







On the Optimization Models for Automatic Grouping of Industrial Products by Homogeneous Production Batches

Guzel Sh. Shkaberina¹ , Viktor I. Orlov^{1,2}, Elena M. Tovbis¹ ,
and Lev A. Kazakovtsev^{1,3}  

¹ Reshetnev Siberian State University of Science and Technology,
prosp. Krasnoyarskiy Rabochiy 31, Krasnoyarsk 660031, Russia
levk@bk.ru

² Testing and Technical Center – NPO PM,
20, Molodezhnaya Street, Zheleznogorsk 662970, Russia

³ Siberian Federal University,
prosp. Svobodny 79, Krasnoyarsk 660041, Russia

Abstract. We propose an optimization model of automatic grouping (clustering) based on the k-means model with the Mahalanobis distance measure. This model uses training (parameterization) procedure for the Mahalanobis distance measure by calculating the averaged estimation of the covariance matrix for a training sample. In this work, we investigate the application of the k-means algorithm for the problem of automatic grouping of devices, each of which is described by a large number of measured parameters, with various distance measures: Euclidean, Manhattan, Mahalanobis. If we have a sample with the composition known in advance, we use it as a training (parameterizing) sample from which we can calculate the averaged estimation of the covariance matrix of homogeneous production batches using the Mahalanobis distance. We propose a new clustering model based on the k-means algorithm with the Mahalanobis distance with the averaged (weighted average) estimation of the covariance matrix. We used various optimization models based on the k-means model in our computational experiments for the automatic grouping (clustering) of electronic radio components based on data from their non-destructive testing results. As a result, our new model of automatic grouping allows us to reach the highest accuracy by the Rand index.

Keywords: K-means · Electronic radio components · Clustering

1 Introduction

The increasing complexity of modern technology leads to an increase in the requirements for the quality, of industrial products reliability and durability.

© Springer Nature Switzerland AG 2020

Y. Kochetov et al. (Eds.): MOTOR 2020, CCIS 1275, pp. 421–436, 2020.

https://doi.org/10.1007/978-3-030-58657-7_33

Determination of product quality is carried out by production tests. The quality of products within a single production batch is determined by the stability of the product parameters. Moreover, an increase in the stability of product parameters in manufactured batches can be achieved by increasing the stability of the technological process.

In order to exclude the possibility of potentially unreliable electronic and radio components (ERC) intended to be installed in the onboard equipment of a spacecraft with a long period of active existence, the entire electronic component base passes through specialized technical test centers [1,2]. These centers carry out operations of the total input control of the ERC, total additional screening tests, total diagnostic non-destructive testing and the selective destructive physical analysis (DPA). To expand the results of the DPA to the entire batch of products obtained, we must be sure that the products are manufactured from a single batch of raw materials. Therefore, the identification of the original homogeneous ERC production batches from the shipped lots of the ERC is one of the most important steps during testing [1].

The k-means model in this problem is well established [1,3–10]. Its application allows us to achieve a sufficiently high accuracy of splitting the shipped lots into homogeneous production batches. The problem is solved as a k-means problem [11]. The aim is to find k points (centers or centroids) X_1, \dots, X_k in a d -dimensional space, such that the sum of the squared distances from known points (data vectors) A_1, \dots, A_N to the nearest of the required points reaches its minimum (1):

$$\arg \min F(X_1, \dots, X_k) = \sum_{i=1}^N \min_{j \in \{1, k\}} \|X_j - A_j\|^2. \quad (1)$$

Factor analysis methods do not significantly reduce the dimension of the space without loss of accuracy in solving problems [12]. However, in some cases, the accuracy of partitioning into homogeneous batches (the proportion of objects correctly assigned to “their” cluster representing a homogeneous batch of products) can be significantly improved, especially for samples containing more than 2 or 3 homogeneous batches. In addition, the methods of factor analysis, although they do not significantly reduce the dimension of the search space, show the presence of linear statistical dependencies (correlations) between the parameters of the ERC in a homogeneous batch.

A slight increase in accuracy is achieved by using an ensemble of models [3]. We also applied some other clustering models, such as the Expectation-Maximization (EM) model and Self-organized Kohonen Maps (COM) [12].

Distance measure used in practical tasks of automatic objects grouping in real space depends on the features of space. Changing distance measures can improve the accuracy of automatic ERC grouping.

The idea of this work is to use the Mahalanobis distance measure in the k-means problem and study the accuracy of clustering results. We proposed a new algorithm, based on k-means model using the Mahalanobis distance measure with an averaged estimation of the covariance matrix.

2 Mahalanobis Distance

In k-means, k-median [13–15] and k-medoid [16–18] models, various distance measures may be applied [19,20]. The use of correlation dependencies can be involved by moving from a search in space with a Euclidean or rectangular distance to a search in space with a Mahalanobis distance [21–24]. The square of the Mahalanobis distance D_M defined as follows (2):

$$D_M(X) = \sum_{i=1}^n (X - \mu)^T C^{-1} (X - \mu), \quad (2)$$

where X is vector of values of measured parameters, μ is vector of coordinate values of the cluster center point (or cluster center), C is the covariance matrix.

Experiments on automatic ERC grouping with the k-medoid and k-median models using the Mahalanobis distance show a slight increase in the clustering accuracy in simple cases (with 2–4 clusters) [25].

3 Data and Preprocessing

In this study, we used data of test results performed in the testing center for the batches of integrated circuits (microchips) [26]. The source data is a set of some ERC parameters measured during the mandatory tests. The sample (mixed lot) was originally composed of data on products belonging to different homogeneous batches (in accordance with the manufacturer’s markup). The total amount of ERC is 3987 devices. Batch 1 contains 71 device, 116 devices for Batch 2, 1867 for Batch 3, 1250 for Batch 4, 146 for batch 5, 113 for Batch 6, 424 for Batch 7. The items (devices) in each batch are described by 205 input measured parameters.

Computationally, the k-means problem, in which the sum of squared distances acts as the minimized objective function, is more convenient than the k-median model using the sum of distances, because when using the sum of the squared distances, the center point of the cluster (the centroid) coincides with the average coordinate value of all objects in the cluster. When passing to the sum of squared Mahalanobis distances, this property is preserved.

Nevertheless, the use of the Mahalanobis distance in the problem of automatic ERC grouping in many cases leads to accuracy decrease in comparison with the results achieved with the Euclidean distance due to the loss of the advantage of the special data normalization approach (Table 1, hit percentage computed as the sum of hits of algorithm (True Positives) in every batch divided by number of products in the mixed lot).

The assumption that the statistical dependences of the parameter values appear in different batches of ERC in a similar way has experimental grounds. As can be seen from Fig. 1, the span and variance of the parameters of different batches vary significantly. Even if the difference in the magnitude of the span and variance of any parameters is insignificant among separate batches, they differ significantly from the span and variance of the entire mixed lot (Fig. 2).

Table 1. Comparison of the clustering results with different measures of distance, number of exact hits (proportion of hits)

Batches	Squared Euclidean distance	Squared Mahalanobis distance	Rectangular (Manhattan) distance	Cosine distance	Correlation distance
Four-batch mixed lot (n = 446)					
Batch 1 (n = 71)	70 (0.99)	47 (0.66)	71 (1.00)	70 (0.99)	70 (0.99)
Batch 2 (n = 116)	78 (0.67)	83 (0.72)	64 (0.55)	78 (0.67)	84 (0.72)
Batch 5 (n = 146)	96 (0.66)	88 (0.60)	105 (0.72)	96 (0.66)	104 (0.71)
Batch 6 (n = 113)	44 (0.39)	91 (0.81)	50 (0.44)	44 (0.39)	38 (0.37)
Average	0.65	0.69	0.65	0.65	0.66
Sum of distances	473.174	26146.350	401.4	0.0012	0.0011
Full mixed lot (n = 3987)					
Batch 1 (n = 71)	67 (0.94)	70 (0.99)	68 (0.96)	67 (0.94)	71 (1.00)
Batch 2 (n = 116)	4 (0.03)	4 (0.03)	4 (0.03)	4 (0.03)	78 (0.67)
Batch 3 (n = 1867)	578 (0.31)	223 (0.12)	558 (0.30)	578 (0.31)	0 (0.00)
Batch 4 (n = 1250)	403 (0.32)	127 (0.11)	446 (0.36)	406 (0.33)	227 (0.18)
Batch 5 (n = 146)	66 (0.45)	81(0.55)	63 (0.43)	64 (0.44)	78 (0.53)
Batch 6 (n = 113)	88 (0.78)	113 (1.00)	82 (0.73)	88 (0.78)	32 (0.28)
Batch 7 (n = 424)	311 (0.73)	404 (0.95)	303 (0.72)	311 (0.73)	314 (0.74)
Average	0.38	0.26	0.38	0.38	0.20
Sum of distances	5008.127	248808.6	1755.8	0.007	0.004

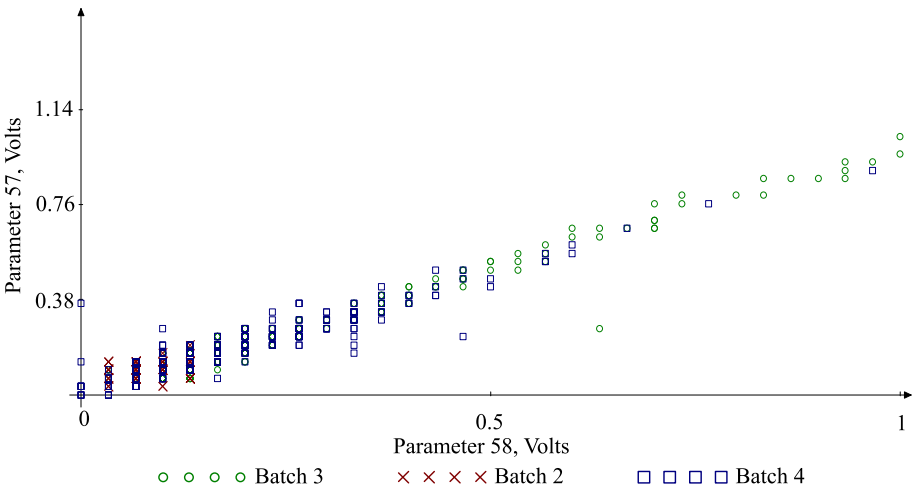


Fig. 1. Statistical dependence of the ERC parameters 57, 58

Thus, it is erroneous to take the variance and covariance coefficients in each of the homogeneous batches equal to the variance and covariance coefficients for the whole sample. Experiments with the automatic grouping model based

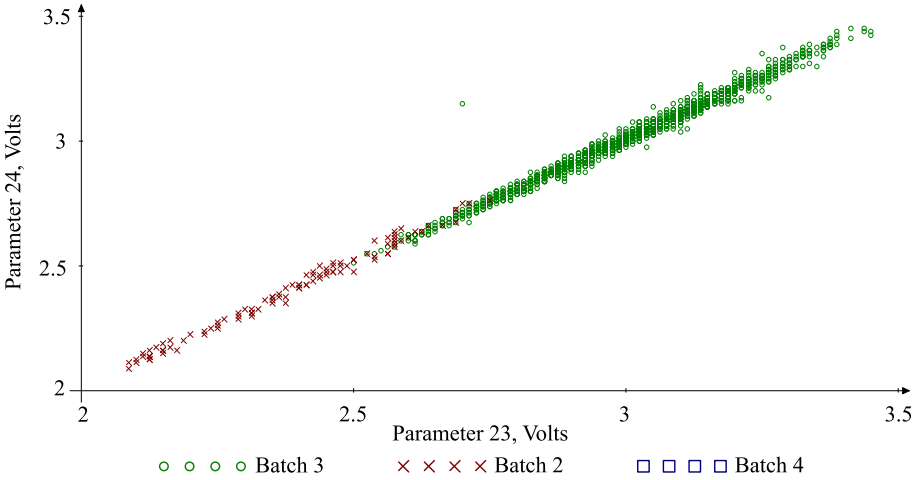


Fig. 2. Statistical dependence of the ERC parameters 23, 24

on a mixture of Gaussian distributions by maximizing the likelihood function by the EM algorithm [27] show a relatively high model adequacy only when using diagonal covariance matrices (i.e. uncorrelated distributions), moreover, equal for all distributions. Apparent correlations between the parameters are not taken into account.

Mahalanobis distance is scale invariant [28]. Due to this property, data normalization does not matter if this distance is applied. At the same time, binding of the boundaries of the parameters to the boundaries, determined by their physical nature, sets a scale proportional to the permissible fluctuations of these parameters under operating conditions, without reference to the span and variance of these values in a particular production batch. The solution to the problem of preserving the scale could be to use the Mahalanobis distance with the correlation matrix R instead of the covariance matrix C (3):

$$D_M(X) = \sum_{i=1}^n (X_i - \mu)^T R^{-1} (X_i - \mu). \tag{3}$$

Each element of the matrix R is calculated as follows (4):

$$r_{XY} = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{(N - 1)S_X S_Y}, \tag{4}$$

where S_X and S_Y are standard deviations of parameters X and Y , \bar{X} and \bar{Y} are their average values.

As shown by experiments, the results of which are given below, this approach does not show advantages compared to other methods.

4 The K-Means Model with Supervised Mahalanobis Distance Measure

The clustering problem is a classical example of the unsupervised learning approach. However, in some cases, when solving the problem of automatic grouping, we have a sample of a known composition. This sample can serve as a training (parameterizing) sample. In this case, a unique covariance matrix C (see (2)) is calculated on this training sample and then used on other data. We call the Mahalanobis distance (2) with the covariance matrix C pre-calculated on a training sample the supervised (or parameterized) Mahalanobis distance.

If there is no training sample, well-known cluster analysis models can be used to isolate presumably homogeneous batches with some accuracy. With this approach, a presumably heterogeneous batch can be divided into the number of presumably homogeneous batches, determined by the silhouette criterion [29–31]. At the same time, a mixed lot can be divided into a larger number of homogeneous batches than it actually is: smaller clusters are more likely to contain data of the same class, i.e. the probability of false assignment of objects of different classes to one cluster reduces. The proportion of objects of the same class, falsely assigned to different classes, is not so important for assessing the statistical characteristics of homogeneous groups of objects.

In the next experiment, there were training sample contains 6 batches: Batch 1 (71 device), Batch 2 (116 devices), Batch 4 (1250 devices), Batch 5 (146 devices), Batch 6 (113 devices), Batch 7 (424 devices). Using covariance matrix C , datasets contain 2 batches in all combinations were clustered with the use of various distance measure. The result was compared with the traditional k-means clustering method with the squared Mahalanobis distance (unsupervised squared Mahalanobis distance, Tables 2, 3, 4 proportion of hits computed as the sum of hits of algorithm in every batch divided by number of products in the batch), and with Euclidean and rectangular distances. For each model, we performed 5 experiments. Average clustering results are shown in Tables 2, 3, 4.

Table 2. Comparison of the clustering results with different measures of distance, number of exact hits (proportion of hits) (Part 1)

Batches	Supervised squared Mahalanobis distance	Unsupervised squared Mahalanobis distance	Squared Euclidean distance	Rectangular (Manhattan) distance
Batch 4 (n = 1250)	850 (0.68)	685 (0.55)	741 (0.59)	895 (0.72)
Batch 7 (n = 424)	390 (0.92)	256 (0.60)	228 (0.54)	423 (1.00)
Average	0.74	0.56	0.58	0.79
Avg. total squared distance	94467	100898	7119	12272
Batch 7 (n = 424)	253 (0.60)	Singular	416 (0.98)	415 (0.98)
Batch 1 (n = 71)	71 (1.00)	Matrix	71 (1.00)	71 (1.00)
Average	0.65	-	0.98	0.98
Avg. total squared distance	17551	-	1233	2795

The experiment showed that the results of solving the k-means problem with a supervised Mahalanobis distance measure are higher in comparison with the results of a model with unsupervised Mahalanobis distance, however, it is still lower than in case of Euclidean and rectangular distances.

5 The K-Means Model with Supervised Mahalanobis Distance Measure Based on Averaged Estimation of the Covariance Matrix

Since the original covariance matrices are of the same dimension, we are able to calculate the average estimation of the covariance matrix among all homogeneous batches of products in the training (parameterizing) sample:

$$C = \frac{1}{n} \sum_{j=1}^k C_j n_j, \tag{5}$$

where n_j is number of objects (components) in j th production batch, n is total sample size, C_j are covariance matrices calculated on separate production batches, each of which can be calculated by (6):

$$C_j = E[(X - EX)(E - EY)^T]. \tag{6}$$

We propose the k-means algorithm using the Mahalanobis distance measure with averaged estimation of the covariance matrix. Convergence of the k-means algorithm using a Mahalanobis distance reviewed in [32]. Optimal k value was found by silhouette criterion [30]:

Algorithm 1

Step 1. Divide randomly initial sample into k clusters.

Step 2. Calculate for each cluster a centroid μ_i . A centroid is defined as the arithmetic mean of all points in a cluster (7):

$$\mu_i = \frac{1}{m} \sum_{j=1}^m X_{ji} \tag{7}$$

where m is number of points, X_j is vector of measured parameter values ($j = 1..m$), $i = 1..n$ (n is a number of parameters).

Step 3. Calculate the averaged estimation of the covariance matrix (5). If the averaged estimation of the covariance matrix is singular, then proceed to Step 4, else proceed to step 5.

Step 4. Increase the number of clusters by $(k + 1)$ and repeat steps 1 and 2. Form new clusters with squared Euclidean distance measure (8):

$$D(X_j, \mu_i) = \sum_{i=1}^n (X_{ji} - \mu_i)^2 \tag{8}$$

where n is a number of parameters.

Return to step 3 with new training sample.

Step 5. Assign each point to the nearest centroid using the squared Mahalanobis distance with averaged estimation of the covariance matrix to form new clusters.

Step 6. Repeat algorithm from step 2 until clusters do not change.

Table 3. Comparison of the clustering results with different measures of distance, number of exact hits (proportion of hits) (Part 2)

Batches	Supervised squared Mahalanobis distance	Unsupervised squared Mahalanobis distance	Squared Euclidean distance	Rectangular (Manhattan) distance
Batch 7 (n = 424)	223 (0.53)	Singular	244 (0.58)	282 (0.67)
Batch 6 (n = 113)	113 (1.00)	Matrix	92 (0.81)	84 (0.75)
Average	0.63	-	0.63	0.68
Avg. total squared distance	18190	-	1396	3300
Batch 7 (n = 424)	216 (0.51)	Singular	217 (0.51)	274 (0.65)
Batch 2 (n = 116)	116 (1.00)	Matrix	95 (0.82)	97 (0.84)
Average	0.62	-	0.58	0.69
Avg. total squared distance	18190	-	1123	3090
Batch 7 (n = 424)	424 (1.00)	218 (0.51)	380 (0.90)	385 (0.91)
Batch 5 (n = 146)	136 (0.93)	85 (0.58)	146 (1.00)	146 (1.00)
Average	0.98	0.53	0.92	0.93
Avg. total squared distance	34385	34282	1250	3202
Batch 1 (n = 71)	71 (1.00)	47 (0.66)	71 (1.00)	71 (1.00)
Batch 4 (n = 1250)	471 (0.38)	653 (0.52)	772 (0.62)	642 (0.51)
Average	0.41	0.53	0.64	0.54
Avg. total squared distance	82458	79599	7237	11120
Batch 4 (n = 1250)	410 (0.33)	648 (0.52)	735 (0.59)	570 (0.46)
Batch 6 (n = 113)	102 (0.90)	59 (0.52)	67 (0.59)	85 (0.75)
Average	0.38	0.52	0.59	0.48
Avg. total squared distance	82649	82054	5452	10014
Batch 4 (n = 1250)	412 (0.33)	622 (0.50)	769 (0.62)	485 (0.39)
Batch 2 (n = 116)	98 (0.85)	69 (0.59)	76 (0.66)	96 (0.82)
Average	0.37	0.51	0.62	0.43
Avg. total squared distance	82693	82318	5410	9996
Batch 4 (n = 1250)	953 (0.76)	772 (0.62)	772 (0.62)	873 (0.70)
Batch 5 (n = 146)	91 (0.62)	91 (0.62)	146 (1.00)	146 (1.00)
Average	0.75	0.62	0.66	0.73
Avg. total squared distance	99605	83963	6689	11619
Batch 1 (n = 71)	71 (1.00)	Singular	71 (1.00)	71 (1.00)
Batch 6 (n = 113)	111 (0.98)	Matrix	113 (1.00)	113 (1.00)
Average	0.99	-	1.00	1.00
Avg. total squared distance	6500	-	354	797

(continued)

Table 3. (continued)

Batches	Supervised squared Mahalanobis distance	Unsupervised squared Mahalanobis distance	Squared Euclidean distance	Rectangular (Manhattan) distance
Batch 1 (n = 71)	71 (1.00)	Singular	71 (1.00)	71 (1.00)
Batch 2 (n = 116)	116 (1.00)	Matrix	112 (0.97)	114 (0.98)
Average	1.00	-	0.98	0.99
Avg. total squared distance	6481	-	325	747
Batch 1 (n = 71)	71 (1.00)	39 (0.56)	70 (0.99)	71 (1.00)
Batch 5 (n = 146)	84 (0.58)	80 (0.55)	99 (0.68)	108 (0.74)
Average	0.71	0.55	0.78	0.83
Avg. total squared distance	22199	13004	223	841
Batch 2 (n = 116)	91 (0.78)	Singular	89 (0.77)	70 (0.60)
Batch 6 (n = 113)	87 (0.77)	Matrix	37 (0.33)	48 (0.42)
Average	0.78	-	0.55	0.52
Avg. total squared distance	7319	-	282	903

Table 4. Comparison of the clustering results with different measures of distance, number of exact hits (proportion of hits) (Part 3)

Batches	Supervised squared Mahalanobis distance	Unsupervised squared Mahalanobis distance	Squared Euclidean distance	Rectangular (Manhattan) distance
Batch 5 (n = 146)	96 (0.66)	81 (0.55)	146 (1.00)	146 (1.00)
Batch 6 (n = 113)	113 (1.00)	66 (0.59)	105 (0.93)	109 (0.75)
Average	0.81	0.57	0.97	0.99
Avg. total squared distance	23172	6564	512	1246
Batch 2 (n = 116)	116 (1.00)	67 (0.57)	108 (0.93)	109 (0.94)
Batch 5 (n = 146)	78 (0.54)	80 (0.55)	146 (1.00)	146 (1.00)
Average	0.74	0.56	0.97	0.97
Avg. total squared distance	23070	15710	458	1175

6 Computational Experiments

A series of experiments was carried out on the data set described above. This mixed lot is convenient due to its composition is known in advance, which allows us to evaluate the accuracy of the applied clustering models. Moreover, this data set is difficult for grouping by well-known models: some homogeneous batches in its composition are practically indistinguishable from each other, and the accuracy of known clustering models on this sample is low [12,33].

As a measure of the clustering accuracy, we use the Rand Index (RI) [34], which determines the proportion of objects for which the reference and resulting cluster splitting are similar.

To train the model with the averaged Mahalanobis distance measure from the components of the mixed lot, new combinations of batches were compiled containing devices belonging to different homogeneous batches. New combinations consists of 2–7 homogeneous batches. Training sample include the entire data from each batch.

Experiments conducted with 5 different clustering models:

Model DM1: K-means with the Mahalanobis distance measure, the estimation of the covariance matrix calculates for the entire training sample. The objective function defines as the sum of the squared distances.

Model DC: K-means with a distance measure similar to the Mahalanobis distance, but using a correlation matrix instead of a covariance matrix (3). The objective function defines as the sum of the squared distances.

Model DM2: K-means algorithm with Mahalanobis distance measure based on averaged estimation of the covariance matrix (4). The objective function defines as the sum of the squared distances.

Model DR: K-means with Manhattan distance measure. The objective function defines as the sum of the distances.

Model DE: K-means with Euclidean distance measure. The objective function defines as the sum of the squared distances.

This paper presents the results of three groups of experiments. In each of the groups of experiments, for each working sample, the k-means algorithm was run 30 times with each of the five studied clustering models. In these groups of experiments the highest RI value was shown by K-means algorithm with Mahalanobis distance measure based on averaged estimation of the covariance matrix.

First Group. The training set corresponds to the working sample for which clustering was carried out. Five series of experiments were carried out. In each series of experiments, the sample is composed of a combination of products belonging to 2–7 homogeneous batches. Table 5 presents the maximum, minimum, mean

Table 5. An experiment of the 1st group

	Rand index					Objective function				
	DM1	DC	DM2	DR	DE	DM1	DC	DM2	DR	DE
Max	0.755	0.66	0.822	0.739	0.745	255921	3843	2645	18902	6008
Min	0.560	0.64	0.732	0.702	0.704	250558	3706	2600	17785	5010
Mean	0.627	0.65	0.771	0.716	0.721	253041	372289	261582	18225	5298
σ	0.051	0.00	0.024	0.010	0.009	1178	261.01	989.3	433.12	290.276
V						0.466	0.701	0.378	2.377	5.479
R						5363	1369	4517	1117	998

value and standard deviation for the Rand index and objective function for the 7-batches sample. For objective function also calculated the coefficient of variation (V) and span factor (R, where $R = Max - Min$).

Second Group. Training and work samples do not match. In practice, the test center can use retrospective data from the supply and testing of products of the same type as a training sample. In this series of experiments, no more than seven homogeneous batches are presented in the training set. The working sample is represented by a new combination of products belonging to different homogeneous batches. In Table 6 represented results for 5-batches working set and 7-batches training set.

Table 6. An experiment of the 2nd group

	Rand index					Objective function				
	DM1	DC	DM2	DR	DE	DM1	DC	DM2	DR	DE
Max	0.7490	0.645	0.8524	0.7337	0.73567	254822	38704	263405	20509	9194.61
Min	0.4312	0.631	0.7470	0.6955	0.68932	249355	37856	257534	19408	6554.1
Mean	0.5660	0.636	0.8117	0.7079	0.71919	251694	37982	259689	19674	7119.85
σ	0.0519	0.003	0.0324	0.0153	0.01002	1462.8	203.55	1502.09	289.63	571.119
V						0.581	0.536	0.578	1.472	8.022
R						5467	848	5871	1102	2641

Third Group. The training and working samples also do not match, but the results of the automatic product grouping were used as the training sample (k-means in multistart mode with Euclidean distance measure). In each series of experiments, the training set consists of 10 batches, which in turn are the result of applying the k-means algorithm to the training set containing the entire sample. The working sample is represented by a new combination of products belonging to different homogeneous batches. In Table 7 showed results for 7-batches working set.

Table 7. An experiment of the 3rd group

	Rand index					Objective function				
	DM1	DC	DM2	DR	DE	DM1	DC	DM2	DR	DE
Max	0.7672	0.6579	0.7489	0.73969	0.73456	255886	379167	281265	18897	6495
Min	0.5618	0.6453	0.6958	0.70286	0.70466	250839	36997	274506	17785	5009
Mean	0.6317	0.6499	0.7246	0.71359	0.71935	252877	37178	277892	18240	5250
σ	0.0468	0.0032	0.0160	0.0081	0.0063	1164.5	152.84	2358.92	452.73	367.5
V						0.461	0.411	0.849	2.482	6.981
R						5047	920	6759	1112	1485

In most cases, the coefficient of variation of the objective function values is highest for the DE model, where the Euclidean distance measure used. The span factor of the objective function, in the opposite, has most high values for the DM2 model, where the Mahalanobis distance measure with the average estimation of the covariance matrix used. Therefore, obtaining consistently good values of the objective function requires multiple attempts to run the k-means algorithm, or using other algorithms based on the k-means model, such as j-means [35] or greedy heuristic algorithms [36] or others.

According to Rand index, DM2 model shows the best accuracy among the presented models (Fig. 3(a)–3(c)) in almost all series of experiments. And in all cases, the DM2 model surpasses the traditional DE model, where Euclidean distance measure used (Fig. 3(b), 3(c)).

Experiments showed that there is no correlation between the values of the objective function and the Rand index in series of experiments with model DM1 in any combinations of training and working samples (Fig. 4(a)). In other models with an increase the volume of training and working samples (n_t and n_w , respectively), the clustering accuracy becomes constant (Fig. 4(b)). For DM2 model there is an inverse correlation between the achieved value of the objective function and the clustering accuracy RI on a small sample (Fig. 5(a)).

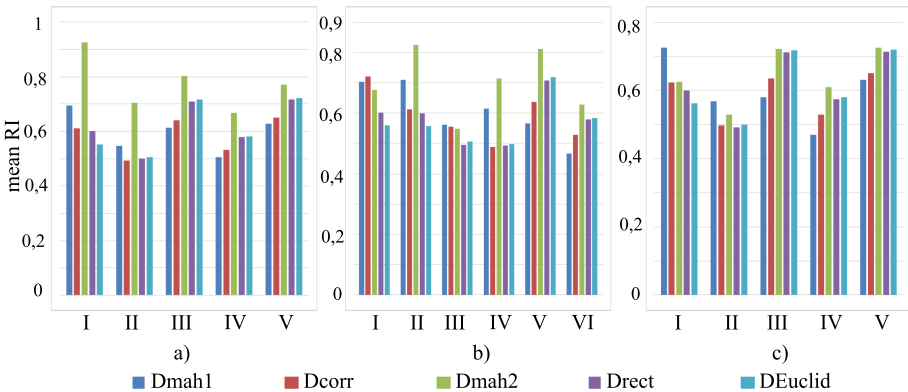


Fig. 3. The mean value of the Rand index for a) 1st group; b) 2nd group; c) 3rd group

In addition, the fact deserves attention that when applying the Euclidean distance measure, the best (smaller) values of the objective function do not correspond to the best (large) accuracy values. (Fig. 5(b)). This fact shows that the model with the Euclidean distance measure is not quite adequate: the most compact clusters do not exactly correspond to homogeneous batches.

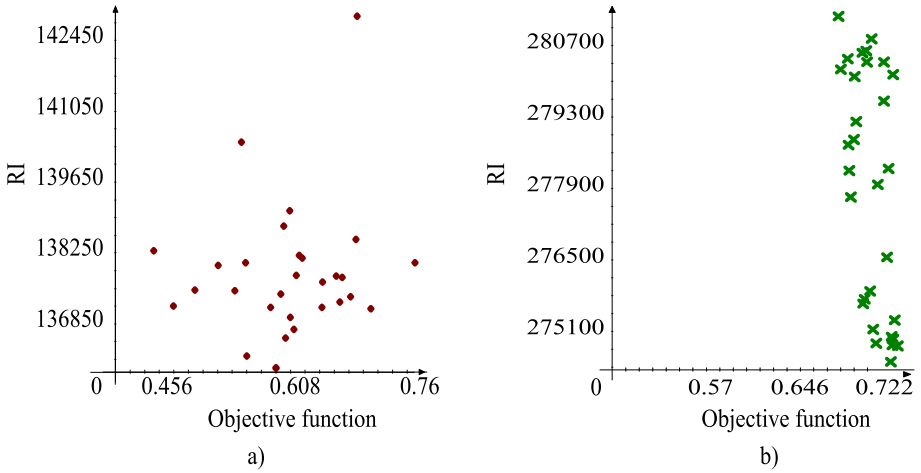


Fig. 4. Dependence of the Rand index on the value of the objective function for a) DM1 model ($n_t = 3987, n_w = 2054$); b) DM2 model ($n_t = 3987, n_w = 3987$)

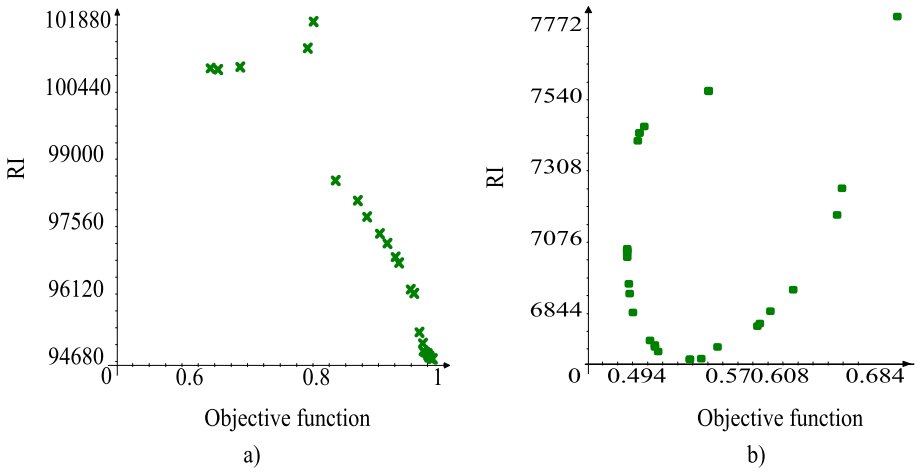


Fig. 5. Dependence of the Rand index on the value of the objective function for a) DM2 model ($n_t = 187, n_w = 187$); b) DE model ($n_t = 187, n_w = 187$)

7 Conclusion

The proposed clustering model and algorithm which uses the k-means model with Mahalanobis distance and an averaged (weighted average) estimation of the covariance matrix was compared with the k-means model with the Euclidean and rectangular distances in solving the problem of automatic grouping of industrial products by homogeneous production batches.

Taking into account the higher average Rand Index value, the proposed optimization model and algorithm applied for the electronic radio components clustering by homogeneous production batches has an advantage over the models with traditionally used Euclidean and rectangular (Manhattan) metrics.

Acknowledgement. Results were obtained within the framework of the State Task FEFE-2020-0013 of the Ministry of Science and Higher Education of the Russian Federation.

References

1. Orlov, V.I., Kazakovtsev, L.A., Masich, I.S., Stashkov, D.V.: Algorithmic support of decision-making on selection of microelectronics products for space industry. Siberian State Aerospace University, Krasnoyarsk (2017)
2. Kazakovtsev, L.A., Antamoshkin, A.N.: Greedy heuristic method for location problems. *Vestnik SibGAU* **16**(2), 317–325 (2015)
3. Rozhnov, I., Orlov, V., Kazakovtsev, L.: Ensembles of clustering algorithms for problem of detection of homogeneous production batches of semiconductor devices. In: 2018 School-Seminar on Optimization Problems and their Applications, OPTASCL 2018, vol. 2098, pp. 338–348 (2018)
4. Kazakovtsev, L.A., Antamoshkin, A.N., Masich, I.S.: Fast deterministic algorithm for EEE components classification. *IOP Conf. Ser. Mater. Sci. Eng.* **94**. <https://doi.org/10.1088/1757-899X/04/1012015>. Article ID 012015
5. Li, Y., Wu, H.: A clustering method based on K-means algorithm. *Phys. Procedia* **25**, 1104–1109 (2012). <https://doi.org/10.1016/j.phpro.2012.03.206>
6. Ansari, S.A., et al.: Using K-means clustering to cluster provinces in Indonesia. *J. Phys. Conf. Ser.* **1028**, 521–526 (2018). 012006
7. Hossain, Md., Akhtar, Md.N., Ahmad, R.B., Rahman, M.: A dynamic K-means clustering for data mining. *Indones. J. Electr. Eng. Comput. Sci.* **13**(521), 521–526 (2019)
8. Perez-Ortega, J., Almanza-Ortega, N.N., Romero, D.: Balancing effort and benefit of K-means clustering algorithms in Big Data realms. *PLoS ONE* **13**(9), e0201874 (2018). <https://doi.org/10.1371/journal.pone.0201874>
9. Patel, V.R., Mehta, R.G.: Modified k-Means clustering algorithm. In: Das, V.V., Thankachan, N. (eds.) *CIIT 2011. CCIS*, vol. 250, pp. 307–312. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25734-6_46
10. Na, S., Xumin, L., Yong, G.: Research on k-means clustering algorithm: an improved k-means clustering algorithm. In: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jingtangshan, pp. 63–67 (2010)
11. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297 (1967)
12. Shkaberina, G.S., Orlov, V.I., Tovbis, E.M., Kazakovtsev, L.A.: Identification of the optimal set of informative features for the problem of separating of mixed production batch of semiconductor devices for the space industry. In: Bykadorov, I., Strusevich, V., Tchemisova, T. (eds.) *MOTOR 2019. CCIS*, vol. 1090, pp. 408–421. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33394-2_32

13. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1981)
14. Bradley, P.S., Mangasarian, O.L., Street, W.N.: Clustering via concave minimization. In: Advances in Neural Information Processing Systems, vol. 9, pp. 368–374 (1997)
15. Har-Peled, S., Mazumdar, S.: Coresets for k-Means and k-Median clustering and their applications. In: Proceedings of the 36th Annual ACM Symposium on Theory of Computing, pp. 291–300 (2003)
16. Maranzana, F.E.: On the location of supply points to minimize transportation costs. IBM Syst. J. **2**(2), 129–135 (1963). <https://doi.org/10.1147/sj.22.0129>
17. Kaufman, L., Rousseeuw, P.J.: Clustering by means of Medoids. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L1-Norm and Related Methods, pp. 405–416. North-Holland, Amsterdam (1987)
18. Park, H.-S., Jun, C.-H.: A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. **36**(2), 3336–3341 (2009). <https://doi.org/10.1016/j.eswa.2008.01.039>
19. Davies, D.L., Bouldin, D.W.: A cluster Separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-1**(2), 224–227 (1979)
20. Deza, M.M., Deza, E.: Metrics on normed structures. In: Encyclopedia of Distances, pp. 89–99. Springer, Heidelberg (2013) https://doi.org/10.1007/978-3-642-30958-8_5
21. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: The Mahalanobis distance. Chem. Intell. Lab. Syst. **50**(1), 1–18 (2000). [https://doi.org/10.1016/S0169-7439\(99\)00047-7](https://doi.org/10.1016/S0169-7439(99)00047-7)
22. McLachlan, G.J.: Mahalanobis distance. Resonance **4**(20), 1–26 (1999). <https://doi.org/10.1007/BF02834632>
23. Xing, E.P., Jordan, M.I., Russell, S.J., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems, vol. 15, pp. 521–528 (2003)
24. Arathiand, M., Govardhan, A.: Performance of Mahalanobis distance in time series classification using shapelets. Int. J. Mach. Learn. Comput. **4**(4), 339–345 (2014)
25. Orlov, V.I., Shkaberina, G.S., Rozhnov, I.P., Stupina, A.A., Kazakovtsev, L.A.: Application of clustering algorithms with special distance measures for the problem of automatic grouping of radio products. Sistemy upravleniia i informacionnye tekhnologii **3**(77), 42–46 (2019)
26. Orlov, V.I., Fedosov, V.V.: ERC clustering dataset (2016). <http://levk.info/data1526.zip>
27. Kazakovtsev, L.A., Orlov, V.I., Stashkov, D.V., Antamoshkin, A.N., Masich, I.S.: Improved model for detection of homogeneous production batches of electronic components. IOP Conf. Ser. Mater. Sci. Eng. **255** (2017). <https://doi.org/10.1088/1757-899x/255/1/012004>
28. Shumskaia, A.O.: Evaluation of the effectiveness of Euclidean distance metrics and Mahalanobis distance metrics in identifying the origin of text. Doklady Tomskogo gosudarstvennogo universiteta system upravleniia i radioelektroniki **3**(29), 141–145 (2013)
29. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
30. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)

31. Golovanov, S.M., Orlov, V.I., Kazakovtsev, L.A.: Recursive clustering algorithm based on silhouette criterion maximization for sorting semiconductor devices by homogeneous batches. *IOP Conf. Ser. Mater. Sci. Eng.* **537** (2019). 022035
32. Lapidot, I.: Convergence problems of Mahalanobis distance-based k-means clustering. In: *IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)* (2018). <https://doi.org/10.1109/icsee.2018.8646138>
33. Shkaberina, G.Sh., Orlov, V.I., Tovbis, E.M., Sugak, E.V., Kazakovtsev, L.A.: Estimation of the impact of semiconductor device parameters on the accuracy of separating a mixed production batch. *IOP Conf. Ser. Mater. Sci. Eng.* **537** (2019). <https://doi.org/10.1088/1757-899X/537/3/032088>. 032088
34. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**(336), 846–850 (1971). <https://doi.org/10.1080/01621459.1971.10482356>
35. Hansen, P., Mladenovic, N.: J-means: a new local search heuristic for minimum sum of squares clustering. *Pattern Recogn.* **34**(2), 405–413 (2001). [https://doi.org/10.1016/S0031-3203\(99\)00216-2](https://doi.org/10.1016/S0031-3203(99)00216-2)
36. Kazakovtsev, L.A., Antamoshkin, A.N.: Genetic algorithm with fast greedy heuristic for clustering and location problems. *Informatica* **38**(3), 229–240 (2014)