ВЫЧИСЛИТЕЛЬНЫЕ АСПЕКТЫ ЦИФРОВОЙ ЭКОНОМИКИ

Добронец Б.С., Попова О.А.
(*Сибирский федеральный университет, Красноярск*)
BDobronets@yandex.ru, OlgaArc@yandex.ru

В работе рассматриваются основные вычислительные проблемы в задачах цифровой экономики, связанные в первую очередь с обработкой и анализом данных больших объемов, организацией вычислительных процессов и повышением точности численных процедур. Подход основан на применении новых подходов к агрегации данных, вычислительного вероятностного анализа, использовании вероятностных расширений и численных операций над кусочно-полиномиальными функциями. Для выявления зависимостей в больших данных предлагается использовать функциональную регрессию на эмпирических распределениях.
Ключевые слова: цифровая экономика, большие данные, вычислительный вероятностный анализ, функциональная регрессия

Введение

Цифровая экономика, основанная на больших данных, является прогностической по своему типу: здесь прогноз, план и факт имеют тенденцию к равенству. Его основным инструментом является прогностическая аналитика, основной вид производства персонализирован для нужд клиента, и конкуренция идет не столько за перераспределение существующих рынков, сколько с образованием новых, где больше не конкурируют товары и технологии, а цифровые системы управления на основе цифровых платформ [3,4,11,17,18]. В тоже время данные становятся коммерческим продуктом.

Остановимся более подробно на вычислительных аспектах, характерных для задач цифровой экономики. Выделим из них три и рассмотрим новые методы и подходы, реализующие их. Первый вычислительный аспект связан с необходимостью обработки данных больших объемов. Для его реализации предлагается использовать процедуры агрегирования данных, основанные на применении математических моделей представления данных.

Второй аспект связан с организацией вычислительного процесса, обеспечивающий необходимую для решения соответствующей практической задачи оперативность получения необходимой информации. Для преодоления этой проблемы предлагается использовать рекурсивную схему организации вычислительного процесса.

Третий аспект отражает требование к надежности полученных результатов моделирования, обеспеченных надежными вычислительными процедурами, адекватными тем типам неопределенности, которые содержатся в сырых данных.

В первую очередь наш подход основан на технологиях Big Data, включая процедуры агрегации данных для входных параметров, и использовании вычислительного вероятностного анализа (ВВА) [1,5−10]. Переход к более обобщенному представлению с помощью агрегирования необходим по нескольким причинам. Во-первых, агрегация существенным образом может снизить объем данных. Во-вторых, детализированные данные часто оказываются очень изменчивыми из-за воздействия различных случайных факторов, разброса значений и, поэтому слабо отражают общие тенденции и свойства исследуемого множества. Агрегация в этом случае позволяет увидеть имеющиеся тенденции и закономерности.

Так, на рисунке 1 представлен пример В. Хардли [2], где слева представлено точечное множество данных, а справа результаты агрегирования этих данных с помощью «цветочного» графика.

Анализируя точечное множество данных, он высказал желание иметь какой-либо метод, позволяющий увидеть места их скопления. В качестве иллюстрации такого метода он привел так называемый «цветочный график» [2]. Цветочный график строится посредством определения сети квадратов, покрывающих плоскость (X,Y), и подсчета числа наблюдений, попадающих в отдельные квадратики. Число «лепестков цветка» соответствует числу наблюдений в квадрате этого «цветка».

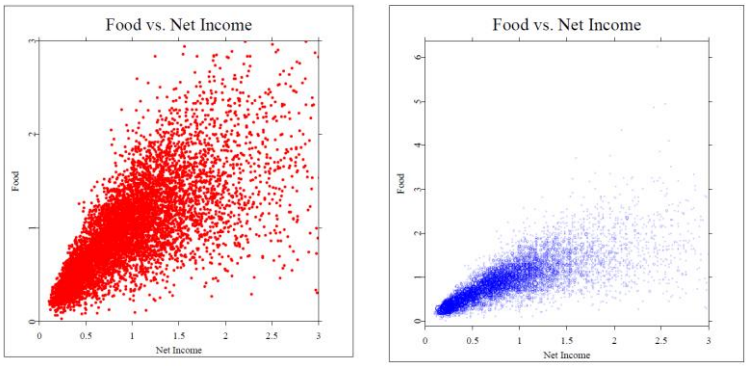


Рис.1. Зависимость расходов на питание от чистого дохода [2]

Левый рисунок 1 – множество точек зависимости расходов на питание от чистого дохода. Правый рисунок 1 – «цветочный график». «Цветочный график» указывает на сосредоточение данных вокруг увеличивающейся группы плотно упакованных «цветков», она не имеет явно выраженных скачков или быстрых локальных флуктуаций. Можно предположить гладкую зависимость кривой отклика.

В статье предлагается использовать кусочно-полиномиальные модели, в том числе сплайн функции. Предлагаемый подход к агрегации данных можно интерпретировать как построение функций распределения. В работах [5−7] обсуж-

даются различные типы математических моделей агрегации данных.

Хотя существует множество способов агрегации данных, включая простое среднее, мы утверждаем, что использование кусочно-полиномиальных функций агрегации будет предлагать более информативное представление о поведении больших данных, чем другие формы агрегации. Разработанные методы уменьшают уровень неопределенности в информационном потоке, позволяют значительно сократить время обработки и использовать численные операции [5,6,13].

В настоящее время доминирующие парадигмы экономических теорий основаны на классической математике и представлены в терминах вероятностных и статистических методов. Следует подчеркнуть, что в практических приложениях вероятностные и статистические методы часто и успешно используются в синтезе с современными методами мягких вычислений. Заметим, что в приложениях мы часто имеем дело со случайной и эпистемической неопределенностью [9].

В последние десятилетия существует много современных методов моделирования неопределенности. Как правило, они не противоречат традиционному вероятностному подходу, поскольку касаются других (не вероятностных) типов неопределенности [8].

В этой статье обсуждается использование вычислительного вероятностного анализа (ВВА) для задач со случайной и эпистемической неопределенностью [9]. Основой ВВА являются численные операции над функциями плотности вероятности случайных значений. Это операции “+”, “−”, “·”, “/”, “↑”, “max”, “min”, а также отношения порядка “≤”, “≥” и некоторые другие. Численные операции над кусочно-полиномиальными представлениями функций плотности вероятности и вероятностные расширения составляют основной компонент ВВА.

Используя подход ВВА, построены численные методы, которые позволяют решать системы линейных и нелинейных алгебраических уравнений со случайными параметрами. В

случае эпистемической неопределенности, мы вводим понятие гистограмм второго порядка. Опираясь на конкретные практические примеры, в работе [9] показано, что использование гистограмм второго порядка может оказаться полезным при принятии решений. В частности, рассмотрена оценка рисков инвестиционных проектов, в которых рассчитываются функции плотности вероятности таких факторов, как чистая текущая стоимость (NPV) и внутренняя норма доходности (IRR).

1. Организация вычислительного процесса

Одной из наиболее важных проблем численного моделирования больших данных является задача вычисления функциональных зависимостей. Для организации вычислительного процесса предлагается использовать одно основных понятий BBA – вероятностное расширение. Данный подход может использоваться как в случае больших данных, так и для различных типов неопределенности.

Под вероятностным расширением $\boldsymbol{f}(\boldsymbol{X}_1,...,\boldsymbol{X}_n)$ функции $f(x_1,x_2,...,x_n)$ мы имеем в виду функцию плотности вероятности случайной величины $z=f(x_1,...,x_n)$. где $(x_1,...,x_n)$ — система непрерывных случайных величин с совместной функцией плотности вероятности $p(x_1,...,x_n)$ [10].

Теорема 1.[10] Пусть $\boldsymbol{f}(\boldsymbol{X}_1,...,\boldsymbol{X}_n)$ вероятностное расширение функции $f(x_1,x_2,...,x_n)$ и для всех вещественных t функция $\boldsymbol{f}(t,\boldsymbol{X}_2,...,\boldsymbol{X}_n)$ вероятностное расширение $f(t,x_2,...,x_n)$. Тогда

$$\boldsymbol{f}(\boldsymbol{X}_1,...,\boldsymbol{X}_n)=\int_{\underline{X}_1}^{\overline{X}_1}\boldsymbol{X}_1(t)\,\boldsymbol{f}(t,\boldsymbol{X}_2,...,\boldsymbol{X}_n)dt. \quad (1)$$

Замечание. Из теоремы 1 вытекает возможность рекурсивного вычисления вероятностных расширений общего вида, сведением их к вычислению одномерных вероятностных расширений.

Рассмотрим вычисление интеграла (1). Для определенности представим (1) в виде квадратуры

$$\int_{\underline{x}_1}^{\bar{x}_1} \boldsymbol{x}_1(t)\, \boldsymbol{f}(t, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) dt \approx \sum_{l=1}^{m} \gamma_l\, \boldsymbol{x}_1(t_l)\, \boldsymbol{f}(t_l, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$$

Для вычисления $\boldsymbol{f}(t_l, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ далее можно использовать численные квадратуры. Теорема позволяет организовать вычислительный процесс в виде рекурсии с использованием процедуры распараллеливания.

Отметим, что на нижнем уровне необходимо вычислить вероятностные расширения для функций только одной переменной. Все вычисления на каждом уровне независимы и могут вычисляться одновременно.

2. Функциональная регрессия на эмпирических распределениях больших данных

Одним из новых направлений для изучения данных является функциональный анализ данных (ФАД) или Function Data Analysis (FDA), который занимается анализом и теорией данных, представленных в виде некоторых функций, изображений или более общих объектов.

Одним из основных понятий ФАД является понятие функциональных данных, которые представлены так, что для каждого субъекта в случайной выборке записывается одна или несколько функций.

Термин FDA был придуман Ramsay, Dalzell [14-16], история этой области намного старше.

Функциональные данные первого поколения обычно состоят из случайной выборки независимых вещественных функций, $X_1(t), ..., X_n(t)$ на компактном интервале $I = [0, T]$ на вещественной прямой. Эти вещественные функции можно рассматривать как реализации одномерного стохастического

процесса, который часто предполагается в гильбертовом пространстве, например, $L_2(I)$.

Функциональные данные следующего поколения — это функциональные данные которые являются частью сложных объектов данных, и они могут быть многомерными, коррелированными или включать изображения, или формы.

В самом общем понимании можно рассматривать функциональные данные как реализации основного стохастического процесса.

Основной процесс в реальных задачах часто не может наблюдаться напрямую, так как данные могут быть собраны дискретно с течением времени, либо на фиксированной или случайной сетке времени. В таких ситуациях основной процесс считается скрытым. Временная сетка, где проводятся наблюдения, может быть плотной, разреженной или пустой, и она может отличаться от предмета к предмету.

Хотя формального определения функциональных данных не существует, соглашение заключалось объявить функциональные данные как плотные (в отличие от разреженных) выборок, Разреженные функциональные данные возникают в исследованиях, для которых измеряются объекты в разные моменты времени и количество измерений n_i для объекта i может быть ограничено, т.е. $\sup_{1\le i\le n} n_i < C < \infty$ некоторой константой

Разреженные и нерегулярно выбранные функциональные данные (которые соответствуют общему типу продольных данных), обычно требуют больше усилий в теории и методологии, чем плотно выбранные функциональные данные, которые записываются непрерывно.

Функциональные данные, которые наблюдаются непрерывно без ошибки являются самым простым типом для обработки в качестве теории для случайных процессов, таких как функциональные законы большие числа и функциональные центральные предельные теоремы легко применимы.

Цели функционального анализа данных в основном такие же, как и у любой другой ветвь статистики. Они включают в себя:

- представлять данные способами, которые помогают дальнейшему анализу;
- отображать данные, чтобы выделить различные характеристики;
- изучать важные источники закономерностей и вариаций среди данных;
- объяснять изменения в результатах с помощью входной информации;
- сравнить два или более набора данных в отношении определенных типов вариации, где два набора данных могут содержать разные наборы дубликатов одних и тех же функций.

Функциональная регрессия — это версия регрессионного анализа, когда ответы или ковариаты включают функциональные данные. Модели функциональной регрессии могут быть классифицированы на четыре типа в зависимости от того, являются ли ответы или ковариаты функциональными или скалярными:

- скалярные ответы с функциональными ковариатами,
- функциональные ответы со скалярными ковариатами,
- функциональные ответы с функциональными ковариатами и
- скалярные или функциональные ответы с функциональными и скалярными ковариатами.

Кроме того, модели функциональной регрессии могут быть линейными, частично линейными или нелинейными. В частности, функциональные полиномиальные модели, функциональные модели с одним и несколькими индексами и функциональные аддитивные модели являются тремя частными случаями функциональных нелинейных моделей.

В статье для описания случайной неопределенности во входных и выходных переменных на этапе преобразования

данных предлагается использовать переменные, которые представляют собой математические модели функций плотности вероятности соответствующих переменных, построенные по эмпирическим данным в классе кусочно-полиномиальных моделей. Для вычисления неизвестных параметров модели предлагается использовать численный вероятностный анализ, в котором имеются соответствующие арифметики и процедуры [5,6,7].

В рамках применения данного подхода рассматриваются новые методы моделирования функциональных зависимостей на основе сплайн аппроксимаций [1]. Для исследования точности вычислений используется метод построения апостериорных оценок [5,6].

Сформулируем функциональную регрессию как регрессию в пространстве распределений.

Пусть известны значения (y_i, x_i) $i = 1, 2, ..., N$. Будем считать, что случайная величина y_i распределена по закону $\boldsymbol{y}(x_i)$, семейство функций плотности вероятности $\boldsymbol{y}(x)$ зависят непрерывно от значений x. Необходимо по данным (y_i, x_i) оценить плотность вероятности $\boldsymbol{y}(x)$ и построить функциональную регрессию.

Для построения оценки $\boldsymbol{y}(x)$ в некоторой точке x_0 зададимся параметром $h > 0$ и построим на отрезке $[x_0 - h, x_0 + h]$ по данным $D_h = \{(y_i, x_i) \mid x_i \in [x_0 - h, x_0 + h]\}$ регрессию $r(x)$. Далее построим выборку $Z_h = \{z_i = y_i - r(x_i) \mid x_i \in [x_0 - h, x_0 + h]\}$. По Z_h используя ядерные оценки построим приближение $y^h(\cdot, x_0) \approx y(\cdot, x_0)$. Заметим, что в данном случае мы строим оценку $y^h(\xi, x_0)$:

$$y^h(\xi, x_0) = \frac{1}{2h} \int_{x_0 - h}^{x_0 + h} y(\xi, x) dx.$$

Несложно видеть, что

$$y(\xi, x_0) = y^h(\xi, x_0) + Ch^2 + O(h^4),$$

где C — константа, независящая от h. В этом случае для повышения точности можно использовать экстраполяцию Ричардсона [13]. Для этого построим оценки для h и $2h$: $y^h(\xi, x_0)$ и $y^{2h}(\xi, x_0)$. Далее

$$y(\xi, x_0) = \frac{4}{3} y^h(\xi, x_0) - \frac{1}{3} y^{2h}(\xi, x_0) + O(h^4).$$

В этом случае будет справедлива оценка

$$y(\xi, x_0) = \frac{4}{3} y^h(\xi, x_0) - \frac{1}{3} y^{2h}(\xi, x_0) + O(h^4).$$

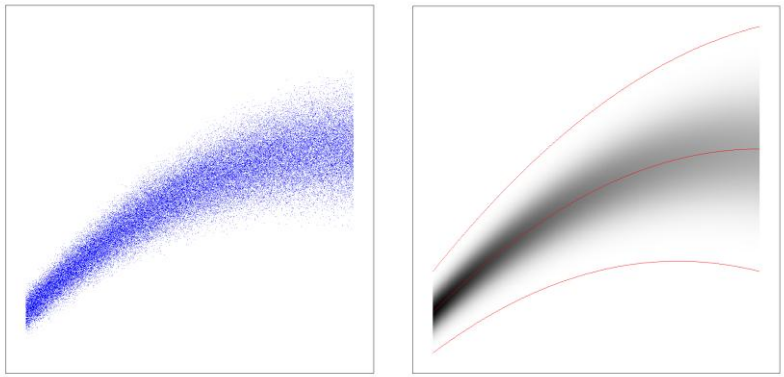


Рис. 2. Модельная задача. Левый рисунок – исходные данные, правый – функциональная регрессия

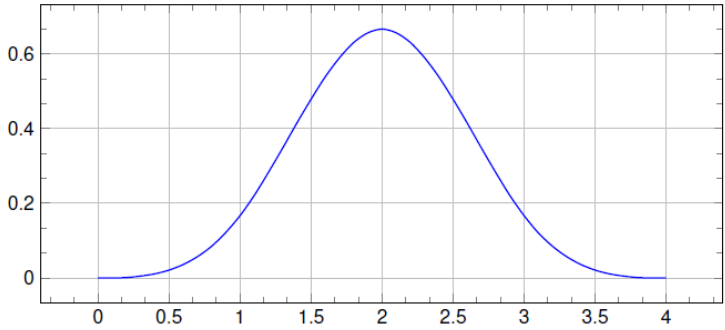


Рис.3. Сплайновая аппроксимация функции плотности вероятности сечения в некоторой точке

Рассмотрим модельную задачу. На левом рисунке 2 представлены данные (размерность выборки 10^5), выборка сгенерирована используя распределение Ирвина-Холла n=4. На правом рисунке 2 оттенками серого цвета представлена восстановленная плотность вероятности.

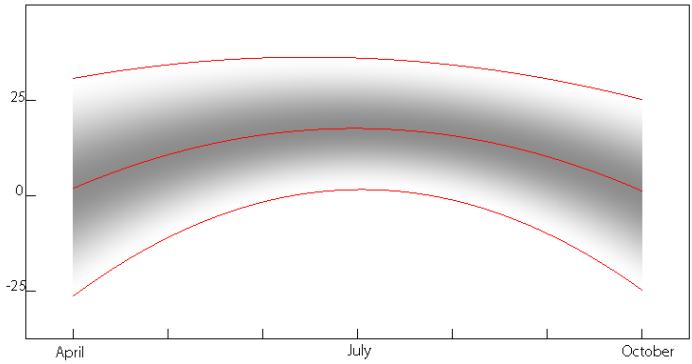


Рис.4. Функциональная регрессия температуры в г. Красноярске с апреля по октябрь за 70 лет.

Рассмотрим функциональную регрессию на данных о температуре в городе Красноярске за последние 70 лет. Для каждого дня с апреля по октябрь данные агрегировались в виде эрми-

товых кубических сплайнов. В этом случае регрессионная модель была представлена в виде

$Y = A\varphi_1(t) + B\varphi_2(t) + C\varphi_3(t)$,

где A, B, C функции плотности вероятности, φ_1, φ_2, φ_3 — квадратичные функции. На рисунке 4 показана функциональная регрессия на данных температуры в г. Красноярске за последние 70 лет. Красные линии — верхняя и нижняя граница области распределений, красная линия внутри области — регрессионная кривая. Оттенками серого показаны значения функции плотности распределения.

3. Оценка рисков инвестиционных проектов

Мы рассматриваем оценку рисков инвестиционных проектов, в которых рассчитываются функции плотности вероятности таких факторов, как чистая текущая стоимость (NPV) и внутренняя норма доходности (IRR).

Чистая приведенная стоимость (NPV) — это формула, используемая для определения текущей стоимости инвестиций по дисконтированной сумме всех денежных потоков, полученных от проекта. Формула для дисконтированной суммы всех денежных потоков может быть переписана как

$$NPV(r) = Cz_1 s_1 \sum_{i=1}^{T} \frac{C_i}{(1+r)^i} - C_0, \quad (2)$$

где $-C_0$ — начальные инвестиции, C_i — денежные потоки, T — время, r — ставка дисконтирования.

IRR определяет максимально допустимую ставку дисконтирования, при которой вы можете инвестировать без каких-либо потерь владельцу: IRR, в котором

$$NPV(r) = 0. \quad (3)$$

В качестве примера, рассмотрим компанию, которая начинает новый проект. Компании необходимо определить инвестиции на развитие своего нового продукта. По оценкам компании, денежный поток будет иметь вид $C_i = c_i \cdot x_i$, где c_i цена и x_i

объем продаж. Заметим, что будущие объемы продаж x_i — неизвестные случайные величины, цена определяется менеджерами и является функцией от объема продаж. Таким образом, для оценки выражения (2) необходимо построить совместную функцию плотности вероятности $p(x_1,x_2,\ldots,x_T)$. Используя Big Date, мы можем построить приведенные объемы продаж фирм-аналогов [3,17,18]. На рисунке 5 показаны приведенные объемы продаж фирм-аналогов. Далее, используя процедуры агрегации [6,7], построена аппроксимация совместной плотности вероятности $p(x_1,x_2,\ldots,x_T)$.

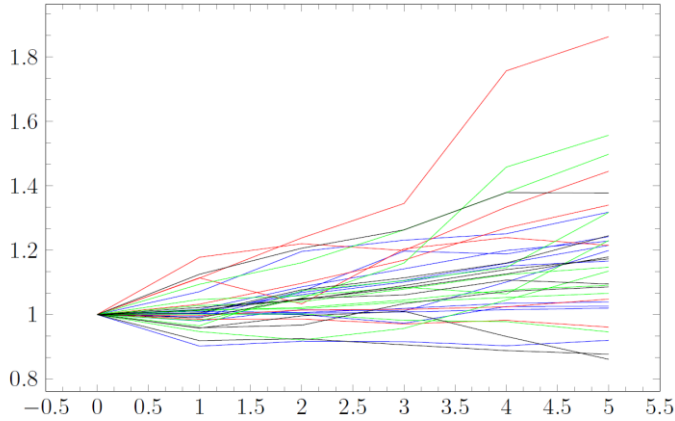
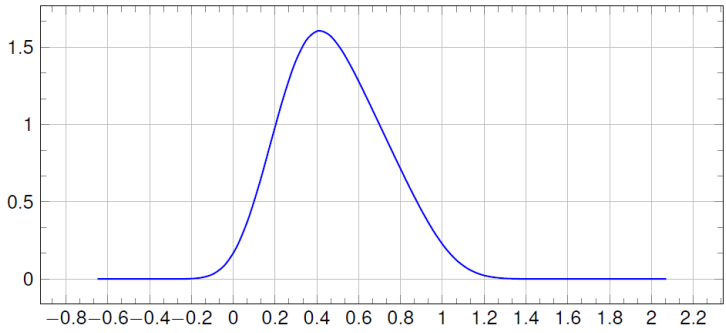


Рис. 5. Приведенные объемы продаж фирм-аналогов



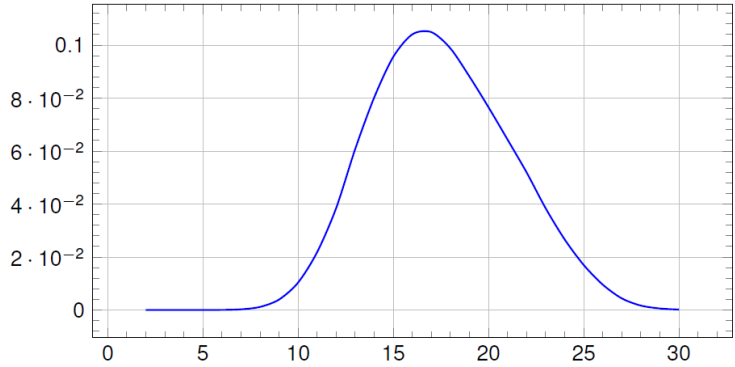


Рис. 6. Функции плотности вероятности NPV и IRR

Далее, на основе вычислительного вероятностного анализа, строим плотности вероятности NPV и IRR. На рисунке 6 приведены оценки плотности вероятности NPV и IRR.

Используя оценки плотности вероятности NPV и IRR в форме кубических сплайнов, мы можем оценить риск того, что инвестиционный проект приносит убытки. Итак, если P_{NPV} функция плотности вероятности NPV, то вероятность того, что инвестиционный проект является убыточным, можно рассчитать по формуле

$$P = \int_{-\infty}^{0} P_{NPV}(\xi)\, d\xi.$$

Заключение

В статье рассмотрены новые методы к обработке и анализу данных больших объемов на основе процедур агрегирования и рекурсивно-параллельной организации вычислительного процесса. Такой подход представляет собой технику быстрых вычислений и позволяет решать актуальную проблему надежности результатов численного моделирования, обеспеченных надежными вычислительными процедурами, адекватными типам неопределенности, которые содержатся в сырых данных.

Сравнение методов ВВА и Монте-Карло показало хорошее согласие результатов. В то же время численные эксперименты

показывают, что ВВА значительно быстрее метода Монте-Карло. В результате подход, основанный на ВВА, может быть успешно применен для решения ряда экономических проблем.

Литература

1. ПОПОВА О.А. *Применение численного вероятностного анализа в задачах интерполяции* // Вычислительные технологии. Т. 22. 2017. №2. С. 99–114. 1.
2. ХАРДЛЕ В. *Прикладная непараметрическая регрессия*: Пер. с англ. М., Мир, 1993. С. 349
3. *Computational economics: a perspective from computational intelligence* / Shu-Heng chen and Lakhmi Jain, editors. London. Idea Group Inc. 2006. p. 339
4. *Digital Economy: Impacts, Influences and Challenges*. Harbhajan Kehal, editor, Varinder P. Singh, editor. Idea Group Publishing. 2005. p. 425.
5. DOBRONETS B.S., POPOVA O.A. *Improving the accuracy of the probability density function estimation* // Journal of Siberian Federal University — Mathematics and Physics. 2017 1 16–21.
6. DOBRONETS B.S., POPOVA O.A. *Improving reliability of aggregation, numerical simulation and analysis of complex system by empirical data //* IOP Conf. Series: Materials Science and Engineering 354 (2018) 012006 doi:10.1088/1757-899X/354/1/012006
7. DOBRONETS B.S., POPOVA O.A. *Piecewise Polynomial Aggregation as Preprocessing for Data Numerical Modeling* // IOP Conf. Series: Journal of Physics: Conf. Series 1015 (2018) 032028 doi :10.1088/1742-6596/1015/3/032028
8. DOBRONETS B.S., POPOVA O.A. *Numerical Probabilistic Approach for Optimization Problems* // Scientific Computing, Computer Arithmetic, and Validated Numerics. Lecture Notes in Computer Science. 2016. Vol. 9553 pp. 43–53

9. DOBRONETS B.S., POPOVA O.A. *Numerical Probabilistic Analysis Under Aleatory and Epistemic Uncertainty//* Reliable Computing. 2014. Vol. 19. № 3, pp. 274-289
10. DOBRONETS B.S., POPOVA O.A. *Computational Aspects of Probabilistic Extensions* // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2019. № 47. С. 41-48.
11. KIM HUA TAN, GUOJUN JI, CHEE PENG LIM & MING-LANG TSENG *Using big data to make better decisions in the digital economy* // International Journal of Production Research, 2017. 55:17, 4998-5000, DOI: 10.1080/00207543.2017.1331051
12. MAYER-SCHONBERGER V., CUKIER K. *Big Data: A Revolution That Will Transform How We Live,* Work, and Think. New York, NY: Houghton Miffin Harcourt. 2013. p. 242
13. POPOVA O.A. *Using Richardson Extrapolation to Improve the Accuracy of Processing and Analyzing Empirical Data* // Measurement Techniques, Vol. 62, No. 2, May, 2019 DOI 10.1007/s11018-019-01594-1
14. RAMSAY J.O. When the data are functions // *Psychometrika*, 47:379–96, 1982.
15. RAMSAY J.O. and DALZELL C. Some tools for functional data analysis // *J. R. Stat. Soc. Ser. B*, 53:539–72, 1991.
16. RAMSAY J.O. and SILVERMAN B.W. *Functional Data Analysis*. Springer 2nd ed., New York, 2005.
17. SHU-HENG CHEN *Computational intelligence in economics and finance: Carrying on the legacy of Herbert Simon* // Information Sciences 170, 2005, p.121–131
18. VALERIU I. *Economic Intelligence* // Journal of Knowledge Management, Economics and Information Technology. Special Issue December. 2013. p. 182–198

COMPUTATIONAL ASPECTS OF DIGITAL ECONOMY

Dobronets B.S., Popova O.A.
(Siberian Federal University, Krasnoyarsk)
BDobronets@yandex.ru, OlgaArc@yandex.ru

The article discusses the main computational problems in the digital economy, which are primarily associated with the processing and analysis of big data, the organization of computational processes and improving the accuracy of numerical procedures. The approach is based on the application of new methods of data aggregation and computational probabilistic analysis, the use of probabilistic extensions and numerical operations on piecewise polynomial functions.
We suggest using functional regression on empirical distributions to identify dependencies in big data.

Keywords: digital economy, big data, computational probabilistic analysis, functional regression