b) Assume that values $\hat{\mathrm{m}}_{**}$, $\hat{\mathrm{y}}_{**}$ were used when quality functional for pair of boundary points $b_{i'}^1$ and $b_{i''-1}^2$ was calculated. To receive value of quality functional for boundary point $b_{i'}^1$ and $b_{i''}^2$ it is sufficiently to take $\hat{\mathrm{y}}_{ll} = \hat{\mathrm{y}}_{ll} + \bar{\mathrm{y}}_{i''}^l$, $\hat{\mathrm{y}}_{lr} = \hat{\mathrm{y}}_{lr} + \bar{\mathrm{y}}_{i''}^r$ $\hat{\mathrm{y}}_{ul} = \hat{\mathrm{y}}_{ul} - \bar{\mathrm{y}}_{i''}^l$ $\hat{\mathrm{y}}_{ur} = \hat{\mathrm{y}}_{ur} - \bar{\mathrm{y}}_{i''}^r$ $\hat{\mathrm{m}}_{ll} = \hat{\mathrm{m}}_{ll} + \bar{\mathrm{m}}_{i''}^l$, $\hat{\mathrm{m}}_{lr} = \hat{\mathrm{m}}_{lr} + \mathrm{m}_{i''}^r$ $\hat{\mathrm{m}}_{ul} = \hat{\mathrm{m}}_{ul} - \mathrm{m}_{i''}^l$ $\hat{\mathrm{m}}_{ur} = \hat{\mathrm{m}}_{ur} - \mathrm{m}_{i''}^r$ and to recalculate quality functional with these new values.

c) Assume that values $\{\bar{\mathrm{y}}_1^l, \ldots, \bar{\mathrm{y}}_{r_2}^l, \bar{\mathrm{y}}_1^r, \ldots, \bar{\mathrm{y}}_{r_2}^r\}$ were used when quality functional values for boundary point $b_{i'-1}^1$ were calculated. To receive values of quality functional for boundary point $b_{i'}^1$ it is necessary to recalculate anew values $\{\bar{\mathrm{y}}_1^l, \ldots, \bar{\mathrm{y}}_{r_2}^l, \bar{\mathrm{y}}_1^r, \ldots, \bar{\mathrm{y}}_{r_2}^r\}$: $\bar{\mathrm{y}}_{i''}^l = \bar{\mathrm{y}}_{i''}^l + \mathrm{y}_{i'i''}$, $\bar{\mathrm{y}}_{i''}^r = \bar{\mathrm{y}}_{i''}^r - \mathrm{y}_{i'i''}$.

d)procedure is repeated until all boundary points are evaluated.

2.4 The validation of partitioning results.

For validation of found optimal partitions the permutation test (PT) is used. In present studies two variants of PT were considered.

The first variant (PT-1) is based on testing basic null hypothesis that variable Y is fully independent on involved explanatory variables. The optimal value of

quality functional F_*^o (it may be F_I^o or F_L^o) is used as PT-1 statistics. Let optimal partition of variable X' admissible interval was found inside families I or II or optimal partition of variables X', X'' joint admissible area was found inside family III for dataset $\tilde{S}_0 = \{(\mathcal{Y}_1, \mathbf{x}_1), \ldots, (\mathcal{Y}_m, \mathbf{x}_m)\}$. Let $F_*^o(\tilde{S}_0)$ is the optimal value of used quality functional. To evaluate statistical validity of discovered regularity set of random permutations $\{\pi_1, \ldots, \pi_{\mathcal{N}}\}$ is calculated with the help of random numbers generator. Initial dataset $\{(\mathcal{Y}_1, \mathbf{x}_1), \ldots, (\mathcal{Y}_m, \mathbf{x}_m)\}$ and permutations $\{\pi_1, \ldots, \pi_{\mathcal{N}}\}$ give rise to permuted datasets $\{\tilde{S}_1^r, \ldots, \tilde{S}_{\mathcal{N}}^r\}$, where $\tilde{S}_j^r = \{(\mathcal{Y}_{\pi_j(1)}, \mathbf{x}_1), \ldots, (\mathcal{Y}_{\pi_j(m)}, \mathbf{x}_m)\}$. For each dataset $\tilde{S}_{\pi_j}^r$ from $\{\tilde{S}_1^r, \ldots, \tilde{S}_{\mathcal{N}}^r\}$ optimal partition is searched inside the same family for the same variable (variables) and by optimizing the same quality functional that were previously used in case of $\tilde{S}_0$. Let $\mathcal{N}_{gt}[F_*^o(\tilde{S}_0)]$ is the number of datasets in $\{\tilde{S}_1^r, \ldots, \tilde{S}_{\mathcal{N}}^r\}$ for which $F_*^o(\tilde{S}_*^r) > F_*^o(\tilde{S}_0)$. The ratio $\mathcal{N}_{gt}[F_*^o(\tilde{S}_0)]/\mathcal{N}$ is used as estimate of PT-1 p-value (see **Appendix**) for regularity discovered in $\tilde{S}_0$ with the help of optimal partitioning. .

The second variant **(PT-2)** is based on testing more complicated null hypothesis that variable Y is independent on involved explanatory variables only inside some apriori defined subregions of X-space. Let explanatory variables admissible region in X-space is partitioned on subregions $q_1^a, \ldots, q_p^a$. This partition produces the partition of dataset $\tilde{S}_0$ on subsets $\tilde{S}_1^a, \ldots, \tilde{S}_p^a$. The following Monte-Carlo procedure of p-values estimating was used in second PT variant. Datasets $\{\tilde{S}_1^{ar}, \ldots, \tilde{S}_{\mathcal{N}}^{ar}\}$ are generated from $\tilde{S}_0$ with the help of permutations $\{\pi_1^{ar}, \ldots, \pi_{\mathcal{N}}^{ar}\}$. As in the first variant only $\mathcal{Y}$-components positions are permuted and the order of X-components remains fixed. Unlike permutations $\{\pi_1^r, \ldots, \pi_{\mathcal{N}}^r\}$ from the first variant permutations $\{\pi_1^{ar}, \ldots, \pi_{\mathcal{N}}^{ar}\}$ do not include transpositions

between $\mathcal{Y}$-components of observations belonging to different subsets from $\tilde{S}_1^a,\ldots,\tilde{S}_p^a$. The procedure of p-values calculating by generated datasets $\{\tilde{S}_1^{ar},\ldots,\tilde{S}_{\mathcal{N}}^{ar}\}$ completely coincides with the procedure of p-values calculating in the first variant. The p-values evaluating the independence of Y inside subregions $q_1^a,\ldots,q_p^a$ and calculated by PT-2 will be referred to as $p_2(q_1^a,\ldots,q_p^a)$-values.

2.5 Forming of sets of output regularities.

The set of output regularities is selected from the set of found optimal partitions using calculated p-values. (To simplify the discussion we shall not differ further between regularity and describing it optimal partition.) The following methods of output set forming were used in OVP procedures considered in present study. The first and simplest way is selecting in output set only regularities with calculated p-values less than previously defined threshold $p_{thr.}$. The OVP procedures using this way of selecting will be referred to as OVP-CIS (complexity independent selecting). The another variant of OVP procedure (OVP-CDS) will be discussed below.

Ill

3 Simulations

Performance of OVP techniques may be evaluated by testing found regularities on new data that was not used in (searching) training procedure. However such testing does not allow to estimate a fraction of really existing but not discovered regularities. Besides it is very difficult (demands huge amount of data for testing) to estimate the exactness of calculated boundaries. So the most reliable approach is evaluating of performance at data where really existing regularities are known. Such situation may arise in artificially simulated tasks.

3.1 The scenario of simulation experiments

The experiments were limited by binary dependent variable $Y \in \{0,1\}$. The OVP procedures were considered with integral quality functional and squared difference as distance function. In other words $F_I(\tilde{Q}, \tilde{S}_0) = \sum_{j=1}^{r} [\nu(\tilde{S}_0) - \nu(\tilde{S}_j)]^2 m_j$ where $\nu(\tilde{S}_*)$ is fraction of "ones" in group $\tilde{S}_*$. Optimal partitions were searched inside families I, II, III. At the first stage of studies OVP-CIS procedure was used.

In each study datasets was generated according 6 specified by scenario basic regularities. The number of explanatory variables was always equal 16. The dependent variable was distributed at {0,1} with equal probabilities 0 and 1. The full number of independent variables was 16. The first 8 variables $X_1, \ldots, X_8$ were generated independently and distributed Uniform [0,1]. The last 8 variables were generated in accordance with regularities described by partitions from families I-III.

The following method of regularities generating was used. Suppose that we want to generate two-dimensional regularity R corresponding to partition $\tilde{Q}$ of variables X_{i_1} and X_{i_2} joint allowable area. Let conditional probability of 1 occurrence in partition $\tilde{Q}$ element q is referred to as $p_1(q)$. One element of partition is chosen (it will be denoted as q_{1d}) where $p_1(q_{1d})$ significantly deviates from p_1 for remaining elements for which p_1 values are equal each other. The regularities are generated according "expression levels" that are specified by scenario and defined as ratio

$$\kappa(R) = \frac{|\, p_1(q_{1d}) - p_1(\bar{q}_{1d})\,|}{\sqrt{\sqrt{p_1(q_{1d})[1-p_1(q_{1d})]}\sqrt{p_1(\bar{q}_{1d})[1-p_1(\bar{q}_{1d})]}}}, \qquad (1)$$

where $\bar{q}_{1d}$ is the union of all elements of $\tilde{Q}$ that do not coincide with q_{1d}. The generating of variables describing regularity was based on probabilities $p^{x}_{\tau}(q_{1d})$ that vector of involved independents belongs to q_{1d} in case dependent variable

is equal τ. Let symmetry constraints $p_\tau(\bar{q}_{1d}) = p_{1-\tau}(q_{1d})$ are introduced. Then taking into account that probabilities of 1 and 0 are equal it is easily to show that symmetry conditions are satisfied only when

$p^x{}_\tau(q_{1d}) = 1 - p^x{}_{1-\tau}(q_{1d})$ or $p^x{}_\tau(q_{1d}) = p^x{}_{1-\tau}(q_{1d})$, where last equality is not of interest because always leads to $\kappa(R) = 0$. It follows from $p^x{}_\tau(q_{1d}) = 1 - p^x{}_{1-\tau}(q_{1d})$ that $p_\tau(q_{1d}) = p^x{}_\tau(q_{1d})$ and $p_\tau(\bar{q}_{1d}) = 1 - p^x{}_\tau(q_{1d})$. So to satisfy equality (1) it is sufficient to choose $p^x{}_1(q_{1d})$ satisfying equality

$$\kappa(R) = \frac{|2p^x{}_1(q_{1d}) - 1|}{\sqrt{p^x{}_1(q_{1d})[1 - p^x{}_1(q_{1d})]}}. \qquad (2)$$

Two levels for probabilities $p^x{}_1(q_{1d})$ and $p^x{}_0(q_{1d})$ were used in simulation experiments

1) $p^x{}_1(q_{1d}) = 0.724$ and $p^x{}_0(q_{1d}) = 0.276$ correspond to $\kappa(R) \approx 1.0$ ($\kappa(R) = 1.0022...$ more exactly);

2) $p^x{}_1(q_{1d}) = 0.854$ and $p^x{}_0(q_{1d}) = 0.146$ correspond to $\kappa(R) \approx 1.0$ ($\kappa(R) = 2.0050...$ more exactly);

Suppose that we want to generate independent variable X, submitted to one-dimensional regularity. At first step random number γ from cut [0,1] was generated. In case $y = \tau$ and $\gamma < p^x_\tau(q_{1d})$ the series of random pairs from [0,1] is generated until some pair (χ_1, χ_2) from the series belongs to q_{1d}, $\tau \in \{0,1\}$. The generated values of variables X_{i_1} and X_{i_2} are set up equal χ_1 and χ_2 correspondingly. In case $y = \tau$ and $\gamma > p^x_\tau(q_{1d})$ the series of random pairs from [0,1] is generated until some pair (χ_1, χ_2) from the series belongs to $\bar{q}_{1d}$. The generated values of X_{i_1} and X_{i_2} again are set up equal χ_1 and χ_2.

The scenario included generating of two regularities for each partitions family with expression level κ equal 1.0 and 2.0. So generating of 6 basic regularities $R_1,\ldots,R_6$ is implied:

a) regularity R_1 with involved variable X_9 is described by partition from family I with boundary 0.5 and had $\kappa \approx 1.0$;

b) regularity R_2 with involved variable X_{10} is described by partition from family I with boundary 0.5 and had $\kappa \approx 2.0$;

c) regularity R_3 with involved variable X_{11} is described by partition from family II with boundary 0.25 and 0.75. and had $\kappa \approx 1.0$;

d) regularity R_4 with involved variable X_{12} is described by partition from family II with boundary 0.25 and 0.75. and had $\kappa \approx 2.0$;

e) regularity R_5 with involved variables X_{13} and X_{14} is described to partitions family III with boundary points for both variables equal 0.707 and had $\kappa \approx 1.0$;

f) regularity R_6 with involved variables X_{15} and X_{16} is described to partitions family III with boundary points for both variables equal 0.707 and had $\kappa \approx 2.0$;

Simulations results

In each study the OVP procedure ability was evaluated to reveal basic regularities specified by scenario. True basic regularity R involving some independent variable X_i (or variables X_i and X_j) was considered uncovered if there was a regularity in output set that involved X_i (X_i and X_j) and was found inside the same partition family that previously was used for R generating. To evaluate the similarity between found regularity and scenario regularity the two parameters b_{ms} and $b_{\max}$ were used, where b_{ms} is root mean square deviation between corresponding boundaries in two regularities, $b_{\max}$ is maximal by absolute value deviation between related boundaries.

In studies probabilities were also evaluated to include in output set mistakenly uncovered (false) regularities that did not correspond to scenario. The two types of false regularities may be distinguished. The first type (completely false) includes output regularities with all involved independent variables belonging to the first group (according the scenario Y was fully independent on this group). The second type includes only regularities found using more complicated partitions families II and III. The second type (partially false) included:

a) all regularities that was found using family II and involved variables from the set $X_9, X_{10}, X_{13}, \dots, X_{16}$;

b) all regularities that was found using family III and involved pairs of variables (X_{i_1}, X_{i_2}) where $i_1 \in \{1,\dots,8\}$ and $i_2 \in \{9,\dots,16\}$.

In other words second type included all optimal partitions where the including of only one of the boundaries or only one of the variables cannot be justified according scenario. It is evident that completely false regularities may involve only variables from the first group. So the maximal number of completely false regularities is 8 for I and II partitions families and 28 for partitions family III in generated according scenario datasets. Scenario also potentially admit occurrences of 6 partially false regularities from family II and 64 partially false regularities from family III .

Besides correctly recognized basic regularities and completely or partially false regularities the output set usually contains partitions corresponding to really existing in data regularities that are accessory elements of scenario. For example joint distribution of variables X_9 and X_{10} may be discussed as regularity described by the partition from family III, distributions of variables $X_{11} - X_{16}$ may be discussed as regularities described by the partition from family I. The accessory regularities are rather numerous, their expression levels varies significantly and cannot be evaluated directly in the same terms as expression levels of basic regularities. So consideration of accessory regularities too

complicates the interpretation of results and the ability of OVP procedures to recognize them was not considered in these studies.

In each experiment OVP procedure performance was evaluated by 10 datasets generated according scenario. The number of permutations in tests was everywhere equal 1000.

Experiment 1. The goal of experiment 1 was to evaluate the performance of OVP-CIS procedure in generated according scenario datasets with 100 observations. The results of experiment are represented in tables 1 and 2. In table I abilities of OVP-CIS to recognize all 6 basic types of regularities are represented as full numbers and fractions (in parentheses) of found regularities. For each type R_* mean values of parameters $b_{\mathrm{m}s}$ and $b_{\max}$ are represented that were calculated by all uncovered basic regularities of R_* type in 10 dataset. The table 2 represents full numbers of completely false and partially false regularities from all 3 partitions families. Fractions of found false regularities among potentially possible variants are given in parentheses.

Table 1. **Detection capabilities of OVP-CIS procedure for all 6 basic regularities**

Regularity	$p_{thr.} = 0.01$	$p_{thr.} = 0.001$	$b_{\mathrm{m}s}$	$b_{\max}$
R_1	9 (90%)	8(80%)	0.055	0.151
R_2	10(100%)	10(100%)	0.022	0.066
R_3	9(90%)	5(50%)	0.041	0.124
R_4	10(100%)	10(100%)	0.042	0.136
R_5	5(50%)	3(30%)	0.077	0.167
R_6	10(100%)	10(100%)	0.024	0.058

Table 2. Completely false and partially false regularities rates for OVP-CIS Procedure in datasets with numbers of observations 100

Used partitions family	$p_{thr.} = 0.01$		$p_{thr.} = 0.001$	
	Completely false	Partially false	Completely false	Partially false
I	0	-	0	-
I	0	42(70%)	0	36(60%)
III	1(0.4%)	386(60.3%)	0	295(46%)

It is seen that OVP-CIS procedure demonstrates good ability to uncover specified by scenario regularities. The only exclusion is type R_5 where 50% regularities were recognized. The found boundaries also turned out to be rather close to specified boundaries. The root mean square deviations b_{ms} nowhere exceeded 0.1 and maximal deviations $b_{\max}$ nowhere exceeded 0.2. Remind that all explanatory variables ranged from 0 to 1.0. The numbers of completely false regularities in output set are also very small and they are in good accordance with expectation. But as it may be seen from table 2 the numbers of partially false regularities from both partitions families II an III were very high. Fractions of found partially false regularities among variables or combinations of variables where they may occur in principle achieved 60-70%. The great amount of partially false regularities may significantly hamper the analysis of results or may be the cause of mistaken conclusions. So the modified OVP procedure (OVP-CDS) was suggested that allows to eliminate partially false regularities.

The OVP-CDS procedure. The basic idea underlying this modification of OVP method is selecting to output set only those optimal partitions from more complicated families II and III where variations between induced groups of observations can not be explained from the viewpoint of previously found regularities from simplest family I. In other words selecting of partitions from

complicated families in OVP-CDS (complexity dependent selecting) is based on testing if Y is independent on explanatory variable (variables) inside subregions belonging to simple regularities involving these explanatory variable (variables). So OVP-CDS includes different selecting modes for optimal partitions from family I and optimal partitions from more complicated families. Selecting of partitions from family I in OVP-CDS always precede selecting of optimal partitions from families II and III. Then the second variant of permutation test is used to evaluate the validity of the last. Assume that uncovered regularities from family I involving variables X' and X'' are contained in the output set. The first from these simple regularities includes subregions q'_1, q'_2 and second regularity includes subregions q''_1, q''_2. Then optimal partition from family II involving variable X' is put to the output set only if $p_2(q'_1, q'_2)$-values is less than threshold $p_{thr.}$. Optimal partition from family III involving variables X' and X'' is placed to the output set only if both inequality $p_2(q'_1, q'_2) < p_{thr.}$ and $p_2(q''_1, q''_2) < p_{thr.}$ are satisfied. In case output regularities from family I do not involve variables used in optimal partitions from more complicated families II and III the selecting procedure for the last partitions are the same as in OPV-CIS.

Experiment 2. The goal of experiment 2 was to evaluate the performance of OVP-CDS procedure in generated according scenario datasets with 100 observations. The results are represented in tables 3 and 4. As it is seen from table 3 the ability of OVP-CDS to recognize basic regularities in 10 generated dataset is only slightly lower than corresponding ability of OVP-CIS procedure. The mean values of $b_{\max}$ for OVP-CDS are the same as for OVP-CIS and mean values of b_{ms} for two procedures are close also. The rate of completely false regularities is as low as in OVP-CIS. But the rate of partially false regularities for OVP-CDS decreased dramatically as compared with OVP-CIS. The number of partially false regularities from partitions family II placed to the output sets at $p_{thr.} = 0.01$ dropped from 42 to 1. The number of partially

false regularities from partitions family III placed to the output sets at $p_{thr.} = 0.01$ dropped from 386 to 5. The fractions of partially false regularities among possible in principle variants became close to used threshold values $p_{thr.}$.

Table 3. Detection capabilities of OVP-CDS procedure for all 6 basic regularities in datasets with numbers of observations 100

Regularity	$p_{thr.} = 0.01$	$p_{thr.} = 0.001$	b_{rms}	$b_{\max}$
R_1	9(90%)	8(80%)	0.055	0.151
R_2	10(100%)	10(100%)	0.022	0.066
R_3	9(90%)	3(30%)	0.049	0.124
R_4	10(100%)	8(80%)	0.033	0.136
R_5	5(50%)	3(30%)	0.072	0.167
R_6	9(90%)	8(80%)	0.025	0.058

Table 4. Completely false and partially false regularities rates for OVP-CDS Procedure in datasets with numbers of observations 100

Used partitions family	$p_{thr.} = 0.01$		$p_{thr.} = 0.001$	
	Completely false	Partially False	Completely false	Partially False
I	0	_	0	_
I	0	1(1.7%)	0	0
III	1(0.4%)	5(0.7%)	1(0.4%)	0

Experiment 3. The goal of experiment 3 was to examine dependence of OVP-CDS performance on datasets size. So the ability of procedure to uncover specified basic regularities and to reject false regularities was additionally

evaluated in datasets with 70, 120 and 170 observations. It is seen from results of experiments 2 and 3 represented in tables 3,4 and 5, 6 that ability to recognize basic regularities increased gradually together with datasets size. The number of correctly recognized at $p_{thr.} = 0.01$ basic regularities from family III with expression level 1 (type R_5 **)** increased from 2 (70 observations in datasets) to 7 (170 observations in datasets). The number of correctly recognized at $p_{thr.} = 0.001$ basic regularities from family II with expression level 2 (type R_4 **)** increased from 6 (70 observations in datasets) to 10 (170 observations in datasets).

Table 5. Detection capabilities of OVP-CDS procedure for all 6 basic regularities in datasets with numbers of observations 70, 120, 170.

Datasets Size	Regularity	$p_{thr.} = 0.01$	$p_{thr.} = 0.001$	b_{ms}	$b_{\max}$
70	R_1	8(80%)	6(60%)	0.066	0.151
70	R_2	10(100%)	10(100%)	0.011	0.028
70	R_3	7(70%)	2(20%)	0.050	0.124
70	R_4	7(70%)	6(60%)	0.042	0.123
70	R_5	2(20%)	0	0.055	0.096
70	R_6	8(80%)	3(30%)	0.077	0.215
120	R_1	10(100%)	10(100%)	0.016	0.036
120	R_2	10(100%)	10(100%)	0.014	0.037
120	R_3	8(80%)	4(40%)	0.032	0.064
120	R_4	10(100%)	10(100%)	0.017	0.045
120	R_5	3(30%)	2(20%)	0.040	0.071
120	R_6	9(90%)	8(80%)	0.039	0.113
170	R_1	10(100%)	10(100%)	0.016	0.032

170	R_2	10(100%)	10(100%)	0,009	0.023
170	R_3	10(100%)	6(60%)	0.034	0.087
170	R_4	10(100%)	10(100%)	0.012	0.036
170	R_5	7(70%)	5(50%)	0.044	0.105
170	R_6	10(100%)	10(100%)	0.022	0.066

Table 6. Completely false and partially false regularities rates for OVP-CDS Procedure in datasets with numbers of observations 70, 120, 170.

Used partitions family	$p_{thr.} = 0.01$		$p_{thr.} = 0.001$	
	Completely false	Partially False	Completely false	Partially false
I	0	–	0	–
II	0	1(1.6%)	0	0
III	1(0.3%)	6(1.0%)	0	0
I	0	–	0	–
II	0	1(1.6%)	0	0
III	1(0.3%)	5(0.8%)	1(0.3%)	0
I	1(1.6%)	–	0	–
II	0	0	0	0
III	0	2(0.3%)	0	0

3. Illustrative Examples

The goal of this section is demonstrating OVP techniques use in different application tasks. The another goal is to compare OVP performance with performance of standard statistical tests. Represented below 5 regularities are related from 5 different tasks with binary dependent variable. Regularities were found using all 3 abovementioned partitions families.

Task 1. The goal of task 1 was to evaluate relationship between migration balance in Russian Federation territorial units and corresponding social and economical indices. In present paper group of 23 territorial units with positive balance is compared with group of 53 territorial units with negative balance by 3 variables.

Task 2. . The goal of task 2 was predicting of long term outcomes in patients with combat psychological trauma by set of psychometric indices. The first group includes 26 patients for which treatment was successful and 14 patients that suffered from trauma consequence more than 2 years after the treatment.

Task 3 The goal of task 3 was predicting from so called "genetic" variables whether or not osteosarcoma patient will survive more than one year after the treatment. "Genetic" variables characterize optical density of tumor cells nuclei. Two groups of patients are compared. The first one (those who was alive after 1 year after the chemiotherapy treatment) includes 25 patients and the second group includes 49 patients.

Task 4 The goal of task 4 was predicting utera mioma relapse from immunological parameters. The group of 6 patients with relapse is compared with 15 patients for which relapse took place before 2 years after operation.

Task 5 Wines that are grown in the same region of Italy but are derived from different cultivars are compared by quantities of chemical constituents**.** Data related to this task was taken from UCI Machine Learning Repository. Here we compared 1-st (59 instances) and 2-nd (71 instances) types.

Datasets that are related to regularities found for these 5 tasks are represented in appendix 2.

Example1. Univariate regularity with one boundary point related to task1

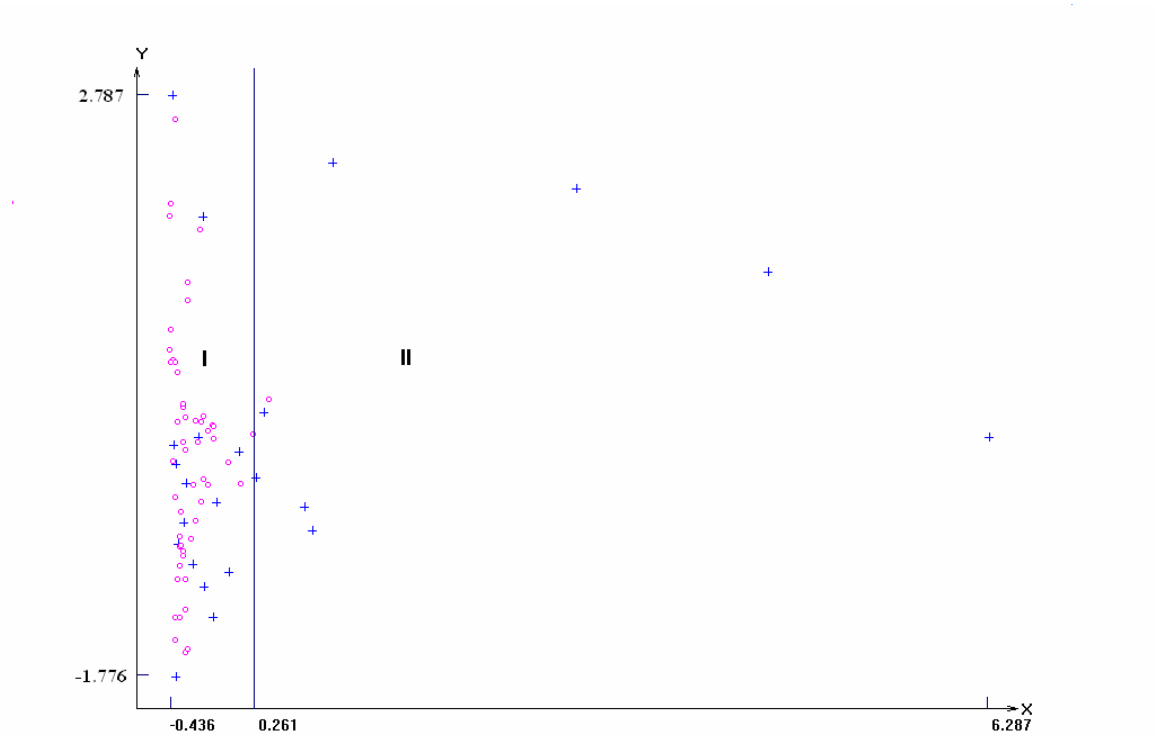


Fig. 1. Optimal 1-dimensional regularity related to dependence of migration balance on variables 1 in Task 1 (see Supplement, table). Var. 1 corresponds to X, var. 2 corresponds to Y.
Quadrant I – number of territorial units with positive balance (+) -15, number of territorial units with negative balance(o) – 52;
Quadrant I I– positive balance -8, negative balance – 1;
It is seen that regions with positive migration balance absolutely prevail when var1>0.261.

Table 1. Validity according standard statistical tests and OVP technique

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value	0.0008	<0.025	0.009777	0.0007(PF-I,PT-1)

So high statistical significance of difference between two groups of regions by variable 1 is evaluated by all univariate tests. Statistical validity according OVP is practically equal to univariate ANOVA significance .

Example 2 . Univariate regularity with one boundary point related to Task 2.

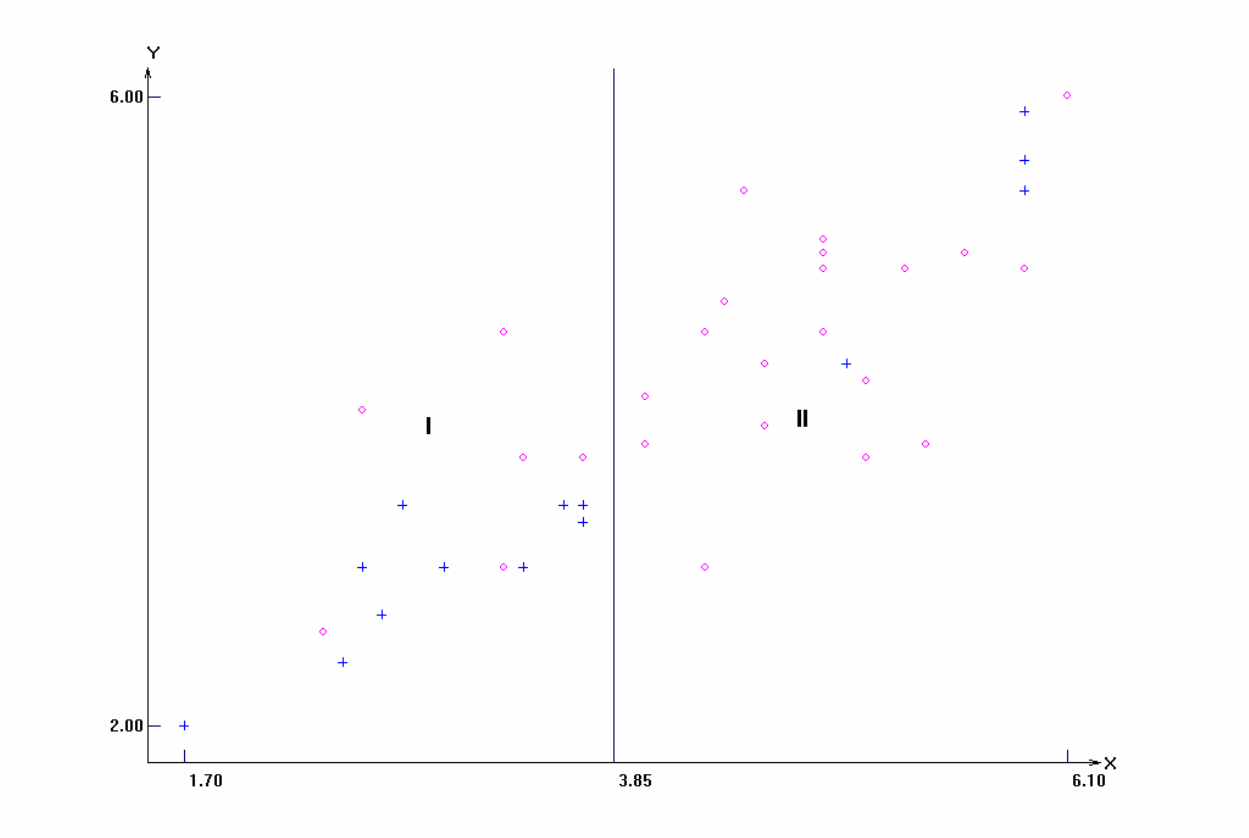


Fig. 1 Optimal 1-dimensional regularity related to dependence long term consequences of psychological trauma on variable 2 in Task 2 (see Supplement, table). Var. 2 corresponds to X, var. 1 corresponds to Y.
Quadrant I – number of patients with long term consequences (+) -10, number of patients with successful treatment(o) – 7;
Quadrant I I– long term consequences -4, successful treatment – 19;
It is seen that fraction of patients with with long term consequences decreases dramatically for cases with var2>3.85.

Table 1. Validity according standard statistical tests and OVP technique

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value	0.105504	p < .10	0.143594	0.0447

Statistical significance at level better 0.05 is evaluated only by OVP method. .

Example 3 . Univariate regularity with one boundary point related to Task 5.

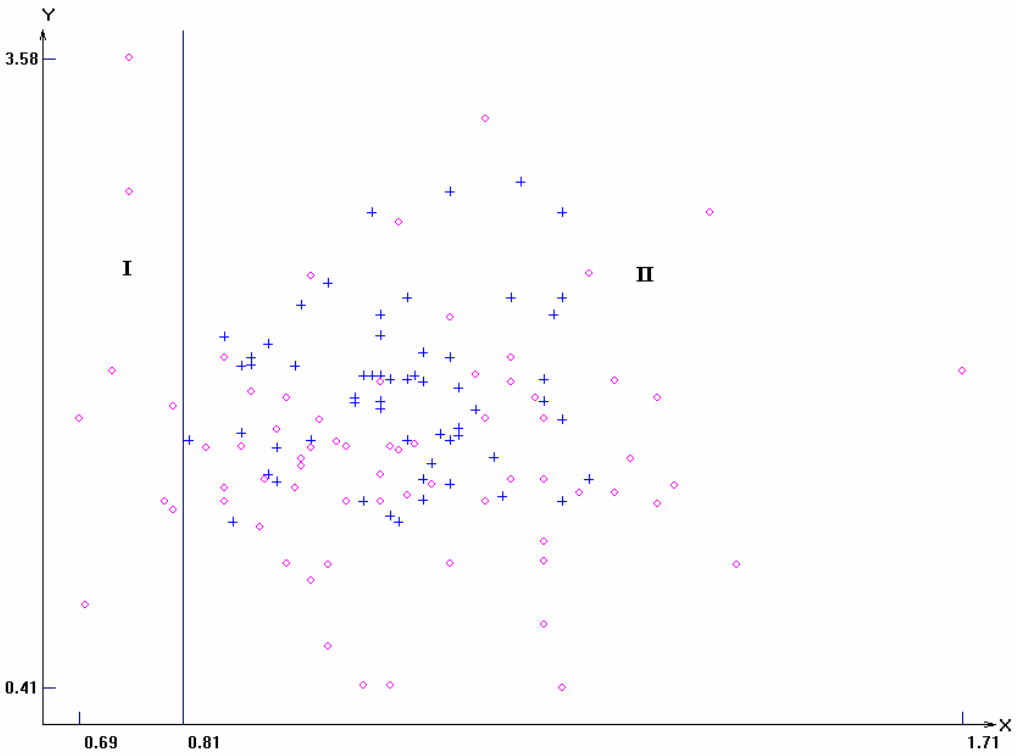


Fig. 1 Optimal 1-dimensional regularity related to relationship between types of wine and chemical constituent corresponding to variable 2 in Task 2 (see Supplement, table). Var.2 corresponds to X, var.1 corresponds to Y.
Quadrant I – instances from 1st type (+) -0, number of instances from 2nd type (o) – 9;
Quadrant I I– instances from 1st type -59, number of instances from 2nd type – 62;

It is seen that second type of wine absolutely prevail when var1<0.261.

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value	0.847388	P>0.1	0.560270	0.045 (PF I, PT-1)

Example 4 . Univariate regularity with two boundary point related to Task 4.

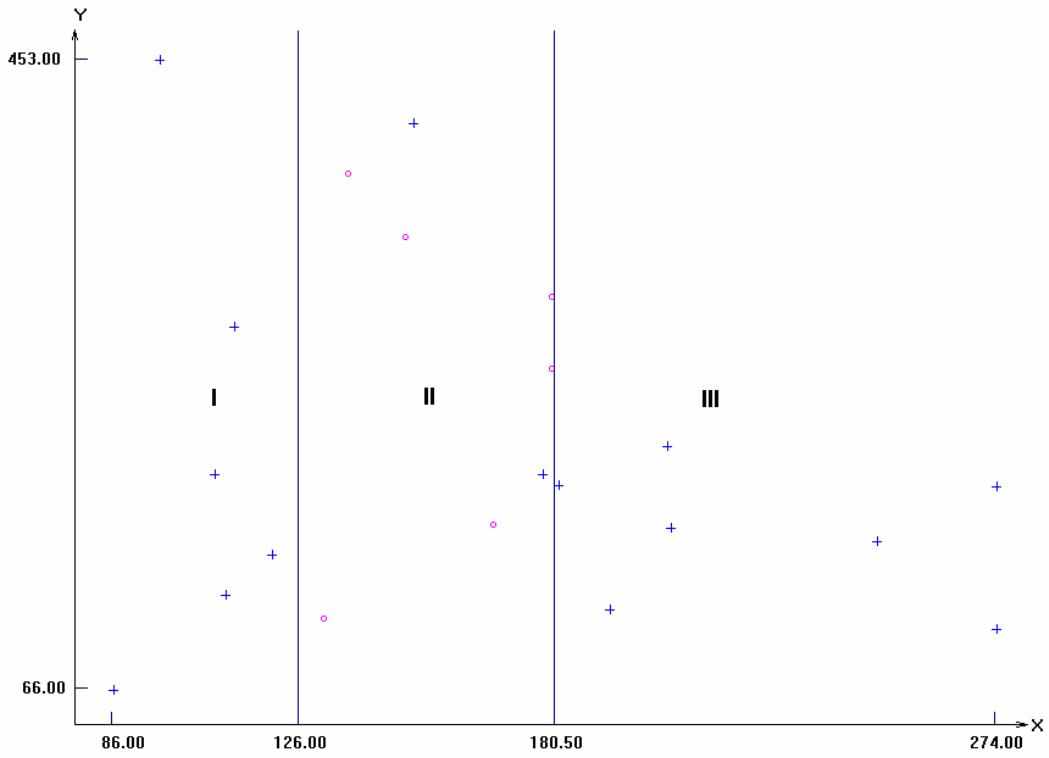


Fig. 1 Optimal 1-dimensional regularity with two boundary points related to dependence of relapse occurrence on variable 1 Task 3 (see Supplement, table). Var. 1 correspond to X, var. 2 correspond to Y, .
Quadrant I – number of patients without relapse(+) -6, number of patients with relapse (o) – 0;
Quadrant I I– without relapse -2, with relapse – 6;
Quadrant III – without relapse -7, without relapse 0;
It is seen from figure 1 that variable 1 values in patients with relapse are concentrated inside middle interval: 126.0 <var1<180.5.

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value	0.672450	>0.1	0.755497	0.013 (PF-II,PT-1)

Standard univariate statistical tests and OVP with one boundary point did not reveal any dependence of relapse on variables 1

Example 5 . Univariate regularity with two boundary point related to Task 5.

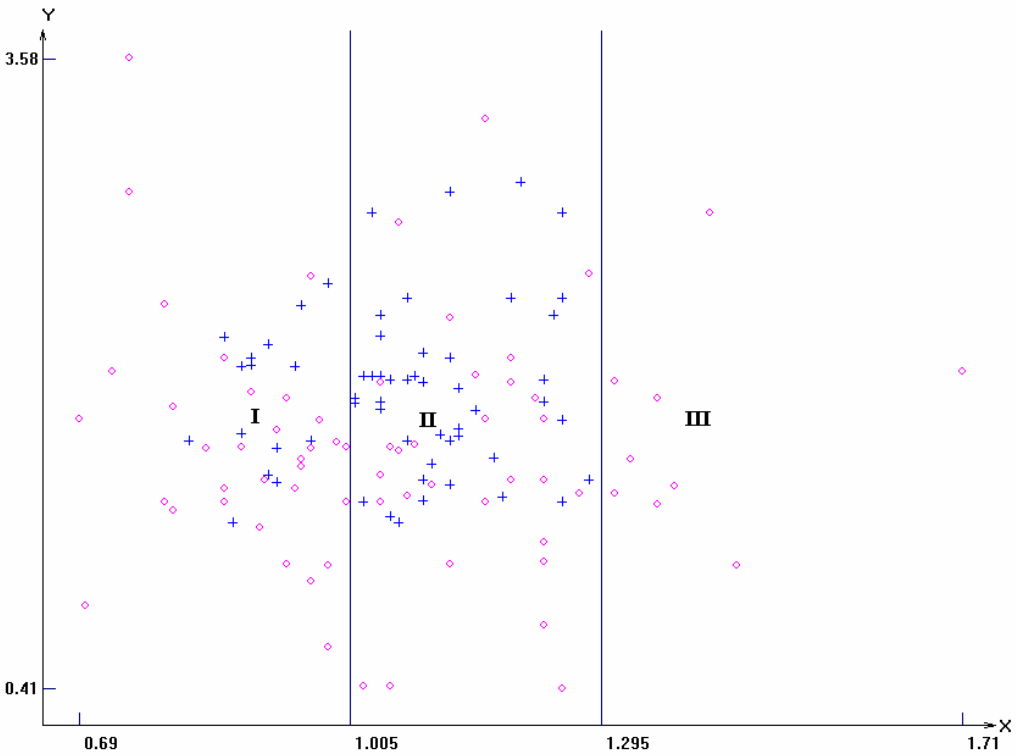


Fig. 1 Optimal 1-dimensional regularity related to relationship between types of wine and chemical constituent corresponding to variable 2 in Task 2 (see Supplement, table). Var.2 corresponds to X, var.1 corresponds to Y.
Quadrant I – instances from 1st type (+) -15, number of instances from 2nd type (o) – 33;
Quadrant I I– instances from 1st type -44, number of instances from 2nd type – 29;
Quadrant I I– instances from 1st type -0, number of instances from 2nd type – 9;
It is seen that fraction of wine type1 in middle quadrant II is significantly greater than fractions of of wine type1 in neighboring quadrants.

Table 1. Validity according standard statistical tests and OVP technique

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value	0.847388	P>0.1	0.560270	0.045 (PF I, PT-1) $p_2(q_1^3, q_2^3)$ =.0675 (PF-II,PT-2) 0.0065 (PF II, PT-1)

Comment, Here $p_2(q_1^3, q_2^3)$ is p-value calculated by permutation test PT-2. Subregions q_1^3 and q_2^3 correspond to Quadrants I and II from **example 3**.

Standard univariate statistical tests did not evaluate difference between two types of wine by variable 2 as significant. However OVP methods reveal rather valid regularities (see example 3 also). The highest level of significance (null hypothesis about independence of wine type on var.2 is rejected at p=0.0065) is receive by partitioning with 2 boundary points. Not so poor $p_2(q_1^3, q_2^3)$ =0.0675 indicates that two boundary regularity from example 5 can be explained by more simple regularity from example 3.

Example 6 . Two-variate regularity with two boundary point related to Task 1.

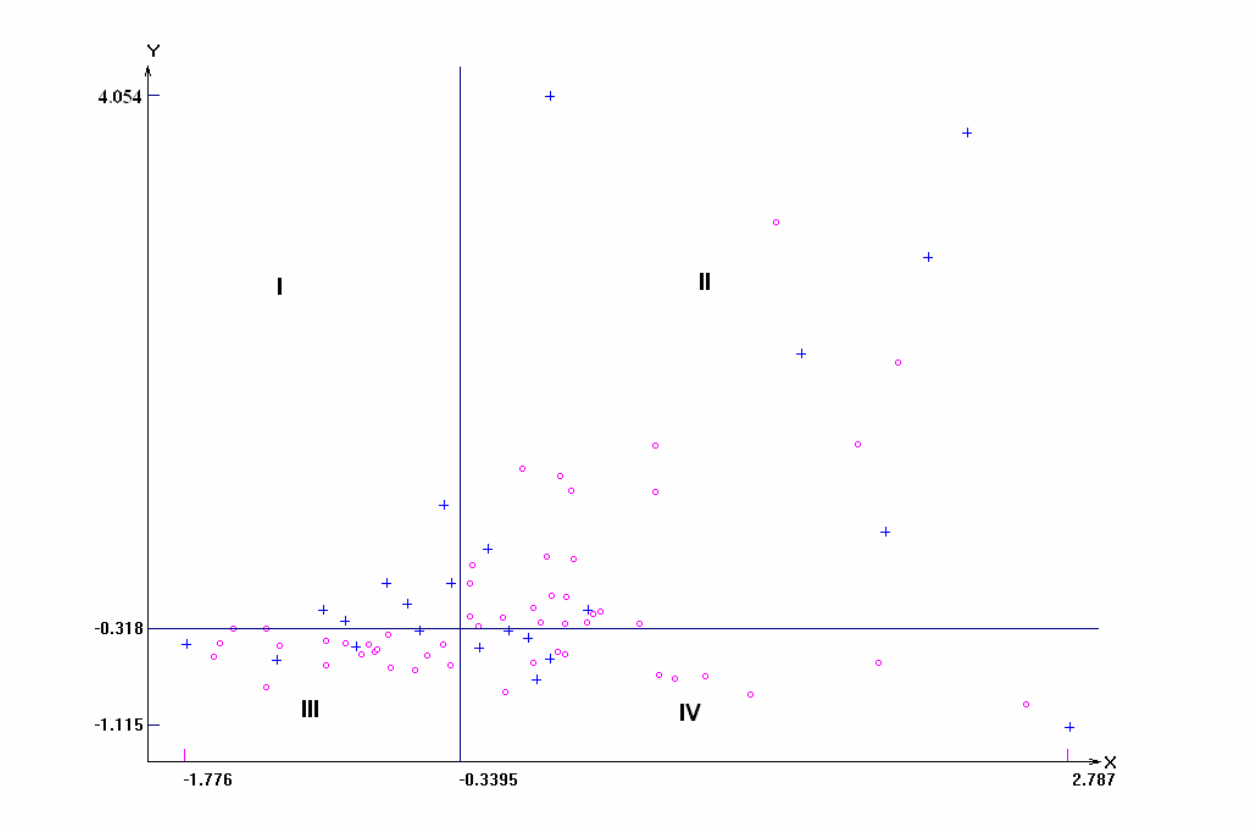


Fig. 1 Optimal 2-dimensional regularity related to dependence of migration balance on variables 2 and 3 in Task 1 (see Supplement, table). Var. 2 correspond to X, var. 3 correspond to Y, .
Quadrant I – number of regions with positive balance (+) -7, number of regions with negative balance(o) – 0;
Quadrant I I– positive balance -7, negative balance – 24;
Quadrant III – positive balance -6, negative balance – 10;
Quadrant IV – positive balance -3, negative balance – 19;
It is seen from figure 1 strong dependence of migration balance on variable 3 in case var2<-.03395, but in case var2>-.03395 a distinct dependence of var2<-.03395 on variable 3 is not observed. Statistical validity of regularity according PT-1 is p=0.014

Table 1. Validity according standard statistical tests and OVP technique

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value var 2	0.686	p>0.1	0.768	0.46 (PF-I, PT-1)
p-value var 3	0.0398	P>0.1	0.062889	0.17(PF-I, PT-1)
2- variate p-value	0.109	-	-	0.014(PF-III, PT-1)

ANOVA F-test reveals valid (p=0.0398) difference between two groups of regions by variable 3 This difference may be related to group of 4 regions in quadrant II with positive balance and high values of variable 3. All univariate tests did not discover any difference between groups of regions by variable 2. No difference was indicated also by 2-variate ANOVA.

Example 7 . Two-variate regularity with two boundary point related to Task 3.

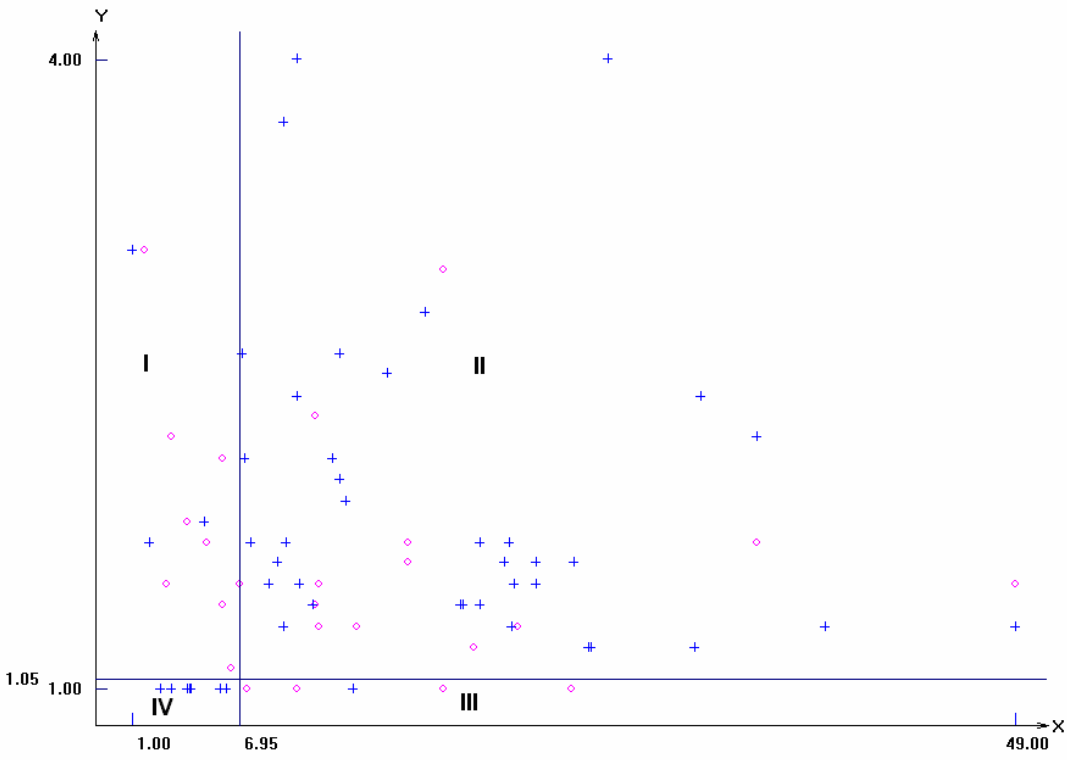


Fig. Optimal 2-dimensional regularity related to dependence of migration balance on variables 1 and 2 in Task 3 (see Supplement, table). Var. 1 corresponds to X, var. 2 corresponds to Y, .
Quadrant I – number of non survivors (+) -3, number of survivors(o) – 9;
Quadrant I I– number of non survivors -37, number of survivors – 12;
Quadrant III – number of non survivors -1, number of survivors – 4;
Quadrant IV – – number of non survivors - 8, number of survivors – 0;
It is seen that in patients for which value of only one of variables is low the survival is better than survival in patients for which values of both variables are high. However in patients for which value of both variables are low the survival is very poor. Statistical validity of regularity according PT-1 is p=0.01

Table 1. Validity according standard statistical tests and OVP technique

	ANOVA	Kolmogorov-Smirnov Test	Mann-Whitney U Test	OVP
p-value var 2	0.502862	p>0.1	0.400850	0.672 (PF-I,PT-1)
p-value var 3	0.264284	p>0.1	0.373187	0.6033
2- variate p-value	0.518	-	-	0.01(PF-III,PT-1)

Standard univariate statistical tests, and 2-variate ANOVA did not reveal any dependence of survival on variables 1 and 2. Univariate OVP methods with one boundary points also did not reveal regularities. So there is no need to use PT-2 and only PT-1

It is seen that OVP using sometimes allows to reveal regularities when commonly used univariate statistical test or ANOVA fail to do it. OVP methods preserve high stability in case existence in data strongly outlying observations.

4 Conclusion

The results of represented studies may be summarized as follows.

The optimal valid partitioning procedure (OVP-CIS) using for selecting of regularities to output set variant of PT testing the null hypothesis about full independence of variable Y on explanatory variables demonstrated good ability to uncover specified by scenario regularities and to reject completely false regularities(regularities involving only variables without any effect on Y). However the serious drawback of OVP-CIS is including to output set great number of partially false regularities that do not fully correspond to scenario but involve variable having effect on dependent Y.

To improve ability of optimal valid partitioning to reject partially false regularities the OVP-CDS procedure was developed that is based on verifying of complicated regularities from the viewpoint of simplest regularities found for the same variables. Experiments demonstrated the good ability of OVP-CDS procedure to uncover specified by scenario regularities and to reject partially false regularities so as completely false. So OVP-CDS procedure may be recommended as the tool of data analysis.

The serious disadvantages of discussed techniques are the great amount of calculating that is necessary to verify regularities by permutation tests. So OVP-CDS calculating in dataset with 170 observations and 16 continuous explanatory variables and 1000 permutations in PT is fulfilled for 14 minutes at PC Pentium(R) 4 CPU 2.60GHz.. One of the ways to increase size of datasets that are accessible for discussed method is using of parallel computing. It must be noted here that initial task of OVP analysis is easily divided on subtasks that can be calculated independently. Another way of datasets size increasing is

preschedule finishing of PT calculating if results for preceding set of permutations strongly testify against validity of tested partition.

References

Abdolell M., LeBlanc M., Stephens D., Harrison R.V. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Statistics in Medicine.2002, 21:3395-3409.

Borovkov A.A. (1983) Mathematical statistics. Evaluating of parameters, hypothesis testing. M: Nauka. (in Russian)

Breiman, L., Freidman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Chapman & HALL/CRC.

Chitchian, R., Safaryan, I. (2001) A Nonparametric Approach to Bivariate Dependence Models Comparison. Computer Science and Information Technologies. Proceedings of the conference. Erevan , Armenia.

T.W. O'Gorman An adaptive permutation test procedure for several common test of significance. Computational Statistics & Data Analysis. 35(2001) 265-281.

Mazumdar, M., Glassman, JR. Tutorial in Biostatistics. Categorizing a prognostic variable: review of methods, coding for easy implementation and applications to decision making about cancer treatment. Statistics in Medicine.2000, 19:113-132.

Senko O.V., Kuznetsova A.V., Kropotov D.A. (2003). The Methods of Dependencies Description with the Help of Optimal Multistage Partitioning. **Proceedings** of the 18th International Workshop on Statistical Modelling Leuven, Belgium, 2003, pp. 397-401.

Sen'ko O.V., Kuznetsova A.V. (1998). The use of partitions constructions for stochastic dependencies approximation. Proceedings of the International conference on systems and signals in intelligent technologies. Minsk (Belarus), pp. 291-297.

Kuznetsova A.V., Sen'ko O.V., Matchak G.N., Vakhotsky V.V., Zabotina T.N.,

Korotkova O.V. The Prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges //J. Theor. Med., 2000, Vol. 2, pp.317-327.

Appendix 1.

Let observations are independent and are elements of probability space $(\Omega,\Sigma,\mathbf{P})$. It is considered that observations in datasets are ordered (according for example to moments of their coming in). **So** datasets of size m are elements of probability space $(\Omega^m,\Sigma^m,\mathbf{P}^m)$ that is the product of m spaces $(\Omega,\Sigma,\mathbf{P})$. We shall consider that Σ consists of finite number of elements. As a matter of fact this point is not essential from practical point of view because descriptions of physical objects are always have finite length.

Let $\tilde{S}_0=\{(\mathcal{Y}_1,\mathbf{x}_1),\ldots,(\mathcal{Y}_m,\mathbf{x}_m)\}$. Assume that $F_*^O(\tilde{S})$ is the value of used quality functional at dataset $\tilde{S}$ that is calculated for optimal partition inside fixed partitions family. The permutation test in OVP method is based on testing of the null hypothesis $\mathbf{H}_o$ that $\mathcal{Y}$-components of observations from $\tilde{S}_0$ do not depend on $\mathbf{x}$ components and probability $\mathbf{P}^m\{F_*^O(\tilde{S})>F_*^O(\tilde{S}_0)\,|\,\mathbf{H}_0,\tilde{S}\in W(\tilde{S}_0)\}$ is used as p-value, $W(\tilde{S}_0)\subseteq\Omega^m$ is defined below family of datasets that are in certain way close to $\tilde{S}_0$.

Assume that $\tilde{\mathcal{Y}}(\tilde{S}_0)=\{(\mathcal{Y}_1^u,\ldots,\mathcal{Y}_{m_y}^u\}$ is the set of all not equal each other values of $\mathcal{Y}$- components in observations from $\tilde{S}_0$ and $\tilde{\mathbf{x}}(\tilde{S}_0)=\{\mathbf{x}_1^u,\ldots,\mathbf{x}_{m_y}^u\}$ is the set of all not equal each other values of $\mathbf{x}$- components in observations from $\tilde{S}_0$, where $m_x,m_y\in\{2,\ldots,m\}$. Let $\hat{l}_i^{\,y}(\tilde{S})$ is number of $\mathcal{Y}_i^u\in\tilde{\mathcal{Y}}(\tilde{S}_0)$ occurrences and $\hat{l}_i^{\,x}(\tilde{S})$ is number of $\mathbf{x}_i^u\in\tilde{\mathbf{x}}(\tilde{S}_0)$ occurrences in some dataset $\tilde{S}$.

Definition 1. The set $\tilde{S}$ belongs to $W(\tilde{S}_0)$ if and only if all following conditions are satisfied:

a) it consists of m observations;

b) the X - components in observations from $\tilde{S}$ coincides with X - components in observations from $\tilde{S}_0$**;**

c) $\forall i \in \{1,\ldots,m_y\}$ equality $\hat{l}_i^{\,y}(\tilde{S}) = \hat{l}_i^{\,y}(\tilde{S}_0)$ is correct. ||

It is evident that arbitrary permuting of $\mathcal{Y}$- components positions in $\tilde{S}_0$ with fixed order of X - components produces dataset belonging to $W(\tilde{S}_0)$.

Definition 2. We shall say that datasets $\tilde{S}' = \{(\mathcal{Y}_1', \mathbf{x}_1), \ldots, (\mathcal{Y}_m', \mathbf{x}_m)\}$ and $\tilde{S}'' = \{(\mathcal{Y}_1'', \mathbf{x}_1), \ldots, (\mathcal{Y}_m'', \mathbf{x}_m)\}$ from $W(\tilde{S}_0)$ belong to the same distinct type if $\mathcal{Y}_i' = \mathcal{Y}_i''$ $\forall i \in \{1,\ldots,m\}$. So distinct type is the set of datasets with the same values of $\mathcal{Y}$- components and X - components for each ordinal position.||

The following theorem is true.

Theorem 1. Let $N_{gt}[F_*^o(\tilde{S}_0)]$ - is the number of such different permutations of $\mathcal{Y}$- components positions in $\tilde{S}_0$ with fixed order of X - components, that for resulting dataset $\tilde{S}$ inequality $F_*^o(\tilde{S}) > F_*^o(\tilde{S}_0)$ is satisfied. Let null hypothesis $\mathbf{H}_o$ is true. Then

$$\mathbf{P}\{F_*^o(\tilde{S}) > F_*^o(\tilde{S}_0) \,|\, \mathbf{H}_0, \tilde{S} \in W^p(\tilde{S}_0)\} = N_{gt}[F_*^o(\tilde{S}_0)]/m!. \;||$$

Proof. It is easily to show that each distinct type from $W(\tilde{S}_0)$ can be received from $\tilde{S}_0$ with the help of the same number of different permutations of $\mathcal{Y}$-components ordinal positions with fixed order of X -components**.** Really, each permutation transforming $\tilde{S}_0$ to dataset of distinct type $\tilde{S}_k^d$ may be represented as product of permutations π_{0k} and π_{gk}, where π_{0k} is permutation transforming $\tilde{S}_0$ to one of datasets of type $\tilde{S}_k^d$ and $\pi_{dk} \in \Pi_{dk}$, where Π_{dk} is group of permutations

of $\mathcal{Y}$- components in $\tilde{S}_k^d$ that preserve this distinct type. The number of permutation in Π_{dk} is the same for all distinct types and equal $\{\prod_{i=1}^{m_y} l_i^y !\}$. Using $\mathbf{H}_o$ and independence of observations we receive that all distinct types have the same probability $\mathbf{P}^m(\tilde{S}_k^d) = \{\prod_{i=1}^{m_y}[\mathbf{P}(\mathcal{Y}_i^u)]^{l_i^y}\}\{\prod_{i=1}^{m_x}[\mathbf{P}(\mathbf{x}_i^u)]^{l_i^x}\}$. The value F_{opt} is unique of course inside each distinct type.

So $\mathbf{P}^m\{F_*^o(\tilde{S}) > F_*^o(\tilde{S}_0) \mid \mathbf{H}_0, \tilde{S} \in W^p(\tilde{S}_0)\}$ may be calculated as ratio $\frac{|\{\tilde{S}_k^d \subset W^p(\tilde{S}_0) \mid F_*^o(\tilde{S}_k^d) > F_*^o(\tilde{S}_0)\}|}{|\{\tilde{S}_k^d \subset W^p(\tilde{S}_0)\}|}$. Taking also into account that numbers of permutations generating distinct types from $\tilde{S}_0$ are the same for all distinct types we receive theorem conclusion. ||

The direct calculating of $N_{gt}[F_*^o(\tilde{S}_0)]$ is possible only for datasets of very limited number size. However ratio $N_{gt}[F_*^o(\tilde{S}_0)]/m!$ can be estimated with the help of Monte-Carlo technique. Set of random permutations $\{\pi_1^r,\ldots,\pi_{\mathcal{N}}^r\}$ is used to generate datasets $\{\tilde{S}_1^r,\ldots,\tilde{S}_{\mathcal{N}}^r\}$ from initial dataset $\tilde{S}_0$ by the technique mentioned in the proof of theorem. To receive dataset $\tilde{S}_i^r$ random permutation π_i^r of $\mathcal{Y}$- components numbers is used with the fixed order of $\mathbf{x}$- components. For each dataset from $\{\tilde{S}_1^r,\ldots,\tilde{S}_{\mathcal{N}}^r\}$ the same optimal partitioning procedure is used as it was previously used in case of initial $\tilde{S}_0$ and the optimal value of quality functional is calculated. The ratio $N_1^p[F_*^o(\tilde{S}_0)]/m!$ is estimated with the help of ratio $\mathcal{N}_{gt}[F_*^o(\tilde{S}_0)]/\mathcal{N}$, where $\mathcal{N}_{gt}[F_*^o(\tilde{S}_0)]$ is the number of datasets in $\{\tilde{S}_1^r,\ldots,\tilde{S}_{\mathcal{N}}^r\}$ for which $F_*^o(\tilde{S}_*^r) > F_*^o(\tilde{S}_0)$.

Appendix 2

Table A1. Data related to regularity from Task 1

Indicates if migration balance is positive	Var . 1	Var . 2	Var . 3	Indicates if migration balance is positive	Var . 1	Var . 2	Var . 3
1	-0.31	-0.255	-0.457	2	-0.376	-1.505	-0.337
1	-0.254	-0.897	-0.445	2	-0.384	2.584	-0.949
1	0.732	-0.627	-0.099	2	-0.413	0.928	-0.72
1	-0.16	-1.066	-0.152	2	-0.277	1.165	-0.878
1	6.287	0.117	4.054	2	-0.425	0.776	-0.741
1	-0.375	-0.728	0.072	2	-0.4	0.692	-0.712
1	-0.326	-0.559	-0.316	2	-0.31	0.32	-0.287
1	-0.096	-1.303	-0.564	2	-0.146	-0.238	-0.307
1	-0.387	-1.776	-0.431	2	-0.108	-0.289	-0.233
1	0.657	-0.441	0.713	2	0.397	0.387	-0.19
1	-0.414	0.049	-0.727	2	-0.394	-0.103	-0.848
1	-0.436	2.787	-1.115	2	-0.203	-0.576	-0.667
1	0.134	-0.001	-0.383	2	-0.161	0.201	-0.543
1	-0.216	0.117	-0.552	2	-0.313	-0.846	-0.545
1	0.324	0.303	-0.155	2	-0.064	0.168	-0.531
1	0.044	-0.948	-0.241	2	-0.343	-0.694	-0.655
1	0.265	-0.204	0.357	2	0.154	-0.272	0.18
1	4.486	1.418	1.954	2	-0.152	-0.424	-0.463
1	2.903	2.06	2.736	2	-0.33	-0.508	-0.555
1	0.899	2.263	3.764	2	-0.379	-0.39	-0.627
1	-0.389	-0.103	-0.32	2	0.066	-0.12	-0.242
1	-0.057	-0.407	0.062	2	-0.366	0.59	-0.289
1	-0.181	1.841	0.497	2	-0.431	1.824	-0.619
2	-0.344	-0.779	-0.531	2	-0.313	0.049	-0.612
2	-0.292	-1.032	-0.629	2	0.257	0.117	0.264
2	-0.369	-1.336	-0.813	2	-0.412	1.925	1.847
2	-0.317	-0.812	-0.463	2	-0.102	0.134	-0.057
2	-0.345	-0.931	-0.452	2	-0.062	0.083	-0.277
2	-0.247	-0.711	-0.388	2	-0.201	0.218	-0.07
2	-0.331	-0.762	-0.503	2	-0.138	0.252	0.23
2	-0.292	-1.269	-0.473	2	-0.307	0.354	-0.214
2	-0.346	-1.336	-0.334	2	-0.173	1.722	1.178
2	-0.365	-1.032	-0.437	2	-0.199	0.049	-0.159
2	-0.296	-1.607	-0.569	2	-0.356	0.201	-0.291
2	-0.283	-1.573	-0.448	2	-0.376	0.675	0.797
2	-0.079	0.185	0.923	2	-0.406	0.675	1.16
2	-0.231	-0.289	0.038	2	-0.292	-0.018	0.99
2	-0.285	0.235	0.808	2	-0.271	1.3	2.995

Table A2. Data related to regularity from task 2

Indicates long term consequences of trauma	Var.1	Var.2	Indicates long term consequences of trauma	Var.1	Var.2
2	5.1	4.9	2	3.8	4
2	4.9	5.3	2	3.7	3.4
2	2.6	2.4	2	4.5	4.9
2	2.6	2.4	2	3	3.3
2	4.7	4.4	2	3.7	3.7
2	6	6.1	2	4.9	4.9
2	5	5.6	1	2.7	2.7
2	3.9	4.6	1	3.4	3.7
2	5	4.9	1	3.4	2.8
2	4.2	5.1	1	2	1.7
2	3	4.3	1	5.4	5.9
2	3.7	5.1	1	5.6	5.9
2	4.3	4.6	1	2.4	2.5
2	5.4	4.5	1	3	3.4
2	4.5	4.3	1	3.4	3.6
2	4	2.6	1	3.3	3.7
2	3.8	5.4	1	4.3	5
2	4.5	3.3	1	3	3
2	4.1	4	1	5.9	5.9
2	4.9	5.9	1	3	2.6

Table A3. Data related to regularity from Task 3

Indicates If patient dead or alive	Var. 1	Var.2	Indicates If patient dead or alive	Var. 1	Var.2	Indicates If patient dead or alive	Var. 1	Var.2
2	22	1.3	1	26	1.2	1	9.4	3.7
2	18	1	2	16	1.7	2	7.3	1
2	16	1.6	1	8.5	1.5	1	10	2.4
2	11	1.4	1	49	1.3	1	7.2	2.1
2	5.2	1.7	1	9	1.6	1	21.8	1.5
2	3.2	2.2	1	15	2.5	1	12.3	2
2	24.9	1	1	4	1	1	4.2	1
2	19.7	1.2	2	3	1.5	2	6.9	1.5
2	6	1.4	2	18	3	1	12.7	1.9
2	4	1.8	1	20	1.4	2	6.5	1.1
2	11.3	1.5	1	17	2.8	2	11.3	1.3
2	13.3	1.3	2	49	1.5	1	4.3	1
1	12	2.1	1	19	1.4	2	1.8	3.1
1	5	1.8	2	35	1.7	1	21.7	1.3
1	2	1.7	1	-999	2.8	1	10.9	1.4
1	-999	1.9	1	35	2.2	1	25.9	1.2
1	7	2.6	1	23	1.5	1	27	4
2	10	1	1	10	4	1	38.7	1.3
2	6	2.1	1	3.2	1	1	12.4	2.6
1	13	1	1	5.8	1	1	21.6	1.7
1	25	1.6	1	9.5	1.7	1	31.6	1.2
2	11	2.3	1	2.6	1	1	21.3	1.6
1	6.2	1	1	7.5	1.7	1	-999	2.4
1	9.4	1.3	1	4.2	1	1	10.2	1.5
1	23	1.6	1	20	1.7	1	18.9	1.4
1	1	3.1	1	32	2.4			

Table A4. Data related to regularity from task 4

Indicates relapse	Var. 1	Var.2	Indicates relapse	Var. 1	Var.2
2	382	137	1	150	120
2	108	132	1	116	192
1	415	150	1	453	96
1	157	249	2	306	180
1	290	112	1	199	178
1	66	86	2	262	180
1	125	110	2	343	149
1	191	274	2	166	168
1	199	108	1	192	181
1	216	204	1	103	274
1	166	205			

Table A5. Data related to regularity from task 5

Indicates type of wine	Var. 1	Var.2	Indicates type of wine	Var. 1	Var.2	Indicates type of wine	Var. 1	Var.2
1	2.29	1.04	1	2.03	0.88	2	1.34	1.36
1	1.28	1.05	1	1.25	0.87	2	1.35	1
1	2.81	1.03	1	2.19	1.04	2	1.38	1.07
1	2.18	0.86	1	2.14	0.91	2	1.64	1.08
1	1.82	1.04	1	2.38	1.07	2	1.63	1.05
1	1.97	1.05	1	2.08	1.12	2	1.62	0.96
1	1.98	1.02	1	2.91	1.12	2	1.99	1.15
1	1.25	1.06	1	2.29	1.24	2	1.35	1.16
1	1.98	1.08	1	1.87	1.01	2	3.28	1.16
1	1.85	1.01	1	1.68	1.13	2	1.56	0.95
1	2.38	1.25	1	1.62	0.92	2	1.77	1.23
1	1.57	1.17	1	2.45	0.98	2	1.95	1.04
1	1.81	1.15	1	2.03	0.94	2	2.81	1.42
1	2.81	1.25	1	1.66	1.07	2	1.4	1.27
1	2.96	1.2	1	2.04	0.89	2	1.35	1.04
1	1.46	1.28	2	0.42	1.05	2	1.31	0.8
1	1.97	1.07	2	0.41	1.25	2	1.42	0.94
1	1.72	1.13	2	0.62	0.98	2	1.48	1.04
1	1.86	1.23	2	0.73	1.23	2	1.42	0.86
1	1.66	0.96	2	1.87	1.22	2	1.63	1
1	2.1	1.09	2	1.03	1.45	2	1.63	0.88
1	1.98	1.03	2	2.08	1.19	2	2.08	0.86
1	1.69	1.11	2	2.28	1.12	2	2.49	0.96
1	1.46	1.09	2	1.04	1.12	2	3.58	0.75
1	1.66	1.12	2	0.42	1.02	2	1.22	0.9
1	1.92	1.13	2	2.5	1.28	2	1.05	1.23
1	1.45	0.92	2	1.46	0.906	2	1.44	1.1
1	1.35	1.02	2	1.87	1.36	2	1.04	0.93
1	1.76	1.25	2	1.03	0.98	2	2.01	1.71
1	1.98	1.04	2	1.96	1.31	2	1.53	0.95
1	2.38	1.19	2	1.65	0.99	2	1.61	1.06
1	1.95	1.09	2	1.15	1.23	2	0.83	0.7
1	1.97	1.23	2	1.46	1.19	2	1.87	0.93
1	1.35	1.25	2	0.95	0.96	2	1.83	0.8

1	1.54	1.1	2	2.76	1.06	2	1.87	0.93
1	1.86	1.04	2	1.95	1.19	2	1.71	0.92
1	1.36	1.09	2	1.43	1.38	2	2.01	0.73
1	1.44	1.12	2	1.77	1.16	2	2.91	0.75
1	1.37	1.18	2	1.4	1.31	2	1.35	0.86
1	2.08	0.89	2	1.62	0.84	2	1.77	0.69
1	2.34	0.95	2	2.35	0.79	2	1.76	0.97
1	1.48	0.91	2	1.46	1.23	2	1.9	0.89
1	1.7	0.88	2	1.56	1.33	2	1.35	0.79
1	1.66	0.82						