Cluster analysis and development of the intellectual cancer monitoring system based on data of the Federal Penitentiary Service of Russia

S B Ponomarev¹, A A Burt¹, E V Dyuzheva², V E Polishchuk¹ and D S Ponomarev¹

¹ Research Institute of the Federal Penitentiary Service of Russia, 125130, Moscow, Russia
² Research Center No.1 of the Research Institute of the Federal Penitentiary Service, 125130, Moscow, Russia

E-mail: ponomarev.dmitry1990@mail.ru

Abstract. The article considers the possibilities of using cluster analysis and modeling of cancer monitoring in inmates. The article presents the analysis based on the hierarchical clustering, the k-means method along with F-test criterion for the main indicators monitored in the penitentiary system. The application of modeling was analyzed using the example of regression analysis of the data. The development of the model was relative to the indicator of the proportion of mortality rate to the total number of malignant neoplasms. According to the statistic data of 2018, the main statistical indicators on oncopathology in inmates reflect a similar situation in the whole country. The researchers consider the possibility to apply the results obtained not only within the penitentiary system, but also to monitor the situation with cancer in the country in general.

Malignant neoplasms (MN) are the second leading cause of death in the world. According to WHO data in 2015 the number of deaths caused by MN was 8.8 million, which puts them in second place in terms of number after deaths caused by cardiovascular diseases [1, 2]. There is an opinion that in many countries with high-income neoplasms have already took the first leading place in depopulation, in the coming decades it will become not only the main cause of death but also of disability in the world [3].

According to the results of 2018, the main statistical indicators on oncopathology in inmates reflect a similar situation in the whole country.

To assess the situation in the regions of the country the researchers have used a number of the most important indicators monitored in institutions of the penitentiary system and have selected 8 parameters with assigned X values, where: 

- \( X_1 \) – relative numbers of MN incidence in territorial bodies of the Federal Penitentiary Service of Russia (FPSR) (per 100 thousand people);
- \( X_2 \) – relative numbers of MN primary incidence (per 100 thousand people);
- \( X_3 \) – relative numbers of primary disability (per 10 thousand people);
- \( X_4 \) – relative numbers of early release from prisons (%);
- \( X_5 \) – early release rate from the number of inmates with MN (per 100 thousand people);
- \( X_6 \) – mortality rate in inmates with MN (%);
- \( X_7 \) – mortality rate (per 100 thousand people);
- \( X_8 \) – relative numbers of MN incidence in women (per 100 thousand people).
The study of the data used in this research is of primary interest. K-means can be used to divide regions into subgroups of the MN spread risk. As for the indicators themselves, here we can apply hierarchical clustering to reveal their division into subgroups and groups (figure 1).

The standardization of values was carried out according to the formula (1):

\[
x_{\text{std}}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}
\]

where: \(\mu_x\) – the empirical average of each individual feature column, \(\sigma_x\) – the corresponding standard deviation. In the study we have applied Squared Euclidian distance to determine the distance between two points \(x\) and \(y\) (2):

\[
d(x, y) = \sum_{i=1}^{m} (x_i - y_i)^2
\]

In the work, the Ward method was applied. This method differs from the others because of the fact that it uses analysis of variance methods to estimate the distances between clusters.

As the distance \(\text{dis} (x, y)\) between clusters \(x\) and \(y\), we take the increment of the sum of the squares of the objects distances to the centers of the clusters, which is obtained as a result of their combination (3):

\[
\text{dis}(x, y) = \frac{n_x n_y}{n_x + n_y} (\bar{x} + \bar{y})^T (\bar{x} + \bar{y})
\]

where: \(x, y\) – radii-defined centers of clusters, \(n_x, n_y\) – the number of elements in them, \(T\) – transposition.

Ward’s method [4] minimizes the sum of the squares for any two (hypothetical) clusters those can be formed. Each step combines two clusters thus it leads to a minimum increase in the objective function or in other words intragroup the sum of the squares. This method combines closely spaced clusters and tends to find (or create) clusters of approximately equal sizes with a hyperspherical shape. In general, the method appears to be very effective but it tends to create clusters of small size.

Let us consider hierarchical clustering for the indicators themselves [5]. The sample consisted of 81 points (the territorial bodies of the FPSR). Figure 1 shows the hierarchical structure for the studied indicators obtained as a result of cluster analysis.

![Figure 1. Hierarchical clustering of data for the main statistical indicators on oncopathology in inmates.](image-url)
As can be seen from figure 2, the combination (the union) of the points occurred as follows: relative numbers of MN incidence in territorial bodies of the Federal Penitentiary Service of Russia (FPSR) (per 100 thousand people) \( (X_1) \) and relative numbers of MN primary incidence (per 100 thousand people) \( (X_2) \); relative numbers of early release from prisons (%) \( (X_3) \) and early release rate from the number of inmates with MN (per 100 thousand people) \( (X_4) \); mortality rate in inmates with MN (%) \( (X_5) \) and mortality rate (per 100 thousand people) \( (X_6) \); relative numbers of MN incidence in women (per 100 thousand people) \( (X_7) \) and relative numbers of primary disability (per 10 thousand people) \( (X_8) \) were separated from the rest into distinct groups, where \( X_3 \) includes the remaining subgroups.

In addition to the hierarchical clustering method, in the research we have used the k-means method, which makes it possible to divide territories into clusters according to indicators \( X_1 \sim X_8 \) [6]. The following parameters were defined: number of variables: 8; number of cases: 81; number of clusters: 3. Figure 2 presents the results of cluster analysis for all sample values.

![Plot of Means for Each Cluster](image)

**Figure 2.** Territory segmentation with K-means clustering.

It is determined that the division was as follows: the significant parameters for the first cluster have become \( X_6 \), \( X_7 \), \( X_8 \) (includes 19 regions), for the second cluster \( X_4 \), \( X_5 \) (includes 23 regions), the third cluster includes all the parameters non considered in the first two clusters (includes 39 regions) (table 1).

**Table 1.** The distribution of regions into three clusters according to the main statistical indicators of cancer incidence in inmates.

| №   | Region                                                   |
|-----|----------------------------------------------------------|
| 1 cluster | Altai Region; Arkhangelsk Region; Volograd Region; Jewish Autonomous Region; Irkutsk Region; Kurgan Region; Nizhny Novgorod Region; Orenburg Region; Penza Region; Primorsky Krai; Komi Republic; Republic of North Ossetia-Alania; The Republic of Khakassia; Ryazan Oblast; Samara Region; Saratov Region; Stavropol Region; Tambov Region; Tula Region. |
| 2 cluster | Amur Region; Belgorod Region; Bryansk Region; Vladimir Region; Vologodskaya Region; Ivanovo Region; Kaliningrad Region; Kemerovo Region; Kirov Region; Kostroma Region; Krasnoyarsk Region; Lipetsk Region; Magadan Region; Moscow; Oryol Region; Pern Region; The Republic of Adygea (Adygea); The Republic of Buryatia; Sverdlovsk Region; Smolensk Region; The Chechen Republic; The Chuvash Republic - Chuvashia; Yaroslavskaya Region. |
To study the results of clustering, we have used the Analysis of Variance. The results of this analysis are presented in table 2, which shows the intergroup and intragroup variance. Table rows present variables (observations), table columns show indicators for each variable: dispersion between clusters, the number of degrees of freedom in interclass dispersion, the dispersion within clusters, the number of degrees of freedom in intraclass dispersion, \( F \) – the criterion to test the hypothesis of inequality of variances.

| Parameter | Between | df | Within | \( F \) | \( p \) |
|-----------|---------|----|--------|--------|--------|
| \( X_1 \) | 25,081  | 1  | 54,918 | 36,080 | 0,000  |
| \( X_2 \) | 11,802  | 1  | 68,198 | 13,671 | 0,0004 |
| \( X_3 \) | 28,959  | 1  | 51,041 | 44,822 | 0,000  |
| \( X_4 \) | 4,971   | 1  | 75,029 | 5,234  | 0,025  |
| \( X_5 \) | 0,096   | 1  | 79,904 | 0,095  | 0,758  |
| \( X_6 \) | 3,066   | 1  | 76,934 | 3,148  | 0,079  |
| \( X_7 \) | 16,997  | 1  | 63,003 | 21,313 | 0,00001|
| \( X_8 \) | 12,552  | 1  | 67,447 | 14,702 | 0,0002 |

Values of \( p \) less than 0.05 indicate that factors contribute to the clustering of data. We can say that the parameters \( X_5, X_6 \) least affected the clustering of data. This fact also confirms the value of the F-criterion (column of values of \( F \)).

The study of the prerequisites for model development. In modern world systems analysis and machine learning methods are of great interest. Such methods as artificial neural networks and regression analysis can be applied in almost any field [7, 8, 9]. Let us consider the development of a model based on regression analysis of data on the example of one indicator - mortality rate in inmates with MN (%) (\( X_6 \)). We take this parameter as the outgoing signal.

The performed cluster analysis for this indicator have identified three main clusters. The first cluster shows a high risk of increasing of mortality rate in inmates with MN from the total number of inmates with MN (4 regions included). The second cluster is medium risk (20 regions included). The third cluster is a low level of risk (57 regions included). Further, the researchers have divided the data into test and training samples. The size of the test sample was 16 points and did not participate in the training. The obtained results (4) of the research base on the regression equation that we have developed.

\[
Y = 2,9963 + 0,0003X_1 + 0,0002X_2 - 0,0014X_3 + 0,0015X_4 - 0,0062X_5 - 0,0102X_7 + 9,5 \cdot 10^{-5} X_8 \tag{4}
\]

where \( Y \) – cluster number (integer equal to 1, 2 or 3), which reflects the position of the region in a certain group of risk of mortality rate increase in inmates. When \( Y \) has negative values, it is equal to 1 cluster. For values above 3, \( Y \) is equal to cluster 3. It was made a comparison of the results obtained.
with the use of the model and actual values of the test sample. The accuracy was 93.75%. It is worth noting that this error is due to the small number of points that entered the 1 cluster and too much data that entered the 3 cluster. In other words, the accuracy of the model can be improved by increasing the sample size (especially if this data is included in clusters 1 and 2).

With the help of hierarchical clustering in determining the significance of the main statistical indicators of MN incidence in inmates there were combined individual parameters in subgroups and groups. Also such indicators as relative numbers of MN incidence in women (per 100 thousand people) ($X_3$) and relative numbers of primary disability (per 10 thousand people) ($X_1$) were separated from the rest into distinct groups. It should be said that relative numbers of primary disability (per 10 thousand people) $X_1$ includes the remaining subgroups.

The results became more interesting when K-means method was applied to segment territories. These parameters were less significant: relative numbers of early release from prisons (%) ($X_3$) ($F=5,234, p=0,025$); early release rate from the number of inmates with MN (per 100 thousand people) ($X_5$) ($F=0,995, p=0,758$); mortality rate in inmates with MN (%) ($X_6$) ($F=3,148, p=0,079$). Following parameters have obtained the greatest impact on data clustering (when dividing into three clusters by levels): relative numbers of MN incidence in territorial bodies of the FPSR (per 100 thousand people) ($X_1$) ($F=36,080$); relative numbers of primary disability (per 10 thousand people) ($X_3$) ($F=44,822$); mortality rate (per 100 thousand people) ($X_5$) ($F=21,313$). Moreover, for all three parameters $p < 0,05$.

A regression model was developed on the base of the parameter of mortality rate in inmates with MN (%) ($X_6$). The accuracy of the results is 93.75%. The accuracy was checked on an independent sample, which was not taken into account in the development of the model.

It should be noted that in both cases of clustering (hierarchical clustering of parameters and regional segmentation using k-means clustering), the indicator of relative numbers of primary disability (per 10 thousand people) ($X_1$) have become one of the determining and most important parameters.

References

[1] Medical and demographic indicators of the Russian Federation in 2017 Stat. reference book (Ministry of Health of Russia) p 264
[2] Kaprina V, Starinsky G and Petrova M 2019 The status of cancer care for the population of Russia in 2018 The branch of the Federal State Budgetary Institution Scientific Research Center for Radiology of the Ministry of Health of Russia 236
[3] Samorodskaya I and Larina S 2016 Contribution of four groups of noncommunicable diseases to the mortality rate in the regions of the Russian Federation in 2015 Preventive medicine 21 236–44
[4] Ward J 1963 Hierarchical Grouping to Optimize an Objective Function Journal of the American Statistical Association 58 236–44
[5] Zhambyu M 1988 Hierarchical cluster analysis and correspondence Finance and Statistics 345
[6] Steinhaus H 1956 Sur la division des corps materiels en parties Bull. Acad. Polon. Sci. 3 801-4
[7] Ponomarev D, Sterlikov S, Ponomarev S and Averyanova E 2019 The use of an artificial neural network to model prognosis in patients with a combination of HIV infection and tuberculosis in prisons Tuberculosis and lung diseases 97 78-9
[8] Ponomarev D, Gorokhov M and Averyanova E 2019 Development of a model based on an artificial neural network to predict the development of AIDS in prisoners Penitentiary medicine in Russia and abroad 98-9
[9] Ponomarev D 2015 Neuregression model of water deodorization based on a multilayer perceptron Bulletin of Kazan National Research Technical University named after A.N.Tupolev 71 16-9