Saint Petersburg State University
Department of mathematical game theory and statistical decisions

Lozkins Aleksejs

Master's thesis

The stability based approach of cluster number determination

Specialization 01.04.02
Applied Mathematics and Informatics
Master's Program Game Theory and Operations Research

Research advisor

D.Sc., Professor

Bure V. M.

Saint Petersburg
2016

Contents

Introduction

The cluster analysis is the data learning tool. There exist some differentiations of data mining tasks: clustering, classification, regression, association rule learning and etc. All this data analysis tools have own specificity. The clustering is subset of data mining tools, which solve the automatic grouping of data without the supervisor. The cluster analysis has two main problems: clustering algorithm and which number of cluster to chose.

The data structure has a different representations and where does not exists the common clustering algorithm for each kind of dataset. There exist lots of clustering algorithms: hierarchical clustering, k-means algorithms, HCS clustering algorithm, biclustering and etc. For this reason are developed set of clustering algorithms for different cases of data.

The supervised learning (classification) has the information about groups and their properties in turn the clustering is the data grouping without the information about clusters properties. In this case the different criteria are introduced [1]. There is no common criterion for each type of dataset and only local-optimum of the clustering can be found. The criterion variety depends on the clustering application area. The clustering analysis is applied in medicine, computational biology, computer vision, economics, genetic and etc.

The proposed methods are related to additional problem in data analysis – data inaccuracy. The both clustering and data clearness problems are employed for development of criterion of cluster number estimation. The goal of the dissertation work is to produce the cluster number validation criterion and to solve the data imprecision problem at the same time. The current work propose the new algorithms of cluster number determination which in certain sense improve the existing algorithm.

The materials are published in 7 printed works, 2 of them are the articles [2, 3] and one paper [4] is indexed in Scopus, 4 conference abstracts [5, 6, 7, 8] and 1 paper is under publication in the "Молодой ученый" journal in geographical section applying the developed method to the real data.

The structure of the dissertation consists of introduction, four sections, conclusion and bibliography.

The introduction presents the relevance of the work, the introduction into the problems of considered topics, the short presentation of goals of research.

The Section 1 presents the existing approaches of considered problem, the description used methodologies, the basic notation is introduced and the literature review conducted.

The Section 2 presents the additional specific notation for proposed methods, the similarity measure between clusterings is introduced and its metric properties are proved, the two algorithms and their descriptions are exposed.

The Section 3 shows the algorithms' work on the artificial data and their operability test.

The Section 4 presents two results of proposed algorithms usage on real data and results interpretation.

In the Conclusion the research results are summed up and the main conclusions are formulated.

Chapter 1

The formal description of existed methods

1.1 The common existed approaches

Nowadays, the problem of "true" cluster number estimation is unsolved. Although it does not mean that there are no alternatives for grouping number estimation in the cluster analysis. A wide spectrum of methodologies exists, which consists of a large amount of criteria for the data natural group number estimation. The criteria is not optimal in general case as it gives only local-optimum.

Stability methodology in cluster analysis has a direct connection with current work, which causes the main interest. Comparing to the mathematical analysis the stability concept in clustering analysis does not have a mathematical interpretation. This concept depends on logical beliefs and in most of the cases of the clustering analysis it looks like a cluster variety estimation. The basis of this variety estimation is considered to have the inverse property — the similarity between the inter results, which are generated under an assumption.

Before the statement of central concepts, the basic notation should be introduced. The common theoretical notation does not exist, because the clustering analysis have been developed in multiple scientific areas (biology, medicine, computer science, statistics etc.). The notation presented here will be simple to use in mathematical descriptions.

Let us denote the finite set $X = \{x_1, x_2, ..., x_n\}$ of d-dimensional Euclidean space R^d, where elements $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$ – d-dimensional vectors $\forall x_i \in X$, further the elements from X will be considered as vectors without additional second index. The clustering method will be denoted as a function from the set of $\alpha(X)$. The data partition and the result of clustering is $\alpha(X) = \{S^1, S^2, ..., S^k\}$, where k is the number of partitions or clusters and $\forall S^j \subset X$ and $S^j \cap S^l = \emptyset$ for $\forall j \neq l$. Additional notations for each of the concept descriptions will be introduced directly in concepts presentation.

Most of stability based concepts use the distances between clusters. A variety of distance measures can be found in literature, the following representation of clusters needs to be introduced:

$$c_{ij} = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ is in the same cluster and } x_i \neq x_j; \\ 0, & \text{otherwise.} \end{cases}$$

The matrix C^q for each clustering $\alpha^q(X)$ its own. The dot product of matrices C^1, C^2 is defined by formula

$$\langle C^1, C^2 \rangle = \sum_{i,j} c^1_{ij} c^2_{ij}.$$

The dot product computes the number of elements which are clustered together.

The dot product have the following property:

$$\langle C^1, C^2 \rangle \leq \sqrt{\langle C^1, C^1 \rangle \langle C^2, C^2 \rangle}.$$

The property is named as Cauchy-Schwartz inequality by its developers. Under this inequality the cosine similarity measure was introduced by Fowlkes and Mallows in [9]:

$$cor(C^1, C^2) = \frac{\langle C^1, C^2 \rangle}{\sqrt{\langle C^1, C^1 \rangle \langle C^2, C^2 \rangle}}.$$

These are normalized correlations between two clustering, which can be considered as “distance” between clusterings or cluster similarity measure.

Significant amount of similarity measures between the clusters is introduced in the cluster analysis, the commonly used measures will be presented in this part of work. The dot product helps us to express the following measures:

$$J(C^1, C^2) = \frac{\langle C^1, C^2 \rangle}{\langle C^1, C^1 \rangle + \langle C^2, C^2 \rangle - \langle C^1, C^2 \rangle},$$

Jaccard coefficient [10] — is the similarity measure, where only "non-negative" matches are considered.

The squared norm of matrix $\| C \|^2 = \langle C, C \rangle$ is proposed to use [11] for the matching coefficient representation:

$$M(C^1, C^2) = 1 - \frac{1}{n^2} \| C^1 - C^2 \|^2,$$

a way of weighed penalty for non-matchings usage in similarity estimation.

The choice of similarity measure depends on the aim of the analysis. In the most of cases the precise measures are excessive and make additional computational load, which is undesirable.

The first concept of "true" cluster number determination which we consider is presented in [12]. The idea consists of the random subsample generation X_1 from general set X. The condition for X_1 is the fixed number of elements from X. This number of elements has been characterized by sampling ratio f, which describes the proportion of elements in subsample X_1 to whole number of elements in X. The fraction of points sampled is $f > 0.5$ and it should be close, but not equal 1. This condition provides the property for two different subsamples X_1 and X_2 with same f:

$$X_1 \cap X_2 \neq \emptyset.$$

The main assumption of this concept is that the clustering structure of X should be "inherent" by subsamples. Under this assumption the clusterings of two different subsamples with same ratio should have the same clusterings on intersection of these subsamples. The comparison of two different clusterings is held by similarity measures. The "true" clustering is stable with respect to sub-sampling.

The second stability concept of cluster number validation is based on supervised learning. The idea is formulated in [13] and the additional modifications are added (modifications of subsample generations) by Lozkins and Bure.

The solution of clustering is presented as a set of labels Y, where $y_i \in Y$ represents the number of the cluster, wherein the element $x_i \in X$ and x_i cluster label is y_i. The supervised learning has the definition of classifier — function ϕ, which under training set makes the conclusion of element label. A lot of classification approaches exists with it's own classification function and in this work the theory of classification is omitted.

The work [13] does not provide the methods how to choose the training set. Therefore the idea of the previous concept of the subsample is used. The training set is the subsample $X_1 \in X$ with fraction level f and the "inheritance" of cluster structure implied.

Let us assume the classifier ϕ_1 is trained on subsample (X_1, Y_1), where Y_1 is the clustering result of X_1. The classification result $Y = \phi_1(X)$ is transformed to matrix C_1. The clustering results of X is represented by C. In this idea the stability estimation is compared to average similarity measures for each considered cluster number. If the variety of clusters for different classifiers for the same fraction ratio for fixed number of clusters is low, then clustering is stable with respect to another number of clusters considered.

The challenge of presented concepts is the fraction ratio and in the last case — the classifier function. In some cases it can cause problems (for example: f is too small and does not include the whole cluster) and the result can be inaccurate. This way a large amount of variants of algorithms needs to be tested.

1.2 Literature review

A lot of methodologies can be found there for "correct" cluster number estimation. Different points of data interpretation makes the fields for disjoint ideas of clustering problems solution. The main existing methodologies in clustering analysis of cluster number validation are described in this section.

The methodology of consideration of intra- and inter-cluster variances as a function of number of clusters is a basis of criteria represented in the following research: [14, 15, 16, 17, 18, 19, 20], [21] (C-index) and [1] (the Gap Statistic method). Methods described in the works mentioned above use the geometrical prospective.

The gap statistic consider the intra clusters dispersion or called the error measure W_k versus number of clusters k. The discrete curve W_k monotonically decrease as the number of clusters increase. The sudden jump down of the curve at certain number of k is an appropriate number of cluster – "elbow" criteria.

The C-index consider the sum of distances over all pairs from the same cluster S, the S_{min} is the sum of l shortest distances between pairs, the S_{max} is the sum of l largest distances in the whole set. The index is defined by formula: $C = \frac{S - S_{min}}{S_{max} - S_{min}}$. The better clustering is determined by small C-index value.

Nonparametric approach defines the "true" number of clusters as the number of areas with high data density. Each cluster corresponds to the "domain of attraction", i.e. area with high data concentration. The Wishart in [22] was the first to suggest the density mode for cluster structures' research. Later the idea was developed by Hartigan in [23].

Multiple methodologies exist apart from the ones mentioned. However, this work is focused on the methodology latter described as it is the closest to the one proposed by the author of the work. Methods involving cluster stability concept are based on consideration of two clustered selections at a time which are taken from initial dataset. The composition of clusters can be interpreted as their reliability [24] at high stability level. The idea described is a basis for criteria of determination of optimal number of clusters used in the work of Levine and Domany [25] and Ben-Hur, Elisseeff and Guyon [12, 26, 27].

Another methodology uses the goodness of fit tests. The paper [28] introduced the X-means algorithm for optimal cluster number determination via Bayessian information criterion. The mean of Hotelling's T-square p-values is considered in [29] for distance estimation between samples. The earlier works by Volkovich, Brazly and Morozensky [30] describe the way for cluster number determination through statistical criterion.

The listed circumstances suggest that “correct” clustering should be stable with respect to the random perturbation of initial data, at least at a low variance. As the value of variance can be varied, mathematical expectation of the random perturbation is always equal to zero, suggesting the absence of regular errors in the initial data. In the papers [31] and [32] similar approaches are being used, these works are focused on a clusters similarity measures and consideration of normal and binomial noise distributions respectively.

The data perturbations in an initial dataset X as a grouping stability estimation are considered in [2, 3, 4]. The main assumption of these methods is inaccuracy of the data studied. If the low perturbation of data is added then the clustering should be the same. The variances of the perturbations are the control parameters for the stability validation. The acceptable clustering is the data sample grouping that is robust to random perturbations of investigated data compared to the other considered variants. These methods used with real data are presented in [3, 7] and will be described in later sections.

Chapter 2

Formalization of the approach

2.1 Methodology description

The idea of the approach comes from the studied data clearness problems. Initial data explored always has inaccuracies, the nature of these errors has a different origin, rate and it has the casual character. Erroneous data arises due to the measurement and/or rounding errors. Also, there are errors in mathematical models. Analyzing the socioeconomic information the situations can be observed when the initial numerical data contains various kinds of inaccuracies. Occasionally, data are intentionally misrepresented, e. g. economical data relating to the income of juridical or natural person, or socio-economical data, which are collected as a result of selective survey of various social groups. Apart from those mentioned other possible sources of observational inaccuracy exist. Consequently, the analyzed data is imprecise. In that kind of clustering tasks the result cannot rely on the "truth" of already existing data. Thus, the "correct" value of observation can differ from the data investigated. The circumstances suggest that "correct" clustering should be stable with respect to the random perturbation of the initial data, at least at the low variance. As the value of the variance can be varied, mathematical expectation of random perturbation is always equal to zero, suggesting the absence of regular errors in the initial data [2, 4].

Two versions of algorithms are presented above which determine the acceptable clustering that is robust to random perturbations of the underlying data. Every option has right for existence. The additional notation has to be introduced.

The new dataset which is generated as a noise addition to each element of X will be denoted as $X_\sigma = \{x_{i,\sigma} | x_i \in X, x_{i,\sigma} = x_i + \varepsilon_{i,\sigma}, i = 1, ..., n\}$, where $\varepsilon_i = (\varepsilon^1_{i,\sigma}, ..., \varepsilon^d_{i,\sigma})^T$ and $\varepsilon^j_{i,\sigma}$ are mutually independent random elements from uniform distribution $U(-\delta, \delta)$, or in a general case, from beta-distribution [33] defined in $(-\sigma, \sigma)$. The noise generation can be performed from other distributions, for example, from truncated normal distribution (i. e. initial normal distribution has a zero mathematical expectation and variance equals to σ^2, $N(0, \sigma^2)$) within a given truncation range $(-\delta, \delta)$. The choice of distribution is contingent on initial supposition on the initial data error nature. In the case of the absence of the distribution information of initial data inaccuracies, as a benchmark distribution is given the uniform distribution.

The result of partition into k disjoint subsets of set X are denoted as $X = \cup_{i=1}^k S_i$, where $S_i \cap S_j, \forall i \neq j \vee i = 1, ..., k; j = 1, ..., k$. The partition algorithm or the approach lets us denote $\alpha_k(X) = \{S_1, ..., S_k\}$.

One of the approaches uses the perturbed set X_σ. The perturbation level σ should be estimated by the expert or the set of $\{\sigma\}$ must be considered. In the second case the complexity of the stability estimation is growing up.The assumption of data inaccuracy is represented by the perturbed data. The grouping of X_σ into k clusters will be denoted as $S_1^\sigma, S_2^\sigma, ..., S_k^\sigma$, which are the k disjoint sets and the union of all $\cup_{i=1}^k S_i^\sigma = X_\sigma$.

The second approach uses the extended perturbed set where the set $X_{\sigma+} = X \cup X_\sigma$ with not more than $2n$ number of elements. The partition of $X_{\sigma+}$ into k disjoint groups will be denoted as $S_1^{\sigma+}, S_2^{\sigma+}, ..., S_k^{\sigma+}$. Lets define the following sets $S_i^{\sigma-} = S_i^{\sigma+} \setminus X_\sigma, i = 1, ..., k$, which are the k disjoint sets and the union of all $S_i^{\sigma-}$ is the set X. The approach of partition is defined as $\alpha_k^{\sigma-}(X) = \{S_1^{\sigma-}, ..., S_k^{\sigma-}\}$.

In this case the stability concept is viewed as group variability under the data extension by the perturbed data. The data extension does not end on one perturbed initial set addition, the number of perturbed sets can differ and noise distribution for each set can have it's own distribution parameters. For example, lets consider the r perturbed sets $X_{\sigma+} = X \cup X_{\sigma_1} \cup ... \cup X_{\sigma_r}$ with no more than rn number of elements. In most cases $r = 1$ is enough. Large number of perturbed sets add complexity but does not give additional precision.

The cluster stability concept implies cluster comparison. The concept of sufficiently rough estimation of cluster consistency comparison is proposed. Consider the partitions: $S_1, S_2, ..., S_k$ and $S_1^\sigma, S_2^\sigma, ..., S_k^\sigma$. If the relevant matrices C and C^σ, of the partition do match, then the distance between partitions is 0. And 1 otherwise. Using the dot product, the introduced (by Lozkins and Bure) rough partition similarity measure can be represented by the following formula:

$$lb(\alpha_k(X), \alpha_k(X_\sigma)) = \begin{cases} 0, & \text{if } \frac{\langle C, C^\sigma \rangle}{\sqrt{\langle C, C \rangle \langle C^\sigma, C^\sigma \rangle}} = 1; \\ 1, & \text{otherwise.} \end{cases}$$

Lemma. The $lb(\alpha_k(X), \alpha_k(X_\sigma))$ similarity measure has distance properties.

Proof. Proposed similarity measure satisfies the non-negative axiom: $lb(\alpha_k(X), \alpha_k(X_\sigma)) \geq 0$, the set of default values is $\{0, 1\}$.

Lets show that identity of indiscernibles axiom holds. The condition $lb(\alpha_k(X), \alpha_k(X_\sigma)) = 0$ holds if and only if $\frac{\langle C, C^\sigma \rangle}{\sqrt{\langle C, C \rangle \langle C^\sigma, C^\sigma \rangle}} = 1$. If $C = C^\sigma$ then $\langle C, C^\sigma \rangle = \langle C, C \rangle = \langle C^\sigma, C^\sigma \rangle$ consequently $\frac{\langle C, C^\sigma \rangle}{\sqrt{\langle C, C \rangle \langle C^\sigma, C^\sigma \rangle}} = 1$. Cauchy-Schwartz inequality $\langle C, C^\sigma \rangle \leq \sqrt{\langle C, C \rangle \langle C^\sigma, C^\sigma \rangle}$ gives that $\frac{\langle C, C^\sigma \rangle}{\sqrt{\langle C, C \rangle \langle C^\sigma, C^\sigma \rangle}} \leq 1$. If $C \neq C^\sigma$ then $\langle C, C^\sigma \rangle < \langle C, C \rangle$ and $\langle C, C^\sigma \rangle < \langle C^\sigma, C^\sigma \rangle$ because $\langle C, C \rangle$ and $\langle C^\sigma, C^\sigma \rangle$ have maximal numbers of matches (from definition of matrix C in the Section 1), consequently $\frac{\langle C, C^\sigma \rangle}{\sqrt{\langle C, C \rangle \langle C^\sigma, C^\sigma \rangle}} < 1$ and $lb(\alpha_k(X), \alpha_k(X_\sigma)) = 1$.

The proposed similarity measure has the symmetry property inherited from the dot product symmetry property: $lb(\alpha_k(X), \alpha_k(X_\sigma)) = lb(\alpha_k(X_\sigma), \alpha_k(X))$.

Let us show that the triangle inequality hold. If $lb(\alpha_k(X), \alpha_k(X_\sigma)) = 0$ consequently the $lb(\alpha_k(X), \alpha_k(X_\sigma)) \leq lb(\alpha_k(X), \alpha_k(X_\beta)) + lb(\alpha_k(X_\sigma), \alpha_k(X_\beta))$ holds as $lb(\alpha_k(X), \alpha_k(X_\beta)) + lb(\alpha_k(X_\sigma), \alpha_k(X_\beta)) \geq 0$ under non-negativity condition. The second case $lb(\alpha_k(X), \alpha_k(X_\sigma)) = 1$, that means $C \neq C^\sigma$. Let us show that $lb(\alpha_k(X), \alpha_k(X_\sigma)) \leq lb(\alpha_k(X), \alpha_k(X_\beta)) + lb(\alpha_k(X_\sigma), \alpha_k(X_\beta))$ holds. Without loss of generality, $C = C^\beta$. Consequently $C^\sigma \neq C^\beta$. In this case $lb(\alpha_k(X), \alpha_k(X_\beta)) + lb(\alpha_k(X_\sigma), \alpha_k(X_\beta)) = 1$ and inequality condition holds. If $C \neq C^\beta$ and $C^\sigma \neq C^\beta$, then $lb(\alpha_k(X_\sigma), \alpha_k(X_\beta)) = 2$ and inequality condition holds too. **Proved.**

The rough similarity measure is chosen to save the computational costs. The other more precise similarity measures can be used, but this would significantly increase computational complexity of the method. In most cases it is excessive and unjustified.

The proposed distance between clusterings is adapted to the algorithms in the later Sections. The main attention is drawn to the perturbation parameter determination using this metric, which allows to reduce the number of settings.

The meaning of the clustering stability in our understanding is defined by the clusterings variety level under random perturbations in data compared with initial clustering. The superior, precise and proper clustering consists of the clusters which have the minimal level of dependence from perturbation level.

Definition. *The variability frequency* – the ratio of the number of observed mismatches to the total number of repetitions, denoted by ν.

The variability frequency is the result of algorithm's work, it shows the stability level for each number of clusters considered. This value represents the normalized number of initial and perturbed data clustering mismatches and consequently the clustering reliability.

2.2 An algorithm based on data inaccuracies

In Section 2.1 the data inaccuracy problem was mentioned as well as some ways of data interpretation in simulation studies for solution. The "true" clustering determination is considered as a main problem, which the algorithm tries to solve. The clustering algorithm and choice of number of clusters are important problems in cluster analysis. The algorithm described in this section finds the local-optimum of cluster number and partially solves the clustering algorithm problem (which clustering algorithm approach is better). The cluster robustness is used for grouping number validation and is the basic assumption for the approach.

The algorithm is based on the stability concept. The stability meaning in this approach is connected to "errors" in data, rather to "errors" representation in the data. In mathematical science this idea is used in stochastic programming.

The algorithm gives a local-optimum, because clustering analysis has lot of clustering methods and there are methods which solutions ore not precise (fuzzy). Different algorithm nature adds the complexity to the cluster analysis problem. Using the expert opinion the clustering algorithm set can be chosen. This set is the parameter of the proposed algorithm. Under this set is related the local-optimum, in real situation all clustering methods can not be considered. In most cases it is not necessary and one or two clustering approaches can be considered (expert opinion).

There are a lot of approaches for cluster number determination (Section 1.2), the majority uses interval of admissible cluster number values and holds in the considered algorithm. The sequence of integers or the integer number interval is enough, the expert opinion is not necessary. Let us denote this set as $[k_{min}, k_{max}]$.

The perturbed set X_σ is the "crown" (the most important part) of the algorithm setting. If the data errors level is known, then fixed σ consideration is enough, otherwise the set of $\{\sigma\}$ should be considered. An important question: how to choose the set of noises distribution variances? This set can be estimated experimentally, starting from small variances values. If for all number clusters the variability frequencies are relatively low then the variances increase. If the variability frequencies are near 1, then the the variances decrease. Let us denote the variances set $\Omega = \{\sigma_i\}_{i=1}^{s}$.

The algorithm works with two sets X and X_σ. The clustering into k clusters of these sets should give the same results on most cases of the noises addition. There is the precision control parameter, which reflects the number of additions to X_σ. In each generation the clusterings are compared. The comparison runs using the *lb* distance between clusterings (another similarity measure can be used, but *lb* distance is more suitable for this algorithm).

The algorithm scheme is presented below:

```
*Input*: $X$, $[k_{min}, k_{max}]$, $\Omega = \{\sigma_i\}_{i=1}^{s}$;
*Output*: $\nu_{\sigma_i,k}$;
*Requires*: a clustering algorithm $\alpha_k(X)$, similarity measure between clusters;
$NumOfDiffers = 0$;
**for** ($k$ in $k_{min}..k_{max}$) **do**
  $\alpha_k(X)$;
  **for** ($\sigma_i$ in $\Omega$) **do**
    **for** ($i$ in $1..NumOfRepetitions$) **do**
      $X_{\sigma_i}$ generation;
      $NumOfDiffers = NumOfDiffers + lb(\alpha_k(X), \alpha_k(X_{\sigma_i}))$;
    **end for**
    $\nu_{\sigma_i,k} = \frac{NumOfDiffers}{NumOfRepetitions}$;
  **end for**
**end for**
```

The $NumOfRepetitions$ is the precision control parameter which was mentioned before.

The result of the algorithm's work is the frequency matrix $\{\nu_{\sigma_i,k}\}$. This result representation is complex and seriously interpreted. The statistics introduction into the result modification is the way of the result matrix simplification. For example, the median or the mean of the variances transform matrix to the vector. The vector represents the variability frequency for each considered number of cluster ν_k.

The first version of proposed algorithm is presented in the work [2] and the practical usage of this approach is considered in [3] and will be discussed in the Sections 3 and 4.1.

2.3 Algorithm based on extended set

In Section 2.2 the algorithm was considered, which is based on the data errors. In this section another algorithm uses a similar assumption, with some differences and significant modifications.

The work [23] describes clusters as the "islands" in the sea, where the heights above the sea level are represented as group densities. There is a similarity between above described concept and the proposed algorithm in this section. The idea offered represents the "islands" artificial uplift. If the "islands" heights above sea level increase beyond uplift, then the clusters are more stable. If square of "islands" grows, then clusters are less stable. The idea is presented in [4] and the application is publishing in the journal "Молодой ученый".

```
*Input*: $X$, $[k_{min}, k_{max}]$, $\Omega = \{\sigma_i\}_{i=1}^{s}$;
*Output*: $\nu_{\sigma_i,k}$;
*Require*: a clustering algorithm $\alpha_k(X)$, similarity measure between clusters;
$NumOfDiffers = 0$;
**for** ($k$ in $k_{min}..k_{max}$) **do**
  $\alpha_k(X)$;
  **for** ($\sigma_i$ in $\Omega$) **do**
    **for** ($i$ in $1..NumOfRepetitions$) **do**
      $X_{\sigma_i+} = X \cup X_{\sigma_i}$;
      $\alpha_k^{\sigma_i-}(X)$;
      $NumOfDiffers = NumOfDiffers + lb(\alpha_k(X), \alpha_k^{\sigma_i-}(X))$;
    **end for**
    $\nu_{\sigma_i,k} = \frac{NumOfDiffers}{NumOfRepetitions}$;
  **end for**
**end for**
```

The “islands” analogy helps to understand the concept of the proposed method. The “islands” uplift procedure are represented by extended dataset $X_{\sigma+}$. The σ has the same meaning as in the previous section. The union of initial and perturbed datasets $X_{\sigma+}$ makes stable clusters more unchangeable and unstable clusters have high variety level comparable with initial clustering. The extended set has an adjustable size $(2n, 3n, ...)$. The size helps to separate the stable clusterings from the unstable accurately, but the computational complexity significantly increases.

The result simplification using statistics takes a place in this approach as well as in previous section. The similarity measure is free to selection and should be simple in computational aspect.

Chapter 3

Computational experiments on artificial data

3.1 Artificial data description

The algorithms mentioned in the previous sections are tested on artificial dataset. This experiment shows the correct work of the proposed approaches and helps to describe parameters of the algorithms.

The artificial dataset consists of five samples generated from multivariate normal distribution with 100 elements in each sample with different means and variances. The parameters of the samples are presented in the Table 3.1.

Table 3.1: Multivariate normal distribution parameters

Sample number	Means vector	Variances matrix
Sample 1	(1,1)	$\begin{bmatrix} 0.15 & 0 \\ 0 & 0.15 \end{bmatrix}$
Sample 2	(1,4)	$\begin{bmatrix} 0.15 & 0 \\ 0 & 0.15 \end{bmatrix}$
Sample 3	(3.5, 2.5)	$\begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$
Sample 4	(6,1)	$\begin{bmatrix} 0.15 & 0 \\ 0 & 0.15 \end{bmatrix}$
Sample 5	(6,4)	$\begin{bmatrix} 0.15 & 0 \\ 0 & 0.15 \end{bmatrix}$

The exposed data represent five samples each can be presented as cluster without special research. In this case, the result of cluster number validation is known. The Figure 3.1 displays the artificial data. The two-dimensional dataset is denoted as X.

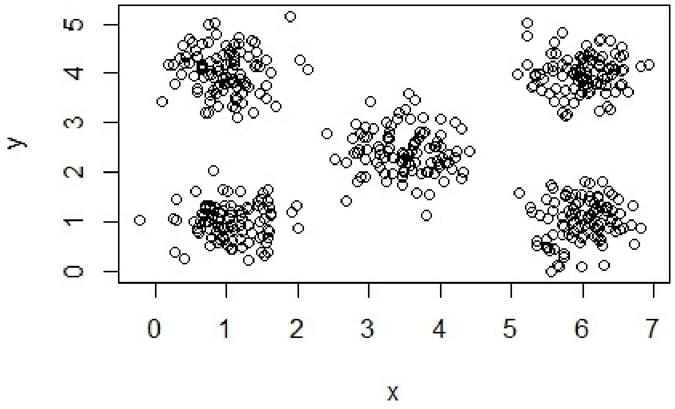

Figure 3.1: Union of 5 samples from multivariate Gaussian distribution

3.2 Algorithms tuning and application on the artificial data

The presented algorithms use similar input parameters. For result comparison the same parameters will be used. The parameters of the algorithms are presented in Table 3.2. The approach for cluster number determination which uses the extended set as union of initial datasets and one perturbed dataset (only one perturbed dataset) is applied.

Table 3.2: Algorithms parameters

Parameter	Value
Cluster number set $[k_{min}, k_{max}]$	[2,10]
Variances sequence	$\sigma_1 = 0.02; \sigma_i = \sigma_{i-1} + 0.02; i = 2..20$
Clustering algorithms	k-mean algorithm
Distance between clusters	$lb(\alpha_k(X), \alpha_k(X_{\sigma_i})$
Number of repetitions/simulations	50
Perturbation distributions	Uniform distribution, Gaussian distribution

The set of perturbation distribution parameters is most important. The correct variances set estimation plays the main role in variability frequency calculation. This set depends on intra data average standard deviation level.

The Figure 3.2 (a) and (b) present the variability frequencies for 2, 3, 5 and 6 number of clusters depending on perturbation variances values. If frequencies for each variance value is low (less than 0.5), then the variances level is not correct and should be increased. If frequencies for each variance value are near one, then the variances level is not correct too and should be reduced.

In Figure 3.2 (a) and (b) is generated on correct variances set, because there exists the cluster number with low level of frequencies (black line corresponds to number of clusters equal 5) which is interpreted as more stable, and other cluster numbers have high variability of frequencies and are not stable.

In both cases the results do not have significant differences. It has to do with main methodology of the algorithms and the simple case of studied data.

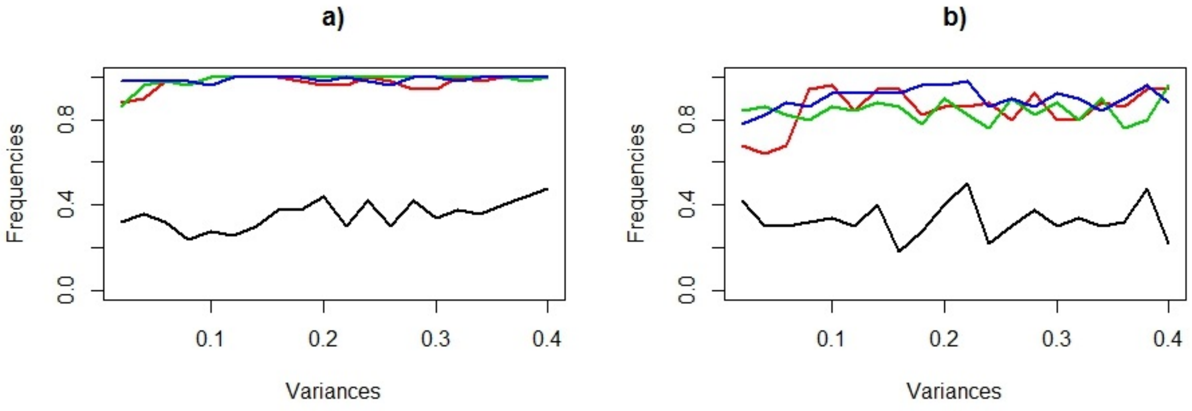

Figure 3.2: The frequencies for both algorithms

(a) The frequencies, resulting from the algorithm described in Section 2.2

(b) The frequencies, resulting from the algorithm described in Section 2.3

The Figure 3.2 (a) and (b) represent the results where variability frequencies $\nu_{\sigma_i,k}$ depend on two parameters: number of clusters and perturbation level. In the case when the set of numbers of clusters consist of 2–5 numbers this results presentation is simple analyze and make conclusions.

When the number of clusters is greater than 5, as in our case, the mean and median are used for result simplification. The "elbow" criterion is used in cluster number validation theory [1]. The statistics introduction to our algorithms makes it possible to use this criterion.

The experimental results for k-means clustering algorithm for the first cluster validation method are presented on Figure 3.4 (a), (b), (c) and (d). There four different cases are considered where two perturbation distributions and two statistics are taken.

The Figures a, c present the medians of frequencies and Figures 3 b, d represent the means of frequencies for different distributions. Two different statistics show similar results and confirm the truthfulness. The "average" of frequencies allows to represent the curves in Figure 3.2 as the point without loss of information.

The Figure 3.4 (a), (b) are generated using uniformly distributed perturbations in comparison with the Figures c,d results for normally distributed perturbation are better with minimal frequency about 0.3, but in the case with normal distribution the minimal frequency is 0.4. The variances set for both distributions was the same then the experiment was carried out.

The Figure 3.6 has the same interpretation as the Figure 3.4. The results of proposed algorithms on artificial dataset obviously are similar and differ insignificantly at non-stable numbers of clusters.

The goal of the cluster number validation algorithms is to obtain the sharp dips of the curve at stable points ("Elbow" criteria). In Figure 3.4 and Figure 3.6 the number of clusters is equal to 5. This is the expected solution thereby proving the confirmation to proposed algorithms. The results can not be interpreted as an optimal solution, because the limited set of number of clusters and number of clustering algorithms were considered.

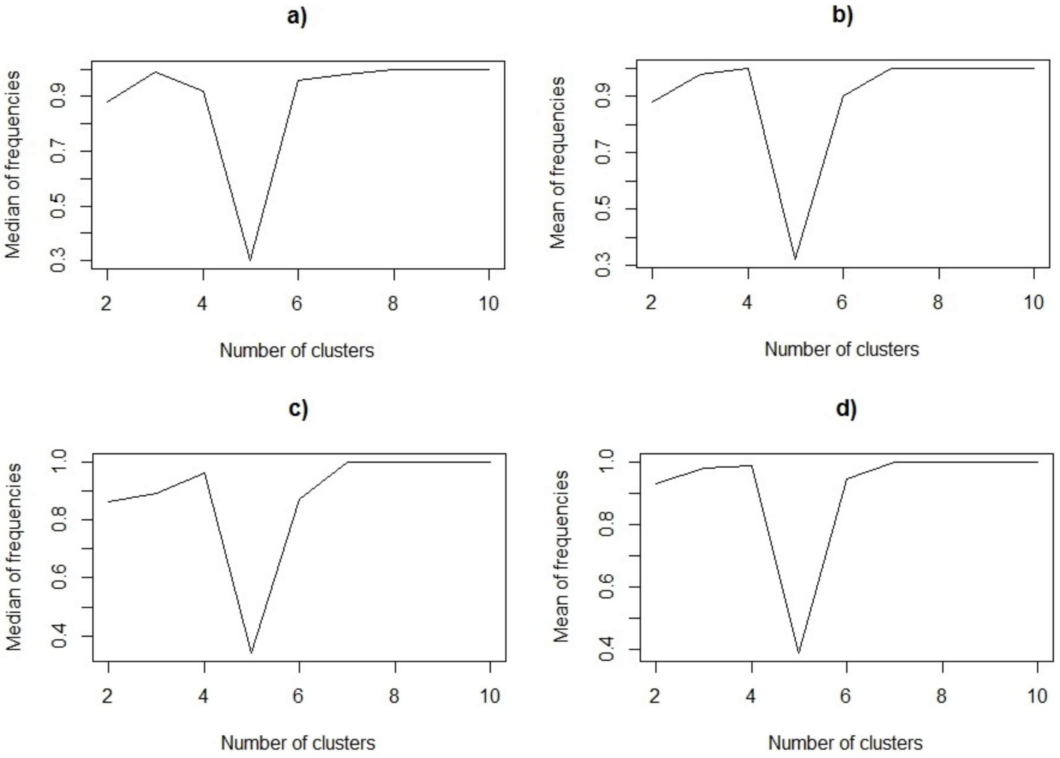

Figure 3.4: The k-mean clustering algorithm is being considered for the cluster number validation algorithm described in Section 2.2.

(a) The median of frequencies for each considered number of cluster the perturbations from uniform distribution

(b) The mean of frequencies for each considered number of cluster the perturbations from uniform distribution

(c) The median of frequencies for each considered number of cluster the perturbations from normal distribution

(d) The mean of frequencies for each considered number of cluster the perturbations from normal distribution

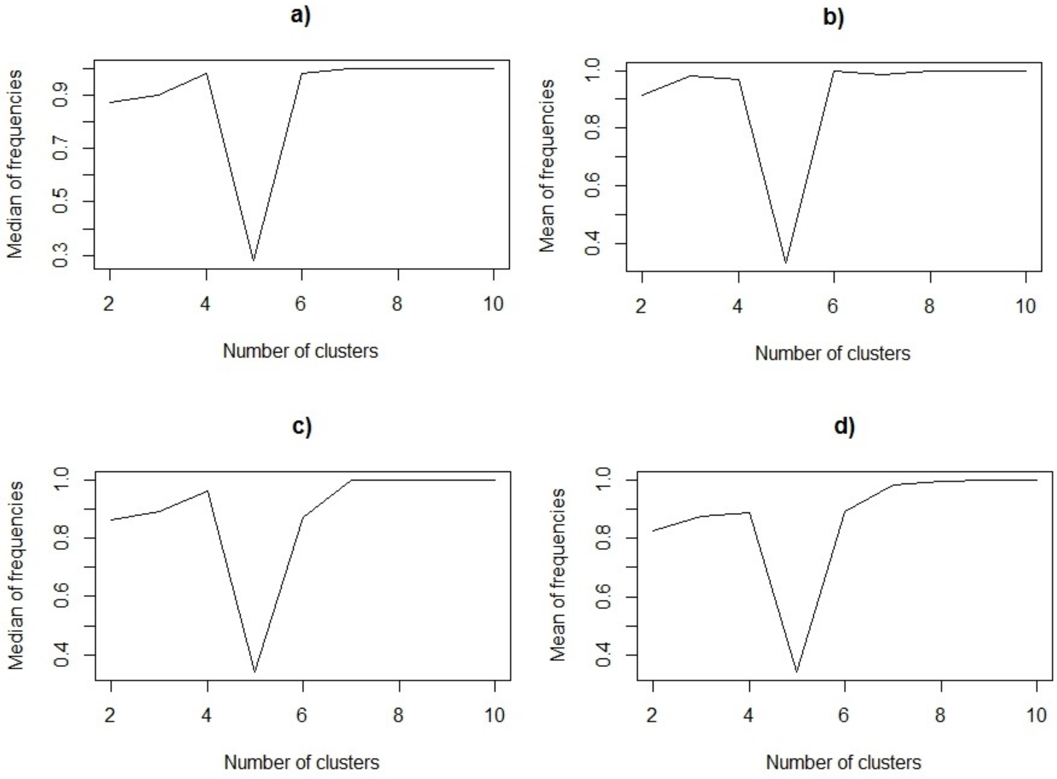

Figure 3.6: The k-mean clustering algorithm is being considered for the cluster number validation algorithm described in Section 2.3.

(a) The median of frequencies for each considered number of cluster the perturbations from uniform distribution

(b) The mean of frequencies for each considered number of cluster the perturbations from uniform distribution

(c) The median of frequencies for each considered number of cluster the perturbations from normal distribution

(d) The mean of frequencies for each considered number of cluster the perturbations from normal distribution

Chapter 4

Applications of algorithms for applied problems

4.1 The cluster analysis of European countries by unemployment rates

The cluster analysis application to data exploration helps in finding dependence or groups of elements with similar properties. The dependence may occur under large amount of parameters and simple tools are insufficient. The results of cluster analysis provide an answer which interpretation may provide new information or discovery.

The Section 4.1 describes the practical application of algorithm from Section 2.2 to economical dataset. The unemployment rates of European countries are considered as underlying data set. The data consist of 3 parameters or features, the unemployment rates of European countries by I, II and III quarters of 2014. The analogous result presentation can be found in the [7] where the inflation rates were considered instead of unemployment rates. The first real application of discussed approaches starts on this data set and is presented in [3]. The dataset of unemployment rates enclosed in Appendix 1. The parameters of the algorithm are presented in the Table 4.1.

The number of elements in investigated data is equal to 42. The maximal tested amount of clusters is 10, consequently the average contents of the clusters is about 4 elements. This groups are finely divided and may add

Table 4.1: Algorithms parameters

Parameter	Value
Cluster number set $[k_{min}, k_{max}]$	[2,10]
Variances sequence	$\sigma_1 = 0.07; \sigma_i = \sigma_{i-1} + 0.07; i = 2..20$
Clustering algorithms	k-mean algorithm, hierarchical clustering
Distance between clusters	$lb(\alpha_k(X), \alpha_k(X_{\sigma_i})$
Number of repetitions/simulations	50
Perturbation distributions	Uniform distribution

noises into research (clustering into 42 clusters would be more stable). The uniform distribution $U(-\delta, \delta)$ is considered only, some practical advantages of mentioned distribution are listed in the Section 3.

Figure 4.1 displays the variability frequencies $\nu_{\sigma_i,k}$ for $[2, 5]$ number of clusters set for k-means clustering method. The variances set is correct: each group number contains frequencies less than 0.1 and greater than 0.5. There are two stable points 2 and 3 with average frequencies 0.15 and 0.3 respectively Figure 4.1 (b) and (c).

Figure 4.3 represents the same results as Figure 4.1 but the clustering algorithm is different. Another clustering algorithm gives closer curves of frequencies for the stable grouping numbers. In the average the 2 and 3 cluster numbers continue to be stable compared to other cluster number cases.

The clusterings into 2 and 3 clusters of underlying set for both clustering algorithm are respectively the same. In the first case, clusters consist of countries presented in Table 4.2.

In the second cluster prevail countries resulting from the collapse of Yugoslavia in 2004. Bosnia and Herzegovina, Croatia, Macedonia and Serbia are post-communistic countries. All countries of second cluster have close geographical location. Greece in 2014 was in economic crisis and had problems with unemployment rate in that year.

The clustering results for 3 clusters have the consistence presented in Table 4.3.

Most of counties from second cluster are the countries which participated in the Soviet Union. After the Union collapse the economies of most

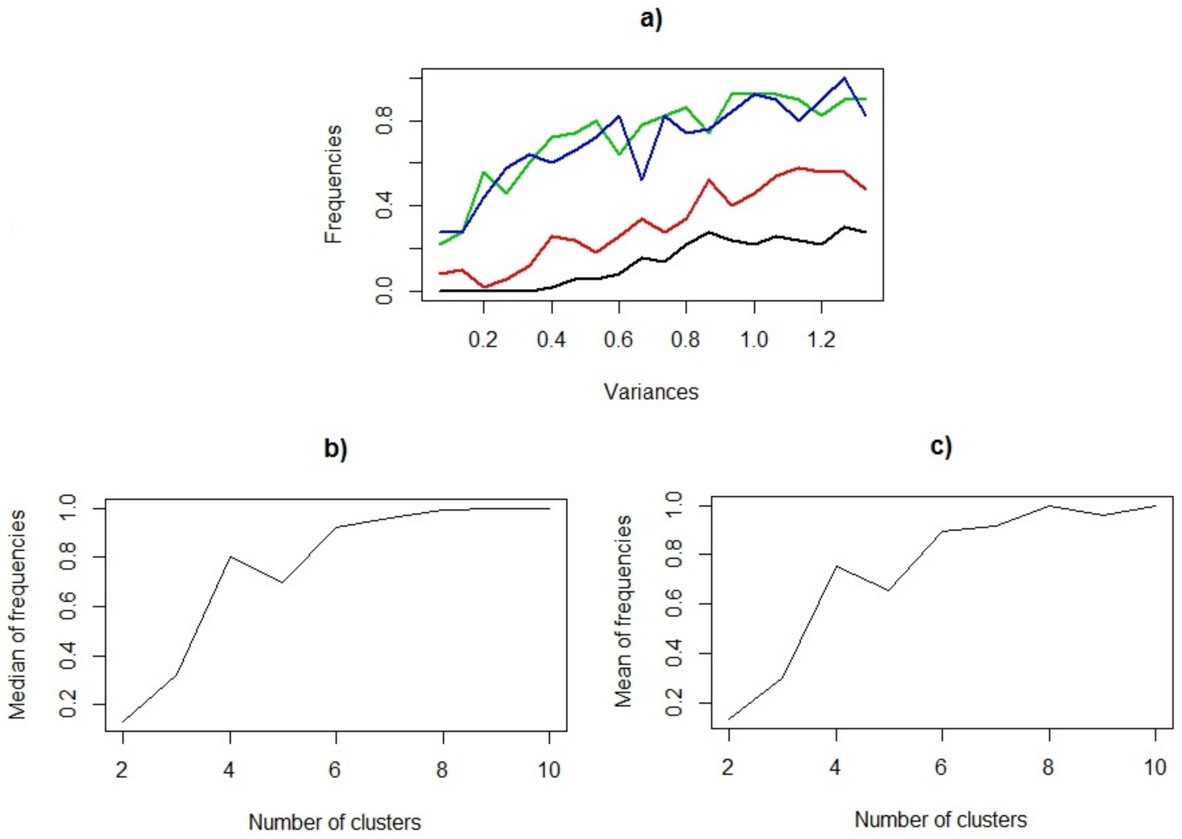

Figure 4.1: The *k*-mean algorithm is considered as clustering algorithm for the variability frequencies calculation

(a) The frequencies for different number of clusters (black – 2, red – 3, green – 4, blue – 5 cluster numbers)

(b) Median of frequencies for each considered number of cluster, perturbations are uniform

(c) Mean of frequencies for each considered number of cluster, perturbations are uniform

Table 4.2: The clustering of European countries by unemployment rate for two clusters

Cluster number	Countries
Cluster Nr.1	Albania, Armenia, Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Iceland, Ireland, Italy, Kazakhstan, Latvia, Lithuania, Luxembourg, Malta, Moldova, Montenegro, Netherlands, Norway, Poland, Portugal, Romania, Russia, Slovakia, Slovenia, Sweden, Switzerland, Turkey, Ukraine, United Kingdom
Cluster Nr.2	Bosnia and Herzegovina, Croatia, Greece, Macedonia, Serbia, Spain

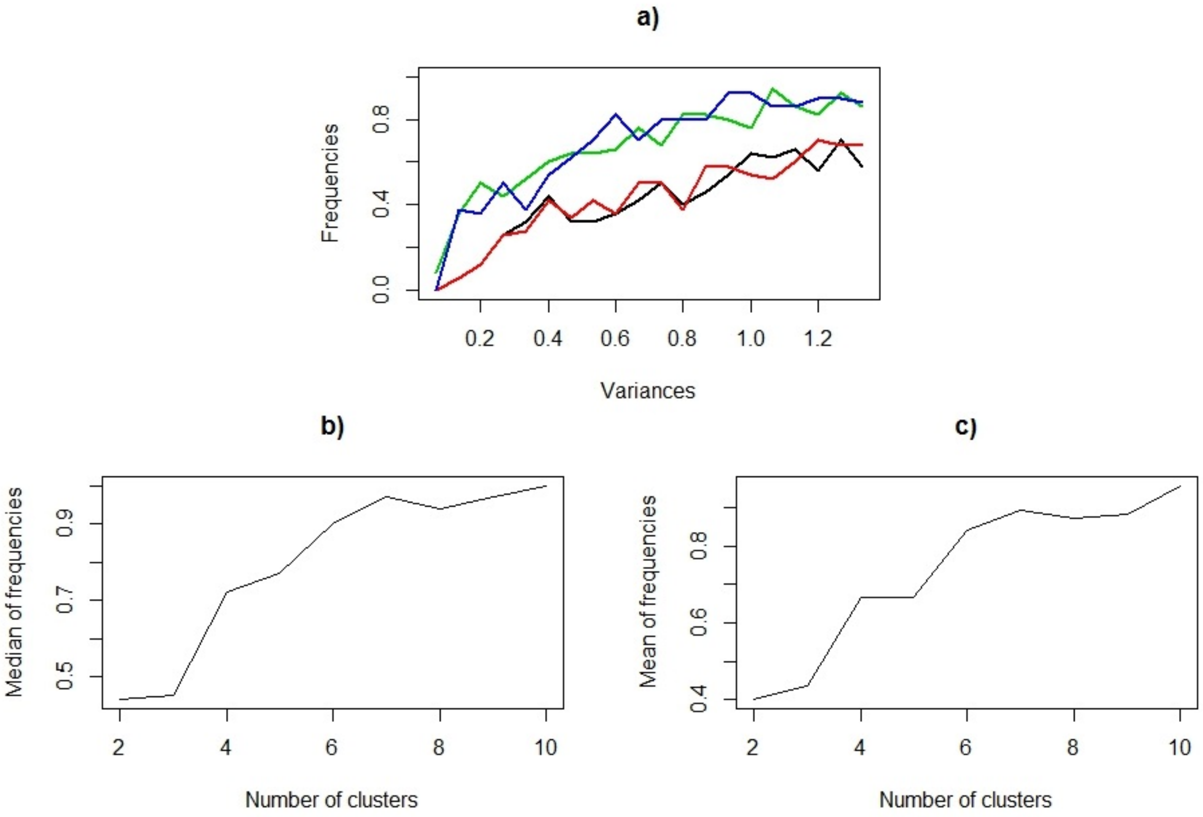

Figure 4.3: The hierarchical clustering algorithm is considered for the variability frequencies calculation

(a) The frequencies for different number of clusters (black – 2, red – 3, green – 4, blue – 5 cluster numbers)

(b) Median of frequencies for each considered number of clusters, perturbations are uniform

(c) Mean of frequencies for each considered number of cluster, perturbations are uniform

Table 4.3: The clustering of European countries by unemployment rate for two clusters

Cluster number	Countries
Cluster Nr.1	Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Iceland, Ireland, Kazakhstan, Luxembourg, Malta, Moldova, Netherlands, Norway, Romania, Russia, Sweden, Switzerland, Turkey, Ukraine, United Kingdom
Cluster Nr.2	Albania, Armenia, Bulgaria, Croatia, Cyprus, Ireland, Italy, Latvia, Lithuania, Montenegro, Poland, Portugal, Slovakia, Slovenia
Cluster Nr.3	Bosnia and Herzegovina, Greece, Macedonia, Spain

of countries had the equal level of development and it affected the todays economics level.

The result interpretation has a lot of points of view, the additional background or analysis of the results by expert in economical area makes this outcome more valuable.

The cluster analysis of European countries by unemployment rates is carried out. The two clustering algorithms are tested with uniformly distributed perturbations. The two stable clusterings are obtained and interpreted in this Section.

4.2 The clustering analysis of European Union countries by demographic parameters

Current demographic trends in Europe are characterized by fairly high mortality rate, low birth rate, low natural population growth and high migration activity.

Cluster analysis reveals significant natural grouping of data. Abstracting from the physical sense, it is possible to determine a division and communication in the tested data at the technical level. This analysis provides non-trivial results.

The heterogeneity of measuring scales demographics are assessed by requires additional analysis tools. According to the initial values the local normalization has been applied.

The following demographic indicators were the basis of the analysis: birth rate, mortality rate, migration, natural population increase and the proportion of the urban population in the total population. The indicators values are enclosed in Appendix 2 [35].

The clustering was carried out on different set of parameters and the most interesting will be presented in this Section. The second algorithm shown in Section 2.3 will be used for stable clustering estimation.

The setting for the algorithm is presented in the Table 4.4.

Table 4.4: Algorithms parameters

Parameter	Value
Cluster number set $[k_{min}, k_{max}]$	[2,10]
Variances sequence	$\sigma_1 = 0.005; \sigma_i = \sigma_{i-1}+0.005; i = 2..10$
Clustering algorithms	hierarchical clustering
Distance between clusters	$lb(\alpha_k(X), \alpha_k(X_{\sigma_i})$
Number of repetitions/simulations	50
Perturbation distributions	Uniform distribution

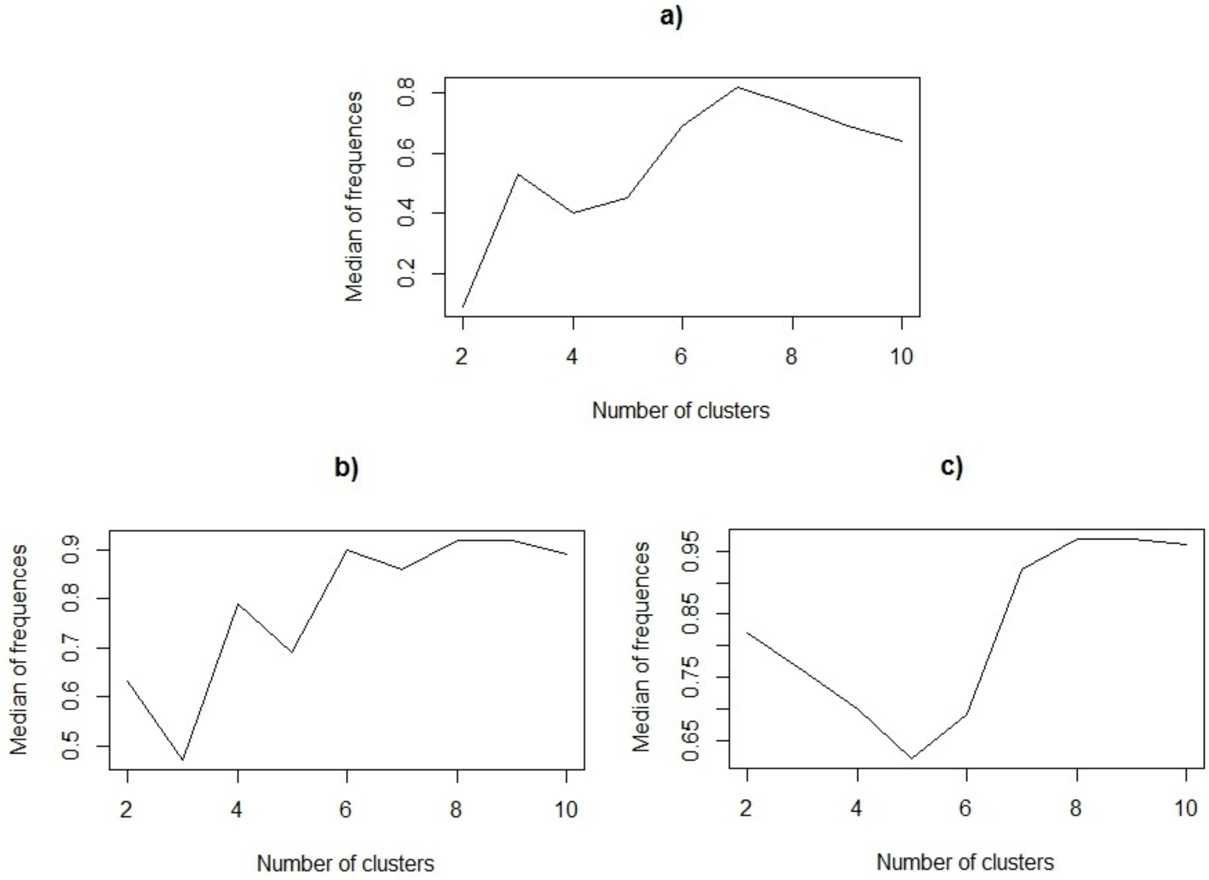

Figure 4.5: The hierarchical clustering algorithm is considered for the calculation of variability frequencies

(a) The five indicators of demography are studied (birth rate, mortality rate, migration, natural population increase and the proportion of the urban population in the total population)

(b) The two indicators of demography are studied (birth rate, natural population increase)

(c) The three indicators of demography are studied (natural growth rate, migration rate and the proportion of the urban population in the total population)

There are five indicators for each country. The all subsets of these indicators are carried out in this experiment. The outcomes with stable points are presented below.

The Figure 4.5 (c) displays the cluster number of five as more stable. The data have the groups presented in Table 4.5.

Table 4.5: Clusters for three demography indicators

Cluster number	Countries
Cluster Nr.1	Austria, Germany, Netherlands, Finland, Greece, Spain, Italy, Portugal, Bulgaria, Bosnia and Herzegovina, Hungary, Latvia, Lithuania, Macedonia, Poland, Romania, Serbia, Slovakia, Slovenia, Croatia, Montenegro, Czech Republic, Estonia
Cluster Nr.2	Belgium, United Kingdom, Luxembourg, France, Switzerland, Denmark, Norway, Sweden, Andorra, Cyprus, Malta, Monaco, San Marino
Cluster Nr.3	Ireland, Albania
Cluster Nr.4	Liechtenstein
Cluster Nr.5	Iceland

The first group, which includes highly Western Europe and Eastern Europe countries with zero natural increase, or close to zero values of natural increase, mostly positive migration rate and rather high proportion of the urban population, is the largest. The second cluster includes the developed countries of Western and Northern Europe, there are positive values of natural growth, high migration increase and high values of the urban population ratio.

The third cluster includes only two countries with high natural population increase, the very low migration increase and the low average values of the proportion of the urban population. The third cluster includes economically moderately developed countries of Europe – the Ireland and the country with a relatively low level of economical development among the other European countries – Albania. In Albania, the highest natural population growth is achieved mainly due to the high proportion of Muslim population. These countries are less attractive to migrants than the developed countries of Western and Northern Europe with higher standards of living. Liechtenstein forms a fourth cluster with approximately equal low values of natural population growth, migration and the proportion of the urban population. The fifth cluster includes only one country — Iceland — with a fairly good indicator of natural population increase among the other European countries, with a negative migration growth and a very high proportion of the urban population.

Five clusters of the level of demographic development are built into the rating. According to the three indicators the best demographic situation is observed in the countries that make up the second cluster. The fourth cluster is in the second place and includes the country with positive indicators that affect the changes in the population, but with a low proportion of the urban population. The fifth cluster takes the third place in the ranking and includes a country with high rates of natural population increase and the proportion of the urban population, but with a low migration increase. The first cluster is in the fourth place in the ranking, and the countries belonging to it have a small positive or negative natural population growth and migration, that is, they have low population dynamics. At the same time in the first cluster sufficiently high proportion of the urban population is observed. The third cluster includes countries with a very uneven distribution of the demographic indicators values that distinguishes it from the other clusters.

Figure 4.5 (a) represents the division into four clusters according to five indicators and demonstrates similar trends to the previous clustering — a favorable demographic situation in the economically developed countries of Western and Northern Europe and less favorable in the countries of Eastern and South-Eastern Europe. The clusters are presented in the Table 4.6.

In the region of Central and Eastern Europe a major factor for birth rate decline is that after the collapse of the world socialist system introduced the change into the socioeconomic order. Earlier, the Central and Eastern European socialist countries belonged to the group and now they belong to the countries which economies are in transition (from a socialist to a market). Under socialism there were many conditions that had a positive effect on the birth rate: the availability of jobs, better housing, benefits for children. As the result of transformation of economical system during early 90-ies of XX century — the beginning of XXI century, countries of Central and Eastern

Table 4.6: Clusters for five demography indicators and partition into 4 groups

Cluster number	Countries
Cluster Nr.1	Austria, Germany, Finland, Greece, Italy, Portugal, Bulgaria, Bosnia and Herzegovina, Hungary, Latvia, Lithuania, Macedonia, Poland, Romania, Serbia, Slovakia, Slovenia, Croatia, Montenegro, Czech Republic, Estonia
Cluster Nr.2	Belgium, United Kingdom, Luxembourg, Netherlands, France, Switzerland, Denmark, Norway, Sweden, Andorra, Spain, Cyprus, Malta, Monaco, San Marino
Cluster Nr.3	Ireland, Albania, Iceland
Cluster Nr.4	Liechtenstein

Europe have lost their social security, social assistance system was not yet so highly developed as in Western and Northern Europe, the people had sought employment in the difficult social and economic conditions. All this has contributed little to increase the birth rate [36].

Table 4.7: Clusters for five demography indicators and partition into 2 groups

Cluster number	Countries
Cluster Nr.1	Austria, Germany, Finland, Greece, Italy, Portugal, Bulgaria, Bosnia and Herzegovina, Hungary, Latvia, Lithuania, Macedonia, Poland, Romania, Serbia, Slovakia, Slovenia, Croatia, Montenegro, Czech Republic, Estonia, Belgium, United Kingdom, Luxembourg, Netherlands, France, Switzerland, Denmark, Norway, Sweden, Andorra, Spain, Cyprus, Malta, Monaco, San Marino
Cluster Nr.2	Ireland, Albania, Iceland, Liechtenstein

In the Table 4.7 the second variant of European countries grouping by five indicators is presented. It is virtually identical to the previous result (Table 4.6): The first and the second clusters are merged into one cluster – the first, the third and the fourth clusters are combined in the second cluster. The first cluster is characterized by the average for the European Region birth rate and mortality, low positive, zero or negative values of natural population and migration growth and relatively high values of the proportion of the urban population. In the second cluster there are countries with a positive natural population growth and negative or low values of the migration population growth. The second cluster is composed of small countries (by area), including those with an moderate level of economic development.

Clustering in terms of birth rate and the coefficient of migration growth allows for grouping of the countries based on the migration gain, and one of the components of natural population increase – birth in Table 4.8, without taking into account mortality rates and other demographic indicators. Clustering result is different from the two previous clustering outcomes. Luxembourg, Cyprus and Monaco – small states (micro-states) of Europe – constitute the third cluster. They have average European migration growth values and a high birth rate compared to other European countries. The first cluster includes the bulk of the studied European countries with average European birth rates and relatively low or negative growth in migration. This demographic situation can be described as stagnant. The second cluster includes only two countries with a relatively high birth rate, but low negative migration increase, so the overall population growth is negative or low positive.

Table 4.8: Clusters for two demography indicators and partition into three groups

Cluster number	Countries
Cluster Nr.1	Austria, Germany, Finland, Greece, Italy, Portugal, Bulgaria, Bosnia and Herzegovina, Hungary, Latvia, Lithuania, Macedonia, Poland, Romania, Serbia, Slovakia, Slovenia, Croatia, Montenegro, Czech Republic, Estonia, Belgium, United Kingdom, Luxembourg, Netherlands, France, Switzerland, Denmark, Norway, Sweden, Andorra, Spain, Malta, San Marin, Iceland
Cluster Nr.2	Ireland, Albania
Cluster Nr.3	Cyprus, Monaco, Liechtenstein

Clustering carried out by different groups of indicators leads to the conclusion that the European countries with higher levels of economic development forms cluster of countries with more stable demographic situation, in which a higher proportion of the urban population is observed and fluctuations in the values of different indicators is small. Most of the countries of Europe make up, as a rule, large clusters, which are the countries with stagnant, but more favorable demographic situation among the rest of the Europe. Small countries with a moderate level of economic development form small clusters.

Thus, it is possible to identify the following demographic trends as result of European countries clustering. Major countries in Western and Northern Europe with a high level of economic development are often grouped into one cluster. The small countries of Western Europe, the countries of Central and Eastern and Southern Europe, as a rule, are grouped into another cluster. Less economically developed European countries, moderate level Western European countries and the small countries of the Western and Northern Europe are grouped in a separate cluster.

Conclusion

In this dissertation we have investigated new approaches to clustering number determination under clustering stability concept. The stability concept was based on assumption of data inaccuracies which gives us the solution to the data inaccuracies problem itself as well as clustering number estimation. The existed and additional notation is developed.

The two different algorithms which use the same assumptions but different data inaccuracies interpretation are proposed. The algorithms are tested on artificial dataset and result of algorithms work is compared with expected result. The algorithms parameters adjustment have been demonstrated and parameters interpretation is given in Section 3.

The two experiments on real data are carried out. The stable clustering under proposed criterion of the European countries unemployment rates found and historical and geographical result interpretation are provided. The cluster analysis of European Union countries demography indices are carried out.

Bibliography

[1] Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters via the gap statistic. *J. Royal Statist. Soc. B*, Vol. 63(2), 411–423.

[2] Ложкинс А., Буре В.М. (2016). Вероятностный подход к определению локально-оптимального числа кластеров // Вестник Санкт-Петербургского университета, сер. 10. Прикладная математика, информатика, процессы управления, вып. 1, С. 28–38.

[3] Ложкинс А. (2015). Кластерный анализ стран Европы по уровню безработицы // Труды XLVI международной научной конференции аспирантов и студентов Процессы управления и устойчивость. СПб.: Издательский Дом Федоровой Г. В., Том 2(18), № 1, С. 641–646.

[4] Lozkins A., Bure V. M. (2015). The method of clusters stability assessing //" Stability and Control Processes" in Memory of VI Zubov (SCP), 2015 International Conference. IEEE, pp. 479–482.

[5] Ложкинс А. (2015). Кластерный анализ стран Европы по уровню безработицы // Сборник тезисов конференции "Процессы управления и устойчивость" 2015.

[6] Ложкинс А., Буре М. В. (2015). Эмпирический подход оценки устойчивости методов кластеризации // Материалы III международной конференции, посвященной 85-летию со дня рождения профессора, чл.-корр. РАН В. И. Зубова Устойчивость и процессы управления. СПб.: Издательский Дом Федоровой Г. В., С. 431–433.

[7] Lozkins A. (2015). Clustering of European Countries by an Inflation Rate and Clusters Research // Abstracts of 20th International Conference of Mathematical Modelling and Analysis, pp. 56.

[8] Lozkins A., Bure V. M. (2016). The approach for estimation of clustering robustness // 12th German Probability and Statistics Days Book of Abstracts. pp. 197–198.

[9] Fowlkes, E. B., Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, Vol: 78, pp. 553–584.

[10] Jaccard P. (1901). Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques regions voisines. Rouge.

[11] Jain, A., Dubes, A. (1988). Algorithms for Clustering Data. *Englewood Cliffs, Prentice-Hall*, New Jersey.

[12] Ben-Hur, A., Elisseeff, A., Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Proceedings of Pacific Symposium on Biocomputing*, pp. 6–17.

[13] Lange T. et al. (2004). Stability-based validation of clustering solutions. *Neural computation*, Vol: 16, № 6, pp. 1299–1323.

[14] Dunn, J. C. (1974). Well Separated Clusters and Optimal Fuzzy Partitions. *Journal Cybern.*, Vol. 4, 95–104.

[15] Calinski, R., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, Vol. 3, 1–27.

[16] Hartigan, J. A. (1985). Statistical theory in clustering. *J. Classification*, Vol. 2, 63–76.

[17] Krzanowski, W., Lai, Y. (1985) A criterion for determining the number of groups in a dataset using sum of squares clustering. *Biometrics*, Vol. 44, 23–34.

[18] Sugar, C., James, G. (2003). Finding the Number of Clusters in a Data Set: An Information Theoretic Approach. *Journal of the American Statistical Association*, Vol. 98, 750–763.

[19] Gordon, A. D. (1994). Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, Vol. 18, 561–581.

[20] Milligan, G., Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, Vol. 50, 159–179.

[21] Hubert, L., Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. *British J. Math. Statist. Psychology*, Vol. 29, 190–241.

[22] Wishart, D. (1969). Mode analysis: A generalization of nearest neighbor which reduces chaining effects. *Numerical Taxonomy*, 282–311.

[23] Hartigan, J. (1975). Clustering Algorithms. New York: John Wiley.

[24] Cheng, R., Milligan, G.W. (1996). Measuring the influence of individual data points in a cluster analysis. *Journal of Classification*, Vol. 13, 315–335.

[25] Levine, E., Domany, E. (2001). Resampling Method for Unsupervised Estimation of Cluster Validity. *Neural Computation*, Vol. 13, 2573–2593.

[26] Ben-Hur, A., Elisseeff, A., Guyon, I. (1998). Statistical learning Theory and randomized algorithms for control. *IEEE Control Systems*, Vol. 12, 69–85.

[27] Ben-Hur, A., Guyon, I. (2003). Methods in Molecular Biology / M. J. Brownstein and A. Khodursky (Ed.). *Humana Press*, 159–182.

[28] Pelleg, D., Moore, A. (2000). X means Extending k-means with efficient estimation of the number of clusters. *In: Proceedings of the 17-th International Conf. on Machine Learning*, San Francisco: Morgan Kaufmann, 727–734.

[29] Volkovich, Z., Brazly, Z., Toledano-Kitai, D., Avros, R. (2010). The Hotelling's metric as a cluster stability measure. *Computer modelling and new technologies*, Vol. 14 (4), 65–72.

[30] Volkovich, Z., Brazly, Z., Morozensky, L. (2008). A Statistical model of cluster stability. *Pattern Recognition*, Vol. 41 (7), 2174–2188.

[31] Barzily, Z., Golani M. and Volkovich Z. (2008). On a simulation approach to cluster stability validation. *Special Issue, Mathematical and Computer Modeling in Applied Problems*, Institute Informatics Problems, RAS, 86–112.

[32] Toledano-Kitai, D., Avros, R., Volkovich, Z., Weber, G.- W. and Yahalom, O. (2013). A binomial noised model for cluster validation. *Journal of Intelligent and Fuzzy Systems, Special, Issue: Recent Advances in Intelligent & Fuzzy Systems*, 417–427.

[33] Буре В.М., Парилина Е.М. Теория вероятностей и математическая статистика. СПб: Лань, 416 с.

[34] Trading Economics [e–resource]: URL: http://tradingeconomics.com/country-list/unemployment-rate (date of treatment: 28.01.2015).

[35] Старкова Н.В. (2015). Экономическая и социальная география стран Европы: Учебно-методическое пособие. СПб.: «СОЛО», 119 с.

[36] Клупт М. (2008). Демография регионов Земли. СПб.: Питер, 347 с.: ил.

Appendix

Appendix 1: The unemployment rates of European countries for 2014.

Country	First quarter	Second quarter	Third quarter
Albania	13,86	13,45	12,91
Armenia	17,80	17,50	17,10
Austria	8,90	7,73	9,00
Belgium	8,50	8,50	8,60
Bosnia and Herzegovina	44,18	43,86	43,66
Bulgaria	13,00	11,40	10,80
Croatia	22,47	19,67	17,67
Cyprus	15,87	16,07	16,27
Czech Republic	8,46	7,60	7,36
Denmark	4,17	3,97	4,00
Estonia	8,50	6,90	7,50
Finland	9,03	9,63	7,53
France	10,10	10,10	10,40
Germany	5,07	5,00	5,00
Greece	27,17	26,83	26,03
Hungary	8,60	8,03	7,63
Iceland	5,43	5,03	4,90
Ireland	11,85	11,70	11,20
Italy	12,67	12,53	12,87
Kazakhstan	5,10	5,03	5,07
Latvia	11,90	10,70	10,60
Lithuania	12,40	11,20	9,10
Luxembourg	7,10	7,23	7,23
Macedonia	28,39	28,22	27,90
Malta	6,00	5,80	5,84

Country	First quarter	Second quarter	Third quarter
Moldova	5,10	3,60	3,30
Montenegro	14,94	14,11	13,48
Netherlands	8,70	8,57	8,07
Norway	2,77	3,27	3,60
Poland	13,70	12,50	11,67
Portugal	15,10	13,90	13,10
Romania	7,00	6,93	6,77
Serbia	20,80	20,30	17,60
Russia	5,53	5,03	4,87
Slovakia	13,46	12,81	12,57
Slovenia	14,10	13,07	12,50
Spain	25,93	24,47	23,67
Sweden	8,55	8,63	7,23
Switzerland	3,40	3,03	2,97
Turkey	10,00	8,97	10,13
Ukrain	9,40	8,60	8,90
UK	6,97	6,50	6,07

Appendix 2: The demography indicators of European Union countries.

Countries	Birth rate	Mortality rate	Natural population growth rate	Migration of population growth rate	Urban population
Austria	9	9	0	5	67
Albania	13	7	6	-15	54
andorra	9	4	5	4	90
Belgium	12	9	3	6	99
Bulgaria	9	15	-6	-1	73
Bosnia and Herzegovina	8	9	-1	0	46
UK	13	9	4	2	80
Hungary	9	13	-4	1	69
Germany	8	11	-3	5	73
Greece	9	10	-1	-1	73
Denmark	10	9	1	4	87
Ireland	16	6	10	-7	60
Iceland	14	6	8	-1	95
Spain	10	8	2	-3	77
Italy	9	10	-1	4	68
Cyprus	12	7	5	14	62
Latvia	10	14	-4	-2	68
Lithuania	10	14	-4	-7	67
Liechtenstein	10	6	4	3	15
Luxembourg	11	7	4	19	83
Macedonia	11	10	1	0	65
Malta	10	8	2	3	100
Monaco	6	6	0	13	100
Netherlands	10	8	2	1	66
Norway	12	8	4	9	80
Poland	10	10	0	0	61
Portugal	9	10	-1	-4	38

Countries	Birth rate	Mortality rate	Natural population growth rate	Migration of population growth rate	Urban population
Romania	9	12	-3	0	55
San Marino	9	7	2	7	84
Serbia	9	14	-5	1	59
Slovakia	10	10	0	1	54
Slovenia	11	9	2	0	50
Finland	11	10	1	3	68
France	13	9	4	1	78
Croatia	10	12	-2	-1	56
Montenegro	12	10	2	0	64
Czech Republic	10	10	0	1	74
Switzerland	10	8	2	8	74
Sweden	12	10	2	5	84
Estonia	11	12	-1	-5	69

Appendix 3: The artificial data generation program code in R.

```
#Open the library
library("mvtnorm", lib.loc="~/R/win-library/3.1")

# Covariance matrix
varcov = matrix(c(0.15, 0, 0, 0.15), 2)

#Generation of four two-dimensional samples from
#the normal distribution
y11 = rmvnorm(100, mean=c(1,1), varcov)
y21 = rmvnorm(100, mean=c(1,4), varcov)
y31 = rmvnorm(100, mean=c(3.5,2.5), matrix(c(0.2, 0, 0, 0.2), 2))
y41 = rmvnorm(100, mean=c(6,1), varcov)
y51 = rmvnorm(100, mean=c(6,4), varcov)

#Union of samples
y1 = c(y11[,1], y21[,1], y31[,1], y41[,1], y51[,1])
y2 = c(y11[,2], y21[,2], y31[,2], y41[,2], y51[,2])

#Making the dataframe
y = data.frame(y1, y2)

#Plot the dataset
plot(y$y1, y$y2, xlab = "x", ylab = "y",
main = "Artificial data set")
```

Appendix 4: The algorithm described in Section 2.2 implementation in R.

```
#Initial clustering
vectOfInitClust = function (numClust, dataSet) {
  kmeanInit = kmeans(dataSet, centers = numClust)
  return (kmeanInit)
}


#Frequencies generation for uniformly distributed noises
kDataClustUnif = function(numClust, lenghts = 500, nRepeat = 50,
dataSet = y, variance, kmeanInit){
  MatrixOfDistance = dist(dataSet, method = "euclidean",
   diag = FALSE)
  hClust1 = hclust(MatrixOfDistance)
  initClust = cutree(hClust1, k = numClust)
  k = 0
  for (i in 1:nRepeat){
    newDataAdding1 = c(dataSet[,1], dataSet[,1] +
     runif(lenghts,min = -delta, max = delta))
    newDataAdding2 = c(dataSet[,2], dataSet[,2] +
     runif(lenghts,min = -delta, max = delta))
    newDataAdding = data.frame(newDataAdding1, newDataAdding2)

    Distance = dist(newDataAdding, method = "euclidean",
             diag = FALSE)
    TreeGeneration = hclust(Distance)
    vect = cutree(TreeGeneration, k = numClust)
    if (mean(abs(initClust - vect[1:lenghts])) != 0)
    {
      k = k + 1
    }
  }
  return(k1/nRepeat)
}

#Set of Frequences for fixed number of Clusters (for uniform)
frequnceUnifSet = function(nClust , variances, nRepeat = 50,
dataSet = y){
  kmeanInit = kmeans(dataSet, centers = nClust)
  i = 0;
  frequences = 0;
  for (sigma in variances){
    i = i + 1
    frequences[i] = kDataClustUnif(numClust = nClust,
     dataSet = dataSet,
     variance = sigma,
     nRepeat = nRepeat,
     kmeanInit = kmeanInit)
  }
  return (frequences)
}

# Iteration through the number of clusters (median)
frequenceNumClustUnifMedian = function(setNumClust = 2:10,
    nRepeat = 50, varSet, dataSet = y){
  i = 0
  medFrequence = 0
  for (k in setNumClust){
    i = i + 1
    medFrequence[i] = median(frequnceUnifSet(nClust = k,
    variances = varSet, nRepeat = nRepeat, dataSet = dataSet))
  }
  return (medFrequence)
}
```