UDC 519.868
MSC 90B15

The distribution centres choice in the facility location problem on the basis of statistical modeling procedures

A. Lozkins

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg,
199034, Russian Federation

For citation: Lozkins A. The distribution centres choice in the facility location problem on the basis of statistical modeling procedures. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2018, vol. 14, iss. 4, pp. 346–351. https://doi.org/10.21638/11702/spbu10.2018.407

The problem of a distribution centres network construction based on the statistical data analysis of the LTL transportation company is considered. The distribution centre network is built on the basis of the demand on terminal services. Statistical criterion for selecting the number of distribution centres in the network based on the application of the network robustness principle to the disturbances in demand for services in each terminal is suggested. Demand distortions are proposed to be carried out taking into account the forecasting of future trends in demand. The simulation study on real data is carried out. The considered task consists of a terminal network where the cargo is generated to deliver other terminals. The goal is to estimate the robust number of hubs in the network which minimizes the total flows costs and is resistant to the possible flows changes in the network. The results on the real dataset are illustrated and discussed.

Keywords: hub location problem, statistical decision making, robustness, networks.

Introduction. The facility location problem became important in the 80's of the 20th century. The main drive was the possibility to solve large linear programming (LP) tasks using the LP solvers and computing power growth. The applied mathematics has gone ahead and the facility location problem has got the expansion, where the problem consists of not only location estimation but of facilities number determination under the some criteria.

Facility location problems are widely studied in computer networks, airlines, transportation, postal delivery systems and telecommunications. The warehouse or hub is the node in the network, which tends to be located by the reason of total costs reduction. The problem is to choose the location and the most appropriate number of hubs.

There are many proposals of modeling the hub network, such the uncapacitated single allocation p-hub median problem (USApHMP), which is formulated by O'Kelly [1, 2] and became the first problem formulation. This problem operates the finite set of possible hub locations as it is described in the works [3, 4]. There exists the opposite model for the continuous hub location problem [5], where the hub locations are not determined and the cluster analysis is applied. There are solutions with introduction of the resilience of the network in articles [6, 7]. The capacity constraints addition to the edges of the network generates the new class of the problems called capacitated single allocation p-hub median problem (CSApHMP). This problem is considered in works [8, 9].

The USApHMP and CSApHMP consider the fixed number of facilities to be allocated. In the case when the hub number is not given there appears the problem — what criterion should be used to find the quantity of hubs in the network. The similar problem exists in cluster analysis, where lots of ideas for cluster numbers determination are proposed. The different approaches are developed in [10–12] — nonparametric cluster number determination based on data density consideration, the ideas based on statistical criteria in [13–15], the stability based algorithms are described in [16–19].

The paper proposes the method of hub number estimation using the stability concept of cluster number validation [18] in cluster analysis, bootstrap data resampling and value at risk idea as variability level metric (similarly to variability frequency in [18]). The goal of the approach is to evaluate the risks in different cases of hub numbers in the network and to suggest the most likely and statistically stable number of hubs.

Proposed method. The proposed approach of finding the hubs' number in the network involves several theories: cluster analysis, statistical simulations and risk theory. The combination of these areas allows us to propose a new result for solving the problem.

The cluster analysis is used to introduce the stability concept in the problem, which is based on initial data perturbations for tracing the hub network changes (hub is interpreted as clusters center). The perturbations represent the possible deviations from initial data (for example, the possible flow, costs changes using forecasting models).

The hub location-allocation problem has high complexity (NP-complete), which does not allow to generate a lot of solutions for the analysis. The bootstrap algorithm provides the possibility to reproduce the number of samples for decision making.

The changes in the network for different hub numbers are represented by variety frequencies introduced in [19]. There are calculated the samples mean and standard deviation, which are the parameters for the value at risk function. The value at risk with the least risk is the most appropriate hub number and the most stable network of hubs in [20] concept.

Simulation algorithm description. The facility location discrete model is the input for the algorithm (example: models described in [1, 6, 20]). Let denote the model $G(W,p)$ with initial commodities flow data W — matrix $n\times m$ and hub number equals p. The f_{ij} is the probability distribution, which represents the forecasting of commodity flow possible changes for $\forall i = 1\dots n, j = 1\dots m$. The matrix F is the result of random values generation from distributions f_{ij} for each element of matrix W respectively. The r is number of generations and $p_{\min},\dots,p_{\max}$ is set of considered number of hubs. The r value depends on time and resources possibilities, the larger value entails more informative results. The possible hubs location amount is assumed to be fixed.

The hub location calculated in the model with the initial parameters $G(W,p)$ is taken as a benchmark. The $G(W+F,p)$ represents the hub location on the perturbed data and shows the changes in the network. Results comparison function $d(G(W,p),G(W+F,p))$ can be written as

$$d(G(W,p),G(W+F,p)) = \sum_{\forall v_i \in G(W,p)} (v_i \neq v_i^q), \tag{1}$$

where v_i and v_i^q are binary variables values, which denote the hub i existence in the network. The function (1) can take only integer values and represents the number of distinct hubs in two different networks.

The similarity table S, where the row number is the simulation number and column number, — amount of hubs p in the network, the table values represents the number of mismatches of the selected hubs in the initial problem and the problem with the disturbed data.

Algorithm of similarity table calculation

```
Input: $W$, $p_{\min}$, $p_{\max}$, $g_{ij}$;
Output: $S = \{s_{kp}\}$;
Input functions: $G(W,p)$, $d(G(W,p), G(W+F,p))$;
for $k$ in $1,\dots,r$
        $F^k$ generation;
        for $p \in p_{\min},\dots,p_{\max}$
            $s_{kp} = d(G(W,p), G(W+F^k,p))$;
        end
end
```

The average amount of network changes or variety frequency is calculated as

$$\nu_p = \frac{\sum_{k=1}^{r} s_{kp}}{r}$$

depending on the number of hubs in the network. The network changes reflect the variety of the network, where the most robust hub number is preferred.

Result resampling. The bootstrap approach in our work is applied to generate the set of ν_p^l values for different numbers of hubs p. The results multiplication are necessary for decision making based on risk assessing.

The ν_p depends on the similarity table. For this reason the similarity tables S_l are generated using the random choice of rows with repetitions from the similarity table S to produce the set of ν_p^l. Let denote the $N_p = \{\nu_p, \nu_p^1, \nu_p^2, ..., \nu_p^b\}$ the set of average network changes for hub number equals to p, where $\bar{N}_p = \frac{\sum_{\nu \in N_p} \nu}{b+1}$ is average value of the sample and $S_p^2 = \frac{1}{b}\sum_{\nu \in N_p}(\nu - \bar{N}_p)^2$ is unbiased sample variance.

There are $\mathbf{p_{max}} - \mathbf{p_{min}}$ amount of samples and samples parameters. It is assumed that the best number of hubs have the minimal sample average value and unbiased sample variance. This is a multicriteria problem which is proposed to be solved using risk theory.

Decision making. The $\bar{N}_p$ and S_p^2 are the characteristics of the network with p hubs, which reflect the variability of the network. We propose to use the criterion for choosing the number of hubs with minimal values of $\bar{N}_p$ and S_p^2. The criterion interpretation is to minimise the risks of network changes or maximise the stability level based on the concept in [19]. The VaR formulation in applied task is following:

$$P(\xi > \bar{N}_p + u_\alpha S_p) = 1 - P(\xi \leqslant \bar{N}_p + u_\alpha S_p) = \alpha. \tag{2}$$

In formula (2) u_α is a standard normal distribution α-quantile and the condition $u_\alpha = -u_{1-\alpha}$ holds. Using this property, we receive expression

$$P(\xi \leqslant \bar{N}_p - u_{1-\alpha} S_p) = 1 - \alpha.$$

The value at risk concept [21] is proposed to be used as a criterion of hub number estimation:

$$p_{\text{opt}} = \text{argmin}_{p \in p_{\min} \cdots p_{\max}} (\bar{N}_p - u_{1-\alpha} S_p).$$

The main assumption that must be made for the application of the criterion — the general distributions of the samples are the normal distributions.

Simulation study. The USApHMP model described in [20] is considered as the hub location-allocation model. The GUROBI Optimizer 7.0.1* solver is applied for MIP to calculate the hub sets to be choose in each iteration in the algorithm of similarity table calculation with restriction GAP < 4 % in time saving goal, the average models optimization time is 178 seconds. The results were obtained with Intel Core i5 2.7GHz processor and 8GB RAM.

The experiment was carried out on the "normalized" dataset granted by "Delovye linii". The data consist of the set of terminals coordinates (178 terminals), the set of possible hub locations (10 potential hubs), the costs of hub construction and transportation costs depend on the direction. The distances between hubs and terminal-hubs were calculated in seconds (driving time by car without traffic jams consideration) using Google Maps Distance Matrix API**.

This test case extends the results presented in [20]. In this example we have considered the hub quantities from 6 to 9 ($p_{\min} = 6$, $p_{\max} = 9$), the repetitions amount for each hubs quantity in the problem were equal to 40 (in total case there were $4 \cdot (40 + 1) = 164$ simulations), the MIP sizes in each iteration were 19 680 variables (17 890 continuous, 1790 binary) and 21 539 constraints, during the presolve stage in optimization process were removed 90 continuous variables.

The flows perturbation were generated by Truncated Normal Distribution with the mean equal to 30 and standard deviation equal to 80 with the same distribution for each direction of flow. The hub construction costs were identical and were equal to 3 million units. This part of data were not granted by company for reasons of confidentiality policy, but the average values and standard deviations were proposed.

The first part of the algorithm results are presented the variety frequencies for each hub number p:

p	6	7	8	9
ν	0.875	0.0	0.1	0.625

There are generations of samples in the second part of the algorithm, where by using bootstrapping procedures, there are produced the 999 variety frequencies for each hub number p. The similarity tables S_p^l are calculated by using random choice of 40 rows with repetition from S_p and the variety frequencies for ν_p^l are estimated. The $\bar{N}_p$, S_p^2 and value at risk VaR_p with $\alpha = 0.05$ for each sample are presented on the Table.

Table. **Value at risk for each hub number**

p	$\bar{N}$	S^2	VaR
6	0.87435	0.00265	0.791
7	0.0	0.0	0.0
8	0.10428	0.00245	0.014
9	0.62343	0.00556	0.504

The study case has an obvious result $p = 7$, where the network of hubs do not get the changes on perturbed data. This is assumed to be the best solution in the proposed concept. The hub number equals to 8 has close results to 0 and could be interpreted as robust. The considered ν_p have a large difference and there is no problem to choose the minimal, but in the cases when the ν_p is close to each other the second criteria should be applied (for example, minimal total costs or maximal revenue).

* URL: http://www.gurobi.com/ (accessed: 29.07.2018).

** URL: https://developers.google.com/maps/documentation/distance-matrix/ (accessed: 29.07.2018).

The results shows that the hub numbers 6 and 9 contains competitive hub locations in the network and the hub numbers 7 and 8 don't contains significant network changes amount. This can be interpreted as settlement of a dispute, where the competitive hub is added in the network or another hub addition resolve the competitive hub dispute.

Conclusion. The work presents the extension of method [20] for hub number estimation in the facilities location problem. The algorithm uses the value at risk measure to find robust solution for possible commodities flow changes. The high complexity problem of USApHMP is solved by using the bootstrap procedures. The paper proposes new results of hub number validation, which can be applied to cluster analysis problem. We have explored the algorithm application on the real case of data, in general, there were simulated 164 mixed integer models and they were solved by using GUROBI Optimizer 7.0.1 software.

References

1. O'Kelly M. E. Activity levels at hub facilities in interacting networks. *Geogr. Anal.*, 1986, vol. 18(4), pp. 343–356.

2. O'Kelly M. E. The location of interacting hub facilities. *Transp. Sci.*, 1986, vol. 20, pp. 92–105.

3. Klincewicz J. G. Heuristics for the p-hub location problem. *European Journal of Oper. Research*, 1991, vol. 53, pp. 25–37.

4. Skorin-Kapov D., Skorin-Kapov J., O'Kelly M. E. Tight linear programming relaxations of uncapacitated p-hub median problems. *European Journal of Oper. Research*, 1996, vol. 94, pp. 582–593.

5. O'Kelly M. E., Miller H. J. Solution strategies for the single facility minimax hub location problem. *Papers in Regional Science*, 1991, vol. 70(4), pp. 367–380.

6. Kim H., O'Kelly M. E. Reliable p-hub location problems in telecommunication networks. *Geogr. Anal.*, 2009, vol. 41(3), pp. 283–306.

7. O'Kelly M. E. Network hub structure and resilience. *Netw. Spat. Econ.*, 2015, vol. 15(2), pp. 235–251.

8. Ebery J., Krishnamoorthy M., Ernst A. T., Boland N. The capacitated multiple allocation hub location problems: formulations and algorithms. *European Journal of Oper. Research*, 2000, vol. 120, pp. 614–631.

9. Lee H., Shi Y., Nazem S. M., Kang S. Y., Park T. H., Sohn M. H. Multicriteria hub decision making for rural area telecommunication networks. *European Journal of Oper. Research*, 2001, vol. 133, pp. 483–495.

10. Wishart D. Mode analysis: A generalization of nearest neighbor which reduces chaining effects. *Numerical Taxonomy*, 1969, pp. 282–311.

11. Hartigan J. *Clustering algorithms.* New York, USA, John Wiley Publ., 1975, 364 p.

12. Hartigan J. Statistical theory in clustering. *J. Classification*, 1985, vol. 2, pp. 63–76.

13. Volkovich Z., Brazly Z., Morozensky L. A statistical model of cluster stability. *Pattern Recognition*, 2008, vol. 41, pp. 2174–2188.

14. Volkovich Z., Brazly Z., Toledano-Kitai D., Avros R. The Hotelling's metric as a cluster stability measure. *Computer Modelling and New Technologies*, 2010, vol. 14, pp. 65–72.

15. Pelleg D., Moore A. X means extending k-means with efficient estimation of the number of clusters. *Proceedings of the 17th Intern. conference on Machine Learning.* San Francisco, USA, Morgan Kaufmann Publ., 2000, pp. 727–734.

16. Barzily Z., Golani M., Volkovich Z. On a simulation approach to cluster stability validation. *Special Issue, Mathematical and Computer Modeling in Applied Problems.* Moscow, Institute Informatics Problems RAS Publ., 2008, pp. 86–112.

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

18. Lozkins A., Bure V. M. The probabilistic method of finding the local-optimum of clustering. *Vestnik of Saint Petersburg University. Series 10. Applied Mathematics. Computer Science. Control Processes*, 2016, iss. 1, pp. 28–37.

19. Lozkins A., Bure V. M. The method of clusters stability assessing. *Stability and Control Processes in memory of V. I. Zubov (SCP), 2015 Intern. conference, IEEE*, 2015, pp. 479–482.

20. Lozkins A. Bure V. M. Single hub location-allocation problem under robustness clustering concept. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2017, vol. 13, iss. 4, pp. 398–406. https://doi.org/10.21638/11701/spbu10.2017.406

21. Lim C., Sherali H. D., Uryasev S. Portfolio optimization by minimizing conditional value-at-risk via nondifferentiable optimization. *Computational Optimization and Applications*, 2010, vol. 46, no. 3, pp. 391–415.

Received: August 7, 2018.
Accepted: September 25, 2018.

Author’s information:

Aleksejs Lozkins — Postgraduate Student; aleksejs.lozkin@gmail.com

Определение количества распределительных центров в сети на основе процедур статистического моделирования

А. Ложкинс

Санкт-Петербургский государственный университет, Российская Федерация,
199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Lozkins A.* The distribution centres choice in the facility location problem on the basis of statistical modeling procedures // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2018. Т. 14. Вып. 4. С. 346–351. https://doi.org/10.21638/11702/spbu10.2018.407

В работе рассматривается решение задачи о размещении распределительных центров в сети, основанное на статистическом моделировании и анализе данных. Расположение распределительных центров зависит от спроса на услуги в терминалах и общих затрат на выполнение услуг. Выбор количества распределительных центров в сети производится посредством статистического критерия, который использует концепцию робастности сети. Изменения сети моделируются случайными возмущениями спроса на услуги в терминалах. Величина случайных возмущений представляет прогнозные значения изменения спроса на услуги. Таким образом, симуляции реализуют потенциально-возможную сеть с некоторой вероятностью. Проведен численный эксперимент на фактических данных логистической компании по перевозке грузов. Решена задача нахождения робастного количества распределительных центров и их расположения, которые минимизируют издержки и устойчивы к потенциальным изменениям спроса на услуги.

Ключевые слова: расположение распределительных центров, статистическое принятие решений, статистическая устойчивость, количество распределительных центров.

Контактная информация:

Ложкинс Алексейс — аспирант, aleksejs.lozkin@gmail.com