ЭКОНОМИКА И МАТЕМАТИЧЕСКИЕ МЕТОДЫ, 2022, том 58, № 2, с. 80–91

МАТЕМАТИЧЕСКИЙ АНАЛИЗ ЭКОНОМИЧЕСКИХ МОДЕЛЕЙ

Определение параметров процесса образования редких событий в экономике для их последующего прогнозирования

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

Поступила в редакцию 05.11.2021

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

Аннотация. В статье представлен метод определения неизвестных параметров процесса, формирующего редкие события в экономике. Редкие события рассматриваются не со статистической точки зрения, а с точки зрения процессов, которые образуют эти события. Причем процесс образования редких событий может быть задан произвольным алгоритмом. Такой процесс также будет использовать неизвестные параметры, которые могут быть не только статическими, но и динамическими. Например, если рассматривать процесс потребления, в результате которого образуются дискретные покупки неподконтрольными нам покупателями, то такими параметрами могут быть максимальный запас и динамически изменяющаяся скорость потребления. В общем виде процесс может быть произвольным и его могут описывать различные параметры. Задача состоит в нахождении этих неизвестных параметров процесса, ориентируясь только на выборку редких событий. Идея метода — минимизация функции потерь, которая определяется на основе различий между событиями, образованными в результате функционирования модели процесса, и событиями из исходной выборки наблюдений. Каждое событие, помимо времени появления, характеризуется еще и дополнительной информацией, например объемом покупки. Необходимо найти такие значения параметров процесса, которые позволяли бы получить очень похожую выборку событий. Динамические параметры процесса задаются в виде кубических сплайнов особой структуры. Для однозначного описания каждого динамического параметра в целевую функцию вносится штраф за чрезмерную гладкость (шероховатость) соответствующих сплайнов. Приведен пример процесса и структура его параметров, подлежащая определению. Оптимизация происходит численными методами, за основу берется алгоритм Нелдера–Мида, который запускается на сетке, чтобы найти глобальный оптимум. Параметры процесса определяются по шагам, в начале только чтобы получить несколько событий, необходимых для продолжения расчетов, потом получаем следующую группу событий и т.д., это позволяет большую оптимизационную задачу разбить на последовательность простых задач, что существенно снижает общую трудоемкость. Описано предположение, которое должно соблюдаться, чтобы такой прием был справедлив. Рассмотрен пример выявления неизвестных параметров на примере процесса потребления. Определив параметры процесса, можно переходить к экстраполяции и осуществлять прогноз будущих событий.

Ключевые слова: редкие события, процесс образования событий, определение параметров процесса, прогнозирование событий, имитационное моделирование, оптимизация, алгоритм Нелдера–Мида.

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

Для цитирования: **Кораблев Ю.А.** (2022). Определение параметров процесса образования редких событий в экономике для их последующего прогнозирования // *Экономика и математические методы*. Т. 58. № 2. С. 80–91. DOI: 10.31857/S042473880020016-6

1. ВВЕДЕНИЕ

Анализ и прогнозирование редких событий в экономике являются актуальной задачей. Данное исследование развивает ранее предложенный подход для анализа и прогнозирования редких событий в экономике (Кораблев, 2020), который основывается на рассмотрении событий с точки зрения процессов, которые формируют эти события.

Другие существующие методы ограничиваются лишь статистическим рассмотрением событий. Одни методы определяют либо вероятности появления событий на определенном участке времени,

моделируют потоки событий как потоки случайных событий, в основном ограничиваясь пуассоновскими потоками без последействия, либо строят простейшие модели марковского процесса, например метод Виллемейна (Willemain et al., 2001, 2004).

Существуют подходы, которые базируются на методах классификации. В них в случае внешних наблюдаемых факторов распознаются закономерности, при которых в следующем периоде можно ожидать появления редкого события. Во всех этих методах не прогнозируется момент времени возникновения события, а дается оценка вероятности возникновения события на интервале времени. В большинстве методов, если предсказанное событие не появилось, то на следующий период опять дается точно такой же прогноз, т.е. все оценки получаются статическими.

Последние обзоры различных подходов к анализу и прогнозированию редких событий можно найти в работах (Kaya, Sahin, Demirel, 2020; Carreno, Inza, Lozano, 2020; Prince, Turrini, Meissner, 2021; Halim, Quaddus, Pasman, 2021). В них рассматриваются сотни работ других авторов, но ничего похожего на предлагаемый здесь метод в них нет.

Основная идея предлагаемого автором подхода заключается в том, чтобы анализировать процессы возникновения событий в источниках этих событий. В предыдущих работах (Кораблев, 2020; Кораблев, Голованова, Кострица, 2020) такими процессами были процессы потребления или накопления возмущения, а источниками — покупатели или клиенты (что позволило прогнозировать будущие покупки клиентов или моменты обращения клиента в парикмахерскую). Однако эти работы ограничивались исключительно процессами, схожими с опустошением/наполнением емкости, и процессы анализировались исключительно математическими методами. В этой работе мы распространим подход на произвольные процессы (для лучшего понимания демонстрационный пример будет основан на процессе потребления, но вместо него можно рассмотреть любой процесс).

Предлагаемый подход анализа редких событий состоит из пяти этапов:

1) события разделяются по разным выборкам — в зависимости от того, в каких источниках событий они были сформированы (выполняется автоматически, если данные от разных источников не смешиваются);

2) выдвигаются предположения о процессе, который формирует события в источнике (построение алгоритмической модели);

3) определяются параметры процесса из имеющейся выборки событий;

4) параметры процесса экстраполируются на будущее (любым известным методом);

5) запускается сам процесс с установленными значениями параметров, в результате которого получают прогноз будущих событий. Точность прогноза будущих событий зависит от точности определения и экстраполяции параметров процесса (при условии, что модель процесса была верной).

Данное исследование направлено на разработку универсального метода анализа и прогнозирования событий, которые могут порождаться произвольными процессами. Стоит отметить, что нас не интересуют случайные процессы, т.е. такие, в которых для образования событий используются случайные числа. Случайность есть мера неопределенности, мера незнания. Нас интересуют процессы, в которых события образуются как бы детерминировано, по некоторым физическим законам. Это предположение можно ослабить, приняв, что если внутри процесса имеется неопределенность, то нас будет интересовать только динамика математического ожидания (усреднение по реализациям) всех нестационарных параметров, когда параметры процессов могут быть описаны в виде аналитических функций. Дополнительно можно допустить присутствие некоторой внешней неопределенности, из-за которой события наблюдаются с погрешностями. В итоге необходимо определить нестационарные параметры процесса по зашумленным данным.

2. ОСНОВНАЯ ИДЕЯ

В первую очередь необходимо описать процесс формирования событий. В общем виде это может быть произвольный алгоритм. Как в эконометрике или анализе данных исследователь формирует математическую модель, так и здесь будет формироваться модель, но не математическая, а алгоритмическая. Ключевым элементом в такой модели является образование событий в результате некоторых операций сравнения. При формировании событий, помимо момента времени, может возвращаться информация об этом событии (например, в процессах потребления такой информацией будет объем покупок).

В модели процесса могут использоваться нестационарные параметры процесса, которые будут изменяться со временем независимо от того, что происходит в самом процессе. Эти параметры ненаблюдаемы, их нельзя назвать экзогенными переменными. Их можно сравнить с коэффициентами эконометрических моделей $(a_0, a_1, \ldots)$, но с допущением, что они будут динамическими, т.е. $a_0(t), a_1(t), \ldots$. Например, в процессах потребления (как в моделях управления запасами) переменной будет являться текущий запас, а параметром — нестационарный спрос. Мы хотим найти эти неизвестные параметры. Определять их будем с помощью соотнесения имеющейся выборки редких событий с выборкой событий, полученной в результате функционирования нашей модели. Для этого можно задать функцию потерь и оптимизировать ее численными методами. В результате такой оптимизации будут подобраны параметры процесса, при которых формируемая выборка будет мало отличаться от имеющейся выборки наблюдений. Основная идея достаточно проста и интуитивно понятна, однако реализация сталкивается с некоторыми трудностями, которые успешно преодолеваются.

3. ФОРМА ПРЕДСТАВЛЕНИЯ ПАРАМЕТРОВ ПРОЦЕССА

Статические параметры процесса будут задаваться одним числом. Динамические параметры будут задаваться в виде кубических сплайнов $g(t)$, представленных через изменяющуюся скачками третью производную $g'''(s_k)$ в каждом узле s_k. Наш сплайн будет немного похож на дифференциальное уравнение. Он будет определяться через начальные значения в стартовом узле, а на всех других узлах будет меняться только его третья производная. Заметим, что такая форма сплайна эквивалента представлению через значения и вторые производные (value-second derivative representation), которая, в свою очередь, эквивалента классическому представлению четырьмя параметрами при степенях переменной[1]. В итоге, чтобы задать значение изменяющегося со временем параметра процесса в виде сплайна $g(t)$, требуется задать:

1) узлы сплайна $s_1 < s_2 < \ldots < s_m$ (точки сочленения полиномов, число узлов m и их местоположение выбирается исследователем априори; самый распространенный способ выбрать узлы, чтобы они совпадали с наблюдениями, или распределить их равномерно на всем интервале наблюдений);

2) начальное значение, а также первую и вторую производную в первом узле $g(s_1)$, $g'(s_1)$, $g''(s_1)$;

3) значения третьей производной $g'''(s_k)$ во всех предшествующих моменту времени t узлах $\forall k : s_k < t$.

Благодаря такой форме представления сплайна необходимо определять только те значения, которые предшествовали интересующему нас периоду времени. Это поможет нам, когда мы начнем заниматься оптимизацией и подгонять параметры процесса (мы сможем подбирать параметры по одному событию последовательно, а не для всей выборки сразу). Заметим, что в последнем узле значение третьей производной можно не задавать.

Значения сплайна в произвольный момент времени t определяется через значения и производные, соответствующие предшествующему узлу s_k (из классической формы сплайна последовательными преобразованиями выводится форма, соответствующая ряду Тейлора, что было ожидаемо). Для времени t и предшествующего этому времени узла s_k выполняем:

$$g''(t) = g''(s_k) + (t - s_k) g'''(s_k), \tag{1}$$

$$g'(t) = g'(s_k) + (t - s_k) g''(s_k) + 0{,}5(t - s_k)^2 g'''(s_k), \tag{2}$$

$$g(t) = g(s_k) + (t - s_k) g'(s_k) + 0{,}5(t - s_k)^2 g''(s_k) + \left[(t - s_k)^3 / 6\right] g'''(s_k). \tag{3}$$

Данная форма представления сплайна гарантирует непрерывность вплоть до второй производной, третья производная меняется скачками в узлах сплайна. Выражения серьезно упрощаются, если время продвигается всегда только на единицу. При этом вычисления опираются на значения не предыдущего узла, а на значения в предыдущий момент времени. В модели процесса будем использовать процедуру продвижения модельного времени, в результате которой продвигаются все значения динамических параметров.

Заметим, что практически во всех работах, посвященных восстановлению неизвестных функций сплайнами, применяются натуральные сплайны, которые в начале и в конце превращаются в прямую линию. Таким образом, вторая производная на концах сплайна обнуляется, в начале

[1] Автор провел все соответствующие аналитические расчеты, чтобы убедиться в этом.

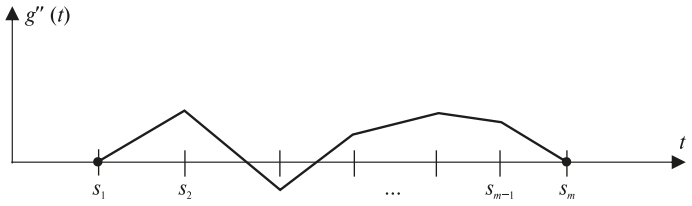

Рис. 1. Поведение второй производной $g''(t)$ в кубическом натуральном сплайне

Примечание. Третья производная изменяется скачками, вторая кусочно-линейно. Вторая производная начинается и заканчивается в нуле.

отсчета она берется равной нулю, и к окончанию выборки она также должна прийти к нулю. Если мы введем такое условие, то вторую производную в начальном узле можно будет не подбирать. Можно не подбирать и значение третьей производной в предпоследнем узле s_{m-1}, так как его можно выразить через ранее определенные значения. Если $h_k = s_{k+1} - s_k$ — расстояние между узлами, $g''(s_1) = 0$, то $g''(s_m) = \sum_{k=1}^{m-1} h_k g'''(s_k) = 0$, откуда

$$g'''(s_{m-1}) = -\sum_{k=1}^{m-2} h_k g'''(s_k) / h_{m-1} \tag{4}$$

(рис. 1).

Ранее было сказано, что наблюдения могут быть с погрешностью (данные зашумлены). Обычно для борьбы с шумом к функции потерь добавляют штраф за чрезмерную гладкость (шероховатость, roughness). В нашей задаче для кубического сплайна штраф на шероховатость рассчитывается как $\int_{s_1}^{s_m} \left(g''(t)\right)^2 dt$. Это можно сделать как во время функционирования процесса, суммируя квадрат второй производной, так и после завершения процесса. Пропуская промежуточные вычисления (следует интегрировать квадрат $g''(t) = g''(s_k) + (t - s_k) g'''(s_k)$), получим, что штраф за шероховатость для нашего сплайна можно определить как

$$\int_{s_1}^{s_m} \left(g''(t)\right)^2 dt = \sum_{k=1}^{m-1} \frac{\left(g''(s_k) + g'''(s_k) h_k\right)^3 - \left(g''(s_k)\right)^3}{3 g'''(s_k)}, \tag{5}$$

где $g''(s_k) = g'''(s_1) h_1 + \ldots + g'''(s_{k-1}) h_{k-1}$.

Далее, возвращаясь к процессу образования событий, вспоминаем, что параметров может быть несколько. И для каждого динамического параметра необходимо предусмотреть все вышеописанное (для разных параметров узлы сплайна удобно располагать одинаковым способом, но в общем случае количество узлов и их расположение может отличаться). Дополнительно позволим некоторым параметрам быть статическими, а не динамическими. В этом случае для них определяется только одно значение. В итоге для модели процесса необходимо предусмотреть схему значений для описания параметров процесса, показанную на рис. 2.

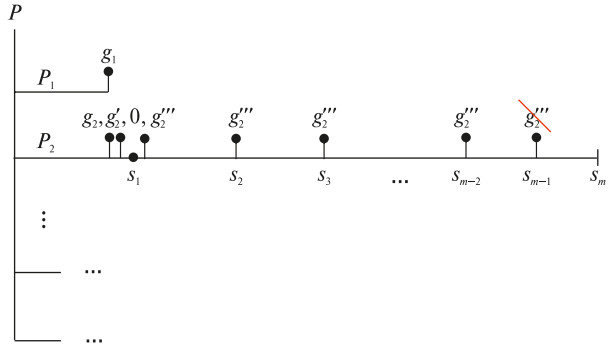

Рис. 2. Пример схемы значений для описания параметров процесса формирования событий; P_1 — первый параметр (статический); P_2 — второй параметр (динамический)

4. ПРИМЕР ПРОЦЕССА И СХЕМЫ ПАРАМЕТРОВ

В качестве примера рассмотрим процесс потребления, как в моделях управления запасами (с пополнением запаса до максимума), но с учетом того, что этот процесс использует параметры. Параметрами будут максимальный запас и спрос. Критический уровень запаса выбирается равным 0, так как он оказывается как бы мультиколлинеарен максимальному запасу, поэтому можно одновременно изменять их на одинаковое число, и результат от этого не изменится. Начальное и конечное время задаются вне модели (так как не являются ни параметрами, ни переменными). Модель процесса будет следующей:

```
1)  Максимальный запас = P[1];
2)  Спрос = P[2];
3)  Критический запас = 0;
4)  Время = Начальное время;
5)  Запас = Максимальный запас;
6)  Пока (Время ≤ Конечное время)
7)  {
8)      Запас = Запас — Спрос;
9)      Если (Запас ≤ Критический запас) то
10)     {
11)     Создать событие (Время, Максимальный запас — Запас);
12)     Запас = Максимальный запас;
13)     }
14)     Продвинуть время и обновить параметры;
15)     Максимальный запас = P[1];
16)     Спрос = P[2];
17) }
```

Функция «Создать событие» добавляет в выборку событие с заданным временем и значением. Процедура «Продвинуть время и обновить параметры» продвигает время на один шаг (в данном примере — на 1 день), обновление параметров происходит, как было описано ранее — похожим на численное интегрирование способом. Схема значений параметров процесса будет такой же, как на рис. 2, но только с двумя параметрами (первый параметр — статический, второй — динамический). Когда происходит обращение к параметрам, то возвращается значение, соответствующее текущему времени. Например, можно хранить обновленное значение в самом первом значении (g_2 — для второго параметра). Статические параметры не изменяются (не обновляются).

Забегая вперед, напомним, что процесс будет запускаться только на время появления заданного числа событий (например, пока не появится четыре события). В этом случае в условие остановки цикла надо добавить проверку того, что сформировалось заданное число событий.

5. ФУНКЦИЯ ПОТЕРЬ

Обозначим моменты появления событий как t_i и t_i — для наблюдений и для событий, полученных в результате функционирования модели процесса (моделирования). Каждое событие несет некоторую информацию/признак/воздействие[2] (например, объем покупки). Обозначим эти признаки как y_i и y_i — соответственно для исходных наблюдений и для событий, полученных в результате моделирования.

[2] Ограничимся случаем, когда каждое событие характеризуется только одним скалярным значением, а не вектором значений (обобщить не составит проблем).

Введем функцию потерь, которая будет показывать различия в двух выборках, исходной и сгенерированной. Рассчитаем квадраты отклонений. Но так как по каждому событию мы имеем две характеристики, то у нас будет несколько сумм:

$$S = \sum_{i=2}^{n} \left(t_i - t_i\right)^2 + \mu \sum_{i=2}^{n} \left(y_i - y_i\right)^2, \tag{6}$$

где нумерация начинается со второго события, так как первое событие служит отправной точкой, с которой начинает функционировать модель процесса; μ — весовой коэффициент, необходимый для того, чтобы смешать в нужных пропорциях отклонения различных величин измерения; n — размер выборки. В качестве альтернативного варианта можно измерять квадраты относительных отклонений, в этом случае значения становятся безразмерными:

$$S = \sum_{i=2}^{n} \left(\frac{t_i - t'_i}{t_i - t_{i-1}} \right)^2 + \mu \sum_{i=2}^{n} \left(\frac{y_i - y'_i}{y_i} \right)^2, \tag{7}$$

где для времени берется отношение к предыдущему интервалу времени между событиями, а для значений — отношение к наблюдаемому значению (можно делить на y'_i, тогда будет определяться относительное отклонение наблюдения от прогнозного значения).

Чтобы восстановить параметры процесса можно было единственным способом, мы вводим штраф за гладкость (шероховатость) для каждого параметра с соответствующим весовым коэффициентом:

$$S = \sum_{i=2}^{n} \left(\frac{t_i - t'_i}{t_i - t_{i-1}} \right)^2 + \mu \sum_{i=2}^{n} \left(\frac{y_i - y'_i}{y_i} \right)^2 + \sum_{z=1}^{N_{\text{дин. пар.}}} \alpha_z \int_{s_1}^{s_m} \left(g''_z\ (t) \right)^2 dt, \tag{8}$$

где z — номер одного из динамических параметров; $N_{\text{дин. пар.}}$ — число динамических параметров; $g''_z(t)$ — вторая производная от функции (сплайна) динамического параметра; α_z — коэффициент, с которым учитывается данный штраф в общей функции потерь (также известен как параметр сглаживания, или коэффициент α-регуляризации Тихонова). Обратим внимание на то, что этот штраф интегрирует абсолютные значения квадрата второй производной. Получается немного странная ситуация, когда при сравнении событий используются относительные отклонения, а штраф дает абсолютные значения. Значение для параметра сглаживания в этом случае надо выбирать намного меньше.

Если модель процесса не дает нужного числа событий, мы можем утверждать, что функция потерь обращается в бесконечность (надо дать шанс сформироваться нужному числу событий, для этого необходимо немного увеличить время моделирования, например, на величину, равную 10 временным интервалам между последними двумя событиями).

6. ОПТИМИЗАЦИЯ

Для оптимизации используем хорошо зарекомендовавший себя алгоритм Нелдера–Мида (Nelder, Mead 1965), который способен двигаться в направлении лучшего значения, даже если расчет целевой функции не всегда возможен. Оптимизацию необходимо производить на сетке, чтобы не застрять в одном локальном оптимуме, причем если при оптимизации мы покидаем заданную ячейку, то оптимизация прекращается и происходит поиск уже из следующей ячейки. Существует несколько других подходов[3] для оптимизации в условиях множества локальных экстремумов, но, по мнению автора, удобнее и понятнее использовать обычный поиск на сетке, тем более что исходные данные и некоторые представления о модели процесса могут подсказать, какой шаг сетки следует выбирать по каждой переменной.

По каждому динамическому параметру в начальном узле сплайна надо определить три значения, а в каждом последующем узле — по одному значению. Если каждое значение перебирается на сетке, то получается огромное число комбинаций значений параметров. Мы будем подбирать значения для такого числа узлов, чтобы получить только несколько событий, а не все сразу. На первом шаге мы подбираем параметры процесса на N событий вперед, и для этих параметров запускаем

[3] Генетические алгоритмы (Goldberg, 1989; Саймон, 2020) и разбросанный поиск (scatter search, поиск по разбросу) (Laguna, Marti, 2006). Разбросанный поиск применяется в оптимизаторе OptQuest, который используется во многих системах имитационного моделирования, включая Anylogic.

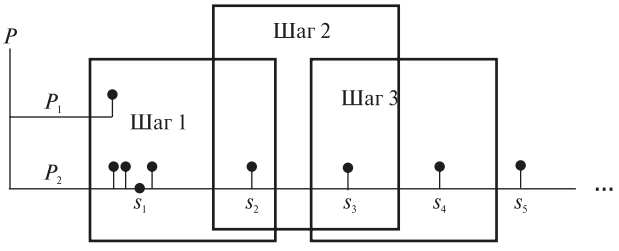

Рис. 3. Подбор параметров на $N = 2$ событий вперед

Примечание. На первом шаге перебирается пять параметров, на втором — два, на третьем — тоже два.

оптимизацию на сетке. На следующем шаге мы как бы сдвигаем скользящее окно на одну позицию (на одно событие) и начинаем подбирать параметры для вошедших N событий. Например, если $N = 2$, то на первом шаге подбираются параметры, отвечающие за образование первого и второго события[4], на втором шаге — второго и третьего, подбирая их на сетке.

Заметим, что в целевой функции учитываются как вошедшие, так и предшествующие, не вошедшие в скользящее окно, события (так как функция потерь строится на сравнении выборок событий). Параметры, отвечающие за образование предшествующих, не вошедших в окно событий, используются в оптимизации, но не на сетке. Мы предполагаем, что их определенные на предыдущем шаге значения лежат в нужном нам локальном оптимуме. На рис. 3 показано, какие параметры перебираются на сетке на каждом новом шаге для случая $N = 2$ (при условии, что узлы сетки s_k совпадают с наблюдениями).

На первом шаге на вход оптимизатора подается вектор из пяти значений (у параметра P_2 третье значение всегда 0), причем в качестве начальных точек берется полная комбинация всех возможных значений параметров. На втором шаге подается шесть значений, но число начальных комбинаций ограничено только комбинациями последних двух значений (в узлах s_2 и s_3 берется полная комбинация возможных значений, но на вход оптимизатора каждый раз подаются все значения, включая значения в узле s_1). То есть первые четыре значения также выбираются оптимизатором. Это позволяет избежать «эффекта бабочки», когда малые отклонения в начале приведут к огромным отклонениям в самом конце. На третьем шаге на вход оптимизатора подается семь значений, но комбинации определяются перебором на сетке только значений в узлах s_3 и s_4. Скользящее окно задает, для каких значений будет происходить полный перебор на сетке для того, чтобы определить начальные точки, из которых запускается оптимизация.

Можно сформулировать следующее предположение.

Предположение. *Параметры процесса, соответствующие событию, не зависят от событий, расположенных в более чем N позициях от текущего события*.

Иными словами, ранее запущенный алгоритм оптимизации нашел такой экстремум, который при добавлении новых событий и соответствующих им значений параметров может по-прежнему оказаться глобальным экстремумом при оптимизации на сетке новых добавленных значений. Это предположение может быть верно, так как восстанавливаемый динамический параметр (как функция) на одном участке может практически не зависеть от поведения этого же параметра на сильно удаленном участке.

Конечно, желательно определять размер N как можно большим, в идеале — равным размеру выборке, но это очень сильно повлияет на объем вычислений. Нужно найти некоторый баланс между временем выполнения всех расчетов и точностью определяемых параметров. Время выполнения первого шага может быть самым большим, так как на нем находится больше всего параметров, а время второго шага может быть наименьшим. С ростом номера шага время будет увеличиваться, так как процесс будет запускаться до тех пор, пока не образуется нужное число событий.

7. ПРИМЕР РАБОТЫ МЕТОДА

Рассмотрим работу метода на примере процесса потребления, описанного выше. В качестве начальных данных возьмем данные, образованные этим же процессом с некоторыми предустановленными параметрами (предполагаем их неизвестными). Нашей целью будет восстановить эти значения параметров, располагая только выборкой событий (табл. 1).

[4] Так как отправной точкой является время первого события, то первое образованное событие на самом деле будет вторым событием.

Таблица 1. Данные редких событий

t_i	y_i	t_i	y_i	t_i	y_i	t_i	y_i
03.01.2020	2429	30.04.2020	2526	17.07.2020	2508	09.10.2020	2418
25.01.2020	2461	14.05.2020	2500	29.07.2020	2466	24.10.2020	2457
22.02.2020	2465	26.05.2020	2556	09.08.2020	2418	12.11.2020	2478
11.03.2020	2528	06.06.2020	2402	21.08.2020	2484	10.12.2020	2409
26.03.2020	2462	19.06.2020	2574	05.09.2020	2517	31.12.2020	2478
12.04.2020	2481	03.07.2020	2437	23.09.2020	2467		

В качестве целевой используем функцию потерь с относительными отклонениями (8). Пусть определение дат событий t_i имеет большее значение, чем характеристика этого события y_i (объем покупки). Пусть $\mu = 0{,}1$ (вес суммы квадратов относительных отклонений y_i). Для штрафа за нелинейность примем коэффициент сглаживания, равный $\alpha = 10^{-6}$.

Прежде чем запускать алгоритм, следует определить, на сколько событий вперед N следует подбирать параметры. Если подбирать параметры на четыре события вперед, то первый шаг получается таким, как на рис. 4а, если на пять — как на рис. 4б (в табл. 2 и 3 показаны числовые результаты первого шага).

Так как алгоритм опирается только на значения целевой функции, рассчитываемые по отклонениям полученных событий от выборки наблюдений, алгоритм не осознает, что на рис. 4а функция потребления оказалась ниже среднего уровня. В случае $N = 4$ алгоритм подобрал максимальный запас, равный 1992, что позволило с меньшим потреблением сформировать события, расположенные близко к имеющимся наблюдениям. Ситуация немного улучшится, если продолжить выполнять алгоритм на следующих шагах (при $N = 4$), но когда скользящее окно сдвинется на одно событие, параметры в первом узле могут оказаться в локальном минимуме; при этом они перестанут перебираться на сетке. Картина заметно улучшается, если определять параметры процесса сразу на $N = 5$ событий вперед. Пятое событие вносит нужную информацию, позволяющую точнее определить начальные параметры. Далее результаты приводятся для случая $N = 5$.

На втором шаге скользящее окно событий сдвигается вправо, и для пяти событий осуществляется подбор параметров, учитывающих значения параметров, найденных для первого сформированного события. Значения параметров, определенные для этого события, подставляются в оптимизатор, «как есть», а для значений параметров, необходимых для образования следующих пяти событий, берется множество возможных комбинаций значений, оптимизация запускается для каждой такой комбинации.

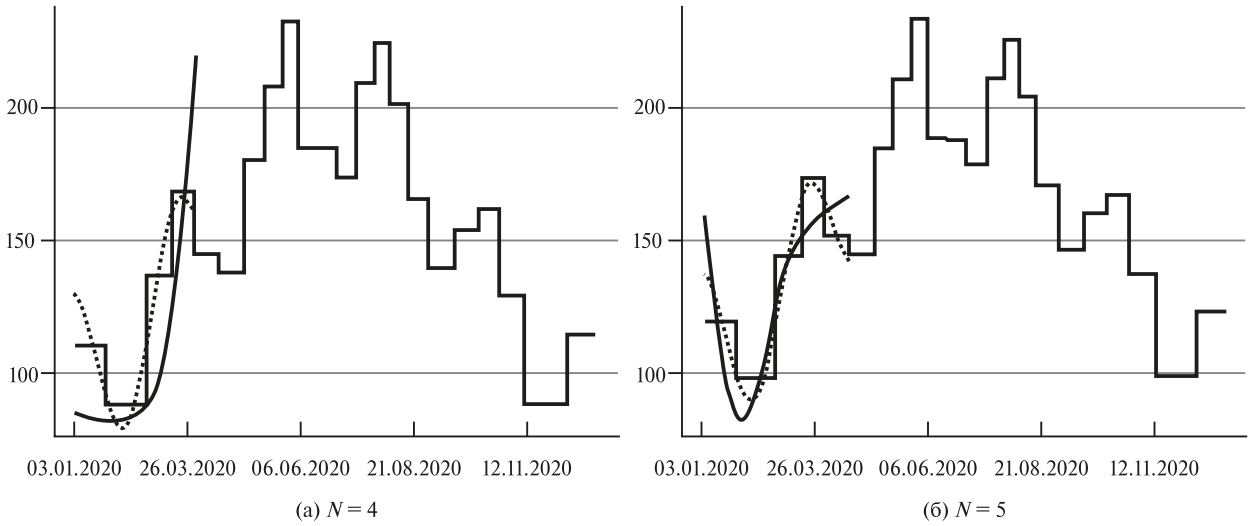

Рис. 4. Первый шаг алгоритма определения параметров на четыре и пять событий вперед

Примечание. Гладкая сплошная линия показывает восстановленное значение скорости потребления $g_2(t)$ (шт./день); пунктирная — искомую функцию потребления, заложенную в модель; ступенчатая — среднее потребление $y_i / (t_i + 1 - t_i)$; площадь под каждой ступенькой будет объемом покупки y_i.

Таблица 2. Значения параметров после первого шага алгоритма

Параметр	Значение	
	$N = 4$	$N = 5$
Максимальный запас (истинное значение 2400)	g_1=1992,141	$g_1 = 2371,628$
Начальная скорость потребления	$g_2(s_1) = 85,084$	$g_2(s_1) = 153,222$
Первая производная скорости потребления	$\dot{g}_2(s_1) = -0,21103$	$\dot{g}_2(s_1) = -4,8323$
Третья производная скорости потребления в узле s_1	$\dddot{g}_2(s_1) = -6,5058 \times 10^{-4}$	$\dddot{g}_2(s_1) = 1,4825 \times 10^{-2}$
Третья производная скорости потребления в узле s_2	$\dddot{g}_2(s_2) = -6,5419 \times 10^{-4}$	$\dddot{g}_2(s_2) = -1,4202 \times 10^{-2}$
Третья производная скорости потребления в узле s_3	$\dddot{g}_2(s_3) = 1,8618 \times 10^{-2}$	$\dddot{g}_2(s_3) = 5,6984 \times 10^{-5}$
Третья производная скорости потребления в узле s_4	$\dddot{g}_2(s_4)$ выражается через предыдущие значения	$\dddot{g}_2(s_4) = 4,0556 \times 10^{-3}$
Третья производная скорости потребления в узле s_5	–	$\dddot{g}_2(s_5)$ выражается через предыдущие значения

Таблица 3. Полученные события после первого шага алгоритма

События	Дата					
Исходные	03.01.2020	25.01.2020	22.02.2020	11.03.2020	26.03.2020	12.04.2020
$N-4$	–	26.01.2020	20.02.2020	12.03.2020	25.03.2020	–
$N-5$	–	24.01.2020	22.02.2020	11.03.2020	27.03.2020	12.04.2020
	Значения (объем покупки)					
Исходные	2429	2461	2465	2528	2462	2481
$N-4$	–	1992,141	2079,106	2116,739	2169,692	–
$N-5$	–	2386,468	2485,890	2378,407	2413,642	2532,705

Это позволяет кардинально пересмотреть ранее полученные значения параметров, попавших в скользящее окно (события со второго по пятое). Значения параметров для первого образованного события также изменяются при оптимизации, но уже только относительно полученных на первом шаге значений. Так, значение максимального запаса изменилось с ранее определенного 2371,628 на 2410,712. На третьем шаге скользящее окно вновь сдвигается на одну позицию и перебором на сетке определяются новые значения параметров, попавшие в это скользящее окно, а также корректируются значения параметров, в него не вошедшие.

Аналогично алгоритм выполняет все последующие шаги и останавливается после 18 шага, когда будут подобраны все значения параметров, необходимые для формирования 22 событий. Результат работы алгоритма после шагов 2 и 3 изображены на рис. 5, а за 18 шагов — на рис. 6. В табл. 4 показаны события, образованные в результате функционирования процесса, с подобранными параметрами. Только в 1 из 22 событий дата события отличается на 1 день. Значения событий (объемы покупок) отличаются в каждом событии от исходных очень незначительно (в большинстве событий меньше, чем на 1%, и только в последних событиях — на 2 и 3%).

Таблица 4. Полученные события после 18 шага

События	Дата					
Исходные	Полное совпадения в 21 из 22 событий; только для события с датой 12.11.2020 процесс определил событие на день раньше (11.11.2020)					
Процесс						
	Значения (объем покупки)					
Исходные	2429	2461	2465	2528	2462	2481
Процесс	–	2468,564	2472,071	2547,122	2464,828	2494,400
Исходные	2526	2500	2556	2402	2574	2437
Процесс	2545,614	2511,836	2566,986	2411,377	2585,020	2445,382
Исходные	2508	2466	2418	2484	2517	2467
Процесс	2498,655	2470,175	2406,726	2462,082	2515,078	2523,396
Исходные	2418	2457	2478	2409	2478	
Процесс	2489,236	2509,046	2522,387	2406,723	2575,255	

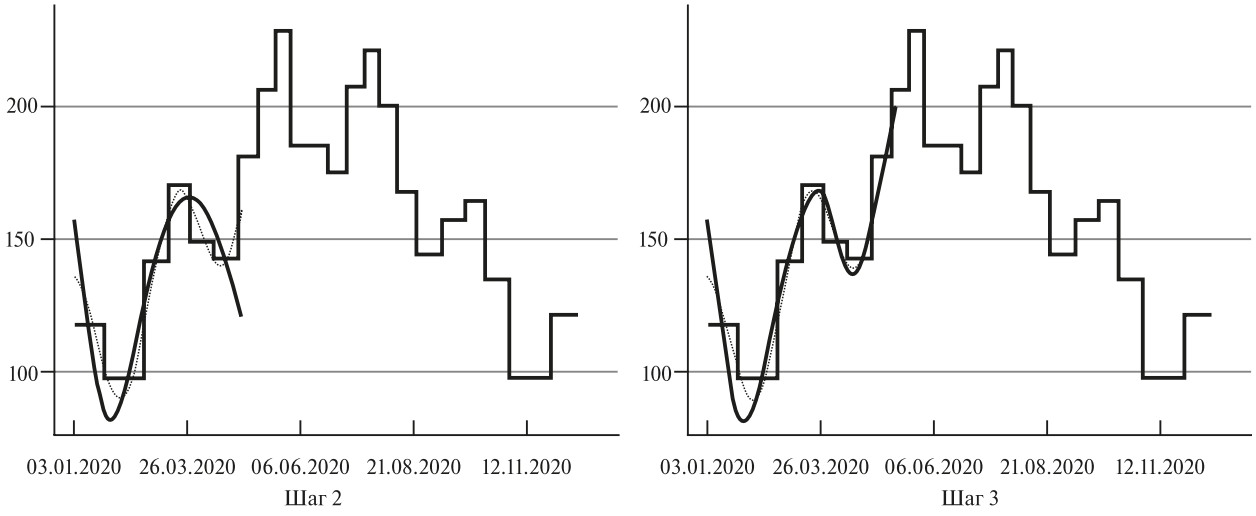

Рис. 5. Результат работы алгоритма после второго и третьего шагов

В результате применения данного метода получилось вполне надежно определить параметры процесса, который был задан лишь обычным алгоритмом, один параметр алгоритма был статическим, а второй динамическим. Напомню, что мы искали скорость расхода продукции не у себя, а у не подконтрольного нам клиента. В (Кораблев, 2020) математическим методом также получалось восстановить скорость потребления (программная реализация в (Korablev, 2022)), но сейчас мы определили даже размер запаса и объемы покупок.

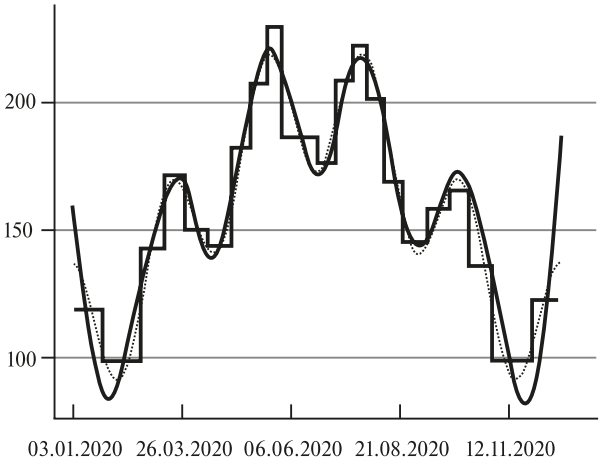

Рис. 6. Результат работы алгоритма после 18 шагов

Стоит обратить внимание на то, что благодаря использованию скользящего окна событий у нас получилось разделить одну большую оптимизационную задачу огромной размерности на небольшое число задач значительно меньшей размерности (трудоемкость немного увеличивается с номером шага). Если бы мы подбирали значения параметров для всех событий из выборки, то трудоемкость задачи была бы астрономической и мы не смогли бы за приемлемое время осуществить необходимые вычисления.

8. ЗАКЛЮЧЕНИЕ

Ключевой особенностью предложенного метода является способность восстанавливать параметры выбранного процесса образования событий, причем не только как статические, а как динамические параметры, которые плавно изменяются со временем. Существуют исследования, где моделируются процессы образования событий (например, в моделях управления запасами), но вот исследований, где восстанавливаются параметры процессов по имеющейся выборке, тем более в виде динамических функций, автор не встречал. В данном исследовании целью является не просто моделировать процесс образования событий, а именно восстановить его параметры.

В нашем примере для более глубокого понимания был взят процесс потребления, но на его место можно было поместить любой другой процесс образования событий с иным набором параметров и с произвольным алгоритмом. В методе сравнивается лишь выходной поток событий (с их информационными характеристиками), ограничений на способ формирования событий мы не делали. Однако модель этого процесса должна позволять однозначно находить значения этих параметров. Так, в нашем примере мы были вынуждены отказаться от определения значения

критического уровня запасов — иначе он бы информационно смешивался с максимальным запасом и нельзя было бы с помощью имеющихся данных отделить один от другого. Исследования на тему, как формировать такие алгоритмические модели, автору не попадались.

В отличие от чисто математических моделей представление процессов в виде алгоритмических моделей дает значительно большую гибкость (практически бесконечную) при исследовании редких событий. Получается некоторая задача регрессии алгоритмов. Если классические регрессионные задачи опираются на математические (алгебраические) модели и определение параметров происходит методом наименьших квадратов аналитически, то сейчас мы используем алгоритмические модели и опираемся на численные методы оптимизации. Интересно, станет ли в будущем такой подход в исследованиях столь же популярным, как эконометрика?

В нашем случае при оптимизации функции потерь мы ограничились обычной оптимизацией на сетке, чтобы не застревать только на одном локальном оптимуме. И этого было вполне достаточно, чтобы мы успешно справились со своей задачей. Однако можно применять и более сложные методы оптимизации, — такие как разбросанный поиск (scatter search) или генетические алгоритмы. Возможно, они позволят быстрее находить глобальный оптимум и, возможно, не будут перебирать лишние неправильные комбинации параметров. Исследователи свободны выбирать любые известные им средства оптимизации.

Как отмечалось в начале статьи, целью определения параметров процесса является дальнейшее прогнозирование будущих событий, но прежде необходимо произвести экстраполяцию параметров процессов на будущее. Экстраполяцию параметров процесса можно производить любым известным методом, можно осуществлять поиск закономерности параметров с такими внешними наблюдаемыми факторами, как ВВП, курс рубля, безработица и т.д. Определив, как будут вести себя параметры процесса в будущем, не составит труда запустить сам процесс и получить выборку будущих событий.

Можно пойти дальше и дополнительно предположить, что процесс является полностью неизвестным и требуется подобрать как процесс, так и его параметры, не опираясь ни на что, кроме как на исходную выборку редких событий. В этом случае можно даже создать особый механизм перебора моделей из некоторого множества операторов. У автора есть некоторые соображения по этому поводу, но данное исследование уже получилось достаточно объемным.

Сам по себе анализ и прогнозирование редких событий в экономике является важной и очень актуальной задачей. Способность разбираться в процессах образования событий позволит влиять на эти процессы и контролировать появление самих событий (если мы имеем возможность управлять этими процессами). В то же время, если мы не можем повлиять на процессы, мы сможем подготовиться к появлению будущих событий, что позволит либо извлечь определенную выгоду, либо уменьшить возможные потери. В русле этого направления исследований можно даже создать специальное направление обучения в экономике, которое бы готовило соответствующих аналитиков, открывать научные лаборатории и проводить научные исследования как для государственных структур, так и для бизнеса.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Кораблев Ю.А. (2020). Метод восстановления функции по интегралам для анализа и прогнозирования редких событий в экономике // *Экономика и математические методы*. М.: ЦЭМИ РАН. Т. 56. № 3. С. 113–124. DOI: 10.31857/S042473880010485-2 [**Korablev Yu.A.** (2020). The function restoration method by integrals for analysis and forecasting of rare events in the economy. *Economics and Mathematical Methods*, 56, 3, 113–124 (in Russian).]

Кораблев Ю.А., Голованова П.С., Кострица Т.А. (2020). Емкостный метод анализа редких событий в сфере услуг // *Экономическая наука современной России*, 90, 3, 132–142. DOI: 10.33293/1609-1442-2020-3(90)-132-142. [**Korablev Yu.A., Golovanova P.S., Kostritsa T.A.** (2020). Capacity method of rare events analysis in the area of services. *Economics of Contemporary Russia*, 90, 3, 132–142 (in Russian).]

Саймон Д. (2020). Алгоритмы эволюционной оптимизации. Биологически обусловленные и популяционно-ориентированные подходы к компьютерному интеллекту. Пер. с англ. А.В. Логунова. М.: ДМК Пресс. 1002 с. [**Simon D.** (2020). *Evolutionary optimization algorithms. Biologically-inspired and population-based approaches to computer intelligence*. Translated from the English by A.V. Logunov. Originally published by Willey, 2013 (in Russian).]

Carreno A., Inza I., Lozano J. (2020). Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, 53, 3575–3594. DOI:10.1007/s10462-019-09771-y

Goldberg D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Boston: Addison-Wesley. 432 p.

Halim S.Z., Quaddus N., Pasman H. (2021). Time-trend analysis of offshore fire incidents using nonhomogeneous Poisson process through Bayesian inference. *Process Safety and Environmental Protection*, 147, 421–429. DOI:10.1016/J.PSEP.2020.09.049

Kaya G.O., Sahin M., Demirel O.F. (2020). Intermittent demand forecasting: A guideline for method selection. *Sadhana — Academy Proceedings in Engineering Sciences*, 45, 1, 45–51. DOI: 10.1007/s12046-020-1285-8

Korablev Yu. A. (2022). Restoration of function by integrals with cubic integral smoothing spline in R. *ACM Transactions on Mathematical Software.* (In print). DOI: 10.1145/3519384

Laguna M., Marti R. (2006). Scatter Search. In: *Metaheuristic procedures for training neutral network*s, 139–152. DOI: 10.1007/0-387-33416-5_7

Nelder J.A., Mead R. (1965). A simplex method for function minimization. *The Computer Journal*, 7, 308–313. DOI: 10.1093/comjnl/7.4.308

Pince C., Turrini L., Meissner J. (2021). Intermittent demand forecasting for spare parts: A critical review. *Omega,* 105, 102513. DOI:10.1016/j.omega.2021.102513

Willemain T.R., Smart Charles N., Schwarz Henry F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, 20, 3, 375–387. DOI: 10.1016/S0169-2070(03)00013-X

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. DOI: 10.1016/S0377-2217(00)00209-5

Determination of parameters for the formation process of rare events in the economy for their subsequent forecasting

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

Received 05.11.2021

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

Abstract. The article presents a method for determining unknown parameters of the process that forms rare events in the economy. The idea behind the study of rare events in economy is to consider these events not just from a statistical point of view, but from the point of view of the processes that form these events, moreover, as well as a process can be an arbitrary algorithm. Such an event generation process will use parameters, which can be static or dynamic. For example, if we consider the consumption process, which forms discrete purchases of an uncontrolled customer, then such parameters can be the maximum stock and a dynamically changing consumption rate. In general, the process can be arbitrary and possess various parameters. The task is to determine these parameters of an unknown process obtaining only a sample of rare events. The idea of the method is to minimize the loss function, which is determined based on the differences between the events generated during the operation of the process and events from the initial sample of observations. Each event, in addition to the time of occurrence, is also characterized by additional information, for example, the purchase volume. We are trying to find out such parameters of such a process that would allow us to get a very similar sample of events. The dynamic parameters of the process are set in the form of cubic splines of a special structure. For an unambiguous determination of each dynamic parameter, a roughness penalty of the corresponding splines is introduced into the objective function. An example of a process and its structure of parameters to be determined is shown. Optimization is performed numerically, based on the Nelder–Mead algorithm, which runs on a grid to determine the global optimum. The process parameters are determined in steps, at the beginning just to get a few events, then the next events. That allows one large optimization task to be divided into a sequence of simple tasks, this significantly reduces the overall complexity. An assumption is described that must be fulfilled for such a technique to be valid. An example of determining unknown parameters is considered on the example of the consumption process. After determining the process parameters, one can proceed to extrapolation of parameters and forecast future events.

Keywords: rare events, process of event formation, determination of process parameters, events forecast, simulation modeling, optimization, Nelder–Mead algorithm.

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

For reference: **Korablev Yu.A.** (2022). Determination of parameters for the formation process of rare events in the economy for their subsequent forecasting. *Economics and Mathematical Methods*, 58, 2, 80–91. DOI: 10.31857/S042473880020016-6