УДК 004.94; 331.2; 519.213

ПРИМЕНЕНИЕ ПАРАБОЛИЧЕСКОГО СПЛАЙНА ДЛЯ АНАЛИЗА РАСПРЕДЕЛЕНИЯ НАСЕЛЕНИЯ ПО ДОХОДАМ

Лебедев Валерий Викторович[1], Лебедев Константин Валерьевич[2], Тюпикова Татьяна Викторовна[3]

[1]Доктор экономических наук, кандидат физико-математических наук, профессор, академик РАЕН;
ФГБОУВО «МИРЭА – Российский технологический университет»
119454, Москва, Проспект Вернадского, д. 78;
e-mail: lebedev.guu@gmail.com.

[2]Кандидат экономических наук;
Институт стратегии развития образования РАО,
105062, Москва, ул. Жуковского, д.16;
e-mail: k.lebedev@instrao.ru.

[3]Кандидат экономических наук, доцент;
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: tanya@jinr.ru.

Целью исследования является эмпирическая проверка гипотезы о возможности использования логнормальной функции в качестве адекватной модели распределения населения по доходам. В качестве основного инструментального средства исследования используется метод сплайн-функций. Значения свободных параметров параболического сплайна определялись из решения задачи о минимизации интегральной гладкости функции распределения при выполнении условия о ее монотонном возрастании. В качестве информационной базы исследования использовались статистические данные о распределении заработной платы работников организаций РФ в 2019 г. Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 19-010-000921).

Ключевые слова: функция распределения, функция плотности, сплайн-функция, аппроксимация, статистические данные.

APPLICATION OF PARABOLIC SPLINE FOR POPULATION INCOME DISTRIBUTIONS ANALYSIS

Lebedev Valery[1], Lebedev Konstantin[2], Tyupikova Tatyana[3]

[1]Doctor of Economics, PhD in Physics and Mathematics, Professor, Academician of RANS;
MIREA – Russian *Technological* University,
Moscow, Vernadsky Avenue, 78;
e-mail: lebedev.guu@gmail.com.

[2]PhD in Economics;
Institute *for* Strategy *of* Education Development *of the* Russian Academy of Education,
105062, Moscow, Zhukovsky st., 16;
e-mail: k.lebedev@instrao.ru.

[3]PhD in Economics, associate Professor;
Dubna State University,
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
e-mail: tanya@jinr.ru.

The aim of the study is an empirical test of the hypothesis about the possibility of using the lognormal function as an adequate model of the distribution of the population by income. The method of spline functions is used as the main research tool. The values of the free parameters of the parabolic spline were determined from the solution of the problem of minimizing the integral smoothness of the distribution function under the condition of its monotonic increase. As the information base of the study, statistical data on the distribution of wages of employees of organizations of the Russian Federation in 2019 were used. This work was supported by the Russian Foundation for Basic Research, grant No. 19-010-000921.

Keywords: distribution function, density function, spline function, approximation, statistical data.

Введение

Вопросам изучения проблем неравенства населения различных регионов России и мира посвящена весьма обширная литература [1-4 и др.]. Количественная оценка степени расслоения населения по доходам опирается на теоретическую кривую Лоренца, которая наглядно иллюстрирует распределение доходов и строится на основе анализа соответствующей статистической информации. К наиболее часто используемым количественным показателям (индикаторам) неравенства относятся индексы Джини, Пальмы, Гувера, коэффициенты фондов и др. Особое место среди количественных показателей неравенства занимает коэффициент Джини, который остается одним из основных индикаторов неравенства.

В результате многочисленных исследований установлено, что количественные оценки показателей неравенства на основе использования теоретических функции распределения и кривой Лоренца существенно зависят от выбора функции распределения и методики оценки точности аппроксимации [5-8 и др.]. Понятно, что выбор функции распределения зависит от конкретной статистической информации. Тем не менее, используемая Росстатом методика расчета показателей дифференциации населения по уровню доходов опирается на использование логнормального распределения [9]. Обоснование такого выбора и соответствующие комментарии можно найти в работах [5, 10, 11]. Так, в работе [11] читаем: «Хорошо известно, что распределение доходов описывается логнормальным распределением». Однако существуют работы, в которых использование функций, опирающихся на нормальное распределение, ставится под сомнение. Приведем мнение А.И. Орлова: «…широко распространена вера в то, что статистические данные часто подчиняются нормальному распределению. Математики думают, что это – экспериментальный факт, установленный в прикладных исследованиях. Прикладники уверены, что математики доказали нормальность результатов наблюдений. Между тем анализ конкретных результатов наблюдений, в частности, погрешностей измерений, приводит всегда к одному и тому же выводу – в подавляющем большинстве случаев реальные распределения существенно отличаются от нормальных» [12].

Целью нашего исследования является эмпирическая проверка гипотезы о возможности использования логнормальной функции в качестве адекватной модели распределения населения по доходам. В качестве основного инструментального средства исследования используется метод сплайн-функций [13, 14]. Значения свободных параметров параболического сплайна определялись из решения задачи о минимизации интегральной гладкости функции распределения при выполнении условия о ее монотонном возрастании. В качестве информационной базы исследования использовались статистические данные о распределении заработной платы работников организаций РФ в 2019 г.

1. Задача построения функции распределения

При анализе неравенства в распределении доходов населения предполагается, что все население ранжировано в порядке возрастания доходов. Ниже используются следующие обозначения: x – величина денежного дохода; x_0 – значение минимального денежного дохода населения; y – доля населения с доходами не более x в общей численности населения (кумулятивная доля населения); z – доля совокупных доходов населения, доходы которых не превышают x, в совокупном доходе всего населения (кумулятивная доля дохода). Упомянутая выше функция логнормального распределения может быть записана так:

$$F(x,\sigma,a)=\int_0^x f(t,\sigma,a)dt, \qquad (1)$$

где

$$f(x,\sigma,a)=\frac{1}{\sigma x\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\ln a)^2}{2\sigma^2}\right). \qquad (2)$$

С формальной точки зрения для определения параметров a и σ кумулятивной функции распределения (1) на основе статистических данных вида (x_k, y_k), где вектор параметров, x_k и y_k – значения переменных x и y, $k=1,2,...,N$, требуется решить задачу минимизации некоторой функции $W(a)$, характеризующей степень близости графика теоретической функции от эмпирических точек $A_k(x_k, y_k)$. Одной из таких функций является следующая часто используемая функция:

$$W(a)=\frac{1}{N}\sum_{k=1}^{N}(F(x_k,a)-y_k)^2. \qquad (3)$$

Однако вместо решения задачи минимизации функции (3) часто используется следующий прием: параметры a и σ функции распределения (1) определяются по двум числовым характеристикам распределения, например по среднему и медиане, или по моде и медиане, или по среднему и среднеквадратическому отклонениям, или по моде и максимальному значению функции плотности (2). «Таким образом, для их вычисления уже не требуется знания всего статистического ряда, а достаточно знания только двух его числовых характеристик» [11].

На рис. 1 пунктирная линия – график функции плотности распределения по заработной плате работников организаций г. Москвы, параметры которой определены на основе кусочно-линейной интерполяции твблично заданной функции распределения [13]. Здесь, как и на других рисунках статьи, на оси абсцисс откладываются приведенные уровни заработной платы. Последние вычисляются как отношение заработной платы к среднему его значению, выраженное в процентах. В рассматриваемом случае соответствующая эмпирическая функция плотности является кусочно-постоянной (она изображена условно на рис. 1 горизонтальными штрихами). Как видим, график теоретической функции плотности (2) достаточно хорошо аппроксимирует эмпирическую функцию.

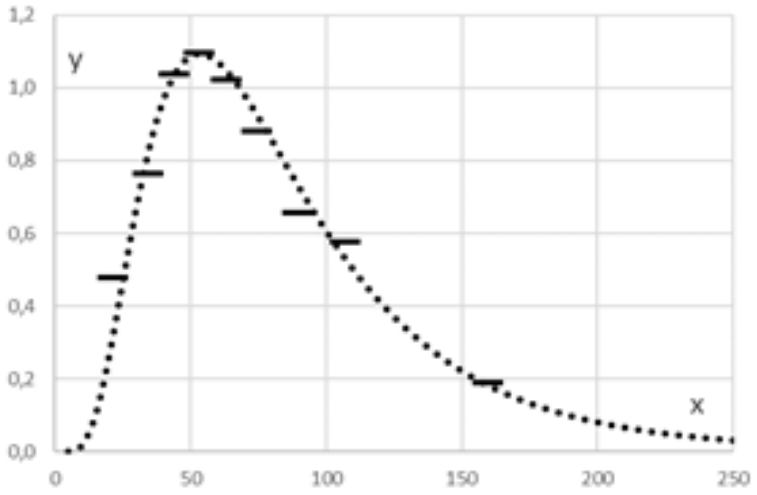

Рис. 1. Эмпирическая и теоретическая (логнормальная) функции плотности распределения работников организаций г.Москвы по заработной плате в апреле 2019 г.
Источник: расчеты авторов по данным Росстата [13]

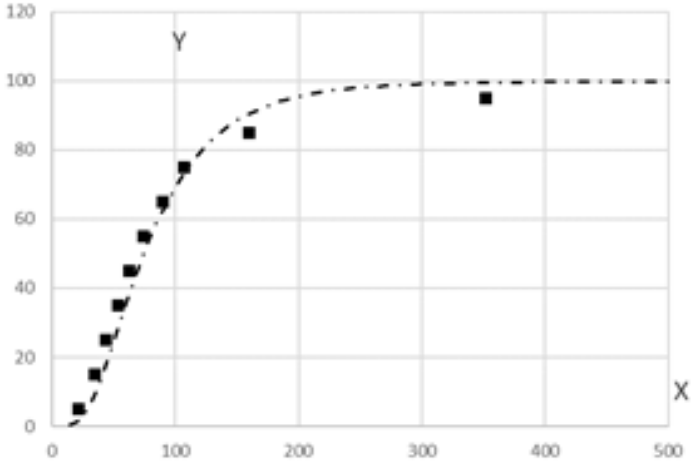

Рис. 2. Эмпирическая и логнормальная функции распределения работников организаций г.Москвы по заработной плате в апреле 2019 г.
Источник: расчеты авторов по данным Росстата [13]

Прямоугольники на рис. 2 – точки $A_k(x_k, y_k)$, где $y_k = 10k - 5$, $x_k = 100\mu_k / \mu$, $\mu_k -$ средний доход в k-ой децильной группе работников, $\mu -$ средний доход всех работников организаций г. Москвы, $k = 1,2,...10.$ Здесь же на рис. 2 приведен график функции распределения по заработной плате работников организаций г. Москвы (штрих-пунктирная линия), которая соответствует функции плотности распределения, изображенной на рис. 1. Как видим, график логнормальной функции распределения (1) достаточно хорошо аппроксимирует эмпирическую функцию распределения при значениях заработной платы, которую получает 80% работников. Однако, для 20% работников с наибольшими заработками график логнормального распределения, соответствующего функции плотности распределения, изображенной на рис. 1, проходит существенно выше эмпирических точек $A_k(x_k, y_k)$, где $k = 1,2,...10.$ Поэтому функция плотности, график которой приведен на рис. 1, не адекватно отражает реальное распределение работников организаций г. Москвы по заработной плате.

2. Применение параболического сплайна

Для повышения точности аппроксимации функции распределения $y = F(x)$ мы использовали квадратичный сплайн; в этом случае соответствующая эмпирическая функция плотности является кусочно-линейной [13, 14]. Приведем основные формулы для следующего случая: задана таблица $(x_k; y_k)$, где $k = 0, 1, ..., n, n+1.$ Здесь $y_0 = 0,\ y_{n+1} = 1,\ x_{k-1} < x_k,\ y_{k-1} < y_k$, где $k = 1, ..., n, n+1.$

3.1. Функция распределения раселения по доходу:

$$F(x) = \begin{cases} 0, \text{ если } x < x_0; \\ y_{k-1} + a_k(x - x_{k-1}) + b_k(x - x_{k-1})^2, \text{ если } x \in [x_{k-1}, x_k),\ k = 1, ..., n+1; \\ 1, \text{ если } x \geq x_{n+1}. \end{cases} \quad (4)$$

Здесь $a_k > 0$, $k = 1, ..., n, n+1.$

3.2. Функция плотности распределения: $f(x) = F'(x).$

$$f(x)=\begin{cases}0, \text{ если } x<x_0;\\ a_k+2b_k(x-x_{k-1}),\ k=1,...,n+1;\\ 0, \text{ если } x\ge x_{n+1}.\end{cases} \qquad (5)$$

$$d_k=y_k-y_{k-1} \qquad h_k=x_k-x_{k-1}$$

3.3. Функция, обратная функции распределения: $x=\varphi(y)$, $\varphi(y)=F^{-1}(y)$, $y\in[0,1]$. Если $y\in[y_{k-1},\ y_k],\ k=1,...,n+1,$ то

$$\varphi(y)=\begin{cases}x_0,\ \text{если } y=0;\\ x_{k-1}+(y-y_{k-1})/a_k, \text{ если } b_k=0;\\ x_{k-1}+(\sqrt{D}-a_k)/(2b_k), \text{ если } b_k\ne 0;\\ x_{n+1},\ \text{если } y=1.\end{cases} \qquad (6)$$

3.4. Средний доход: $\mu=\int_0^1\varphi(y)dy.$ Интегрируя функцию (6), получаем:

$$\mu=\sum_{k=1}^{n+1}r_k\,, \qquad (7)$$

где при $k=1,...,n+1$ имеем: $r_k=d_k\mu_k$, $d_k=y_k-y_{k-1}$,

$$\mu_k=\begin{cases}x_{k-1}-\dfrac{a_k}{2b_k}+\dfrac{\left(a_k^2+4b_kd_k\right)^{3/2}-a_k^3}{12b_k^2d_k}, \text{ если } b_k\ne 0;\\ \dfrac{x_{k-1}+x_k}{2}, \text{ если } b_k=0.\end{cases}$$

3.5. Вычисление параметров сплайна. Отметим, что функции (4) и (5) должна быть непрерывны. Поэтому справедливы равенства

$$y_k=y_{k-1}+a_k(x_k-x_{k-1})+b_k(x_k-x_{k-1})^2,\ a_{k+1}=a_k+2b_k(x_k-x_{k-1}).$$

Обозначим $h_k=x_k-x_{k-1},\ u_k=d_k/h_k,$ где $k=1,...,n+1.$ Тогда из условий непрерывности функций (4) и (5) получаем: $u_k=a_k+b_kh_k$, $a_{k+1}=a_k+2b_kh_k$. Известно, что интерполяционный параболический сплайн имеет один свободный параметр [14]. Будем считать, что значение параметра a_1 известно. Тогда полученные соотношения позволяют вычислить все остальные значения параметров сплайна:

$$b_1=(u_1-a_1)/h_1,\ a_{k+1}=2b_kh_k+a_k,\ b_{k+1}=(u_{k+1}-a_{k+1})/h_{k+1},\ \textit{где}\ k=1,...,n.$$

Таким образом, сплайн и все соответствующие функции зависят от параметра a_1. Средний доход также является функцией параметра a_1: $\mu=\mu(a_1).$

Выбор значения свободного параметра a_1 зависит от имеющейся информации. Если средний доход известен, то значение параметра a_1 находим из решения уравнения $\mu = \mu(a_1)$. Если средний доход не известен, то значения параметра a_1 находим из условия минимизации интегральной гладкости функции распределения, которая вычисляется по формуле $W(a_1) = \sum_{k=1}^{n+1} b_k{}^2$, при выполнении условия $a_k > 0$, где $k = 1,...,n+1$.

На рис. 3 приведен график интерполяционного параболического сплайна (линия 1), который отражает распределение по заработной плате работников организаций г. Москвы (пунктирная линия). Узлами сплайна служат те же точки $A_k(x_k, y_k)$, где $k = 1,2,...10$, что и точки на рис. 2. Штрих-пунктирная линия 2 на рис. 3 – график логнормальной функции распределения по заработной плате работников организаций г. Москвы, параметры которой определены в результате минимизации функции (3).

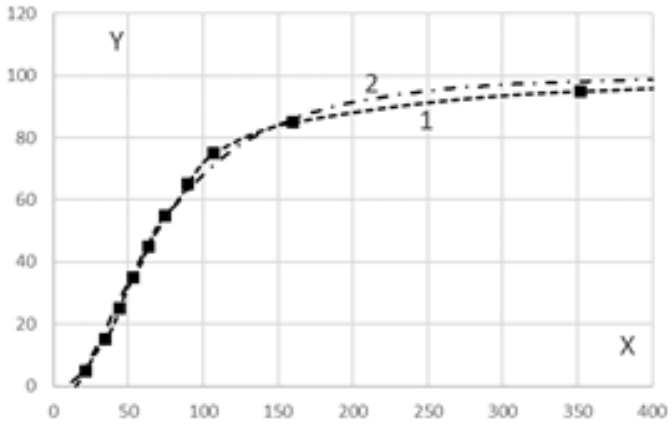

Рис. 3. Аппроксимация распределения работников организаций г.Москвы по заработной плате в апреле 2019 г. параболическим сплайном (линия 1) и логнормальной функцией (линия 2). Источник: расчеты авторов по данным Росстата [13]

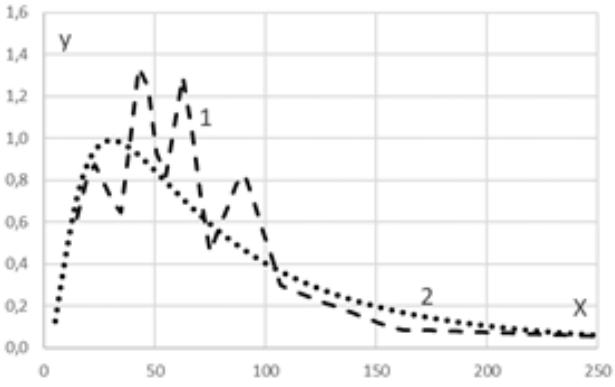

Рис. 4. Графики функций плотности распределения работников организаций г.Москвы по заработной плате в апреле 2019 г. (линия 1 – линейный сплайн, линия 2 – логнормальная функция). Источник: расчеты авторов по данным Росстата [13]

Как видим, график логнормальной функции распределения (1) достаточно хорошо аппроксимирует эмпирическую функцию распределения при значениях заработной платы, которую получает около 85% работников. Однако, для 15% работников с наибольшими заработками график логнор-

мального распределения на рис. 3 проходит выше интерполяционного параболического сплайна (линия 1). Если принять (на основании рис. 3), что сплайн более адекватно отражает распределение работников по заработной плате, чем логнормальное распределение, то использование поседнего в рассматриваемом случае приведет к занижению среднего значения заработной платы и искажению показателей дифференциации (коэффициента Джини и др.).

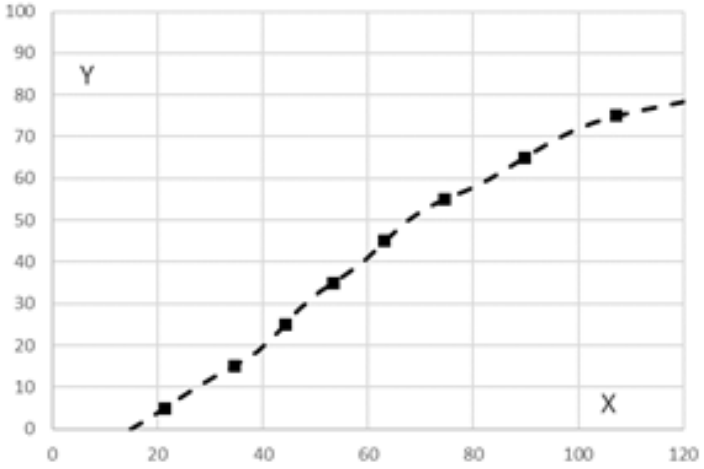

Рис. 5. Гафик сплайновой функции распределения работников организаций г.Москвы по заработной плате при значениях заработной платы ниже 120% от ее среднего значения.
Источник: расчеты авторов по данным Росстата [13]

Отметим еще одну особенность рассматриваемых статистических данных, которая иллюстрируется на рис. 4. На этом рисунке приведены графики функций плотности распределения работников организаций г.Москвы по заработной плате в апреле 2019 г., соответствующие функциям распределения, графики которых приведены на рис. 3. Здесь линия 1 – линейный сплайн, линия 2 – логнормальная функция плотности распределения. Как видим, график сплайновой функции плотности (линия 1) имеет несколько локальных максимумов и, следовательно, сплайновая функция распределения работников организаций г.Москвы по заработной плате в апреле 2019 г. (линия 1 на рис. 3) имеет несколько точек перегиба. Это хорошо видно на рис. 5, где приведен начальный участок сплайновой функции распределения при значениях заработной платы ниже 120% от ее среднего значения. В заключение обратим внимание на то, что, согласно рис. 3 и рис. 5, заработная плата более 70% работников организаций г. Москвы не тпревышает значения средней заработной платы всех работников.

Выводы

Использование сплайновой функции распределения достаточно адекватно отражает распределение населения по доходам. Выполненные расчеты с использованием сплайновой модели дают основание сделать вывод о том, что логнормальное распределение не является универсальным законом распределения. Показано, что особенностью функции распределения работников всех организаций г. Москвы в апреле 2019 г. по заработной плате является наличие у нее нескольких точек перегиба, в результате чего соответствующая функция плотности имеет несколько локальных максимумов. Для повышения адекватности модели распределения населения по доходам предполагается в дальнейшем модифицировать использованный здесь подход, представив «хвост» распределения экспоненциальной функцией $y = 1-(1-y_n)\exp[(x_n - x)/h]$, где $x > x_n$.

Список литературы

1. Айвазян С.А. Анализ качества и образа жизни населения (эконометрический подход). — М.: Наука. 2012. — С. 402.
2. Лившиц В.Н. Бедность и неравенство денежных доходов населения в России и за рубежом: системный анализ некоторых важных фрагментов проблемы. — М.: Институт экономики РАН, 2017. — С. 292.
3. Стиглиц Дж. Цена неравенства. Чем расслоение общества грозит нашему будущему. Перевод с англ. — М.: ЭКСМО, 2015. — С. 512.
4. Шевяков А.Ю., Кирута А.Я. Неравенство, экономический рост и демография: неисследованные взаимосвязи. — М.: Ин-т соц.-эконом. проблем народонаселения РАН. М-студия, 2009. — С. 192.
5. Колмаков И.Б. Метод и модели прогнозирования показателей дифференциации денежных доходов населения. — М.: Институт микроэкономики, 2004. — С. 168.
6. Abdalla I. M., Hassan M. Y. Maximum likelihood estimation of Lorenz curves using alternative parametric model. Metodoloski Zvezki, 1(1), 2004. — Pp. 109-118.
7. Chang A.C., Li P., Martin S.M. Comparing cross-country estimates of Lorenz curves using a Dirichlet distribution across estimators and datasets. Journal of Applied Econometrics. 2018. — Pp. 473-478. — [Электронный ресурс]. URL: https://doi.org/10.1002/jae.2595 (дата обращения: 01.10.2019).
8. Chotikapanich, D., Griffiths, W. E. Averaging Lorenz curves. Journal of Economic Inequality, 2005. — 3(1). — Pp. 1-19.
9. Методологические положения по статистике. Методика расчета показателей распределения и дифференциации по уровню доходов населения. Федеральная служба государственной статистики. — [Электронный ресурс]. URL: https://www.gks.ru/bgd/free/b99_10/isswww.exe/stg/d000/i000150r.htm (дата обращения: 01.10.2019).
10. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. Учебник для вузов. — М.: Издательство «Юнити», 1998. — С. 1022.
11. Золотухина Л.А., Тихоненко Д.В., Фридман Г.М. Исследование зависимостей между показателями дифференциации населения по доходам //Финансы и бизнес — 2015. — 3. — Стр. 39-54.
12. Орлов А.И. Публикации за полвека (1970-2019): Комментарии к списку научных и методических трудов. Изд. 3, испр. и доп. — М.: Институт высоких статистических технологий и эконометрики МГТУ им. Н.Э. Баумана, 2020. — С. 501.
13. Сведения о распределении численности работников по размерам заработной платы за апрель 2019 года (статистический бюллетень). Федеральная служба государственной статистики (РОССТАТ), Главный межрегиональный центр (ГМЦ), 2019. — [Электронный ресурс]. URL: https://gks.ru/compendium/document/13268 (дата обращения: 01.10.2019).
14. Стечкин С.Б., Субботин Ю.Н. Сплайны в вычислительной математике. — М.: Наука, 1976. — С. 248.
15. Иванилов Ю.П., Лебедев В.В. Применение сплайнов для сглаживания динамических рядов // Сообщения по прикладной математике. — М.: ВЦ АН СССР, 1990. — С. 48.