Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Демин Сергей Станиславович

Методы оболочечного анализа данных для интервальных оценок и их применение

РЕЗЮМЕ ДИССЕРТАЦИИ
на соискание ученой степени кандидата наук
по прикладной математике

Научный руководитель:
доктор технических наук,
старший научный сотрудник
Алескеров Фуад Тагиевич

Москва – 2023

Постановка проблемы

В последнее время в большинстве сфер нашей жизни стараются в первую очередь обращать внимание на эффективность работы. Обычно этот показатель оценивается на основе конкретных данных, характеризующих исследуемый объект.

Один из основных подходов для решения этой задачи – оболочечный анализ данных. Этот метод позволяет сравнивать эффективность схожих объектов, при этом вклад каждого параметра оценивается автоматически, учитывая особенности каждого объекта в виде весовых коэффициентов каждого параметра. Благодаря этому нет необходимости делать какой-либо дополнительный анализ с учётом внешних источников для оценки весовых коэффициентов для разных параметров.

Однако, также как и большинство методов, оболочечный анализ данных требует точных данных, которые на практике не всегда оказываются доступны. Поэтому для практического применения приходится использовать специальные модификации оболочечного анализа данных, которые дают лишь приблизительные оценки эффективности.

Степень разработанности проблемы

Метод оболочечного анализа данных был сформулирован А. Чарнсом, У. Купером и Э. Родсом в 1978 г. (Charnes et al., 1978) на основе идей М. Фаррелла о технической эффективности объекта (Farrell M.J., 1957), который предложил оценивать эффективность k-го объекта e_k через отношение взвешенных сумм выходных (y_{ik}) и входных (x_{jk}) параметров:

$$e_k = \frac{\sum_{i=1}^{M} u_i y_{ik}}{\sum_{j=1}^{N} v_j x_{jk}},$$

где M и N – количество выходных и входных параметров соответственно, а u_i и v_j – неотрицательные коэффициенты, показывающие важность каждой конкретной характеристики объекта.

В результате в основе оболочечного анализа данных лежит следующая оптимизационная задача:

$$\max_{u_i, v_j} \left(e_k = \frac{\sum_{i=1}^{M} u_i y_{ik}}{\sum_{j=1}^{N} v_j x_{jk}} \right)$$

с ограничениями (L – общее количество объектов в выборке):

$$\begin{cases} \frac{\sum_{i=1}^{M} u_i y_{il}}{\sum_{j=1}^{N} v_j x_{jl}} \leq 1 & l \in \{1, \dots, L\} \\ u_i \geq 0 & i \in \{1, \dots, M\} \\ v_j \geq 0 & j \in \{1, \dots, N\} \end{cases}$$

В дальнейшем этими же авторами была предложена модель оболочечного анализа данных, при которой исследуется ситуация с переменной отдачей от масштаба, что расширило область применения оболочечного анализа данных (Banker et al., 1984):

$$\min_{\theta_k, \lambda} \theta_k$$
$$\begin{cases} -y_k + Y\lambda \geq 0 \\ \theta_k x_k - X\lambda \geq 0 \\ \lambda \geq 0 \end{cases}$$

где θ_k – это искомая эффективность k-го объекта, X и Y – матрицы входных и выходных параметров соответственно, а λ – вектор весов, показывающих ближайшие оптимальные элементы выборки, с помощью которых можно понять, как надо изменить значения параметров для того, чтобы тоже иметь 100% эффективности функционирования.

В 2000-х годах была обозначена проблема неточности в данных для оценки эффективности с помощью методов оболочечного анализа данных. В

ряде работ было предложено использовать различные способы предобработки данных для учёта возможных изменений в данных. Так, например, в (Entani et al., 2002) предлагается рассматривать два сценария и оценивать их по отдельности – оптимистичный и пессимистичный вариант. В первом случае все входные характеристики (используемые ресурсы) объекта предполагаются равными минимальным значениям внутри интервалов неопределённости, а выходные параметры (достигаемые результаты) – максимальным. Во втором случае происходит обратная ситуация – минимальные значения выходных параметров и максимальные значения входных.

Альтернативным подходом было использование теории нечётких множеств. Например, в (Kao & Liu, 2000) предлагается заменить все входные и выходные параметры в оптимизационной задаче по подбору весовых коэффициентов на треугольные нечёткие числа, оставив при этом веса обычными действительными числами. А в (Lertworasirikul et al., 2003) авторы несколько обобщают эту идею, заменяя параметры объекта уже на трапециевидные нечёткие числа.

Однако в основе всё равно оставалась классическая модель оболочечного анализа данных, которая имеет свои недостатки. В результате на данный момент практически не существует моделей, которые учитывали бы в своей структуре отсутствие точных данных, при этом оставались модификацией оболочечного анализа данных и сохраняли положительные свойства, такие как автоматическое определение относительной важности параметров.

Цели и задачи исследования

Целью данной работы является изучение свойств различных моделей оболочечного анализа данных, а также разработка и последующая апробация

новых моделей для оценки эффективности, позволяющих учитывать неточные данные.

Задачи диссертационного исследования:

1. провести анализ российских и иностранных работ, посвященных изучению различных моделей оболочечного анализа данных;
2. исследовать свойства и области применения существующих методов оболочечного анализа данных для работы с неточными данными;
3. разработать новые модификации оболочечного анализа данных для учёта погрешности в данных;
4. исследовать свойства полученных модификаций оболочечного анализа данных;
5. провести анализ эффективности в различных областях с использованием сгенерированных, а также реальных данных.

Актуальность

Исследование существующих методов оболочечного анализа данных, а также разработка новых методов с последующим изучением их свойств имеет огромное значение. Во-первых, эта информация даёт более глубокое понимание основных особенностей, преимуществ и недостатков самих методов оболочечного анализа данных, что позволит подбирать для решения каждой отдельной задачи наиболее подходящую модификацию оболочечного анализа данных

Во-вторых, изучение свойств существующих методов даёт информацию о том, в каких ситуациях предпочтительнее применять определённую модификацию оболочечного анализа данных для оценки эффективности с учётом особенностей конкретной задачи, а в каких

ситуациях лучше воспользоваться альтернативными подходами к оценке эффективности функционирования.

В-третьих, разработка новых методов оболочечного анализа данных позволит более продуктивно учитывать доступную информацию в условиях максимально приближенных к реальным.

Наконец, оболочечный анализ данных сейчас начинает применяться во многих областях. Однако, из-за использования базовых моделей результаты не всегда удовлетворяют поставленным задачам. Разработка новых моделей, учитывающих неточности в данных, позволит расширить применимость данной методики для оценки эффективности во всех сферах, тем самым повысив универсальность данного инструмента.

Личный вклад автора диссертационного исследования

Автором лично поставлена задача и разработан ряд новых математических моделей. Также сформулированы и доказаны утверждения о свойствах разработанных интервальных модификаций оболочечного анализа данных. Автором лично выполнена программная реализация разработанных моделей оболочечного анализа данных и проведены расчёты.

Также в работах по оценке эффективности противопожарных мероприятий в регионах Российской Федерации и оценке эффективности противокоронавирусных карантинных мер автором лично проводился сбор и обработка исходных данных, применение разработанных интервальных моделей оболочечного анализа данных и сравнение результатов с классической моделью оболочечного анализа данных.

Теоретическая значимость работы заключается в

1. разработке новых интервальных моделей оболочечного анализа данных;

2. исследовании теоретических свойств предложенных моделей, которые могут повлиять на область применения.

Практическая значимость диссертационного исследования заключается в широкой области применения предложенных интервальных моделей оболочечного анализа данных при сохранении интерпретируемых результатов с точки зрения полученных результатов. Помимо разобранных в рамках диссертационного исследования, предложенные модели могут быть использованы в аналогичных задачах оценки эффективности вне зависимости от масштаба и технических характеристик объектов.

Основные результаты исследования и положения, выносимые на защиту:

1. разработаны новые интервальные модели оболочечного анализа данных;
2. исследованы свойства предложенных моделей, которые влияют на область применения;
3. осуществлена программная реализация предложенных интервальных моделей оболочечного анализа данных;
4. проведена оценка эффективности превентивных мер противопожарной безопасности в регионах Российской Федерации;
5. проведена оценка эффективности противокоронавирусных карантинных ограничений в разных странах мира.

Методологическая основа исследования

В рамках диссертационного исследования используются методы линейной алгебры, методы теории оптимизации, а также методы программирования и компьютерного моделирования.

Научная новизна

В рамках диссертационного исследования получены следующие новые научные результаты:

1. построены новые интервальные модели оболочечного анализа данных, использующие в своей основе идеи об интервальных шкалах значений, предложенные Винером (Wiener, 1914, 1921);
2. исследованы свойства предложенных моделей;
3. апробированы предложенные интервальные модели для оценки эффективности в разных сферах и на разном масштабе сравниваемых объектов (страны, регионы).

Краткое содержание работы

Глава 1 носит обзорный характер, в ней приведена общая постановка задачи оценки эффективности функционирования схожих объектов, напоминается история оболочечного анализа данных, а также приводится обзор литературы в данной области. Кроме того, описываются существующие модификации оболочечного анализа данных, учитывающие гетерогенность выборки, а также внутреннюю структуру объектов для более качественного анализа данных.

Глава 2 посвящена разработке новых интервальных моделей оболочечного анализа данных. В начале главы рассказываются предлагаемые в литературе модели для анализа эффективности в случаях неточных данных.

Далее вводится две новые модели оболочечного анализа данных, которые работают с объектами, характеристики которых имеют интервальные значения.

Первая из них – интервальная модель оболочечного анализа данных с оптимальной трубкой (best tube Interval DEA). Согласно ей на первом этапе строится классическая граница оптимальной эффективности. Затем все объекты, которые находятся на границе эффективности, а также несравнимые

с оптимальной границей эффективности получают 100% оценку эффективности. В результате строится так называемая "оптимальная трубка", что и даёт название новому методу. Все остальные объекты оцениваются согласно классическому подходу на основе расстояния от идеальной границы (Aleskerov & Demin, 2021).

Вторая предлагаемая модель, использующая интервальную шкалу критериев, основана на идее, что любой параметр (как входной, так и выходной) может быть наиболее важным. Следовательно, если один из объектов имеет наилучшее значение хотя бы по одному признаку, его следует считать оптимальным. Для того, чтобы это учесть, в модели предлагается использовать принцип оптимальности по Парето. Благодаря этому все объекты, имеющие хотя бы один параметр с оптимальным значением попадает в Парето-оптимальное множество и получает максимальную оценку эффективности.

Этот принцип обладает полезными свойствами и может быть эффективно применен для сравнения объектов с интервальными значениями параметров, что было исследовано в (Aleskerov, 1994). Согласно принципу оптимальности по Парето, набор наилучших объектов строится из всех объектов, которые не доминируемы по Парето.

Именно с построения Парето-оптимального множества и начинается процедура оценки эффективности согласно Парето-версии интервального оболочечного анализа данных (Pareto IDEA). Все объекты, попадающие в него, автоматически получают максимальную эффективность. Для всех остальных же элементов выборки используется классическая версия оболочечного анализа данных (Aleskerov & Demin, 2021).

В разделе 2.3 проводится исследование свойств разработанных процедур. В результате были сформулированы и доказаны два утверждения:

Утверждение 1. В случае применения классического и новых интервальных методов оболочечного анализа данных (best tube IDEA и Pareto IDEA) для одних и тех же данных оценка эффективности любого объекта новыми методами всегда будет не ниже, чем оценка эффективности классическим методом.

Доказательство. Согласно интервальному оболочечному анализу данных с оптимальной трубкой, на первом шаге алгоритма группа наилучших объектов $B(X)$ и несравнимых с ними, исключается со 100% оценкой эффективности. В случае применения Парето-версии интервального оболочечного анализа данных на первом шаге алгоритма все оптимальные по Парето объекты исключаются со 100% оценкой эффективности. В результате могут появиться некоторые объекты, которые получат 100%-ную эффективность вместо более низких значений.

Кроме того, ориентир для оценки эффективности остальных объектов снижается из-за возможного расширения множества $B(X)$. Следовательно, расстояние от всех неэффективных объектов до границы эффективности не может стать больше, а значит, оценка эффективности не может снизиться. ■

Утверждение 2. Увеличение показателя неопределенности в данных (параметр ε, характеризующий ширину интервалов (x_i^-, x_i^+) и (y_i^-, y_i^+)) не приводит к уменьшению оценки эффективности объектов по обоим интервальным методам оболочечного анализа данных.

Доказательство. Рассмотрим значения интервальных параметров как (x_i^-, x_i^+) и (y_i^-, y_i^+), которые эквивалентны $(x_i - \varepsilon_i, x_i + \varepsilon_i)$ и $(y_i - \delta_i, y_i + \delta_i)$.

Увеличение неопределенности данных ε приводит к увеличению ширины интервалов параметров ε_i и δ_i.

Для примера рассмотрим ε_i и δ_i и увеличенные значения $\varepsilon_i{}'$ и $\delta_i{}'$.

$$\forall i,j,k \quad x_{ij} - \varepsilon_i \le x_{ik} - \varepsilon_i \le x_{ij} + \varepsilon_i \le x_{ik} + \varepsilon_i \quad \Rightarrow$$
$$\Rightarrow x_{ij} - \varepsilon_i{}' \le x_{ik} - \varepsilon_i{}' \le x_{ij} + \varepsilon_i{}' \le x_{ik} + \varepsilon_i{}'$$
$$\forall i,j,k \quad y_{ij} - \delta_i \le y_{ik} - \delta_i \le y_{ij} + \delta_i \le y_{ik} + \delta_i \quad \Rightarrow$$
$$\Rightarrow y_{ij} - \delta_i{}' \le y_{ik} - \delta_i{}' \le y_{ij} + \delta_i{}' \le y_{ik} + \delta_i{}'$$

Следовательно, все пары пересекающихся интервалов остаются пересекающимися. В результате,

$$\forall x \in X \quad x \in B(X) \Rightarrow x \in B'(X),$$

где $B(X)$ и $B'(X)$ – множества наилучших (оптимальных по Парето в Парето-версии и объектов внутри оптимальной трубки в интервальном оболочечном анализе с оптимальной трубкой) объектов с двумя значениями ε (для $B'(X)$ параметр ε больше). Другими словами, $B(X) \subseteq B'(X)$.

Между тем, могут возникнуть новые пары пересекающихся интервалов, что означает, что могут возникнуть новые объекты, которые будут включены в $B(X)$ и получат оценку эффективности 100% вместо более низкого показателя.

Кроме того, возможное расширение $B(X)$ сократит $X \backslash B(X)$ $\left(X \backslash B'(X) \subseteq X \backslash B(X)\right)$. В результате эффективность объектов из $X \backslash B(X)$ может только вырасти за счет снижения для них ориентира границы эффективности. ■

Также в разделе 2.3 описывается группа численных экспериментов на сгенерированных данных, подтверждающих утверждения о свойствах разработанных интервальных методов оболочечного анализа данных. Причём вышеупомянутые свойства сохраняются даже в случае определенных модификаций алгоритмов (например, при добавлении учёта гетерогенности

выборки как в работах Алескерова и Петрущенко (2015, 2016)). Тем самым, это подтверждает, что эти специфические особенности являются основополагающими для предлагаемых интервальных моделей оценки эффективности функционирования схожих объектов.

В Главе 3 рассмотрены приложения разработанных моделей к решению ряда прикладных задач. В разделе 3.1 исследуется эффективность противопожарных превентивных мер в различных регионах Российской Федерации. Для этого проанализированы бюджетные траты регионов на охрану окружающей среды и лесоводство и произошедшие в этих регионах в 2020 году пожары. На основе результатов применения различных интервальных модификаций оболочечного анализа данных получено несколько рейтингов, которые имели незначительные различия с точки зрения порядка регионов. При этом согласно интервальным методам разница между наименее эффективными субъектами Российской Федерации увеличивается, благодаря чему становится проще выделить регионы, где в первую очередь надо оптимизировать организацию противопожарных мер.

В разделе 3.2 производится оценка эффективности противокоронавирусных карантинных мер, проводимых в разных странах мира. При этом помимо самих карантинных мер и количества заболевших коронавирусом в стране для оценки эффективности используется степень законопослушности жителей. Кроме того, проводится анализ весовых коэффициентов, благодаря чему удаётся выделить наиболее важные направления карантинных мер, что позволит в будущем организовывать противокоронавирусные ограничения ещё эффективнее.

Основные выводы исследования

В рамках настоящего диссертационного исследования реализованы следующие научные задачи:

1. разработаны новые интервальные модификации оболочечного анализа данных;
2. доказаны утверждения, позволяющие сделать выводы о свойствах и применимости разработанных модификаций оболочечного анализа данных;
3. продемонстрированы возможности использования предложенных интервальных моделей оболочечного анализа данных для оценки эффективности функционирования схожих объектов вне зависимости от масштаба объектов.

Список опубликованных статей, где отражены основные научные результаты диссертации

Все публикации входят в журналы/издания, индексируемые в международных базах Scopus и Web of Science:

1. Aleskerov F., Demin S., An Assessment of the Impact of Natural and Technological Disasters Using a DEA Approach, Dynamics of Disasters — Key Concepts, Models, Algorithms, and Insights / Ed.: P. M. Pardalos, A. Nagurney, I. S. Kotsireas, Springer, 2016, pp. 1-14.
2. Aleskerov F.T., Demin S.S., DEA for the Assessment of Regions' Ability to Cope with Disasters, Dynamics of Disasters. Impact, Risk, Resilience, and Solutions / Ed.: P. M. Pardalos, A. Nagurney, I. S. Kotsireas, A. Tsokas, Springer, 2021, Ch. 2. pp. 31-37.
3. Aleskerov F., Demin S., Myachin A., Yakuba V. Short-Term Covid-19 Incidence Prediction in Countries Using Clustering and Regression Analysis. 9th International Conference on Computers Communications and Control (ICCCC) 2022, Springer, 2023, Vol. 1435. pp. 333-342.

4. Demin S., COVID-19 Quarantine Measures Efficiency Evaluation by Best Tube Interval Data Envelopment Analysis, Operations Research Forum, Springer, 2023, 4(21).

Список литературы

1. Aleskerov F. (1994). Multicriterial Interval Choice Models. *Information Sciences*, Elsevier Inc., 80, 25-41.

2. Aleskerov F.T., Demin S.S. (2021). DEA for the Assessment of Regions' Ability to Cope with Disasters. *Dynamics of Disasters. Impact, Risk, Resilience, and Solutions*, Springer, 1(2), 31-37.

3. Aleskerov F.T., Petrushchenko V.V. (2015). An Approach to DEA for Heterogeneous Samples. Modelling, Computation and Optimization in Information Systems and Management Sciences, Advances in Intelligent Systems and Computing, Springer, 15-21.

4. Aleskerov F.T., Petrushchenko S. (2016). DEA by sequential exclusion of alternatives in heterogeneous samples. International Journal of Information Technology and Decision Making, World Scientific Publishing Co. Pte Ltd, 15(1), 5-22.

5. Banker R.D., Charnes A., Cooper W.W. (1984). Some Models for Estimating Technical and Scale Inefficies in Data Envelopment Analysis. *Management Science*, INFORMS Institute for Operations Research and the Management Sciences, 30(9), 1078-1092.

6. Charnes A, Cooper WW, Rhodes E. (1978). Measuring the efficiency of decisionmaking units. *European Journal of Operations Research*, Elsevier, 2(6), 429-444.

7. Entani T., Maeda Y., Tanaka H. (2002). Dual models of interval DEA and its extension to interval data. *European Journal of Operational Research*, Elsevier, 136(1), 32-45.

8. Farrell M.J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society. Series A (General)*, Wiley-Blackwell Publishing Ltd, 120(3), 253-290.

9. Kao C., Liu S.T. (2000). Fuzzy efficiency measures in data envelopment analysis. *Fuzzy Sets and Systems*, Ed.: I. Couso, B. De Baets, L. Godo, Elsevier, 113(3), 427-437.

10. Lertworasirikul S., Fang S.-C., Joines J.A., Nuttle H.L.W. (2003). Fuzzy data envelopment analysis (DEA): a possibility approach. *Fuzzy Sets and Systems*, Ed.: I. Couso, B. De Baets, L. Godo, Elsevier, 139(2), 379-394.

11. Wiener N. (1914). A contribution to the theory of relative position. *Proceedings of the Cambridge Philosophical Society*, Cambridge Philosophical Society, 17, 441-449.

12. Wiener N. (1921). A new theory of measurement: a study in the logic of mathematics. *Proceedings of the London Mathematical Society*, John Wiley and Sons Ltd, 19, 181-205.