KANTOROVICH-RUBINSTEIN DISTANCE MINIMIZATION: APPLICATION TO LOCATION PROBLEMS

Viktor Kuzmenko,[*] Stan Uryasev[†]

Current Version: August 28, 2018

Correspondence should be addressed to: Stan Uryasev

Abstract

The paper considers optimization algorithms for location planning, which specifies positions of facilities providing demanded services. Examples of facilities include hospitals, restaurants, ambulances, retail and grocery stores, schools, and fire stations. We reduced the initial problem to approximation of a discrete distribution with a large number of atoms by some other discrete distribution with a smaller number of atoms. The approximation is done by minimizing the Kantorovich-Rubinstein distance between distributions. Positions and probabilities of atoms of the approximating distribution are optimized. The algorithm solves a sequence of optimization problems reducing the distance between distributions. We conducted a case study using Portfolio Safeguard (PSG) optimization package in MATLAB environment.

Keywords: location planning; distribution; Kantorovich-Rubinstein distance; optimization; Portfolio Safeguard, PSG

1 Introduction

An optimal location of facilities is an important part of supply chain planning. Various approaches are available for this problem, see for instance [9]. The location of facilities defines the chain structure and flow of goods through a chain. The location depends on a type of goods shipped in the chain [17]. If demands of customers are fixed, then the location problem can be reduced to an approximation of one discrete probability distribution with a large number of atoms (customers) by some other discrete distribution with a smaller number of atoms (facilities).

[*]Glushkov Institute of Cybernetics, Kyiv, Ukraine.
[†]University of Florida, USA. Email: uryasev@ufl.edu.

An approximation of a distribution by some other distribution is a popular topic in academic literate, see for instance [19]. Various metrics measuring distance between distributions are used in probability and risk theory, including the Kolmogorov, Levy, L_p metrics (see Rachev et al [20]) and other metrics [4]. A new CVaR distance between distributions is used in Pavlikov and Uryasev [18].

This paper considers Kantorovich-Rubinstein metric [12]. It is closely related with a continuous transportation problem on a compact metric space, which was first formulated by Kantorovich [10] in 1942 and further developed in [13]. Kantorovich associated this problem with the excavation and embankment Monge problem, which is a transportation problem on a Euclidean plane [11]. The history of the Kantorovich-Rubinstein metric is discussed in Vershik [21]. A brief survey of recent applications of the Kantorovich-Rubinstein metric in computer scienceis is in [6].

On the other hand, the Kantorovich-Rubinstein metric for finite discrete distributions is related to the k-clustering problem, see [15]. A distance function is important for defining a problem and for solution methods [22]. Depending on the distance function and usage, the k-clustering problem is called k-median, k-medoids, k-means, k-center problem, see for instance [7]. The k-clustering problem is NP-hard, therefore various heuristics were developed, see [16] for the k-median problem.

Approximate algorithms and guaranteed estimates of solutions are based on metric in a point space. This paper minimizes the weighted sum of distances between points and centers - every distance is multiplied by a probability of a point. Such objective does not allow for direct using of the triangle inequality. Recent papers consider approximate algorithms when the triangle inequality for a metric does not hold [2]. Also, several papers considered k-median clustering for a weighted set of points [3] or weighted distance [14], which is similar to the problem in this paper.

The considered problem can be classified as a k-median problem with a weighted distance and continues positions for k centers. This classification reveals relation between clustering problem, location planning, and approximation of a distribution. We used a distribution approximation approach for the location optimization in supply chain planning.

This paper proposes an algorithm for an approximate solving of an un-capacitated location problem when distance function is defined in multidimensional space by l_p-norm and location of objects is continues. In this case, the problem may be normalized by dividing transportation variables by a sum of consumptions and can be reduced to the minimization of Kantorovich-Rubinstein distance.

Below we define the Kantorovich-Rubinstein metric for a discrete finite distribution and formulate the problem of approximation of one multivariate distribution by another one. We propose an approximation algorithm minimizing the Kantorovich-Rubinstein metric. We solve several test problems for the Euclidean distance and placed the case study codes, data and calculation results to web (see section 5).

2 Kantorovich-Rubinstein metric

Further we define the Kantorovich-Rubinstein metric for two finite multidimensional discrete distributions. Let X, Y be discrete random vectors with atoms $\{x_1, ..., x_n\}$, $\{y_1, ..., y_m\}$ and corresponding probabilities $\{p_1, ..., p_n\}$, $\{q_1, ..., q_m\}$. Let c_{ij} be a nonnegative distance between points x_i and y_j. The Kantorovich-Rubinstein metric between distributions is defined as an optimal value of the following transportation problem,

$$D(X,Y) = \min_{w_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} w_{ij} \tag{1}$$

subject to

$$\sum_{j=1}^{m} w_{ij} = p_i, \ i = 1, ..., n, \tag{2}$$

$$\sum_{i=1}^{n} w_{ij} = q_j, \ j = 1, ..., m, \tag{3}$$

$$w_{ij} \geq 0, i = 1, ..., n, j = 1, ..., m. \tag{4}$$

According to the location terminology, problem (1-4) corresponds to the transportation part of a location-allocation problem. Values p_i and q_j are normalized supplies and demands, and variables w_{ij} are normalized transportation volumes.

Objective (1) is the total transportation cost in a normalized transportation model.

Constraints (2), (3) describe a balance between supply and consumption. These constraints imply

$$\sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_j \,.$$

Therefore, the system of linear equations (2), (3) is consistent but it is not independent. So, one equation can be removed from this system or one system of equalities (2) or (3) can be replaced by inequalities, for example,

$$\sum_{j=1}^{m} w_{ij} \leq p_i, \ i = 1, ..., n \,.$$

These properties provide some flexibility in formulation of the transportation problem.

3 Approximation of a Multivariate Distribution

Let us consider two k-dimensional multivariate probability distributions. We suppose that a random vector $\vec{Y} \in R^k$ is discretely distributed with fixed atoms $\vec{y}_1, ..., \vec{y}_m$ and probabilities $q_1, ..., q_m$. Every atom j is a k-dimension vector $\vec{y}_j = (y_{j1}, ..., y_{jk})$. We want to approximate this random vector $\vec{Y}$ by some other discrete random vector $\vec{X}$ with atoms $\vec{x}_1, \ldots, \vec{x}_n$, and probabilities $p_1, \ldots, p_n$, where $n < m$. Every atom i is defined by a vector $\vec{x}_i = (x_{i1}, ..., x_{ik})$. Coordinates, x_{il}, and probabilities, p_i, are variables of the approximation problem. Positions of atoms of the random variable $\vec{X}$ with the smallest Kantorovich-Rubinstein distance to $\vec{Y}$ can be found by solving the following optimization problem,

$$\min_{\vec{x}_i, w_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} dist(\vec{x}_i, \vec{y}_j) w_{ij} \tag{5}$$

subject to

$$\sum_{i=1}^{n} w_{ij} = q_j,\ j = 1, ..., m, \tag{6}$$

$$w_{ij} \geq 0, i = 1, ..., n, j = 1, ..., m. \tag{7}$$

With the location terminology, this problem finds locations of points $\vec{x}_i$ in k-dimensional space and transportation volumes w_{ij} to consumers with demands q_j. This is an uncapacitated allocation problem in k-dimensional space without fixed costs of location.

Supply volumes are defined after solving the problem (5-7). Let $\vec{x}_i^*, w_{ij}^*, i = 1, ..., n, j = 1, ..., m$, be an optimal solution vector of the problem (5-7). For the optimal locations $\vec{x}_i^*$, the optimal probabilities are equal to $p_i^* = \sum_{j=1}^{m} w_{ij}^*, i = 1, ..., n$.

Usually, l_p norm is used to define the distance between points in function (5),

$$dist(\vec{x}_i, \vec{y}_j) = \sqrt[p]{\sum_{l=1}^{k} |x_{il} - y_{jl}|^p}, \quad p \geq 1.$$

The case p=2 corresponds to the Euclidian norm.

4 Algorithm for Distribution Approximation

This section describes the algorithm for finding an approximating random vector $\vec{X}$ for the random vector $\vec{Y}$, as described in the previous section. The algorithm includes two stages. Stage 1 finds an initial approximation for the following Stage 2.

Stage 1. We consider the problem (5-7) with the assumption that points $\vec{x}_i$ are located only at points $\vec{y}_j$. This problem is further formulated as a mixed-integer optimization problem with a cardinality constraint (the number of positive p_i in an optimal solution should not exceed n):

$$\min_{\delta_i, w_{ij}} \sum_{i=1}^{m} \sum_{j=1}^{m} dist(\vec{x}_i, \vec{y}_j) w_{ij} \tag{8}$$

subject to

$$\sum_{i=1}^{m} w_{ij} = q_j,\ j = 1, ..., m, \tag{9}$$

$$\sum_{j=1}^{m} w_{ij} \le \delta_i,\ i = 1, ..., m, \tag{10}$$

$$\sum_{j=1}^{m} \delta_i \le n, \tag{11}$$

$$w_{ij} \ge 0,\ j = 1, ..., m, \quad \delta_i \in \{0, 1\},\ i = 1, ..., m. \tag{12}$$

Problem (8-12) provides an initial approximation for the problem (5-7). We consider only n vectors $\vec{x}_i$ with $\delta_i = 1$ and start with a feasible solution of the problem (5-7). δ_i are technical Boolean variables for formulating cardinality constraint (11) on the number of positive p_i.

Stage 2. Further, problem (5-7) is approximately solved with a sequence of pairs of optimization problems. Solving of every pair of sub-problems decreases the distance between the fixed and the approximating distributions. We stop iterating pairs of sub-problems when the distance stops decreasing. Further we formulate these two sub-problems.

Optimization Problem 1

$$\min_{\vec{x}_i} \sum_{i=1}^{n} \sum_{j=1}^{m} dist(\vec{x}_i, \vec{y}_j) w_{ij}\ . \tag{13}$$

Optimization Problem 2

$$\min_{w_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} dist(\vec{x}_i, \vec{y}_j) w_{ij} \tag{14}$$

subject to

$$\sum_{i=1}^{n} w_{ij} = q_j,\ j = 1, ..., m\ , \tag{15}$$

$$w_{ij} \ge 0,\ i = 1, ..., n,\ j = 1, ..., m\ . \tag{16}$$

The first sub-problem (13) changes positions of n vectors $\vec{x}_i$ while keeping the fixed values w_{ij}. This problem is known as a Weber problem [8]. The second problem (14-16) changes values w_{ij} for the obtained fixed positions $\vec{x}_i$.

Solving pairs of these sub-problems in cycle monotonically decreases objective function of initial problem (5-7): 1) both sub-problems have the same objective; 2) the distance is decreased by every solved sub-problem; 3) the solution point of one sub-problem is used as an initial point of the following sub-problem.

The described optimization process stops after a finite number of steps. This follows from the fact that an optimal vector w_{ij} of the problem (14-16) has the following property: for every $j = 1, ..., m$ there exist an index i_j such that $w_{i_j j} = q_j$ and $w_{ij} = 0$ for $i \neq i_j$. Because the number of such combinations of values w_{ij} is finite and process monotonically decreases the objective function, the optimization process will stop after a final number of steps.

5 Case Study

Data, codes, and solutions for this case study are posted at this link[1]. We solved several test problems with different space dimension k. This case study reports calculations results with $k = 2$. For this two-dimensional case we used the Euclidian norm,

$$dist(\vec{x}_i, \vec{y}_j) = \sqrt{(x_{i1} - y_{j1})^2 + (x_{i2} - y_{j2})^2}. \tag{17}$$

Further we present one solved problem. We select the number of fixed and the number of approximating atoms and simulate coordinates of the fixed atoms.

In the considered example, the vector $\vec{Y}$ has 100 equally probable atoms. The approximating vector $\vec{X}$ has 10 atoms. Table 1 presents coordinates of approximating atoms and sum of probabilities of "attached" fixed atoms. Approximating atoms in the table are ordered by vertical coordinate of the points. A special feature of this example is that the fixed points are not "normally" distributed and exhibit some "heavy tail" behavior. Figure 1 graphically presents the dataset and solution. Bold black criss-crosses show atoms of the approximating random vector $\vec{X}$. Atoms of $\vec{Y}$ "attached" to a criss-cross have the same color. Axes of the graph are rescaled to improve clarity of the image. The highest green point in the graph coincides with an individual approximating point. Table 1 shows that the highest in the graph approximating point with ID=1 has "attached" probability 0.01, i.e., only one point of the fixed distribution is "attached" ("transported") to this ID=1 point. The approximating point with ID=2 approximates three "red" points and with ID=3 three "pink" points, see Table 1 and Figure 1.

The case study is done in MATLAB environment. The MATLAB code reads initial data, obtains an initial approximating solution, organizes cycles, prepares and modifies data for Optimization Problems 1 and 2. `kmeans` MATLAB standard function is used to get an initial approximating solution. This function finds centroid seeds for k-means clustering [5]. Optimization Problem 1, see (13), is modified on every step and is solved by Portfolio Safeguar (PSG) optimization package [1] called from the MATLAB code. PSG has different precoded functions, including the Square Root Quadratic function

[1] http://www.ise.ufl.edu/uryasev/research/testproblems/advanced-statistics/approximation-of-a-discrete-distribution-by-some-other-discrete-distribution-in-euclidean-space-by-minimizing-k-r-distance/

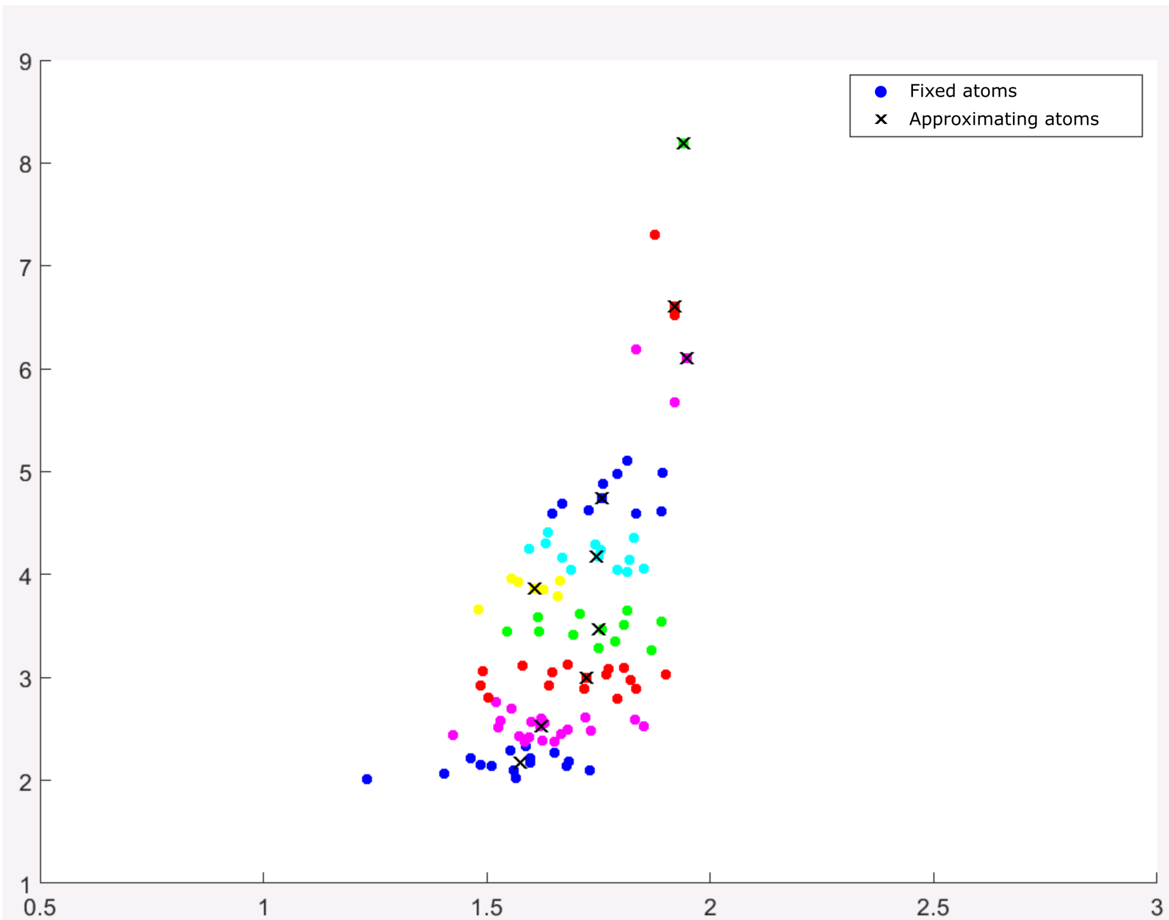

Figure 1: The random vector $\vec{Y}$ with 100 equally probable atoms (colored dots) is approximated by the random vector $\vec{X}$ with 10 atoms (crosses).

Table 1: Optimal Probabilities and Positions of Approximating Atoms

Approximating Atom ID	Attached Probability	Horizontal Coordinate	Vertical Coordinate
1	0.01	1.939	8.188
2	0.03	1.921	6.602
3	0.03	1.948	6.105
4	0.10	1.756	4.740
5	0.13	1.745	4.179
6	0.07	1.606	3.865
7	0.12	1.751	3.467
8	0.16	1.723	2.995
9	0.20	1.621	2.526
10	0.15	1.575	2.171

called `sqrt_quadratic`. For every pair i, j, the `sqrt_quadratic` function calculates distance (17). Further we provide beginning and ending parts of Problem Statement generated by MATLAB code for one instance of problem (13) for solving this problem by PSG. MATLAB code generates such Problem Statement and matrices `matrix_i_j` using solution of problem (14-16) and packs them into input structures for PSG. Here we present one instance of the problem in Text (RunFile Environment). So it is possible to solve just this individual problem without understanding how MATLAB code is organized. This problem with data is placed separately at the case study webpage (in addition to the MATLAB code, which generates and solves many similar problems).

```
   minimize
0.01*sqrt_quadratic(matrix_1_10)
+0.01*sqrt_quadratic(matrix_1_15)
+0.01*sqrt_quadratic(matrix_2_1)
+0.01*sqrt_quadratic(matrix_2_26)
+0.01*sqrt_quadratic(matrix_3_3)
+0.01*sqrt_quadratic(matrix_3_29)
+0.01*sqrt_quadratic(matrix_4_2)
+0.01*sqrt_quadratic(matrix_4_9)
+0.01*sqrt_quadratic(matrix_5_8)
.............................
+0.01*sqrt_quadratic(matrix_10_100)
Solver:  van, init_point = point_initial_x
```

This Problem Statement minimizes the sum of the norms (17) with coefficients w_{ij} obtained after solving problem (14-16). Every row in the Problem Statement contains a `sqrt_quadratic` function, which calculates the norm with a non zero coefficient, w_{ij}. In this Problem Statement $w_{ij} = 0.01$ for included functions (terms with $w_{ij} = 0$

are skipped from the sum). Every `sqrt_quadratic` function is defined by the matrix `matrix_i_j`. The first row of the matrix `matrix_i_j` contains names of the optimization variables x_{i1}, x_{i2}. The vector $\vec{y}_j = \{y_{i1}, y_{i2}\}$, which is the position of atom j, defines the elements of the matrix. Variables of the optimization problem are not listed separately in the PSG code. Every function has a set of variables and the solver assembles the optimization problem using the functions included in the analytic problem statement. `Van` is one of PSG optimization solvers. The vector `point_initial_x` contains an initial point (which is an optimal point of the previously solved problem (13) or an initial approximating point).

Optimization Problem 2 (14)-(16)) is solved with standard MATLAB capabilities: for every $j = 1, ..., m$, a nearest i_j is found with formula: $i_j = argmin_{s=1,...,n} dist(\vec{x}_s, \vec{y}_j)$. Then, Problem 2 solution vector equals: $w_{i_j j} = q_j$ and $w_{ij} = 0$ for $i \neq i_j$, $j = 1, ..., m$.

6 Conclusion

The proposed new approach for the location problem finds coordinates and capacities of "servers" in a multidimensional space with distance function defined by metric in this space. We considered an uncapacitated location problem when transportation cost is the main factor for formulation of the objective function. The aim of this paper to show that numerical approaches from probability theory can be used for some types of location and supply chain problems. The proposed algorithm includes two parts: 1) initial approximation; 2) two-step cycling procedure for finding a local minimum of the optimization problem. We implemented and tested the algorithm with several problem instances. One instance of the problem (codes, data, and calculation results) are posted at web and are available for benchmarking. Further research in this area will be focused on solving capacitated location problems and on theoretical precision bounds.

References

[1] American Optimal Decisions (AORDA), *Portfolio Safeguard (PSG)*, http://www.aorda.com

[2] Ahmadian S., Norouzi-Fard A., Svensson O., Ward, J. (2017) *Better Guarantees for k-Means and Euclidean k-Median by Primal-Dual Algorithms*, 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), DOI: 10.1109/FOCS.2017.15.

[3] Chen K. (2006) *On k-Median Clustering in High Dimensions*, Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 1177–1185.

[4] Chung J.K., Kannappan P.L., Ng C.T., Sahoo P.K. (1989) *Measures of Distance between Probability Distributions*, Journal of mathematical analysis and applications, Vol. 138, pp. 280–292.

[5] David A., Vassilvitskii S. (2007) *K-means++: The Advantages of Careful Seeding*, Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 07, 1027–1035.

[6] Deng Y., Du W. (2009) *The Kantorovich Metric in Computer Science: A Brief Survey*, Electronic Notes in Theoretical Computer Science, Elsevier, Vol. 253, Iss. 3, 73–82. DOI: 10.1016/j.entcs.2009.10.006.

[7] Fayed A., Atiya A. (2013) *A Mixed Breadth-depth First Strategy for the Branch and Bound Tree of Euclidean k-Center Problems*, Computational Optimization and Applications, Springer, Vol. 30, Iss. 2, DOI: 10.1007/s10589-012-9503-x.

[8] Fernandes I.F., Aloise D., Aloise D.J., Hansen P., Liberti L. (2014) *On the Weber facility location problem with limited distances and side constraints*, Optimization Letters, Springer, Vol. 8, Iss. 2, 407–424. DOI: 10.1007/s11590-012-0538-9

[9] Geunes J., Pardalos P.M. (Editors) (2005) *Supply Chain Optimization. Applied optimization*, Springer, Vol. 98, 414.

[10] Kantorovich L.V. (1942) *On the Translocation of Masses*, Dokl. Akad. Nauk SSSR, Vol. 37, N-s. 7–8, 227–229.

[11] Kantorovich L.V. (1948) *On a Problem of Monge*, Uspekhi Mat. Nauk, Vol. 3, No. 2, 225–226.

[12] Kantorovich L.V., Rubinstein G.Sh. (1958) *On a Space of Totally Additive Functions*, Vestn. Lening. Univ., Vol. 13, No. 7, 52–59.

[13] Kantorovich L. (1958) *On the Translocation of Masses*, Management Science, Vol. 5, No. 1, 1–4.

[14] Kumar A. (2016) *Capacitated k-Center Problem with Vertex Weights*, 36th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2016), No. 8, 1–14.

[15] Lattanzi S., Vassilvitskii S. (2017) *Consistent k-Clustering*, Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70.

[16] Li S., Svensson O. (2016) *Approximating k-Median via Pseudo-approximation*, SIAM J. Comput., 45(2), 530–547.

[17] Meloa M.T., Nickelab S., Saldanha da Gamac F. (2006) *Dynamic Multi-commodity Capacitated Facility Location: a Mathematical Modeling Framework for Strategic Supply Chain Planning*, Computers and Operations Research, Vol. 33, Iss. 1, 181–208.

[18] Pavlikov K., Uryasev S. (2018) *CVaR Distance Between Univariate Probability Distributions and Approximation Problems*, Annals of Operations Research, Vol. 262, Iss. 1, 67–88. DOI: 10.1007/s10479-017-2732-8

[19] Pflug G.Ch., Pichler A. (2011) *Approximations for Probability Distributions and Stochastic Optimization Problems*, Stochastic Optimization Methods in Finance and Energy, Springer, Vol. 163, 343–387.

[20] Rachev S.T., Stoyanov S.V., Fabozzi F.G. (2011) *A Probability Metrics Approach to Financial Risk Measures*, New York: John Wiley & Sons.

[21] Vershik A.M. (2006) *Kantorovich Metric: Initial History and Little-Known Applications*, Journal of Mathematical Sciences, Springer, Vol. 133, Iss. 4, 1410–1417.

[22] Zarinbal M. (2009) *Distance Functions in Location Problems*, Facility Location. Contributions to Management Science, Springer, 5–17. DOI: 10.1007/978-3-7908-2151-2_1.