ЭКОНОМИКА И МАТЕМАТИЧЕСКИЕ МЕТОДЫ, 2020, том 56, № 3, с. 113–124

МАТЕМАТИЧЕСКИЙ АНАЛИЗ ЭКОНОМИЧЕСКИХ МОДЕЛЕЙ

Метод восстановления функции по интегралам для анализа и прогнозирования редких событий в экономике

Ю.А. Кораблев,
Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Москва; e-mail: yura-korablyov@yandex.ru

Поступила в редакцию 18.07.2019

Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект 19-010-00154).

Аннотация. В статье рассматривается метод анализа редких событий, который базируется на изучении процессов, порождающих эти события. В экономике самым распространенным процессом образования событий являются процессы потребления или накопления возмущения, которые можно моделировать как процесс опустошения/наполнения «емкости». Параметром процесса образования событий будет нестационарная функция скорости опустошения/наполнения емкости, которую можно восстановить из имеющихся данных. После этого с ней можно проводить необходимые для исследователя действия (анализировать и экстраполировать функцию, построить модель, получить прогноз будущих событий) и снова запустить процесс образования событий. Такой метод исследования редких событий мною был назван емкостным методом. Для восстановления функции скорости опустошения/наполнения/ емкости в статье приведена оптимизационная задача в виде нахождения специального сглаживающего интегрирующего кубического сплайна. Получены формулы в матричном виде для восстановления (регрессии) искомой функции. Так как интервалы между событиями, как правило, могут быть разными, следует переходить к базисным сплайнам (B-сплайнам), которые не зависят от исходных данных. Получены формулы в матричном виде для построения соответствующего B-сплайна. Подробно показано, как следует заполнять такие матрицы. Приведен пример использования математического метода восстановления функции по данным редких событий и получения прогноза будущих событий.

Ключевые слова: редкие события, емкостный метод, скорость потребления, восстановление, регрессия, сплайн, B-сплайн, интегрирующий сплайн, интегро-дифференциальный сплайн, штраф на нелинейность.

Классификация JEL: C1, C15, C4, C5, C53.

DOI: 10.31857/S042473880010485-2

1. ВВЕДЕНИЕ

Анализ и прогнозирование событий позволяет должным образом к ним подготовиться, что способствует уменьшению возможных потерь или увеличению прибыли. Для этого могут использоваться различные математические методы, среди которых выделяют методы работы с *редкими событиями*. Редкие события отличаются от частых, как правило, тем, что представляются в виде потоков дискретных событий, возникающих через случайные периоды времени, а не в виде числа событий за период времени (или временного ряда). Время между событиями может быть произвольным (дни, года, микросекунды, при этом события будут по-прежнему относиться к редким). Важным является способ представления данных.

Представление редких событий в виде временного ряда приведет к тому, что такой временной ряд будет содержать множество нулей. Тем не менее некоторые методы работают и с такими рядами. Иногда для этого адаптируются методы классификации. Например, метод «ближайших соседей» (Altman, 1992; Cover, Hart, 1967) ищет в наблюдениях подпоследовательности, похожие на вектор предшествующих значений фиксированной длины, после чего возвращает прогноз как значение, следующее за наиболее похожей подпоследовательностью.

Если временной ряд состоит из нулей и единиц, то иногда применяют метод логистической регрессии (Walker, Duncan, 1967), когда по набору данных внешних признаков строится классификационная модель, которая показывает, что при заданных признаках должна появиться единица или ноль. Иногда используют нейронные сети (Барцев, Охонин, 1986; Rumelhart, Hinton, Williams, 1986), которые строят модель, но уже скрытым от исследователя способом. Метод Кростона (Croston, 1972; Johnston, Boylan, 1996) предполагает разделение исходного ряда данных на два — ряд из ненулевых значений и ряд длительности между ненулевыми значениями,— после чего проводится экспоненциальное сглаживание каждого ряда, а прогнозное значение получается как ожидаемое ненулевое значение через ожидаемое число нулевых значений.

В логистике, когда надо определить размер запаса, достаточного для удовлетворения спроса для заданного числа периодов, иногда используется метод бутстрэппинга (Виллемейна) (Efron, Tibshirani, 1993; Willemain, Park, Kim, Shin, 2001). Для этого из имеющихся наблюдений случайным образом извлекают число значений, соответствующих числу периодов, и суммируют их; эту процедуру многократно повторяют, а затем строят функцию распределения для этой суммы значений. Размер запаса устанавливается на уровне, который обеспечит удовлетворение спроса с заданной доверительной вероятностью. Иногда для этого применяют селективные методы (Иванько, 2005), которые переключают модели прогнозирования по значению ошибки прогноза на предыдущем шаге.

Перечисленные методы работают с временными рядами, содержащими большое число нулевых значений. Однако наиболее обоснованным является представление событий в виде потоков дискретных событий, которые появляются через произвольные периоды времени. Для работы с данными в виде потоков событий используется теория случайных процессов (Вентцель, Овчаров, 2000). Потоки событий представляются в виде пуассоновского потока, когда время между событиями подчиняется экспоненциальному распределению, или в более сложном варианте — потоком Пальма с ограниченным последействием (здесь время между событиями соответствует произвольному закону распределения). Иногда для моделирования сверхредких событий вводят модифицированные пуассоновские процессы (Дзанагова, Хугаева, 2015). На практике чаще всего применяют классические пуассоновские процессы, когда на основе статистических данных редких продаж определяют параметры потока событий, после чего рассчитывают размер собственных запасов, зная вероятности возникновения определенного числа событий за выбранный период времени (Лукинский, Замалетдинова, 2015; Вожжов А., Луняков, Вожжов С., 2015). С помощью пуассоновских потоков можно определить вероятность появления заданного числа событий на выбранном интервале времени, а с помощью потоков Пальма — ожидаемое оставшееся время до следующего события (однако потоки Пальма являются стационарными и подходят только для случаев с постоянной интенсивностью). Использование нестационарных непуассоновских потоков не встречается.

У каждого метода есть своя область применения, в которой он может дать хорошие результаты. Причем для одних и тех же задач иногда можно применять разные методы, но их эффективность будет разной. Также существуют условия, для которых методы еще не разработаны. Разработка новых методов, которые дадут новые возможности либо будут более эффективными, — есть цель науки.

2. ОСНОВНАЯ ИДЕЯ

Почему процесс возникновения событий представляется случайным? Почему интервалы между событиями должны быть случайными числами? Неужели нет информации о том, как возникают эти события? Почему из статистических данных определяется закон распределения случайных интервалов времени, а не процесс, который порождает эти события? Используя знания о характере процесса, определяя из статистических данных его параметры и закономерности, а затем экстраполируя параметры процесса на будущее время, можно получить более точный прогноз возникновения будущих событий (рис. 1). Информация о процессе формирования событий способна избавить нас от неопределенности при их появлении. События формируются уже не случайным образом, не через абсолютно случайные периоды времени, а по определенному механизму, параметры которого стали известны из статистических данных.

Самыми распространенными причинами появления событий в экономике могут быть процессы потребления (запас ведет себя как опустошающаяся емкость) и процессы накопления некоторого возмущения до определенного уровня, вследствие чего возникает некоторое событие. В обоих

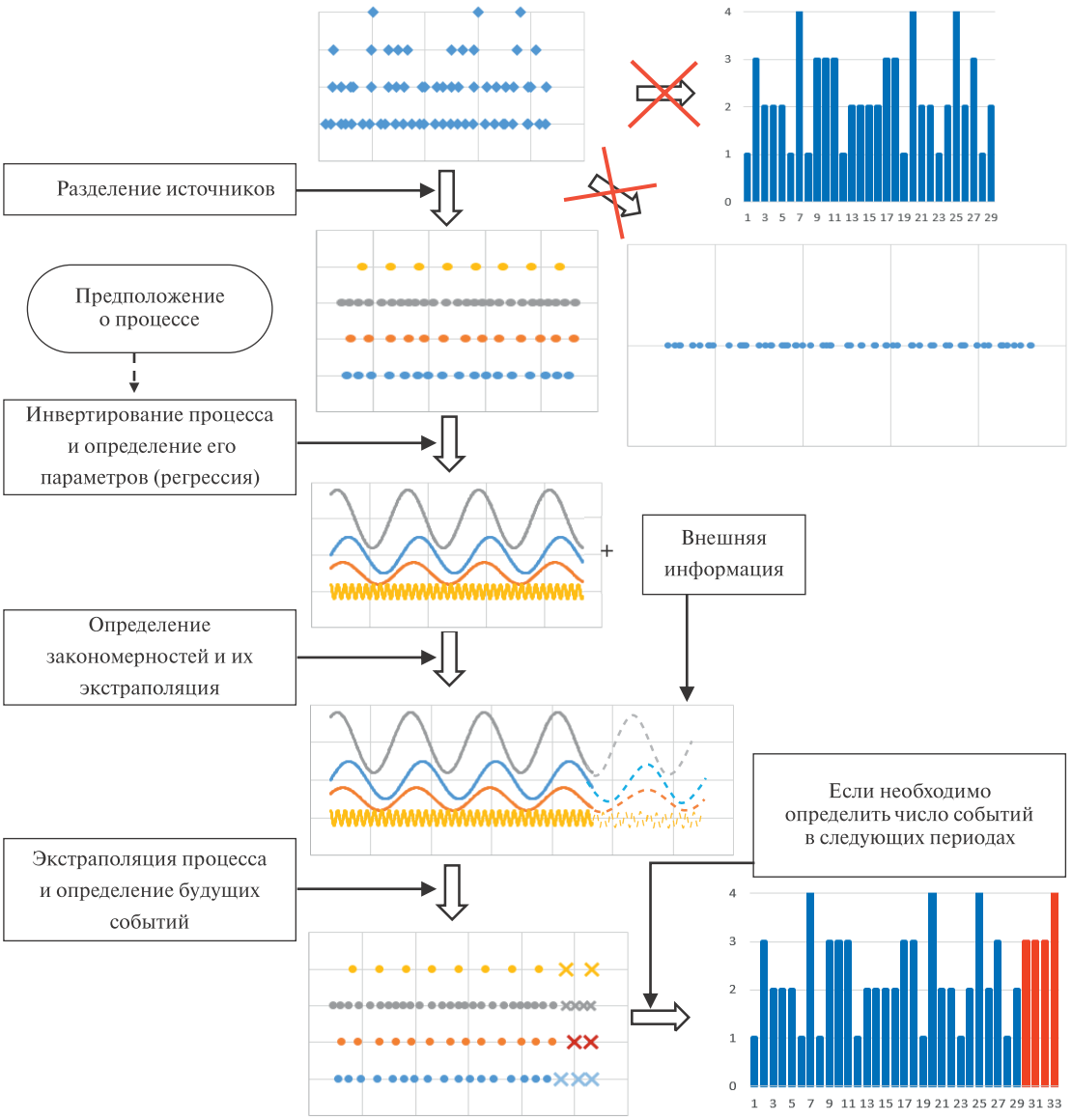

Рис. 1. Схема анализа и прогнозирования редких событий

вариантах источники событий[1] можно моделировать как емкости. Предложенный метод анализа и прогнозирования редких событий получил название «емкостный метод» (Кораблев, 2015а, 2015б, 2018, 2019а, 2019б). Согласно этому методу параметром процесса образования событий является нестационарная функция скорости расхода запаса или накопления воздействия $f(t)$, подлежащая определению. Такой функцией может быть, например, спрос, зависящий от времени, индивидуальная скорость потребления продукции, интенсивность покупок у выбранного не подконтрольного нам оптового покупателя (ненаблюдаемые значения).

Оказывается, что из данных редких событий можно легко восстановить функцию $f(t)$. Для этого инвертируем процесс потребления продукции и получаем задачу, обратную к задаче управления запасами (алгоритм в минус первой степени), когда по имеющимся данным о моментах времени и величинах воздействия события (покупок) (t_i, y_i) определяется скорость воздействия $f(t)$. Для этого используем основное предположение.

[1] Под источниками события понимаются некоторые объекты или системы, в которых происходят какие-то процессы, приводящие к возникновению этих событий.

Предположение. Величина совершенного события y_i есть интеграл функции $f(t)$ за время от момента возникновения этого события t_i до момента совершения следующего события t_{i+1}.

Для процессов потребления или накопления возмущения это предположение справедливо, оно также негласно применяется в теории управления запасами при моделировании собственных запасов (Бауэрсокс, Клосс, 2008). Изменение предпочтений потребителей не нарушает этого предположения, а выражается в изменении функции $f(t)$. Конечно, для отдельных видов товаров или услуг это предположение может выполняться не всегда, а при определенном поведении потребителей оно может нарушаться. Но в данной работе мы будем считать, что предположение в большей степени выполняется, пусть и с погрешностью $y_i = \int_{t_i}^{t_{i+1}} f(t)dt + \varepsilon_i$.

Используя это предположение, задача определения (регрессии) $f(t)$ превращается в оптимизационную задачу восстановления неизвестной функции, для которой известна последовательность интегралов за непересекающиеся периоды времени, с дополнительным штрафом на нелинейность (C — параметр, влияющий на степень сглаживания, n — размер выборки):

$$\sum_{i=1}^{n-1}\left(y_i - \int_{t_i}^{t_{i+1}} f(t)dt\right)^2 + C\int_{t_1}^{t_n}\left(f''(t)\right)^2 dt \to \min. \qquad (1)$$

Нам необходимо найти решение этой оптимизационной задачи и продемонстрировать работу метода для событий, которые образуются процессами, схожими с процессами опустошения емкости.

3. ПОСТРОЕНИЕ ИНТЕГРАЛЬНОГО СПЛАЙНА

Наиболее подробно задачи восстановления функций по интегралам изучены в работах (Киреев, 1994; Киреев, Бирюкова, 1998, 2014; Бирюкова, Киреев, Гершкович, 2016). Исследования этих авторов посвящены сплайнам, построение которых зависит одновременно от интегралов и дифференциалов. Такие сплайны получили название *интегро-дифференциальные сплайны*, или ИД-сплайны. Однако в этих работах сплайн строится на основе системы уравнений, состоящей из условий согласования $y_i = \int_{t_i}^{t_{i+1}} f(t)dt$ в виде точных равенств, что, по сути, является интерполяцией интегралов, а не их аппроксимацией. Кроме того, в них строятся параболические сплайны, а не кубические. В работе (Boor, 2001, p. 79) также рассматривается интерполяционный параболический сплайн, а не сглаживающий кубический. В работах (Федорова, 2008, 2016) строится одномерный и двумерный сплайны по известной площади под кривой закона распределения, однако в этой работе сплайн также является интерполяционным, а не аппроксимирующим. Готового решения нашей задачи мне найти не удалось, поэтому его пришлось разрабатывать самостоятельно.

Мой метод базируется на методе аппроксимации кубическими сплайнами обычных функций (не интегралов функции) со штрафом на нелинейность[2], но модифицируется для работы с интегралами функции. Решение ищется в виде $f(t) = g(t)$, где $g(t)$ — кубический сплайн[3], причем на каждом участке функция записывается не как полином с четырьмя неизвестными коэффициентами, а выражается только через две переменные — значение функции в точке $g_i = g(t_i)$ и ее вторую производную в этой точке $\gamma_i = g''(t_i)$. Значение сплайна в произвольной точке определяется по формуле

$$g(t) = \frac{(t-t_i)g_{i+1} + (t_{i+1}-t)g_i}{t_{i+1}-t_i} - \frac{1}{6}(t-t_i)(t_{i+1}-t)\left\{\left(1+\frac{t-t_i}{t_{i+1}-t_i}\right)\gamma_{i+1} + \left(1+\frac{t_{i+1}-t}{t_{i+1}-t_i}\right)\gamma_i\right\}, \qquad (2)$$

$$i: \quad t_i \le t \le t_{i+1}.$$

Набор всех значений $g = (g_1,\ldots,g_n)^T$, $\gamma = (\gamma_2,\ldots,\gamma_{n-1})^T$ (в начальной и последней точке $\gamma_1 = \gamma_n = 0$) полностью задает весь сплайн. Условия непрерывности первой производной в точках сочленения

[2] В великолепно написанной работе (Green, Silverman, 1994) представлено необходимое объяснение всей теории.

[3] Сочленение кусочков из полиномов третьей степени в точках t_i с условием непрерывности как самой функции, так и ее производной в точках сочленения.

$g'(t_i+0)=g'(t_i-0)$, $i=2,\ldots,n-1$ дают систему из $n-2$ уравнений, которая может быть записана в матричном виде через матрицы коэффициентов Q, R при неизвестных g_i, γ_i:

$$\begin{aligned}&\frac{g_{i+1}-g_i}{t_{i+1}-t_i}-\frac{g_i-g_{i-1}}{t_i-t_{i-1}}=\left\{(t_{i+1}-t_i)(\gamma_{i+1}+2\gamma_i)+(t_i-t_{i-1})(2\gamma_i+\gamma_{i-1})\right\}/6,\quad i=2,\ldots,n-1,\\&Q^Tg=R\gamma,\end{aligned}\tag{3}$$

где матрица Q размерностью $n\times(n-2)$ и R размерностью $(n-2)\times(n-2)$ имеют вид:

$$Q=\begin{pmatrix}h_1^{-1} & 0 & \ldots & 0\\ -h_1^{-1}-h_2^{-1} & h_2^{-1} & \ldots & 0\\ h_2^{-1} & -h_2^{-1}-h_3^{-1} & \ldots & 0\\ 0 & h_3^{-1} & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & \ldots & h_{n-2}^{-1}\\ 0 & 0 & \ldots & -h_{n-2}^{-1}-h_{n-1}^{-1}\\ 0 & 0 & \ldots & h_{n-1}^{-1}\end{pmatrix},$$

$$R=\begin{pmatrix}(h_1+h_2)/3 & h_2/6 & 0 & \ldots & 0\\ h_2/6 & (h_2+h_3)/3 & h_3/6 & \ldots & 0\\ 0 & h_3/6 & (h_3+h_4)/3 & \ldots & 0\\ 0 & 0 & h_4/6 & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & h_{n-2}/6\\ 0 & 0 & 0 & \ldots & (h_{n-2}+h_{n-1})/3\end{pmatrix},$$

$$h_i=t_{i+1}-t_i,\ \ i=1,\ldots,n-1.$$

Штраф на нелинейность $\int_a^b\left(g''(x)\right)^2dx$ упрощается (Green, Silverman, 1994, p. 24–25):

$$\int_a^b\left(g''(x)\right)^2dx=\gamma^TQ^Tg=\gamma^TR\gamma=g^T\left[QR^{-1}Q^T\right]g=g^TKg.\tag{4}$$

Для решения задачи (1), где $f(t)=g(t)$, найдем интеграл $\int_{t_i}^{t_{i+1}}g(t)dt$, где $g(t)$ определяется через искомые неизвестные g_i, γ_i по формуле (2). После преобразований получаем формулу:

$$\int_{t_i}^{t_{i+1}}g(t)dt=\frac{g_{i+1}h_i}{2}+\frac{g_ih_i}{2}-\frac{\gamma_{i+1}h_i^3}{24}-\frac{\gamma_ih_i^3}{24}.\tag{5}$$

Тогда оптимизационная задача (1) для искомых g и γ может быть записана в виде

$$S(g)=(Y-Vg+P\gamma)^T(Y-Vg+P\gamma)+\alpha g^TKg\to\min,\tag{6}$$

где $Y=(y_1,\ldots,y_{n-1})^T$; V — матрица размера $(n-1)\times n$ и P — матрица размера $(n-1)\times(n-2)$ являются матрицами коэффициентов при неизвестных g и γ:

$$V=\frac{1}{2}\begin{pmatrix}h_1 & h_1 & 0 & \cdots & 0 & 0\\ 0 & h_2 & h_2 & \cdots & 0 & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & 0 & h_{n-1} & h_{n-1}\end{pmatrix},\quad P=\frac{1}{24}\begin{pmatrix}h_1^3 & 0 & 0 & 0 & \cdots & 0 & 0\\ h_2^3 & h_2^3 & 0 & 0 & \cdots & 0 & 0\\ 0 & h_3^3 & h_3^3 & 0 & \cdots & 0 & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & 0 & 0 & h_{n-2}^3 & h_{n-2}^3\\ 0 & 0 & 0 & 0 & 0 & 0 & h_{n-1}^3\end{pmatrix}.$$

Далее, благодаря тому, что условия непрерывности по-прежнему дают систему уравнений $Q^T g = R\gamma$, выражая $\gamma = R^{-1}Q^T g$, перепишем формулу (6) так, чтобы в ней была только одна неизвестная:

$$S(g) = \left(Y - \left(V - PR^{-1}Q^T\right)g\right)^T \left(Y - \left(V - PR^{-1}Q^T\right)g\right) + \alpha g^T Kg = \left(Y - Cg\right)^T \left(Y - Cg\right) + \alpha g^T Kg, \qquad (7)$$

где $C = V - PR^{-1}Q^T$ матрица размера $(n-1)\times n$. Для нахождения минимума выражения (7) раскроем скобки, перегруппируем слагаемые и приравняем производную по g к нулю ($d\left(x^T b\right)/dx = b$, $d\left(bx\right)/dx = b^T$, а если матрица симметрична (что у нас выполняется), то $d\left(x^T Ax\right)/dx = \left(A + A^T\right)x = 2Ax$):

$$S(g) = g^T\left(C^T C + \alpha K\right)g - 2Y^T Cg + Y^T Y, \quad S'(g) = 2\left(C^T C + \alpha K\right)g - 2\left(Y^T C\right)^T = 0,$$
$$g = \left(C^T C + \alpha K\right)^{-1} C^T Y. \qquad (8)$$

На этом сплайн полностью построен (значения g и $\gamma = R^{-1}Q^T g$ задают сплайн $g(t)$).

Заметим, что исходные матрицы Q, R, V, P (из которых также получаются $K = QR^{-1}Q^T$ и $C = V - PR^{-1}Q^T$) зависят только от интервала между наблюдениями $h_i = t_{i+1} - t_i$, но не зависят от значений в этих наблюдениях y_i, а значения $Y = \left(y_1, \ldots, y_{n-1}\right)^T$ участвуют только в выражении (8).

Пример использования интегрального сплайна. Пусть нам известны данные $\left(t_i, y_i\right)$ о датах и объемах поставок полуторалитровых бутылок кваса в универсам (табл. 1). По ним можно построить график (рис. 2), на котором ступенчатой линией показано среднее число проданных за день бутылок. Гладкая линия обозначает аппроксимирующий сплайн, который минимизирует разницу между интегралами функции и объемом поставки (площадь под ступенькой). Большое расхождение в ширине интервалов наблюдений (куски сплайнов имеют разную ширину) и неудачный выбор параметра α могут влиять на сглаживающие свойства сплайна (местами функция становится отрицательной, что противоречит физическому смыслу). Также при очень больших наборах данных, когда кусков сплайна, привязанных к точкам наблюдения, становится очень много, вычисления могут быть очень трудоемкими. Желательно, чтобы участки сплайна не были привязаны к точкам наблюдения.

Таблица 1. Данные о поставках бутылок кваса в универсам

Дата	Поставки	Дата	Поставки	Дата	Поставки
02.02.2018	12	28.05.2018	60	12.11.2018	18
12.02.2018	12	18.06.2018	18	17.12.2018	42
26.02.2018	24	29.06.2018	60	27.12.2018	18
12.03.2018	12	16.07.2018	54	14.01.2019	12
26.03.2018	18	30.07.2018	24	11.02.2019	18
09.04.2018	36	06.08.2018	30	04.03.2019	18
23.04.2018	18	20.08.2018	30	11.03.2019	6
07.05.2018	60	03.09.2018	48		
14.05.2018	60	29.10.2018	24		

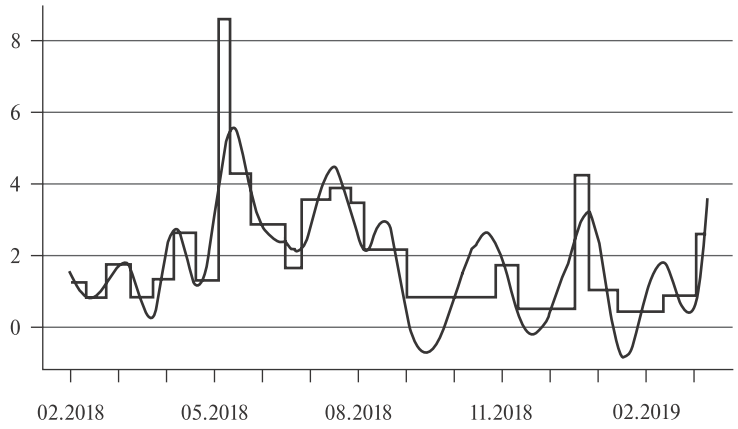

Рис. 2. Скорость расхода бутылок кваса универсамом, шт. в день

4. ПЕРЕХОД К БАЗИСНОМУ СПЛАЙНУ

Чтобы куски сплайнов не были привязаны к точкам наблюдения, следует перейти к базисному сплайну (B-сплайну), состоящему из набора m базисных функций $\beta_k(t)$, которые, как правило, тоже будут полиномами, но определенными в произвольных точках $s_1 < \ldots < s_m$ (чаще всего распределенными равномерно). Каждая функция $\beta_k(t)$ берется с некоторым коэффициентом δ_k, который является некоторым индикатором, принимающим значение 0 или 1 в зависимости от того, какая функция соответствует текущему моменту времени, $g(t) = \sum_{k=1}^{m} \delta_k \beta_k(t)$.

Дополнительно добавим возможность задавать вес каждого наблюдения w_i. Тогда оптимизационная задача примет вид

$$S_W(g) = \sum_{i=1}^{n-1} w_i \left\{ y_i - \int_{t_i}^{t_{i+1}} \sum_{k=1}^{m} \delta_k \beta_k(t)\, dt \right\}^2 + \alpha \int_{t_1}^{t_n} \left(\left(\sum_{k=1}^{m} \delta_k \beta_k(t) \right)'' \right)^2 dt \to \min. \tag{9}$$

Для ее решения надо найти значения сплайна $g = (g_1, \ldots, g_m)^T$ и его вторых производных $\gamma = (\gamma_2, \ldots, \gamma_{m-1})^T$, но уже в новых точках $s_1 < s_2 < \ldots < s_m$.

Штраф на нелинейность по-прежнему будет выражаться как $\alpha g^T K g$, где $K = QR^{-1}Q^T$, но при этом размерность матриц Q и R будет зависеть не от n, а от m, а элементы — от расстояния между новыми точками, где $h_k = s_{k+1} - s_k$, $k = 1, \ldots, m-1$.

Рассчитаем интеграл $\int_{t_i}^{t_{i+1}} \sum_{k=1}^{m} \delta_k \beta_k(t)\, dt$. В зависимости от того, где появятся точки наблюдений (рис. 3) и как будут заданы новые точки сплайна, возможно несколько способов расчета.

Для того чтобы получить универсальное выражение, подходящее для всех трех случаев, представим интеграл в виде

$$\int_{t_i}^{t_{i+1}} \sum_{k=1}^{m} \delta_k \beta_k(t)\, dt = \sum_{l=0}^{L} \int_{s_{k+l}}^{s_{k+l+1}} \beta_{k+l}(t)\, dt - \int_{s_k}^{t_i} \beta_k(t)\, dt - \int_{t_{i+1}}^{s_{k+L+1}} \beta_{k+L}(t)\, dt, \tag{10}$$

$$L:\ s_{k+L} < t_{i+1} \le s_{k+L+1}.$$

Первая часть выражения (10) есть интеграл от всех L участков сплайна; вторая — интеграл от начала первой базисной функции k до текущего наблюдения i; третья — интеграл от наблюдения $i + 1$до конца последнего интервала $k + L$, на который попало следующее наблюдение. Значения k и L определяются в зависимости от того, куда попало текущее и следующее наблюдение.

Первая часть выражения находится из полученной ранее формулы, но границами интервала стали новые точки:

$$\sum_{l=0}^{L} \int_{s_{k+l}}^{s_{k+l+1}} \beta_{k+l}(t)\, dt = \sum_{l=0}^{L} \left[\frac{h_{k+l}}{2} g_{k+l+1} + \frac{h_{k+l}}{2} g_{k+l} - \frac{h_{k+l}^3}{24} \gamma_{k+l+1} - \frac{h_{k+l}^3}{24} \gamma_{k+l} \right]. \tag{11}$$

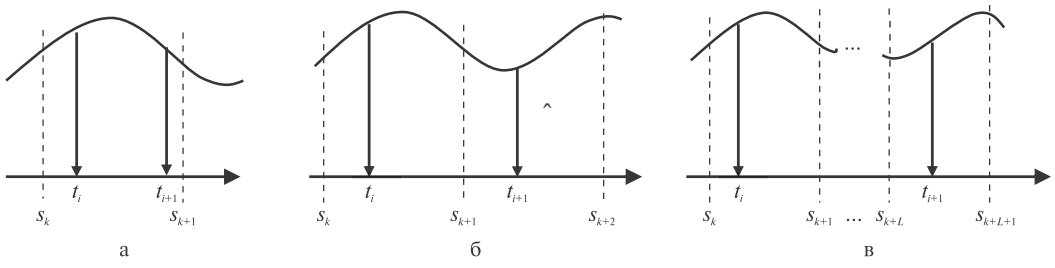

Рис. 3. Расположение соседних наблюдений на разных участках сплайна: а) в одном интервале; б) в двух соседних интервалах; в) в L интервалах друг от друга

После некоторых преобразований вторую часть можно представить в компактной форме

$$\int_{s_k}^{t_i}\beta_k(t)\,dt=\frac{\left(h_k^{-i}\right)^2}{2h_k}g_{k+1}+\frac{\left(h_k\right)^2-\left(h_k^{+i}\right)^2}{2h_k}g_k+\frac{\gamma_{k+1}}{24h_k}\left(h_k^{-i}\right)^2\left(\left(h_k^{-i}\right)^2-2\left(h_k\right)^2\right)-$$
$$-\frac{\gamma_k}{24h_k}\left(h_k^{-i}\right)^2\left(h_k^{+i}+h_k\right)^2,\quad h_k^{-i}=t_i-s_k,\ h_k^{+i}=s_{k+1}-t_i,\ h_k=s_{k+1}-s_k; \tag{12}$$

третья часть —

$$\int_{t_{i+1}}^{s_{k+L+1}}\beta_{k+L}(t)\,dt=\frac{g_{k+L+1}\left(\left(h_{k+L}\right)^2-\left(h_{k+L}^{-(i+1)}\right)^2\right)}{2h_{k+L}}+\frac{g_{k+L}\left(h_{k+L}^{+(i+1)}\right)^2}{2h_{k+L}}-\frac{\gamma_{k+L+1}\left(h_{k+L}^{+(i+1)}\right)^2\left(h_{k+L}^{-(i+1)}+h_{k+L}\right)^2}{24h_{k+L}}+$$
$$+\frac{\gamma_{k+L}\left(h_{k+L}^{+(i+1)}\right)^2\left(\left(h_{k+L}^{+(i+1)}\right)^2-2\left(h_{k+L}\right)^2\right)}{24h_{k+L}},\quad h_{k+L}^{-(i+1)}=t_{i+1}-s_{k+L},\ h_{k+L}^{+(i+1)}=s_{k+L+1}-t_{i+1}. \tag{13}$$

Подставляя выражения (11)–(13) в (10), можем найти $\int_{t_i}^{t_{i+1}}\sum_{k=1}^{m}\delta_k\beta_k(t)\,dt$. Как и раньше, форма этого выражения будет линейной по отношению к неизвестным g и γ. В результате оптимизационную задачу для нахождения искомого B-сплайна интегралов можно записать в знакомом виде $S_W(g)=(Y-Vg+P\gamma)^T W(Y-Vg+P\gamma)+\alpha g^T Kg\to\min$. Заполнение матриц V и P происходит на основе наблюдений о моментах времени возникновения текущего и следующего событий, в зависимости от того, на интервал какой базисной функции выпало это наблюдение.

Возможно, будет удобно воспользоваться следующим представлением: $V=G^{\mathrm{I}}-G^{\mathrm{II}}-G^{\mathrm{III}}$, $P=\Gamma^{\mathrm{I}}-\Gamma^{\mathrm{II}}-\Gamma^{\mathrm{III}}$, где матрицы G^{I}, G^{II}, G^{III} имеют размерность $(n-1)\times m$, Γ^{I}, Γ^{II}, Γ^{III} — размерность $(n-1)\times(m-2)$ (так как $\gamma_1=\gamma_m=0$ не участвуют). Элементы этих матриц заполняются по формулам:

$$G_{i,k}^{\mathrm{I}}=0{,}5h_k,\quad t_k\le t_i<t_{k+1};\quad G_{i,k+l}^{\mathrm{I}}=0{,}5\left(h_{k+l-1}+h_{k+l}\right),\ l=1,\ldots,L:t_k\le t_i,\ t_{k+L}\le t_{i+1}<t_{k+L+1};$$
$$G_{i,k+L+1}^{\mathrm{I}}=h_{k+L}/2,\ L:t_{k+L}\le t_{i+1}<t_{k+L+1};$$
$$G_{i,k}^{\mathrm{II}}=h_k/2-\left(h_k^{+i}\right)^2/2h_k,\ t_k\le t_i<t_{k+1};\quad G_{i,k+1}^{\mathrm{II}}=\left(h_k^{-i}\right)^2/2h_k,\ t_k\le t_i<t_{k+1};$$
$$G_{i,k+L}^{\mathrm{III}}=\left(h_{k+L}^{+(i+1)}\right)^2/2h_{k+L},\ t_{k+L}\le t_{i+1}<t_{k+L+1};\quad G_{i,k+L+1}^{\mathrm{III}}=h_{k+L}/2-\left(h_{k+L}^{-(i+1)}\right)^2/2h_{k+L},\ t_{k+L}\le t_{i+1}<t_{k+L+1};$$
$$\Gamma_{i,k}^{\mathrm{I}}=h_k{}^3/24,\ t_k\le t_i<t_{k+1};\quad \Gamma_{i,k+l}^{\mathrm{I}}=\left(h_{k+l-1}{}^3+h_{k+l}{}^3\right)/24,\ l=1,\ldots,L:t_k\le t_i,\ t_{k+L}\le t_{i+1}<t_{k+L+1};$$
$$\Gamma_{i,k+L+1}^{\mathrm{I}}=h_{k+L}{}^3/24,\ L:t_{k+L}\le t_{i+1}<t_{k+L+1};$$
$$\Gamma_{i,k}^{\mathrm{II}}=\left(h_k^{-i}\right)^2\left(h_k^{+i}+h_k\right)^2/24h_k,\ t_k\le t_i<t_{k+1};\quad \Gamma_{i,k+1}^{\mathrm{II}}=-\left(h_k^{-i}\right)^2\left(\left(h_k^{-i}\right)^2-2\left(h_k\right)^2\right)/24h_k,\ t_k\le t_i<t_{k+1};$$
$$\Gamma_{i,k+L}^{\mathrm{III}}=-\left(h_{k+L}^{+(i+1)}\right)^2\left(\left(h_{k+L}^{+(i+1)}\right)^2-2\left(h_{k+L}\right)^2\right)/24h_{k+L},\ t_{k+L}\le t_{i+1}<t_{k+L+1};$$
$$\Gamma_{i,k+L+1}^{\mathrm{III}}=\left(h_{k+L}^{+(i+1)}\right)^2\left(h_{k+L}^{-(i+1)}+h_{k+L}\right)^2/24h_{k+L},\ t_{k+L}\le t_{i+1}<t_{k+L+1}.$$

Обозначим $C=V-PR^{-1}Q^T$, где матрица C будет иметь размерность $(n-1)\times m$. Тогда оптимизационная задача примет знакомый вид $S_W(g)=(Y-Cg)^T W(Y-Cg)+\alpha g^T Kg\to\min$, решение которой дает искомые значения $g=\left(C^TWC+\alpha K\right)^{-1}C^TWY$, $\gamma=R^{-1}Q^Tg$, определяющие сплайн $g(t)$ в любой точке по формуле (2).

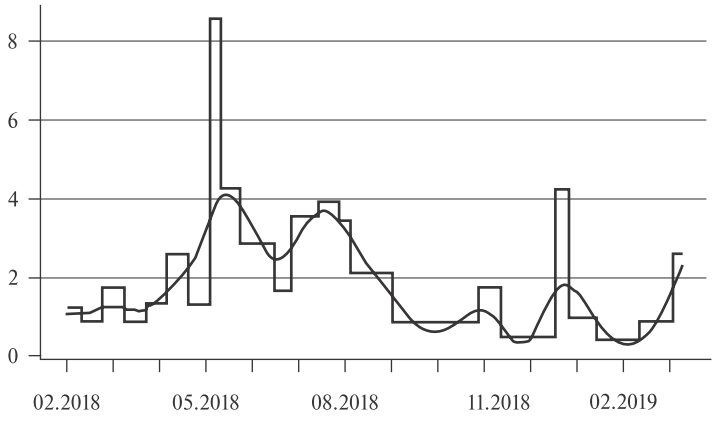

Рис. 4. Скорость расхода бутылок кваса универсамом, шт. в день

На рис. 4 для примера с поставками кваса в универсам показана построенная функция, которая в значительной степени лишена недостатка способа без использования базисных функций (когда узлами сплайна являлись точки наблюдений). Так как разница в сумме квадратов считается между значениями интегралов, которые значительно превосходят значение самой функции, квадрат второй производной у которой достаточно мал; параметр α, отвечающий за сглаживание, должен быть взят достаточно большим, например $\alpha = 10^5$.

5. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЯ

Представленный математический аппарат позволяет восстанавливать функцию по последовательности ее интегралов, причем в условиях, когда эти интегралы наблюдаются с погрешностью. По данным редких событий, таких как дискретные продажи и поставки, которые образуются в результате процесса потребления, схожего с опустошением емкости, можно определить, с какой скоростью заканчивался запас продукта у клиентов (в этом примере клиентом был сам универсам). В свою очередь, если со стороны универсама применить описанный метод, можно определить, с какой интенсивность расходуется квас у каждого конечного потребителя.

Определить точность восстановления функции на реальных данных не получится, так как неизвестна исходная функция, т.е. не с чем сравнивать. Мы можем самостоятельно заложить исходную функцию (спрос), моделируя процесс потребления (модели управления запасами) и получая данные покупок (табл. 2). Восстановление исходной функции происходит с очень хорошей точностью (рис. 5). Далее можно переходить к следующему этапу: определять закономерность и проводить экстраполяцию любыми известными методами.

На этапе поиска закономерности ответственность за результат экстраполяции полностью ложится на плечи исследователя, который, как предполагается, является специалистом в соответствующей прикладной области. На этом шаге можно использовать экспертное мнение и информацию из внешних источников, например пробовать искать зависимость от таких внешних признаков, как ВВП, уровень безработицы, курс рубля и др. В последнем примере внешней информацией является знание того, что исходная функция являлась гармонической, с помощью алгоритма Куинна–Фернандеса (Quinn–Fernandes algorithm) (Quinn, Fernandes, 1991; Quinn, Hannan, 2001) происходит определение соответствующей закономерности как разложение на фиксированное количество гармонических функций.

Стоит заметить, что наибольшая погрешность восстановления наблюдается на концах интервала, так как в этих точках сплайн не знает, куда стремиться, поэтому можно улучшить качество модели,

Таблица 2. Данные моделирования системы управления запасами

t_i	y_i	t_i	y_i	t_i	y_i	t_i	y_i	t_i	y_i
01.01.2018	1444,92	02.06.2018	1431,26	27.09.2018	1423,71	29.03.2019	1409,63	26.07.2019	1423,52
07.02.2018	1419,99	28.06.2018	1447,22	01.11.2018	1405,42	22.04.2019	1421,73	16.08.2019	1463,59
22.03.2018	1405,61	23.07.2018	1460,58	08.12.2018	1427,89	14.05.2019	1425,66	06.09.2019	1419,05
18.04.2018	1420,30	13.08.2018	1418,59	09.01.2019	1418,25	07.06.2019	1423,06	03.10.2019	1415,66
10.05.2018	1415,2	03.09.2018	1467,09	21.02.2019	1421,34	03.07.2019	1435,58	11.11.2019	1427,14

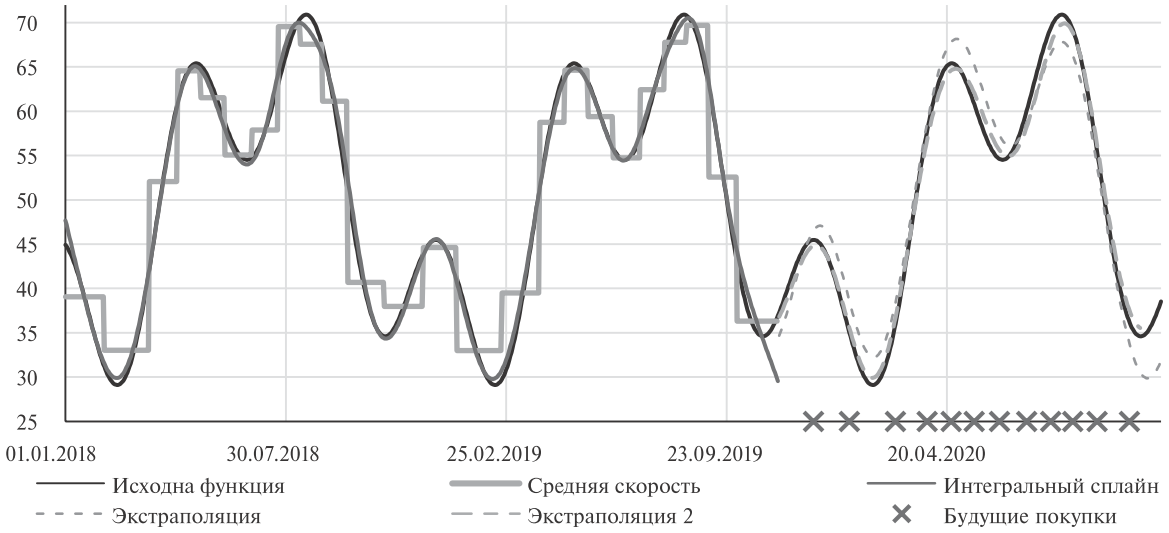

Рис. 5. Пример анализа и прогнозирования редких событий

Таблица 3. Сравнение моментов времени прогнозных и фактических событий

Прогноз	15.12.2019	18.01.2020	02.03.2020	01.04.2020	24.04.2020	16.05.2020
Факт	15.12.2019	18.01.2020	03.03.2020	02.04.2020	25.04.2020	18.05.2020
Прогноз	09.06.2020	05.07.2020	28.07.2020	18.08.2020	10.09.2020	11.10.2020
Факт	12.06.2020	07.07.2020	29.07.2020	18.08.2020	10.09.2020	12.10.2020

если отбросить часть значений с обоих концов восстановленной функции. На рис. 5 линия «Экстраполяция» построена по модели, оцененной по всей выборке, а «Экстраполяция 2» — по выборке после отбрасывания 20 точек с каждого конца. После экстраполяции функции скорости расхода запаса определяем моменты будущих событий, моделируя процесс потребления как в системах управления запасами (величина заказа определяется из данных редких событий) (табл. 3).

Полученные прогнозные значения моментов времени возникновения будущих событий очень близки к моментам фактических событий (если продолжать моделирование). Ни один другой метод анализа редких событий не в состоянии дать прогноз с такой точностью. Однако из-за того что восстановление функции было неидеальным и параметры модели (частота, амплитуда и фаза колебаний) определялись с погрешностью, расхождение может со временем нарастать и прогнозирование на очень далекую перспективу будет неточным. Стоить отметить, что способ восстановления функции, приводящей к событиям, может иметь большое значение для науки в соответствующей прикладной области.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Барцев С.И., Охонин В.А. (1986). Адаптивные сети обработки информации. Красноярск: Институт физики СО АН СССР. Препринт № 59Б. [**Bartsev S.I., Okhonin V.A.** (1986). *Adaptive information processing networks.* Krasnoyarsk: Institute of Physics, Siberian Branch of the Academy of Sciences of the USSR. Preprint No. 59B (in Russian).]

Бауэрсокс Д. Дж, Клосс Д. Дж. (2008). Логистика: интегрированная цепь поставок. 2-е изд. Пер. с англ. Н.Н. Барышниковой, Б.С. Пинскера. М.: ЗАО «Олимп-Бизнес». [**Bowersox D.J., Closs D.J.** (2008). *Logistical management: The integrated supply chain process.* 2th ed. Translated from the English N.N. Baryshnikova, B.S. Pinsker. Moscow: Olimp-Biznes. Originally published by McGraw-Hill Higher Education, 1996 (in Russian).]

Бирюкова Т.К., Киреев В.И., Гершкович М.М. (2016). Методы численного дифференцирования и восстановления сеточных функций по интегралам, основанные на интегродифференциальных сплайнах. В сб.: Системы компьютерной математики и их приложения. Материалы XVII Международной научной конференции. Вып. 17. С. 106–112. Смоленск: Издательство СмолГУ. [**Biryukova T.K., Kireev V.I., Gershkovich M.M.** (2016). Methods of numerical differentiation and recovery of grid functions by integrals, based on integro-differential splines. *Computer mathematics systems and their applications. Materials of the XVII International scientific conference.* Issue 17. Smolensk: Izdatel'stvo SmolGU, 106–112 (in Russian).]

Вентцель Е.С., Овчаров Л.А. (2000). Теория случайных процессов и ее инженерные приложения. Учеб. пособие для втузов. 2-е изд. М.: Высшая школа. [**Wentzel E.S., Ovcharov L.A.** (2000). *The theory of random processes and its engineering applications. Training manual for technical colleges.* 2nd ed. Moscow: Vysshaja Shkola (in Russian).]

Вожжов А.П., Луняков О.В., Вожжов С.П. (2015). Формирование страховых запасов предприятия при пуассоновском характере поступающих и выдаваемых потоков. В сб.: Экономика и управление: теория и практика. Т. 1. № (1). С. 30–35. [**Vozhzhov A.P., Lunyakov O.V., Vozhzhov S.P.** (2015). Safety stock determination with application of the Poisson processes to incoming and outgoing flows. In: *Economics and management: Theory and practice,* 1 (1), 30–35 (in Russian).]

Дзанагова И.Т., Хугаева Л.Т. (2015). Информационно-статистические методы построения экстремальных моделей редких событий // *Фундаментальные исследования*. № 11 (6). С. 1081–1084. [**Dzanagova I.T., Khugaeva L.T.** (2015). Method of operator series for constructing extremal models of rare events. *Fundamental Research*, 11 (6), 1081–1084 (in Russian).]

Иванько Р.С. (2005). Краткосрочное прогнозирование нестационарного спроса в оптовой торговле: дисс. ... канд. эконом. наук. Москва. [**Ivanko R.S.** (2005). *Short-term forecasting of non-stationary demand for wholesale.* Abstract of thesis for Cand. Sc. (Economics). Moscow (in Russian).]

Киреев В.И. (1994). Интегральный метод приближения функций алгебраическими многочленами и биквадратными сплайнами // *Вестник Московского авиационного института*. Т. 1. № 1. **С**. 48–58. [**Kireev V.I.** (1994). Integral method of approximation of functions by algebraic polynomials and biquadratic splines. *Vestnik Moskovskogo aviatsionnogo institute*, 1, 1, 48–58 (in Russian).]

Киреев В.И., Бирюкова Т.К. (1998). Полиномиальные интегродифференциальные одномерные и двумерные сплайны // *Вычислительные технологии*. Т. 3. № 3. **С**. 19–34. [**Kireev V.I, Biryukova T.K.** (1998). Polynomial integro-differential one-dimensional and two-dimensional splines. *Computational Technologies*, 3, 3, 19–34 (in Russian).]

Киреев В.И., Бирюкова Т.К. (2014). Интегродифференциальный метод обработки информации и его применение в численном анализе. М: ИПИ РАН. [**Kireev V.I, Biryukova T.K.** (2014). *Integro-differential information processing method and its application in numerical analysis.* Moscow: IPI RAS (in Russian).]

Кораблев Ю.А. (2015а). Емкостный метод определения функции скорости потребления // *Экономика и менеджмент систем управления*. Т. 15 (1.1). С. 140–150. [**Korablev Yu.A.** (2015a). Capacity method determination consumption rate function. *Economics and Management Systems*, 15 (1.1), 140–150 (in Russian).]

Кораблев Ю.А. (2015б). Обоснование емкостного метода определения спроса // *Экономика и статистика*. № 5. С. 96–101. [**Korablev Yu.A.** (2015b). Argumentation of capacity method demand determination. *Statistics and Economics*, 5, 96–101 (in Russian).]

Кораблев Ю.А. (2017а). Емкостный метод анализа редких продаж в Excel // *Экономика и управление: проблемы, решения*. № 6. Т. 3 (66). С. 224–230. [**Korablev Yu.A.** (2017a). Capacity method for analyzing rare sales in Excel. *Ekonomika i Upravlenie: Problemy, Resheniya*, 6, 3 (66), 224–230 (in Russian).]

Кораблев Ю.А. (2017б). Разбор причин и оценка погрешности аномальных картин в емкостном методе анализа редких событий // *Экономика и управление: проблемы, решения*. Т. 8 (6). С. 8–12. [**Korablev Yu.A.** (2017b). The causes analysis and error estimation of the anomalous pictures in the capacity method for the analysis of rare events. *Ekonomika i Upravlenie: Problemy, Resheniya,* 8 (6), 8–12 (in Russian).]

Кораблев Ю.А. (2018). Исследование точности емкостного метода от позиции в цепочке распространителей // *Экономика и управление: проблемы, решения*. Т. 7 (5). С. 106–121. [**Korablev Yu.A.** (2018). The study of the capacitive method accuracy from the position in the chain of distributors. *Ekonomika i Upravlenie: Problemy, Resheniya,* 7 (5), 106–121 (in Russian).]

Кораблев Ю.А. (2019а). Погрешность емкостного метода анализа редких событий, удаленность от конечного потребителя // *Известия Кабардино-Балкарского научного центра РАН*. № 3 (89). С. 48–77. DOI: 10.35330/1991-6639-2019-3-89-48-77 [**Korablev Yu.A.** (2019a). Error of the capacity method of rare events analysis, remoteness from the end user. *The News of KBSC of RAS*, 3 (89), 48–77. DOI: 10.35330/1991-6639-2019-3-89-48-77 (in Russian).]

Кораблев Ю.А. (2019б). Емкостный метод анализа редких событий в торговле различными товарами // *Бизнес. Образование. Право. Вестник Волгоградского института бизнеса*. № 3(48). С. 121–131. DOI: 10.25683/VOLBI.2019.48.313 [**Korablev Yu.A.** (2019b). Capacity method of analyzing rare events in the trade of various goods. *Business. Education. Law. Bulletin of Volgograd Business Institute*, 3, 121–131. DOI: 10.25683/VOLBI.2019.48.313 (in Russian).]

Лукинский В., Замалетдинова Д. (2015а). Методы управления запасами: расчет показателей запаса для товарных групп, относящихся к редким событиям (часть I) // *Логистика*. № 1 (98). С. 28–33. [**Lukinsky V., Zamaletdinova D.** (2015a). Methods of inventory management: The calculation of inventory indicators for product groups related to rare events (Part I). *Logistics*, 1 (98), 28–33 (in Russian).]

Лукинский В., Замалетдинова Д. (2015б). Методы управления запасами: расчет показателей запаса для товарных групп, относящихся к редким событиям (часть II) // *Логистика*. № 2 (99). С. 24–27. [**Lukinsky V., Zamaletdinova D.** (2015b). Methods of inventory management: the calculation of inventory indicators for product groups related to rare events (Part II). *Logistics*, 2 (99), 24–27 (in Russian).]

Федорова О.П. (2008). Об одном подходе к приближению функции сплайнами // *Вестник Томского государственного университета. Математика и механика.* № 2 (3). С. 61–66. [**Fedorova O.P.** (2008). One variant of spline approximating a function. *Tomsk State University Journal of Mathematics and Mechanics*, 2 (3), 61–66 (in Russian).]

Федорова О.П. (2016). Метод построения сплайна, сохраняющего интеграл функции двух переменных по области ее задания // *Научный альманах*. № 1–3 (15). С. 31–35. [**Fedorova O.P.** (2016). Method of creation of a spline with the integral equal to integral of function of two variables on area of its definition. *Science Almanac*, 1–3 (15), 31–35 (in Russian).]

Altman N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46 (3), 175–185. DOI:10.1080/00031305.1992.10475879

Boor C. de (2001). *A Practical Guide to Splines*. Revised Edition. New-York: Springer.

Cover T., Hart P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory,* 13 (1), 21–27.

Croston J.D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly (1970–1977),* 23 (3), 289–303.

Efron B., Tibshirani R.J. (1993). *An introduction of the bootstrap*. New York: Chapman & Hall.

Green P.J., Silverman B.W. (1994). *Nonparametric regression and generalized linear models. A roughness penalty approach*. New York: Chapman & Hall.

Johnston F.R., Boylan J.E. (1996). Forecasting intermittent demand: A comparative evaluation of Croston's method. Comment. *International journal of forecasting,* 12 (2), 297–298.

Quinn B.G., Fernandes J.M. (1991). A fast efficient technique for the estimation of frequency. *Biometrika*, 78, 3 (Sep.), 489–497.

Quinn B.G., Hannan E.J. (2001). *The estimation and tracking of frequency*. Cambridge: Cambridge University Press.

Rumelhart D.E., Hinton G.E., Williams R.J. (1986). Learning internal representations by error propagation. In: *Parallel Distributed Processing*. 1, 318–362. Cambridge: MIT Press.

Willemain T.R., Park D.S., Kim Y.B., Shin K.I. (2001). Simulation output analysis using the threshold bootstrap. *European Journal of Operational Research*, 134 (1), 17–28.

Walker S.H., Duncan D.B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54 (1/2), 167–178. DOI: 10.2307/2333860. JSTOR2333860

The function restoration method by integrals for analysis and forecasting of rare events in the economy

Yu.A. Korablev,
Financial University under the Government of the Russian Federation, Moscow, Russia;
email: yura-korablyov@yandex.ru

Received 18.07.2019

This study was supported by the Russian Foundation for Basic Research (project 19-010-00154).

Abstract. The article discusses a rare events analysis method, which is based on the study of the processes that generate these events. In the economy the most common process of event formation is the process of consumption or the disturbance accumulation, which can be modeled as a process of emptying or filling a capacity. The consumption process parameter will be the unsteady capacity emptying / filling rate function, which can be recovered from the available data. After restoring this function, you can analyze it, build a model and extrapolate it, then get a forecast of future events by starting again the process of event formation. I call this rare events research method the capacity method. To restore the emptying / filling rate function, an optimization problem has been solved, which is represented in the form of finding a special smoothing integrating cubic spline. Formulas are obtained in matrix form for the restoration (regression) of the desired function. Since the intervals between events can be different, it is necessary to proceed to basic splines (B-splines), which do not depend on the initial data. Formulas in matrix form for constructing the corresponding B-spline are obtained. Details are given of how to fill all the matrices. A mathematical method of restoring a function from rare events and example of future events forecast obtaining are given.

Keywords: rare events, sparse events, capacity method, consumption rate, recovery, regression, spline, B-spline, integrating spline, integro-differential spline, nonlinearity penalty.

JEL Classification: C1, C15, C4, C5, C53.

DOI: 10.31857/S042473880010485-2