ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Р.А. Бобров, А.Е. Лепский

РАНЖИРОВАНИЕ ВУЗОВ ПО БАЛЛАМ ЕГЭ МЕТОДАМИ СРАВНЕНИЯ НЕЧЕТКИХ ЧИСЕЛ

Препринт WP7/2014/01
Серия WP7
Математические методы
анализа решений в экономике,
бизнесе и политике

Москва
2014

УДК 519.25
ББК 22.172
Б

Редакторы серии WP7
«Математические методы анализа решений в экономике,
бизнесе и политике»
Ф.Т. Алескеров, В.В. Подиновский, Б.Г. Миркин

Бобров, Р.А., Лепский, А.Е. Ранжирование вузов по баллам ЕГЭ методами сравнения нечетких чисел [Текст]: препринт WP7/2014/01 Р.А. Бобров, А.Е. Лепский; Нац. исслед. ун-т «Высшая школа экономики». – М.: Изд. дом. Высшей школы экономики, 2014. – 18 с. – (Серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике»). – 20 экз.

Дан обзор основных подходов к ранжированию гистограмм с подробным анализом нечеткостных методов ранжирования. Приведены результаты численного анализа применения нечеткостных методов сравнения к ранжированию вузов по гистограммам средних баллов ЕГЭ зачисленных в вузы абитуриентов. Показано, что результаты ранжирования нечеткостными методами могут довольно сильно отличаться от ранжирования по среднему баллу. Указаны сильные и слабые стороны нечеткостного ранжирования большого числа гистограмм.

УДК 519.25
ББК 22.172

Бобров Р.А. – бакалавриат Отделения прикладной математики и информатики НИУ ВШЭ, Москва, Россия.
Лепский А.Е. – Международная лаборатория анализа и выбора решений; Департамент математики факультета экономики НИУ ВШЭ, Москва, Россия.

Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу:
http://www.hse.ru/org/hse/wp

1. Введение

Во многих прикладных задачах появляется необходимость сравнения двух гистограмм. Такая задача возникает, например, при необходимости сравнения (и, в частности, ранжирования) результатов различных опытов (например, в известных работах под руководством С.Э. Шноля [Панчелюга и др. 2006, Shnoll & Zenchenko & Udaltsova 2004] сравниваются формы гистограмм результатов измерений процессов разной природы, полученных в одно и то же местное время и в разных географических пунктах), при сравнении показателей функционирования каких-либо однородных (организационных, технических и пр.) систем [Алескеров и др. 2009], при принятии решений в условиях нечеткостной неопределенности [Vanegas & Labib 2001], при моделировании нечетких предпочтений [Fodor & Roubens 1994], при сравнении распределений доходов в рамках социально-экономического анализа [Rothschild & Stiglitz 1973, Shorrocks 1983, Sen 1973], при ранжировании учащихся по результатам-гистограммам их оценок и т.д.

Сравнивать гистограммы можно по форме, по мере близости и т.д. В ряде задач необходимо сравнивать гистограммы по отношению типа «больше-меньше», т.е. надо ранжировать гистограммы. Например, речь может идти о построении отношения полного предпорядка (рефлексивного, полного и транзитивного отношения) $\succ$. Поскольку в этой работе в качестве базового приложения рассматривается задача ранжирования вузов по баллам ЕГЭ, то будем предполагать, что вектор аргументов (баллы ЕГЭ) гистограмм упорядочен по возрастанию их важности, т.е. если две гистограммы отличаются друг от друга только в двух «столбцах», то гистограмма, которая имеет «больший столбик» с большим номером, для нас более предпочтительна, чем гистограмма, имеющая «больший столбик» с меньшим номером (точное определение см. ниже).

В задаче ранжирования гистограмм можно выделить несколько подходов. Первый (в порядке перечисления, но не исторически) – вероятностный подход. В этом подходе сравниваются некоторые числовые характеристики случайных величин, связанных со сравниваемой парой гистограмм [Sevastjanov & Róg 2003]. Кроме того, к вероятностному подходу можно отнести и принцип стохастического доминирования, который широко используется в теории риска [Ватник 2009, Wolfstetter 1999].

Другой подход связан с применением методов ранжирования распределений доходов в теории коллективного выбора [Shorrocks 1983]. В этом

случае сравниваются гистограммы доходов вида $U = (i, u_i)_{i=1}^{n_U} = (u_i)_{i=1}^{n_U}$, где $u_1 \le u_2 \le ... \le u_{n_U}$ с помощью функций благосостояния $W(U)$, удовлетворяющих определенным условиям (симметричности, монотонности, вогнутости и др.). Этот подход в случае, когда размерности векторов-гистограмм одинаковы, равносилен ранжированию упорядоченных по возрастанию векторов. В этом случае можно использовать, например, методы теории важности критериев [Подиновский 2007], методы некомпенсаторного выбора [Aleskerov, Chistyakov, Kaliaguine 2010] и др.

Третий подход к ранжированию гистограмм связан с применением инструментария сравнения нечетких чисел и будет подробно рассмотрен ниже.

Все эти подходы будут проанализированы в данной работе применительно к задаче ранжирования гистограмм вузов, построенных по результатам приема по баллам ЕГЭ.

2. Постановка задачи

Под гистограммой в этой работе будем понимать пару двух упорядоченных наборов чисел $U = (x_i, u_i)_{i \in I}$, $(x_i)_{i \in I}$ – упорядоченный по возрастанию вектор различных аргументов гистограммы (т.е. $x_i < x_{i+1}$, $i \in I$), $(u_i)_{i \in I}$ – вектор неотрицательных значений гистограммы, I – некоторое индексное множество.

Предположим, что задано множество гистограмм $\mathcal{U} = \{U\}$, где $U = (x_i^{(U)}, \tilde{u}_i)_{i \in I_U}$, $x_i^{(U)} < x_{i+1}^{(U)}$, $\tilde{u}_i \ge 0$ для всех $i \in I_U$ и для всех $U \in \mathcal{U}$. Необходимо задать ранжирование элементов множества $\mathcal{U}$, т.е. построить отношение полного предпорядка (рефлексивного, полного и транзитивного отношения) R. Если гистограммы U и V находятся в отношении R (т.е. $(U, V) \in R$)), то будем обозначать это так: $U \succcurlyeq V$ и говорить, что «U больше V». Если же $U \succcurlyeq V$ и $V \succcurlyeq U$, то будем называть эти гистограммы «равными» и обозначать это так: $U \sim V$. Это отношение также должно удовлетворять условию упорядоченности аргументов гистограмм по возрастанию их важности: если $U' = (x_i, u_i')$, $U'' = (x_i, u_i'')$ две такие гистограммы, что $u_i' = u_i''$ для всех $i \ne k, l$ и $u_l' - u_l'' = u_k'' - u_k' \ge 0$, то $U'' \succcurlyeq U'$ при $k > l$ и $U' \succcurlyeq U''$ при $k < l$.

Без ограничения общности можно считать, что все гистограммы «выровнены по числу столбцов», т.е. $I_U = I$ для всех $U \in \mathcal{U}$ и $\{x_i^{(U)}\}_{i \in I_U} = \{x_i\}_{i \in I}$. Тогда $U = (x_i, u_i)_{i \in I} = (u_i)_{i \in I}$ для всех $U \in \mathcal{U}$. Действительно, для выравнивания гистограмм надо объединить все множества аргументов гистограмм: $X^{(U)} = \{x_i^{(U)}\}_{i \in I_U}$, $X = \bigcup_U X^{(U)} = \{x_i\}$ и применить какую-либо процедуру заполнения пробелов в данных. Например, можно использовать следующее правило: $u_i = \tilde{u}_k$, если $x_k^{(U)} \le x_i < x_{k+1}^{(U)}$.

Пример 1. Пусть $\mathcal{U} = \{U_1, U_2\}$, где $U_1 = (x_i^{(U_1)}, \tilde{u}_i^{(1)})_{i \in I_{U_1}} = \{(1,2),(2,5),(4,6),(6,3)\}$, $U_2 = (x_i^{(U_2)}, \tilde{u}_i^{(2)})_{i \in I_{U_2}} = \{(1,3),(3,4),(4,5),(5,3)\}$. Тогда $X^{(U_1)} = \{1,2,4,6\}$, $X^{(U_2)} = \{1,3,4,5\}$. Следовательно, $X = X^{(U_1)} \cup X^{(U_2)} = \{1,2,3,4,5,6\}$ – объединённое множество аргументов гистограмм, а $(u_i^{(1)})_{i \in I} = \{2,5,5,6,6,3\}$, $(u_i^{(2)})_{i \in I} = \{3,3,4,5,3,3\}$ – новые значения гистограмм, выровненных по числу столбцов.

3. Нормирование данных

Многие методы ранжирования могут быть применены только к определенным образом нормированным гистограммам. И наоборот, некоторые способы нормирования определяют и класс возможных методов ранжирования. Например, если гистограммы выровнены по высоте, т.е. $\overline{u} = \max\limits_{i \in I} u_i = 1$ для всех $U = (u_i)_{i \in I} \in \mathcal{U}$ и удовлетворяют определенным условиям «выпуклости» (см. ниже), то для ранжирования возможно использование методов сравнения нечетких чисел. Если же гистограммы выровнены по площади, т.е. $\sum_{i \in I} u_i = 1$ для всех $U \in \mathcal{U}$, то возможно использование методов сравнения вероятностных распределений.

Выбранный способ нормирования должен быть интерпретируем с точки зрения ранжирования данного вида гистограмм. Например, если мы сравниваем гистограммы средних баллов ЕГЭ абитуриентов, принятых в вузы, то нормирование по площади означает, что мы сравниваем доли (относительно общего числа абитуриентов принятых в вуз) числа студентов, имеющих определенный балл.

4. Вероятностные методы сравнения

В этом случае множество гистограмм $\mathcal{U} = \{U\}$, $U = (x_i, u_i)_{i \in I} = (u_i)_{i \in I}$ должно быть выровнено по площади, т.е. $\sum_{i \in I} u_i = 1$ для всех $U \in \mathcal{U}$. Тогда $\mathcal{U} = \{U\}$ – множество вероятностных распределений случайных величин U (случайные величины и их распределения будем обозначать одинаково), принимающих значения из множества $\{x_i\}_{i \in I}$ с вероятностями $\{u_i\}_{i \in I}$ соответственно. В приложениях рассматриваются различные способы сравнения вероятностных распределений. Перечислим лишь некоторые из них.

1) $U \succcurlyeq V$, если $E[U] \geq E[V]$ (сравнение по среднему значению). Обобщением этого способа является: $U \succ V$, если $E[f(U)] \geq E[f(V)]$, где f – некоторая функция (функция полезности).

2) $U \succcurlyeq V$, если $F_U(x) \leq F_V(x)$ для всех $x \in \mathbb{R}$, где F_U – функция распределения случайной величины U. Это принцип стохастического доминирования 1-го порядка, который широко используется в теории риска [Ватник 2009, Wolfstetter 1999].

3) $U \succcurlyeq V$, если $P\{U \geq V\} \geq P\{U \leq V\}$. Такой способ сравнения рассматривался, например, в [Шахнов 2008]. Если считать, что случайные величины $U = (u_i)_{i \in I}$ и $V = (v_j)_{j \in I}$ независимы, то

$$P\{U \geq V\} = \sum_{(i,j): x_i \geq x_j} u_i v_j \,. \qquad (1)$$

Пример 2. Пусть $\mathcal{U} = \{U, V\}$, $U = (u_i)_{i=1}^5$, $V = (v_i)_{i=1}^5$, где $u_1 = u_5 = 0.1$, $u_2 = u_4 = 0.2$, $u_3 = 0.4$, $v_1 = v_5 = 0.15$, $v_2 = v_3 = 0.25$, $v_4 = 0.2$ (см. Рис. 1). Тогда $P\{U \geq V\} = 0.605$, а $P\{U \leq V\} = 0.595$. Следовательно, $U \succ V$.

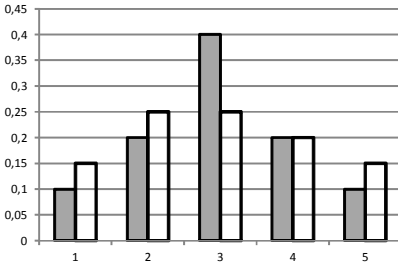

Рис. 1. Вероятностное сравнение двух гистограмм

5. Ранжирование нечетких чисел

В этом случае все гистограммы должны быть выровнены по высоте, т.е. $\overline{u} = \max_{i \in I} u_i = 1$ для всех $U = (u_i)_{i \in I} \in \mathcal{U}$. Теперь каждой гистограмме $U \in \mathcal{U}$ можно поставить в соответствие нечеткое множество [Zadeh 1965] с функцией принадлежности $U = (u_i)_{i \in I}$, определенной на универсальном множестве $X = (x_i)_{i \in I}$. Величина u_i может интерпретироваться как степень доверия к тому, что студент с оценкой x_i учится в вузе U. В теории нечетких множеств в большей степени развит математический аппарат сравнения так называемых нечетких чисел [Wang & Ruan & Kerre 2009], которые являются обобщением обычных (неразмытых) действительных чисел. В нашем случае гистограмма-нечеткое множество $U = (u_i)_{i \in I}$ будет нечетким числом, если все так называемые α-срезы $U_\alpha = \{x_i : u_i \geq \alpha\}$ являются непустыми и выпуклыми множествами для любого $\alpha \in (0,1]$. Заметим, что некоторые методы ранжирования нечетких множеств (например, центроидный метод, см. ниже) нечувствительны к этому требованию, другие (в основном – те, где используются α-срезы) – чувствительны. Это требование будет заведомо выполняться, если гистограммы являются унимодальными. В случае ранжирования гистограмм приема в вузы по результатам ЕГЭ, такие гистограммы будут близки к унимодальным при условии, что рассматривается прием только на одну специальность, один тип приема (например, по конкурсу) и выборка является достаточно большой. Абсолютной унимодальности гистограмм можно добиться путем применения определенных процедур сглаживания. Простейшей из них является метод группировки соседних разрядов. При этом, конечно, часть (иногда и значительная часть) информации о распределении баллов ЕГЭ абитуриентов вуза

теряется. Кроме того, в методе группировки соседних разрядов результат будет существенно зависеть от порядка группировки разрядов. Более щадящими с точки зрения потери количества информации о гистограмме являются процедуры приведения к унимодальному виду с помощью минимальных преобразований. В данной работе при необходимости приведения к унимодальному виду использованы только простейшие процедуры группировки разрядов. Дальше будем считать, что $\mathcal{U}$ – это множество нечетких чисел. На Рис. 2 приведен пример выравнивания гистограммы распределения результатов ЕГЭ студентов, поступивших в 2012 году на специальность «Экономика» по конкурсному набору в Национальный исследовательский университет «Высшая школа экономики» (г. Москва). Выравнивание осуществлено путем группировки двух соседних разрядов (отмечены фигурной скобкой). В результате этой процедуры два разновысотных столбца (отмечены пунктиром) были заменены двумя столбцами одинаковой «средней» высоты.

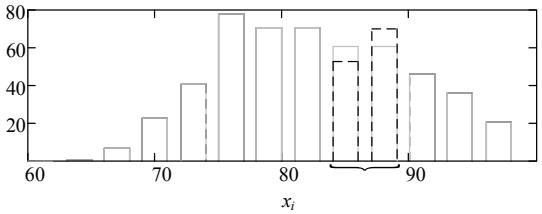

Рис. 2. Выравнивание гистограммы распределения результатов ЕГЭ студентов, поступивших в 2012 году на специальность «Экономика» по конкурсному набору в Национальный исследовательский университет «Высшая школа экономики» (г. Москва)

Существует много способов ранжирования нечетких чисел, но все их условно можно разделить на три группы.

Первую группу составляют методы сравнения с помощью построения функции (индекса) ранжирования. В этом случае определяется некоторая функция $F : \mathcal{U} \to \mathbb{R}$, которую также называют операцией дефаззификации, и считают, что $U \succcurlyeq V$ ($U \sim V$), если $F(U) \geq F(V)$ ($F(U) = F(V)$).

Во вторую группу включают методы, основанные на вычислении меры близости между нечетким числом U и некоторым эталонным нечетким числом U_0. В качестве эталонного числа U_0 чаще всего рассматриваются:

– максимизация нечетких чисел из $\mathcal{U}$ (т.е. число $\bar{U} = \max\{U : U \in \mathcal{U}\}$);

– минимизация нечетких чисел из $\mathcal{U}$ (т.е. число $\underline{U} = \min\{U : U \in \mathcal{U}\}$);

– некоторое «среднее» между максимальным и минимальным нечеткими числами.

Третью группу составляют методы, основанные на вычислении индексов парного сравнения всех нечетких чисел из $\mathcal{U}$ и последующего нахождения наилучшего элемента на графе парных сравнений.

Рассмотрим более подробно некоторые методы сравнений нечетких чисел в рамках указанных подходов.

5.1. Построение функции (индекса) ранжирования

В качестве индексов ранжирования чаще всего рассматриваются либо некоторые функции границ отрезков α-срезов $U_\alpha = [u_1(\alpha), u_2(\alpha)]$ нечеткого числа U, либо некоторые средние значения нечетких чисел. Примерами индексов ранжирования являются следующие функции.

1. Индекс Адамо [Adamo 1980] $F_\alpha(A) = u_2(\alpha)$ для некоторого фиксированного $\alpha \in (0,1]$. Другими словами, сравниваются только правые концы α-срезов для определенного значения α. Уровень α при этом является мерой риска неправильного решения: чем больше α, тем меньше риск неправильного решения.

2. Индекс Ягера [Yager 1981] $F(U) = \frac{1}{2}\int_0^1 (u_1(\alpha) + u_2(\alpha))d\alpha$, который представляет собой среднеинтегральное значение середин интервалов α-срезов. С другой стороны, в ряде задач ранжирования более информативными для установления отношения «больше» между нечеткими числами являются правые концы α-срезов, чем левые. Например, при ранжировании гистограмм поступления в вузы по результатам ЕГЭ правые концы α-срезов информируют о максимальных баллах поступивших на этом α-срезе. Поэтому в некоторых индексах ранжирования используется весовой коэффициент, регулирующий приоритет использования в индексе левых и правых концов α-срезов.

3. Обобщенный индекс Ягера: $F_\lambda(U) = \int_0^1 ((1-\lambda)u_1(\alpha) + \lambda u_2(\alpha))d\alpha$. Этот индекс совпадает с индексом Ягера при $\lambda = 0.5$. Большему значению λ соответствует меньший риск неправильного решения при сравнении двух нечетких чисел.

И в индексе Ягера, и в обобщенном индексе Ягера все α -срезы интегрируются с одинаковыми весами. Если же мы хотим учесть с большим приоритетом баллы больших групп поступивших, то это можно сделать с помощью неубывающей неотрицательной весовой функции $r(\alpha)$, удовлетворяющей условию нормировки $\int_0^1 r(\alpha)d\alpha = 1$. Тогда $F_{r,\lambda}(U) = \int_0^1 r(\alpha)((1-\lambda)u_1(\alpha) + \lambda u_2(\alpha))d\alpha$.

4. Центроидный индекс ранжирования. В качестве индекса ранжирования нечеткого числа U рассматривается алгебраическое (с учетом знака) расстояние между геометрическим центром фигуры $\{(x, y) : 0 < y < \mu_U(x)\}$ и осью Oy: $F_c(U) = \sum_i x_i u_i \Big/ \sum_i u_i$.

5.2. Методы, основанные на использовании меры близости

В этом случае по множеству нечетких чисел $\mathcal{U}$ вычисляется некоторое нечеткое число U_0 (которое условно назовем эталонным). Далее вычисляется некоторая мера близости между найденным эталонным нечетким числом U_0 и каждым числом из $\mathcal{U}$. Исходя из значений этой меры близости между U_0 и множествами из $\mathcal{U}$ принимается решение о ранжировании элементов множества $\mathcal{U}$. В качестве эталонного множества U_0 чаще всего рассматриваются число, равное максимизации нечетких чисел из $\mathcal{U}$, т.е. $\overline{U} = \max\{U : U \in \mathcal{U}\}$, или минимизация нечетких чисел из $\mathcal{U}$, т.е. $\underline{U} = \min\{U : U \in \mathcal{U}\}$ или некоторое «среднее» число. В качестве меры близости, как правило, рассматривают некоторое расстояние $d(U, U_0)$ от нечеткого числа U до эталонного нечеткого числа U_0, например, манхэттенское расстояние: $d_1(U, V) = \sum_i |u_i - v_i|$.

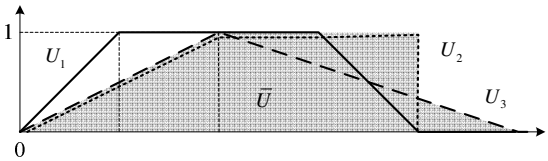

Рис. 3. Максимизация трех нечетких чисел

Широко известный индекс Керри [Kerre 1982] вычисляется, как $F(U) = d_1(U, \bar{U})$, и нечеткое число $\bar{U} = \max\{U : U \in \mathcal{U}\}$ вычисляется с помощью принципа обобщения Заде [Wang & Ruan & Kerre 2009]. На Рис. 3 приведены графики функций принадлежности трех нечетких чисел U_1, U_2, U_3 и их максимизации $\bar{U} = \max\{U_1, U_2, U_3\}$.

Заметим, что понятия максимального нечеткого числа и функции максимума от нечетких чисел (максимизации чисел) существенно (в отличие от неразмытой арифметики) различны. Чтобы найти максимальное нечеткое число из данного множества чисел, нужно произвести их ранжирование. Функция же максимума от нечетких чисел вычисляется и без ранжирования, но, в отличие от неразмытой арифметики, может не совпадать ни с одним из чисел данного множества. Нетрудно видеть, что максимум нечетких чисел является нечетким числом, у которого левая граница равна правой огибающей всех левых границ, а правая граница равна правой огибающей всех правых границ. На Рис. 4 приведена гистограмма-нечеткое число $\bar{U}$, равное нечеткому максимуму всех гистограмм-нечетких чисел распределения результатов ЕГЭ студентов, поступивших в 2012 году в вузы РФ на специальность «Экономика» по конкурсному набору. Бледным цветом показана гистограмма распределения U результатов ЕГЭ вуза из середины списка ранжирования. Видно, что носители этих гистограмм имеют небольшое пересечение.

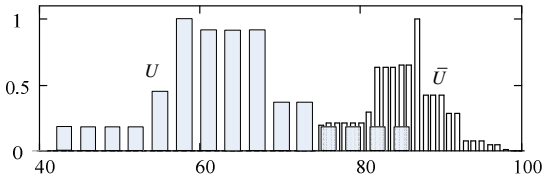

Рис. 4. $\bar{U}$ – гистограмма-нечеткое число, равное нечеткому максимуму всех гистограмм-нечетких чисел распределения результатов ЕГЭ студентов, поступивших в 2012 году в вузы РФ на специальность «Экономика» по конкурсному набору; U – гистограмма распределения результатов ЕГЭ вуза из середины списка ранжирования

При использовании индекса Керри полагают, что $U \succcurlyeq V$, если $F(U) \le F(V)$ и $U \sim V$, если $F(U) = F(V)$.

Метод Керри и аналогичные методы, основанные на измерении расстояния до эталонного нечеткого числа, хорошо зарекомендовали себя при

сравнении небольшого числа нечетких чисел. Если же необходимо ранжировать много нечетких чисел, то применение метода Керри и аналогичных ему оказывается менее эффективным. Действительно, если в качестве эталонного нечеткого числа рассматривается максимальное нечеткое число $\bar{U} = \max\{U : U \in \mathcal{U}\}$, причем количество чисел в $\mathcal{U}$ довольно велико, то число $\bar{U}$ может иметь «маленький» носитель $\mathrm{supp}\bar{U} = \{i : u_i > 0\}$, который будет иметь непустое пересечение с носителями только тех нечетких чисел-гистограмм, которые «близки» к гистограмме максимального числа. В этом случае расстояние от гистограммы U, носитель которой не пересекается с носителем числа $\bar{U}$ ($\mathrm{supp}\bar{U} \cap \mathrm{supp}U = \varnothing$), не будет характеризовать «близость» U к $\bar{U}$. Другими словами, метод Керри при большом числе гистограмм будет эффективно ранжировать только «близкие» к $\bar{U}$ гистограммы.

5.3. Методы, основанные на вычислении индекса парного сравнения нечетких чисел

В этой группе методов вводится некоторый индекс $r(U,V)$ парного сравнения нечетких чисел U и V. В результате вычисления этого индекса для всех пар нечетких чисел из множества $\mathcal{U}$ мы получим матрицу $R = (r_{ij})$, где $r_{ij} = r(U_i, U_j)$, $U_i, U_j \in \mathcal{U}$. Матрицу $R = (r_{ij})$ можно рассматривать как матрицу смежности взвешенного графа или как матрицу нечеткого отношения на $\mathcal{U}$ (если $r_{ij} \in [0,1]$). Далее рассматривается некоторая процедура поиска наилучших элементов на этом графе. Затем найденные элементы исключаются из графа, процедура повторяется и т.д.

В качестве индекса парного сравнения нечетких чисел в литературе рассматривается, например, так называемый индекс ранжирования Бааса – Квакернаака [Baas & Kwakernaak 1977]:

$$r(U,V) = \sup_{x_i \geq x_j} \min\left\{u_i, v_j\right\}. \qquad (2)$$

Этот индекс представляет собой «нечеткий» аналог формулы (1) вычисления вероятности $P\{U \geq V\}$, если U и V – независимые случайные величины, в которой операции умножения и сложения заменяются на «min» и «max» (sup) соответственно.

Индекс Бааса – Квакернаака обладает следующими свойствами:

а) если $\mathrm{Ker}U \cap \mathrm{Ker}V \neq \varnothing$, то $r(U,V) = r(V,U) = 1$, где $\mathrm{Ker}U = \{x_i : u_i = 1\}$ – так называемое ядро нечеткого числа;

б) если найдутся два таких числа $u' \in \mathrm{Ker}U$ и $v' \in \mathrm{Ker}V$, что $u' > v'$, то $r(U,V) = 1$ и $r(V,U) = hgt(U \cap V)$, где $hgt(U) = \sup\limits_i u_i$ – высота нечеткого множества.

Из этих свойств следует, что индекс Бааса – Квакернаака $r(U,V)$ принимает максимальное значение, равное единице только в том случае, когда промежуток наиболее достоверных значений множества U расположен на числовой оси правее промежутка наиболее достоверных значений множества V (при этом считаем, что первый промежуток U_1 расположен «правее» второго U_2, если найдутся таких два числа $u_1 \in U_1$ и $u_2 \in U_2$, что $u_1 > u_2$). Кроме того, используя свойства а) и б) можно вычислить индекс Бааса – Квакернаака, не прибегая к вычислению его по определению (2), как решению оптимизационной задачи.

Другой способ попарного сравнения был предложен в [Sevastjanov & Róg 2003], а именно, каждой паре гистограмм-нечетких чисел U и V ставилось в соответствие нечеткое множество $S_{U,V}$, определенное на универсальном множестве $[0,1]$ с функцией принадлежности $\mu_{S_{U,V}}(\alpha) = P\{U_\alpha > V_\alpha\}$, равной вероятности события $U_\alpha > V_\alpha$ при условии, что две случайные величины независимо и равномерно распределены на α-срезах (четких отрезках) U_α и V_α. Тогда в качестве индекса (строгого) парного сравнения гистограмм-нечетких чисел U и V рассматривалась величина $\overline{P}(U > V) = F(S_{U,V})$, где F – некоторая операция дефаззификации (см. пункт 5.1).

В [Bronevich & Rozenberg 2013] в случае вероятностных нечетких чисел U и V с функциями принадлежности $\mu_U(x) = P_U[x, +\infty)$, $\mu_V(x) = P_V[x, +\infty)$, где P_U, P_V – вероятностные меры, рассматривался следующий индекс включения ψ_β: $\psi_\beta\{U \subseteq V\} = P_U\left\{(V)_\beta \mid (U)_\beta\right\}$, $(U)_\beta = \{x : F_U(x) < \beta\}$, F_U – функция распределения случайной величины U. Тогда $U \succcurlyeq V$, если $\psi_\beta\{U \subseteq V\} \geq \psi_\beta\{V \subseteq U\}$ для любого $\beta \in [0,1]$. Ес-

ли $F_U(x) > F_V(x)$ для всех $x \in \mathbb{R}$, то $\psi_\beta\{U \subseteq V\} = 1$ для любого $\beta \in [0,1]$ и $V \succcurlyeq U$. Таким образом, с помощью индекса включения обобщается понятие стохастического доминирования.

В [Dubois & Prade 1983] дана интерпретация индекса Бааса – Квакернаака (2) в терминах теории возможностей [Дюбуа и Прад 1990]. А именно, с каждым нечетким числом U с функцией принадлежности $U = (u_i)_{i \in I}$ можно связать так называемую меру возможности

$$\Pi_U(A) = \sup_{x_i \in A} u_i, \ A \subseteq X, \tag{3}$$

которая характеризует возможность того, что истинная альтернатива принадлежит множеству A, если известно, что эта альтернатива находится в нечетком множестве U. Мера возможности удовлетворяет условиям: а) $\Pi_U(\varnothing) = 0$, $\Pi_U(X) = 1$; б) $\Pi_U(A \cup B) = \max\{\Pi(A), \Pi(B)\}$, $A, B \subseteq X$. Меру возможности можно определить и на множестве всех нечетких подмножеств данного универсального множества X по формуле

$$\Pi_U(V) = \sup_i \min\{u_i, v_i\}, \ V = (v_i)_{i \in I}, \tag{4}$$

которая является обобщением формулы (3). По аналогии с неразмытым промежутком $[x, +\infty)$, $x \in \mathbb{R}$, для нечеткого числа $V = (v_i)_{i \in I}$ вводится нечеткий промежуток $[V, +\infty)$ с функцией принадлежности

$$\mu_{[V,+\infty)}(r) = \sup_{r \geq x_i} v_i = \Pi_V((-\infty, r]). \tag{5}$$

Тогда индекс Бааса – Квакернаака (2) с учетом (4) и (5) можно переписать так:

$$r(U,V) = \sup_{x_i \geq x_j} \min\{u_i, v_j\} = \sup_{x_i} \min\left\{u_i, \sup_{x_i \geq x_j} v_j\right\} =$$

$$\sup_{x_i} \min\{u_i, \mu_{[V,+\infty)}(x_i)\} = \Pi_U([V, +\infty)).$$

Таким образом, величину $r(U,V) = \Pi_U([V, +\infty))$ можно интерпретировать как меру возможности того, что истинная альтернатива принадлежит нечеткому промежутку $[V, +\infty)$, если известно, что эта альтернатива находится в нечетком множестве U.

В то же время индекс Бааса – Квакернаака $r(U,V)=\Pi_U([V,+\infty))$ не различает, какое нечеткое число «больше», если $\mathrm{Ker}U \cap \mathrm{Ker}V \neq \varnothing$. Хотя может оказаться, что в этом случае одно число расположено правее другого (см. Рис. 5, где для наглядности показаны функции принадлежности двух «непрерывных» нечетких чисел), т.е. $u_i \geq v_i$ для всех $i \in I$.

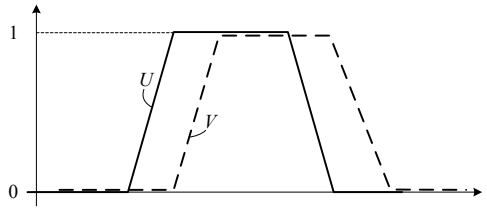

Рис. 5. Пример двух нечетких чисел, для которых $r(U,V)=\Pi_U([V,+))=1$

Поэтому в [Dubois & Prade 1983] предложено вместе с индексом $\Pi_U([V,+\infty))$ рассматривать и такие связанные с ним величины, как

$$\Pi_U(]V,+\infty))=\sup_{x_i}\inf_{x_j:x_j\geq x_i}\min\left\{u_i,1-v_j\right\},$$

$$N_U([V,+\infty))=\inf_{x_i}\sup_{x_j:x_j\leq x_i}\max\left\{1-u_i,v_j\right\},$$

$$N_U(]V,+\infty))=1-\sup_{x_i\leq x_j}\min\left\{u_i,v_j\right\},$$

где $N_U(A)=1-\Pi_U(\neg A)$, $A\subseteq X$ – двойственная к мере возможности Π_U, так называемая мера необходимости. Четверка значений $\Pi_U([V,+\infty))$, $\Pi_U(]V,+\infty))$, $N_U([V,+\infty))$, $N_U(]V,+\infty))$ будет точнее характеризовать расположение нечетких чисел U и V относительно друг друга. Так, например, для чисел, изображенных на Рис. 5 $\Pi_U([V,+\infty))=N_U([V,+\infty))=1$, $\Pi_U(]V,+\infty))=N_U(]V,+\infty))=0$.

Ранжирование нечетких чисел-гистограмм из $\mathcal{U}$ на основе значений индексов парных сравнений $r(U,V)$ можно получить с помощью некоторых правил коллективного выбора [Алескеров, Хабина, Шварц 2012]. В этом случае граф парных сравнений с матрицей смежности $R=(r_{ij})$, где

$r_{ij} = r(U_i, U_j)$, $U_i, U_j \in \mathcal{U}$, можно считать взвешенным мажоритарным графом в задаче коллективного принятия решения и использовать правила выбора на таком графе. Примерами таких правил являются правила, использующие мажоритарное отношение и правила, использующие вспомогательную числовую шкалу.

В первом случае, например, может быть использовано так называемое правило выбора минимального недоминируемого множества [Алескеров, Хабина, Шварц 2012] относительно отношения γ: $U \gamma W \Leftrightarrow r(U,W) < r(W,U)$. Множество H называется недоминируемым относительно γ, если $H = \{U : r(U,W) \geq r(W,U) \; \forall W \in \mathcal{U}\}$. Недоминируемое множество называется минимальным, если оно не содержит никакое собственное недоминируемое подмножество. Тогда процедура ранжирования нечетких чисел-гистограмм с помощью выбора минимальных недоминируемых множеств будет следующей. Пусть H_1 – минимальное недоминируемое множество на $\mathcal{U}$. Если $\mathcal{U}_1 = \mathcal{U} \setminus H_1 \neq \varnothing$, то на $\mathcal{U}_1$ найдем минимальное недоминируемое множество H_2 и т.д., пока на очередном k-м шаге не окажется, что $\mathcal{U}_k = \varnothing$. В результате мы получим k множеств $H_1, ..., H_k$. Тогда считаем, что $U \succcurlyeq V$, если $U \in H_m$, $V \in H_n$ и $m < n$. Если же $U, V \in H_m$ для некоторого m, то $U \sim V$. Заметим, что такая процедура применима для любого индекса парного сравнения r, который определяет ацикличное нечеткое отношение на $\mathcal{U}$.

Примерами правил выбора, использующих вспомогательную числовую шкалу, являются следующие процедуры.

1. Пусть $\rho_1(U) = \sum_W r(U,W)$. Тогда $U \succcurlyeq V$, если $\rho_1(U) \geq \rho_1(V)$. Заметим, что если вместо взвешенного мажоритарного графа с матрицей смежности $R = (r_{ij})$ использовать ориентированный граф с матрицей смежности $\tilde{R} = (\tilde{r}_{ij})$, $\tilde{r}_{ij} = \begin{cases} 1, r_{ij} \geq r_{ji}, \\ 0, r_{ij} < r_{ji} \end{cases}$, то $\rho_1(U) = \left|\{V : \tilde{r}(U,V) \geq \tilde{r}(V,U)\}\right|$ и указанная процедура ранжирования фактически совпадает со вторым правилом Коупленда [Алескеров, Хабина, Шварц 2012].

2. Пусть $\rho_2(U) = \sum_W r(U,W) - \sum_W r(W,U)$. Тогда $U \succcurlyeq V$, если $\rho_2(U) \geq \rho_2(V)$. Это правило обобщает так называемое первое правило Коупленда [Алескеров, Хабина, Шварц 2012].

3. Пусть $\rho_3(U) = \sum_W r(W,U)$. Тогда $U \succcurlyeq V$, если $\rho_3(U) \leq \rho_3(V)$ (обобщение третьего правила Коупленда).

6. Ранжирование вузов по результатам ЕГЭ абитуриентов, поступивших на специальность «Экономика», методами сравнения нечетких чисел

Базой исследования были результаты ЕГЭ абитуриентов, поступивших в 2012 году на специальность «Экономика» в один из вузов РФ и только по конкурсному набору. При этом рассматривались только те вузы, в которых число поступивших на данную специальность по конкурсу было не меньше 20 человек. Таким образом, в базу исследований попали данные 298 вузов РФ. Предварительно в этой базе были выполнены следующие процедуры: 1) построены для каждого вуза унимодальные гистограммы (путем группировки соседних разрядов гистограммы); 2) гистограммы выровнены по числу аргументов (столбцов); 3) гистограммы выровнены по высоте: $\overline{u} = \max\limits_{i \in I} u_i = 1$ для всех $U = (u_i)_{i \in I} \in \mathcal{U}$.

Результаты ранжирования первых 10 вузов, отобранных по убыванию среднего балла (M), с помощью обобщенного индекса Ягера (Y_λ) с $\lambda = 0.25$, 0.5 и 0.75, центроидным методом (Cen) и методом Керри (Ker) приведены в Табл. 1. В каждом столбце соответствующего метода указан порядок вуза в новом ранжировании.

Таблица 1. Итоговая таблица результатов ранжирования

Вузы	M	$Y_{0.25}$	$Y_{0.5}$	$Y_{0.75}$	Cen	Ker
1) ВШЭ-М	1	2	2	4	2	4
2) МГИМО	2	1	1	1	1	1
3) ПермГНИУ	3	11	11	10	16	11
4) ФУ	4	18	16	16	13	18
5) ВШЭ-СП	5	8	7	7	9	6
6) СПГПУ	6	3	3	3	3	2
7) МГУ	7	4	4	2	6	3
8) ЮУ НИУ	8	10	12	14	10	9

9) РЭА	9	24	21	17	23	25
10) СПбГУ	10	12	9	5	11	10

Примечание. Использованы следующие аббревиатуры вузов: 1) ВШЭ-М – Национальный исследовательский университет «Высшая школа экономики», г. Москва; 2) МГИМО – Московский государственный институт международных отношений; 3) ПермГНИУ – Пермский государственный национальный исследовательский университет; 4) ФУ – Финансовый университет при Правительстве Российской Федерации, г. Москва; 5) ВШЭ-СП – Национальный исследовательский университет «Высшая школа экономики», г. Санкт-Петербург; 6) СПГПУ – Санкт-Петербургский государственный политехнический университет; 7) МГУ – Московский государственный университет им. М.В. Ломоносова; 8) ЮУ НИУ – Национальный исследовательский Южно-Уральский государственный университет, г. Челябинск; 9) РЭА – Российская экономическая академия им. Г.В. Плеханова, г. Москва; 10) СПбГУ – Санкт-Петербургский государственный университет.

Стандартный способ сравнения двух ранжирований осуществляется с помощью вычисления ранговых корреляций – коэффициента корреляции Спирмена, коэффициента Кендалла [Кобзарь 2006]. Коэффициент корреляции Спирмена между двумя ранжированиями вычисляется по формуле

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2 ,$$

где d_i – разность между рангами двух ранжирований i-го вуза. Коэффициент корреляции Спирмена принимает значения из отрезка $[-1,1]$ и характеризует степень линейной зависимости между двумя ранжированиями. Результаты вычисления коэффициента корреляции Спирмена приведены в Табл. 2.

Таблица 2. Коэффициента корреляции Спирмена

r_S	M	$Y_{0.25}$	$Y_{0.5}$	$Y_{0.75}$	Cen	Ker
M	1	0.773	0.745	0.551	0.763	0.785
$Y_{0.25}$	0.773	1	0.988	0.944	0.975	0.994
$Y_{0.5}$	0.745	0.988	1	0.971	0.973	0.986
$Y_{0.75}$	0.551	0.944	0.971	1	0.922	0.953

Cen	0.763	0.975	0.973	0.922	1	0.97
Ker	0.785	0.994	0.986	0.953	0.97	1

Из Табл. 2 видно, что наиболее близким к ранжированию по среднему значению являются ранжирования по методу Керри и по методу Ягера с небольшим весом λ. Кроме того, нечеткостные методы ранжирования гистограмм вузов дают очень близкие между собой результаты ранжирования относительно коэффициента корреляции Спирмена.

Другим стандартным способом сравнения двух ранжирований является коэффициент Кендалла. Коэффициент корреляции Кендалла вычисляется по формуле:

$$r_K = 1 - \frac{4}{n(n-1)} R,$$

где R – количество инверсий одного ранжирования, расположенных в порядке возрастания другого ранжирования. Коэффициент корреляции Кендалла также принимает значения из отрезка $[-1,1]$ и характеризует степень линейной зависимости между двумя ранжированиями. Но, если в коэффициенте Спирмена используется евклидова метрика между векторами-рангами, то коэффициент Кендалла основан на вычислении минимального преобразования, преобразующего одно ранжирование в другое. Результаты вычисления коэффициента корреляции Кендалла приведены в Табл. 3.

Таблица 3. Коэффициент корреляции Кендалла

r_K	*M*	$Y_{0.25}$	$Y_{0.5}$	$Y_{0.75}$	*Cen*	*Ker*
M	1	0.378	0.378	0.2	0.333	0.289
$Y_{0.25}$	0.378	1	0.867	0.689	0.911	0.867
$Y_{0.5}$	0.378	0.867	1	0.822	0.867	0.867
$Y_{0.75}$	0.2	0.689	0.822	1	0.689	0.822
Cen	0.333	0.911	0.867	0.689	1	0.867
Ker	0.289	0.867	0.867	0.822	0.867	1

Из Табл. 3 видно, что ранжирование по среднему значению слабо коррелирует с ранжированиями нечеткостными методами относительно коэффициента Кендалла, но довольно сильно связаны между собой.

7. Выводы и заключение

Данная работа носит обзорно-аналитический характер. Цель, с одной стороны, состояла в том, чтобы дать обзор основных методов ранжирования гистограмм с подробным анализом нечеткостных методов ранжирования. С другой стороны, приведены результаты численного анализа применения нечеткостных методов сравнения к ранжированию вузов по гистограммам средних баллов ЕГЭ зачисленных в вузы абитуриентов. Приведенный анализ показал, что результаты ранжирования нечеткостными методами могут довольно сильно отличаться от ранжирования по среднему баллу. В то же время ранжирования, осуществленные совершено разными нечеткостными методами, зачастую ближе друг к другу, чем к ранжированию по среднему значению. Это говорит о том, что учет всей полноты информации, отраженной в гистограммах средних баллов ЕГЭ зачисленных в вузы абитуриентов, довольно сильно меняет результат ранжирования.

Оценивая применимость методов нечеткостного сравнения к ранжированию большого числа гистограмм, можно сделать вывод, что в целом эти методы дают более адекватное ранжирование, чем метод средних значений, поскольку учитывают всю полноту информации о гистограммах. Вместе с тем были выявлены и определенные трудности, связанные с: а) необходимостью для ряда методов приведения гистограмм к унимодальному виду с потерей части информации; б) неэффективностью применения некоторых методов именно к ранжированию большого числа гистограмм. Основной же трудностью «внедрения» этих методов в практику ранжирования является, по нашему мнению, невозможность для ряда методов дать простую наглядную интерпретацию сравнения (в баллах, абитуриентах и т.д.).

Дальнейшие исследования могут проходить в нескольких направлениях:

1) адаптировать «срезовые» методы нечеткостного сравнения для ранжирования неунимодальных гистограмм;

2) разработать эффективные и интерпретируемые процедуры приведения гистограмм к унимодальному виду с оценкой потери информации;

3) разработать интерпретируемые методы сравнения гистограмм.

Благодарности. Авторы выражают благодарность Ф.Т. Алескерову, В.В. Подиновскому, А.Г. Броневичу за полезные замечания и обсуждения результатов работы. Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2013 году.

Литература

Алескеров Ф.Т., Белоусова В.Ю., Солодков В.М., Сердюк М.Ю. Динамический анализ стереотипов поведения крупнейших российских коммерческих банков // В кн.: Модернизация экономики и глобализация: В 3 кн. Кн.3. / Отв. ред.: Е.Г. Ясин. М.: Издательский дом ГУ-ВШЭ, 2009, с.371-381.

Алескеров Ф.Т., Хабина Э.Л., Шварц Д.А. Бинарные отношения, графы и коллективные решения. М.: Физматлит, 2012.

Ватник П.А. Теория риска: учеб. Пособие. СПб.: С.-Петерб. гос. инж.-экон. ун-т., 2009.

Дюбуа Д., Прад А. Теория возможностей. Приложения к представлению знаний в информатике. М.: Радио и связь, 1990.

Кобзарь А.И. Прикладная математическая статистика. М.: Физматлит, 2006.

Панчелюга В.А., Коломбет В.А., Панчелюга М.С., Шноль С.Э. Исследование эффекта местного времени на малых пространственно-временных масштабах // Гиперкомплексные числа в геометрии и физике. 2006. Т.3. №1(5). С.116-121.

Подиновский В.В. Введение в теорию важности критериев в многокритериальных задачах принятия решений. М.: Физматлит, 2007.

Шахнов И.Ф. Задачи ранжирования интервальных величин при многокритериальном анализе сложных систем // Известия РАН. Теория и системы управления. 2008. №1. С.37-44.

Adamo J.M. Fuzzy decision trees // Fuzzy Sets and Systems. 1980. No.4. P.207-219.

Aleskerov F.T., Chistyakov V.V., Kaliaguine V.A. Social threshold aggregations // Social Choice and Welfare. 2010. Vol.35. No.4. P.627-646.

Baas S.M., Kwakernaak H. Rating and ranking of multiple-aspect alternatives using fuzzy sets // Automatic. 1977. No.13. P.47-58.

Bronevich A.G., Rozenberg I.N. Ranking probability measures by inclusion indices in the case of unknown utility function // Fuzzy Optimization and Decision Making. 2013. No.12(2).

Dubois D., Prade H. Ranking fuzzy numbers in the setting of possibility theory // Information Science. 1983. No.30 P.183-224.

Fodor J., Roubens M. Fuzzy preference modelling and multicriteria decision support. Dordrecht: Kluwer Academic Publishers, 1994.

Kerre E. The use of fuzzy set theory in electrocardiological diagnostics. In: Gupta, M., Sanchez, E. (eds.) Approximate reasoning in decision-analysis. North-Holland Publishing Company, Amsterdam. 1982. P.277-282.

Rothschild M., Stiglitz J.E. Some further results on the measurement of inequality // Journal of Economic Theory. 1973. No.6, P.188-204.

Shnoll S.E., Zenchenko K.I., Udaltsova N.V. Cosmophysical Effects in the Structure of Daily and Yearly Periods of Changes in the Shape of Histograms Constructed from the Measurements of 239P u alpha-Activity // Biophysics. 2004. Vol.49, Suppl.1. P.155.

Shorrocks A.F. Ranking Income distributions // Economica. 1983. Vol.50. February. P.3-17.

Sen A.K. On economic inequality. Oxford: University Press, 1973.

Sevastjanov P.V., Róg P. A Probabilistic Approach to Fuzzy and Crisp Interval Ordering // Task Quarterly. 2003. Vol.7. No.1. P.147-156.

Vanegas L.V., Labib A.W. Application of new fuzzy-weighted average (NFWA) method to engineering design evaluation // International Journal of Production Research. 2001, Vol.39. P.1147-1162.

Wang X., Ruan D., Kerre E.E. Mathematics of Fuzziness – Basic Issues. Berlin Heidelberg: Springer-Verlag, 2009.

Wolfstetter E. Topics in microeconomics: industrial organization, auctions, and incentives. Cambridge: Cambridge University Press, 1999.

Yager R.R. A procedure for ordering fuzzy sets of the unit interval // Information Sciences. 1981. Vol.24. P.143-161.

Zadeh L.A. Fuzzy sets // Information and Control. 1965. Vol.8. P.338-353.

Bobrov, R.A., Lepskiy A.E. Ranking universities according to the results of USE by means of fuzzy numbers comparison methods [Text]: Working paper WP7/ R.A. Bobrov, A.E. Lepskiy; National Research University “Higher School of Economics”. – Moscow: Publishing House of the Higher School of Economics, 2014. – 18 p. – (Series WP7 “Mathematical methods for decision making in economics, business and politics”). – 20 copies (in Russian).

A survey of the main approaches to ranking histograms is given with a detailed analysis of fuzzy ranking methods. The results of fuzzy comparison methods to the ranking of universities are given. It is shown that the results of fuzzy ranking methods can be quite different from the average score ranking. The strengths and weaknesses of fuzzy ranking of a large number of histograms are presented.

Bobrov, R.A. – School of Applied Mathematics and Information Science NRU HSE, Moscow, Russia.

Lepskiy A.E. – International Laboratory of Decision Choice and Analysis; Department of Mathematics for Economics NRU HSE, Moscow, Russia.

Р.А. Бобров, А.Е. Лепский

РАНЖИРОВАНИЕ ВУЗОВ ПО БАЛЛАМ ЕГЭ МЕТОДАМИ СРАВНЕНИЯ НЕЧЕТКИХ ЧИСЕЛ

Препринт WP7/2014/01
Серия WP7
Математические методы
анализа решений в экономике,
бизнесе и политике

Отпечатано в типографии
Национального исследовательского университета
«Высшая школа экономики» с представленного оригинал-макета
Формат 60×84 1/16. Тираж 20 экз. Уч.-изд. л. 1,1.
Усл. печ. л. 1,1. Заказ № . Изд. № 1571
Национальный исследовательский университет
«Высшая школа экономики»
125319, Москва, Кочновский проезд, 3
Типография Национального исследовательского университета
«Высшая школа экономики»